Bayesian Clustering and Modeling Approaches for the Analysis of Brain-Imaging Data

by

Haoyi Fu

B.S. in School of Life Sciences, Nanjing Agricultural University, China, 2015

M.S. in Department of Biostatistics, University of Pittsburgh, USA, 2017

Submitted to the Graduate Faculty of the

School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Haoyi Fu

It was defended on

December 8, 2022

and approved by

Robert T. Krafty, PhD, Professor, Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University

Lu Tang, PhD, Assistant Professor, Department of Biostatistics, School of Public Health, University of Pittsburgh

Ori Rosen, PhD, Professor, Department of Mathematical Sciences, University of Texas at El Paso

Andriy I. Bandos, PhD, Associate Professor, Department of Biostatistics, School of Public Health, University of Pittsburgh

Alison E. Hipwell, PhD, PsyD, Professor, Department of Psychiatry, University of Pittsburgh

Bayesian Clustering and Modeling Approaches for the Analysis of Brain-Imaging Data

Haoyi Fu, PhD

University of Pittsburgh, 2022

With the rapid development of modern techniques to measure functions and structures of the brain, statistical methods for analyzing brain-imaging data have become increasingly important to the advancement of science. My dissertation focuses on developing Bayesian clustering and modeling approaches for brain-imaging data with application to a brain-imaging technique in particular: functional near-infrared spectroscopy (fNIRS).

In the first part, I propose a group-based approach to clustering univariate time series via a mixture of smoothing splines experts. Time-independent covariates are incorporated via the logistic weights of a mixture-of-experts model. I formulate the approach in a fully Bayesian framework and conduct inference via reversible jump Markov chain Monte Carlo (RJMCMC) where the number of mixture components is assumed unknown. The superior performance of the approach in terms of subgroup detection and estimation is demonstrated through both simulation studies and applications to the analysis of fNIRS data.

In the second part, the approach proposed in the first part is extended to the multivariate time series setting. Parameter estimation and inference are performed by Gibbs sampling, and the number of multivariate components is selected based on the deviance information criterion (DIC). The superior performance of the approach in terms of subgroup detection and estimation is demonstrated by simulation studies and applications to the analysis fNIRS data.

In the final part, I propose a horseshoe prior-based generalized lasso for interpretable scalar on function regression. The approach is able to penalize regression coefficients with selected orders of differences by specifying appropriate prior structures. The horseshoe prior is able to control both the global and local shrinkage levels of each coefficient simultaneously. The proposed method is demonstrated to have superior performance in terms of signal detection and prediction accuracy through simulation studies, and is applied to the analysis of

fNIRS data.

Public Health Significance:

Developing model-based clustering and modeling approaches provides innovative statistical methods for the analysis of brain-imaging data, which overcome the challenges of heterogeneity and high dimensionality. My proposed methods facilitate the discovery of underlying patterns of brain-imaging signals, as well as associations between these functional signals and clinical outcomes.

Table of Contents

List of Tables

List of Figures

Preface

This dissertation is a summary of my PhD career at the University of Pittsburgh. I have a feeling of both happiness and sadness when I am writing the preface of my dissertation. The happiness, of course, comes from the achievements I have made during my PhD career. All my hard work will pay off if I successfully defend my dissertation later and am awarded with the Doctoral degree. The sadness comes from a feeling that it marks the end of my student career. When I look back into the past 29 years of my life, I spent 22 years as a student. During the long journey from primary school in China to graduate school in the United State, I have always been enjoying gaining new knowledge and skills. However, my PhD career is a totally different learning experience compared to my previous student career. At Pitt, I have learned the so-called PhD way of thinking, which allows me to think and learn totally by myself, with less dependence on others. From my PhD career, I have learned how to conduct research by myself, solve different types of problems by myself and think about my future more comprehensively by myself. I would say, this period is the most unforgettable so far in my life. I will always remember the achievements I have made, the hardship I have gone through, and all the happy moments I have spent with my friends.

There are many people who gave me many valuable suggestions and help during my PhD career. Without their help, it would have been impossible for me to complete my PhD. First, I would like to give a huge thank to my dissertation advisor, Dr. Robert Krafty. I really appreciate those times I had worked with you before my PhD and as your student for my PhD. I will always remember your help when I applied for the PhD program several years ago. You are such a great and warm-hearted advisor, which not only gave me lots of ideas and directions for my dissertation research but also provided me with tremendous support. You are one of the greatest mentors I have ever met and I hope to be a person like you in the future.

In addition, I want to thank my dissertation advisor, Dr. Lu Tang, for your priceless help on my dissertation projects in the past two years. You gave me lots of useful suggestions for my second and last projects. I really appreciate those conversations with you and you

truly helped me a lot to formulate solid ideas for my last project. For me, it is an enjoyable experience to work with you since we are in similar ages and I feel like we are friends each time I have talked with you.

I also would like to thank Dr. Ori Rosen, who is one of my committee members with whom I have had weekly meetings over the past two and half years. You have given me lots of useful ideas, suggestions, and comments on my first and second projects. I could not have done those things well without your guidance and feedback.

I also want to extend my thanks to Dr. Andriy Bandos and Dr. Alison Hipwell for joining my committee, giving me permission to use their data, and guiding me in my dissertation. Moreover, I would like to thank my collaborators, Dr. Theodore Huppert and Dr. Kate Keenan, for their help in the analysis and processing of the data. Furthermore, I would like to give my thanks to Dr. Daniel Weeks, Dr. Ryan Minster, Dr. Theodore Huppert, and Dr. Andriy Bandos, for being my GSR advisors and giving me valuable support. I also would like to thank all the friends I have met during my PhD. Your help with research, studies and life means a lot to me.

Lastly, I want to give special thanks to my family members. I am so lucky and proud to be the son of my parents, Ming Fu and Hong Xu, who gave me the largest spiritual and financial support. Thank you for giving me the strength and power to overcome the difficulties I met during my PhD career. Thank you for supporting me and providing me with useful suggestions when I made every decision. I hope you both will be healthy and enjoy your life. I also would like to give thanks to my girlfriend, Linchen He, who gave me tremendous support and warmth since last year. I hope our story will continue and begin our new chapter in the following year.

1.0 Introduction

In this chapter, the background information will be introduced. Section 1.1 gives an overview of functional near-infrared spectroscopy (fNIRS). Section 1.2 discusses related statistical issues in the analysis of fNIRS data. Section 1.3 presents an introduction to the motivating study. Finally, an overview of the dissertation will be introduced in Section 1.4.

1.1 Overview of functional near-infrared spectroscopy (fNIRS)

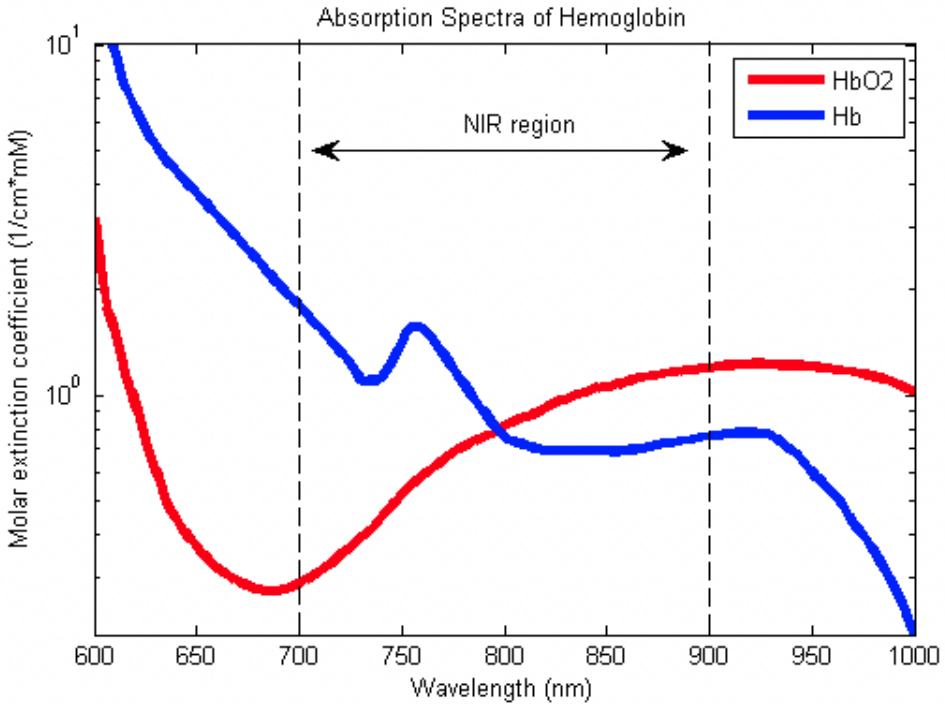
Due to the rapid development of modern neuroimaging techniques, many types of brain-imaging techniques have emerged and been used in a wide range of biomedical areas. Popular brain-imaging techniques include but are not restricted to: Functional Magnetic Resonance Imaging (fMRI), Computed Tomography (CT) scan, Positron Emission Tomography (PET), Magnetoencephalography (MEG), Electroencephalography (EEG) and Functional near-infrared spectroscopy (fNIRS). My dissertation focuses on the analysis of fNIRS brain-imaging data.

1.1.1 Description of fNIRS

Functional near-infrared spectroscopy (fNIRS) is a noninvasive brain imaging technique that measures changes in both oxy- and deoxy-hemoglobin using the near-infrared light (Jobsis, 1977; Villringer et al., 1993). fNIRS is sensitive to hemodynamic changes in localized cerebral blood flow, which is similar to functional magnetic resonance imaging (Aarabi and Huppert, 2016). When near-infrared light goes through the head, it can be either scattered or absorbed by the brain tissue.

Blood carries oxygen and glucose to the brain, which provides the necessary substances for the normal operation of brain functions. Hemoglobin is a specific protein that transports oxygen from the respiratory organs to the rest of the body, including the brain (Wells and

Figure 1. Absorption spectra of oxy- and deoxy-hemoglobin (Reprinted from Wikipedia).



Hung, 1990; Sidell and O'Brien, 2006). When the brain uses up oxygen, the hemoglobin in the red blood cells changes from oxygenated to de-oxygenated. By measuring the amount of oxygenated and de-oxygenated blood, we have a proxy measurement for brain activity. Typically, the oxygenated blood is red while the de-oxygenated blood is blue. Since hemoglobin is a significant absorber of near-infrared light while tissues and bones are transparent to near-infrared light, changes in absorbed light can be used to reliably measure changes in hemoglobin concentration. Figure 1 shows the absorption spectra for oxy- and deoxy-hemoglobin. They have different absorption coefficients in the near-infrared (NIR) region, which allows us to measure the amount of oxy- and deoxy-hemoglobin by colors.

Due to the different absorption coefficients of oxy- and deoxy-hemoglobin (Figure 1), one can measure changes in hemoglobin concentration at different wavelengths. Two wavelengths, one below 810 nm and one over 810 nm are selected to represent the measurement of two hemoglobin, where 810 nm is the wavelength that oxy- and deoxy-hemoglobin have the same absorption coefficients (Cope et al., 1988; Villringer and Chance, 1997; Aarabi

and Huppert, 2016). fNIRS device measures the change in the light intensity, which can be converted to the measurement of changes in relative hemoglobin concentration through the Modified Beer–Lambert law (mBLL) (Wyatt et al., 1986; Kocsis et al., 2006; Baker et al., 2014).

fNIRS has the advantage of low cost and high portability compared to other brain-imaging methods such as fMRI and PET. Since fNIRS is non-confining and noninvasiveness, it is the most appropriate brain-imaging technique for children and infants (Barker et al., 2013). Previously, the fNIRS technique has been widely used to assess and study the relationship between different types of cognitive tasks and cerebral activation (Ferrari and Quaresima, 2012; Boas et al., 2004). The capability to measure brain activity during moderate movement of participants has expanded the use of fNIRS to different motion (Miyai et al., 2001; Suzuki et al., 2008) and balance tasks (Karim et al., 2013b, 2012, 2013a).

1.1.2 Statistical analysis tools for fNIRS data

fNIRS has been used in a variety of study fields over the past several decades, as described in Section 1.1.1. With the development and availability of advanced fNIRS techniques and devices, the need for more sophisticated statistical analysis methods, which aim to address different scientific questions, has increased dramatically.

Currently, many statistical analysis methods for fNIRS data come directly from the analysis of fMRI data, which contain first-level (single scan, subject or trial level) and second-level (group level) analysis (Ashby, 2019). However, direct use of those methods fails to consider the unique features and properties of fNIRS data such as complex motion and physiological noise artifacts (Huppert, 2016; Santosa et al., 2018).

1.1.2.1 HOMER and NIRS-SPM toolbox

Many statistical analysis tools and packages have been developed to uniquely address the analysis of fNIRS data. HOMER is a set of MATLAB scripts used to analyze fNIRS data, with functions to obtain estimates and map brain activation (Aasted et al., 2015; Cui et al., 2010). The current versions are HOMER3 and AtlasViewer, which are MATLAB applications

that provide data analysis for fNIRS including calculation of hemodynamic changes, signal processing, and basic first and second-level statistical analysis (Huppert et al., 2009). Ye et al. 2009 developed a new statistical toolbox called NIRS-SPM where a general linear model was used to build the model at the first level. Pre-whitening (Worsley and Friston, 1995) and pre-coloring methods (Bullmore et al., 1996; Friston et al., 2002) have been implemented to estimate temporal correlations.

1.1.2.2 NIRS AnalyzIR toolbox

Santosa et al. 2018 developed the NIRS Brain AnalyzIR Toolbox. This toolbox is an open-source Matlab-based analysis package, which integrates several tools such as HOMER3, AtlasViewer, and NIRS-SPM including new functions in both signal processing and statistical analysis. This toolbox incorporates multiple functions and is designed for fNIRS data management, pre-processing, statistical analysis, image reconstruction, and region-of-interest (ROI) statistics (Santosa et al., 2018).

Statistical modules in the AnalyzIR toolbox contain different statistical methods for both first-level (single scan, subject, or trial level) and second-level (group level) models. First-level model is usually used to test whether the task condition is significantly different from the reference level for each scan. Different options for statistical methods are offered in the AnalyzIR toolbox, including ordinary least-squares (OLS) used in HOMER2, autoregressive and iteratively reweighted least-squares (AR-IRLS), pre-whitening with AR(1) model in NIRS-SPM and other nonlinear GLM methods. The default first-level method in the toolbox is AR-IRLS, which is able to control type-I errors. More details about the AR-IRLS model can be found in Barker et al. 2013. Second-level models focus on investigating the effect of a pre-specified group (e.g. tasks, disease vs. non-disease group) for a collection of subjects, while possibly adjusting for other covariates. Some widely-used methods for repeated measurements such as mixed models, ANOVA, and fixed-effect models are provided in the toolbox, using coefficients from the first-level as the outcomes. More details about second-level models are given in Santosa et al. 2018.

1.2 Statistical issues in the analysis of fNIRS data

Many statistical issues have arisen with the increased need for the analysis of fNIRS data. Those issues include correlated noise, heterogeneity of data across subjects, and high dimensionality for modeling.

1.2.1 Noise in fNIRS

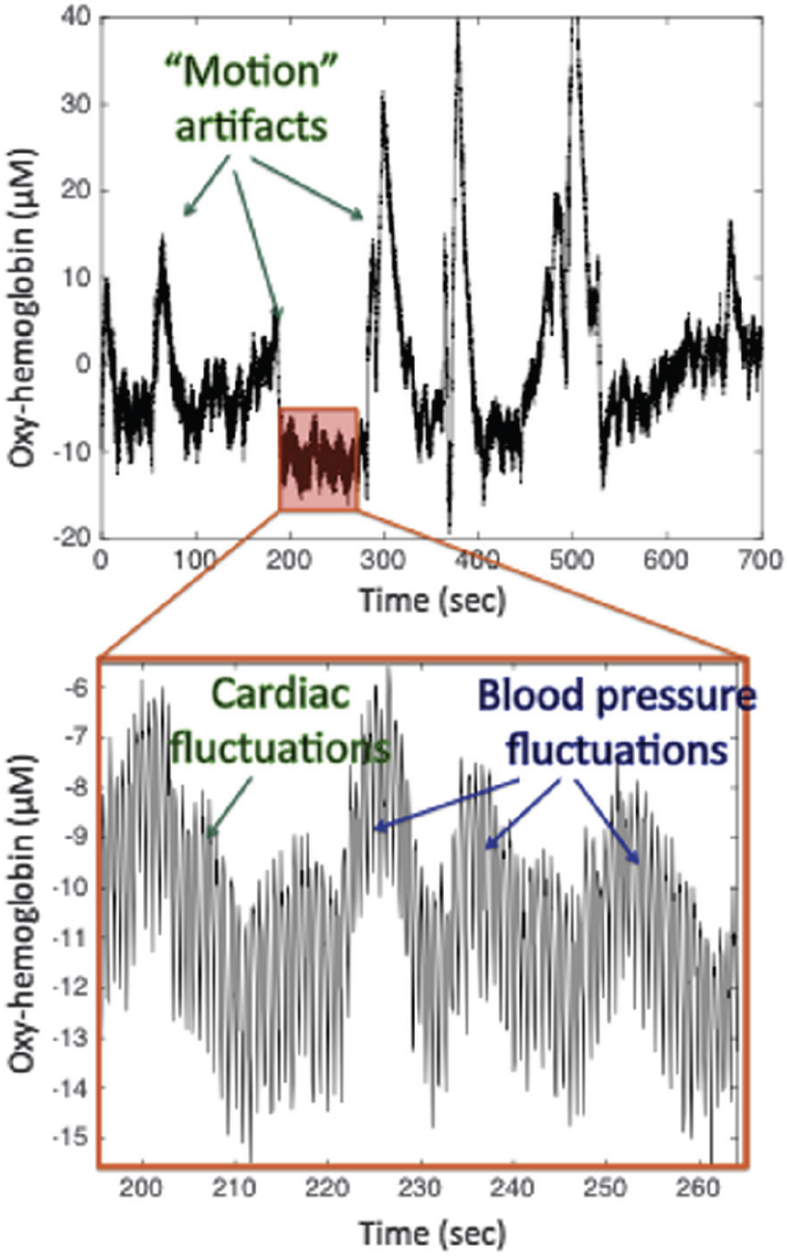
Ordinary least squares (OLS) is the most common way to estimate the regression parameters in linear regression assuming zero mean and common variance for the error term. However, for fNIRS data, those assumptions are violated. Two major issues with respect to noise in fNIRS are the serial correlation of errors caused by systematic physiology and heavy-tailed noise distributions caused by motion artifacts. Figure 2 shows an example of physiological noise and motion artifacts existing in the fNIRS time series.

1.2.1.1 Serially-correlated errors

The noise in fNIRS data is correlated. Due to the existence of strong physiological noise such as cardiac, respiratory, and blood pressure variation, noise in fNIRS exhibits serial correlation within each channel (Naseer and Hong, 2015). Physiological noise is usually slower than the sampling rate of fNIRS data, leading to stronger serial correlation compared to that in fMRI. This physiological noise can be found with the presence of specific frequencies in fNIRS temporal data (Huppert, 2016). Serial correlation in fNIRS noise masks the true signal and results in systematic bias when using traditional statistical models. Thus, preprocessing steps are needed before fitting a model.

Many approaches have been applied to solve the serial correlation issue triggered by physiological noise. Solutions include applying a specific filter to the model, which aims at transforming correlated noise into independent normally distributed random variables (Friston et al., 1994, 1995; Plichta et al., 2007; Jang et al., 2009). The purpose of these approaches is to allow the use of general linear models (GLM) with different noise structures. Common filters include: low-pass filter (Izzetoglu et al., 2005), high-pass filter (Huppert

Figure 2. An example of motion artifacts and physiological noise (Reprinted from Huppert 2016).



et al., 2009), wavelet filter (Chiarelli et al., 2015), principal component analysis (PCA) filter (Li et al., 2017) and independent component analysis (ICA) filter (Robertson et al., 2010; Aarabi and Huppert, 2016). Each filtering approach aims at eliminating or reducing a certain type of noise with particular frequencies or components.

Pre-whitening is another technique used to remove noise from the error term in the model. Pre-whitening in linear regression is given by:

$$\boldsymbol{WY} = \boldsymbol{WX\beta} + \boldsymbol{W\epsilon},$$

where a pre-whitening matrix W multiplies both sides of the equation. By using pre-whitening, data, model and errors are re-weighted such that new error vector $\boldsymbol{W\epsilon}$ meets the assumptions of ordinary least squares. Pre-whitening solves the issue of biased estimation resulting from the filtering, which only applies the filtering to the measurement data. In general, pre-whitening is preferred over filtering because pre-whitening can give unbiased estimators unless we are confident that removed parts are not related to the model part ($\boldsymbol{X\beta}$). Different choices of the pre-whitening matrix can be found in (Ye et al., 2009; Jang et al., 2009; Schroeter et al., 2004; Purdon and Weisskoff, 1998; Barker et al., 2013).

1.2.1.2 Heavy-tailed noise distributions

Another issue in fNIRS noise is its heavy-tailed noise distribution induced by motion artifacts. Motion artifacts arise from the movement of the head since fNIRS sensors are placed on the surface level of participants' heads. The heavy-tailed noise distribution refers to outliers originating from motion artifacts, which are usually stronger than physiological and other types of noise (Huppert, 2016; Cooper et al., 2012). The upper panel of Figure 2 shows several outliers resulting from motion artifacts. The existence of these outliers violates the common variance assumption of OLS, indicating that the variance of the noise is heteroscedastic. Since motion artifacts are not part of the model, including them will result in estimates with non-ignorable bias. Different statistical methods have been developed to eliminate or reduce motion artifacts.

Robust regression is widely used to solve the issue of heavy-tailed outliers by iteratively down-weighting each outlier (Ruppert and Wand, 1994; Holland and Welsch, 1977). Barker et al. 2013 developed the autoregressive and iteratively reweighted least-squares (AR-IRLS) method to solve the issue of motion artifacts, which is the default first-level statistical method used in the AnalyzIR toolbox of Section 1.1.2.2. This method first uses an auto-regression

filter to pre-whiten both sides of the model. Robust regression, aimed at down-weighting outliers iteratively, is then applied to the pre-whitened data.

1.2.2 Heterogeneity across subjects

Section 1.2.1 introduces serially correlated noise which violates the assumptions of linear regression. These methods described in Section 1.2.1 transform heteroscedastic noise to follow the normal distribution by applying a filter or a pre-whitening method. Coefficients are estimated using the proposed methods, reflecting the brain activation level as opposed to the reference or the baseline period. However, these methods focus on obtaining robust estimates of coefficients, without looking at the estimation of time series trajectories.

The fNIRS time series itself is important since we can observe the level of brain activation throughout multiple periods with different tasks. However, due to the heterogeneity across different subjects, different patterns of brain activation levels could be observed for subjects who belong to different unobserved groups. In contrast, subjects who are in the same unobserved group may have similar cerebral responses with regard to a certain stimulus event. Thus, there is a need to develop reliable statistical methods to cluster subjects and discover underlying time series patterns related to certain stimulus events.

1.2.3 High-dimensionality

High dimensionality is a common issue for any types of functional imaging data. fNIRS data are usually observed at a large number of time points, where the number of time points is largely greater than the number of subjects. Functional regression (Ramsay and Silverman, 2013) is a statistical tool for the modeling of functional data, where either the outcome or predictors, or both the outcome and predictors, can be functional. However, functional data analysis gives complicated coefficient curve estimates, where sometimes is lack of a good interpretation. Hence, for the modeling using fNIRS data, it is desirable to assume sparsity and find an approach to produce interpretable trends for large coefficients and shrink other coefficients towards zero.

1.3 Motivating study

Due to portability and mobility compared to fMRI, fNIRS has a wide range of use, from the population of infants to the elderly, as well as the area of movement (e.g. walking and gait speed) and cognitive tasks with various study designs. Our motivating study is called face-to-face still-face (FFSF) study, which aims to understand patterns of an infant's brain activity before, during and after an emotionally stressful probe (Tronick et al., 1978).

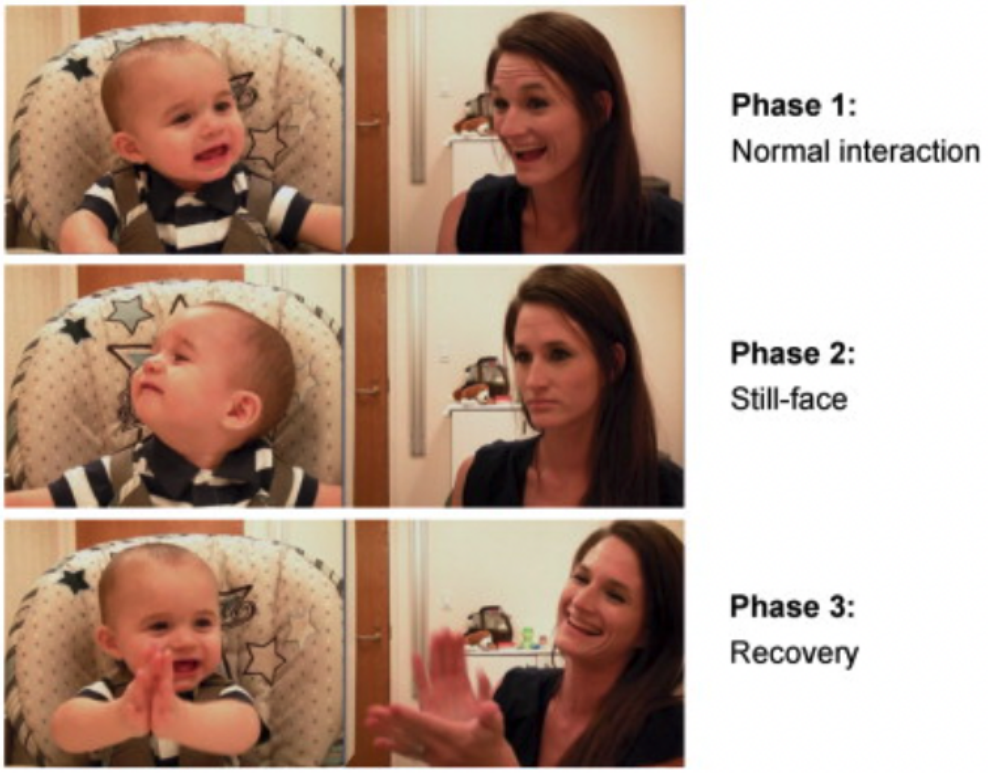
1.3.1 Overview of the Still-face study

Face-to-face interactions between mothers and infants are essential to the development of infants with respect to communication and social skills, as well as the regulation of emotion and temperament (Northrup et al., 2019; Hipwell et al., 2019; Ainsworth et al., 2015; Feldman et al., 1999). To be more specific, Ainsworth and Bell 1972 and Bigelow et al. 2010 demonstrated that positive interactions between mothers and infants are associated with infants' language and cognitive development. In contrast, negative interactions between parents and infants are related to the changes in behavioral and emotional problems (Belsky et al., 1998; Edwards and Hans, 2015; Levendosky et al., 2006).

The FFSF paradigm is a widely used stress task (a violation of the expectation of social interaction) that allows for biobehavioral measurement of individual differences in infant response and recovery (Tronick et al., 1978; Northrup et al., 2019). The still-face paradigm comprises of three phases: interact or baseline, still-face and recovery (Adamson and Frick, 2003). A picture of three phases of the still-face study is shown in Figure 3. In phase 1, mothers perform normal interactions with infants without the use of toys; this phase serves as the baseline. In phase 2, mothers adopt a neutral facial expression (still-face with no facial or oral communication) to infants, followed by phase 3, where mothers resume normal interactions with their infants (Mesman et al., 2009; Sravish et al., 2013). Prior to the start of the FFSF, an fNIRS cap is fitted on the infant's head to measure the level of and change in brain activation across the three phases.

Many studies have revealed that the increase in negative affect, along with a decrease

Figure 3. Three-phase still-face paradigm (Reprinted from Kim et al. 2014).



in positive affect in infants has been found as a result of their mothers' still-face. These responses reflect infants' ability to communicate with their mothers intentionally after sensing a negative emotion or expression (Mesman et al., 2009; Carter et al., 1990; Tarabulsy et al., 2003). However, although many studies have discovered common responses from infants in different designs of still-face studies, there still exists a large amount of heterogeneity at the individual level (Mesman et al., 2013). There are wide individual differences in response, which may be related to the quality of the mother-infant relationship as well as temperamental features. Due to its portability and non-confining property, fNIRS is a good brain-imaging tool to investigate the dynamic level of brain activation during a still-face study. However, to the best of our knowledge, there are few studies that have focused on applying fNIRS to still-face studies. Thus, it is critical to apply fNIRS technology to still-face studies, while taking subject heterogeneity into consideration.

1.3.2 PGS-ECHO fNIRS still-face study

Participant mothers in the motivating study were recruited from the longitudinal Pittsburgh Girls Study (PGS), a population-based study of 2,450 girls who were recruited in the city of Pittsburgh between the ages of 5 and 8 (Keenan et al., 2010). In 2016, a large-scale sub-study of the PGS was initiated to investigate how environmental factors, such as psychological stressors experienced during childhood and adolescence, affect later maternal pregnancy and child health. The study is part of the National Institutes of Health Environmental Influences of Child Health Outcomes (ECHO) program, which examines different impacts of prenatal environmental exposures across biological, chemical, physical, and social domains on offspring health and development (Gillman and Blaisdell, 2018).

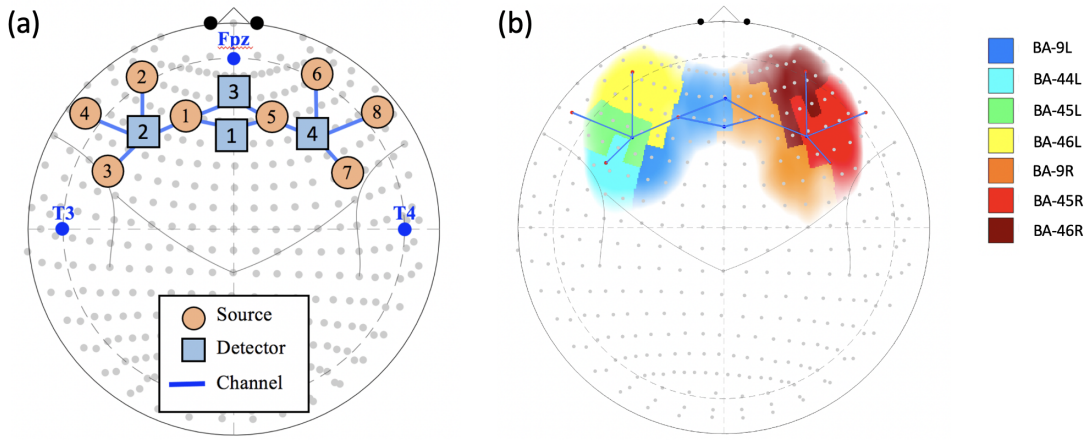
The PGS-ECHO study enrolls PGS participants as they become pregnant or recently deliver a live birth. Participants complete multiple prenatal lab visits and the children are followed from ages 6 to 36 months. The lab protocol includes interviews and interaction tasks to assess contextual stressors, health, mood, lifestyle behaviors, and offspring behavioral and emotional development.

In the current study, we measured infant brain activity using the above fNIRS probe (roughly 120 seconds of measurements for each phase). At the end of 2021, recorded fNIRS still-face data had been collected from 155 infant subjects. Demographic variables of infants and mothers such as gestational age, infant age, sex, birth weight, head circumference, along with parent reports on the Infant Behavior Questionnaire-Revised (IBQ-R) (Gartstein and Rothbart, 2003) were also collected. Until the end of 2021, we have collected fNIRS data from 155 subjects, with still-face data that pass quality control. Some demographic variables of infants and mothers such as gestational age, infant age, sex, birth weight, head circumference, along with several parent-reported scores from the Infant Behavior Questionnaire (IBQ) are also provided.

1.3.3 fNIRS probe configuration

PGS-ECHO fNIRS still-face data are recorded using a continuous NIRS imaging system (NIRScout; NIRx Medical Technologies, Berlin, Germany) at the sampling rate of 7.8125

Figure 4. fNIRS probe configuration. (a) Positioning of 8 sources, 4 detectors and 12 channels. (b) Brodmann areas covered by the fNIRS probe.



Hz and NIRStart acquisition software. The data are measured simultaneously at two wavelengths (760 nm and 850 nm). As shown in Figure 4(a), this fNIRS probe comprises of 12 channels from 8 sources and 4 detectors. Brodmann area is a region of the cerebral cortex in the human brain defined by its cytoarchitecture or histological structure and organization of cells (Brodmann, 1909; von Economo and Koskinas, 1925; Garey, 1999). Figure 4(b) shows that a total of seven Brodmann areas are covered by the fNIRS probe.

1.3.4 fNIRS data pre-processing

Before performing any desired statistical analysis, data need to be pre-processed including rescaling of data, dealing with subjects with incomplete data and outliers.

1.3.4.1 Sample size

Until the end of 2021, a total of 155 infants have participated in the PGS-ECHO fNIRS still-face study resulting in recorded still-face data that have passed the manual quality control check. However, some experiments were terminated due to extremely negative responses (e.g. crying, excessive movements, etc.). In some experiments, phases were too short or in-

complete, or severe outliers were present. Thus, by removing infants who did not complete three phases of the still-face paradigm, who had strong outliers based on leverage and who had very short period measurements of any of the three still-face phases, we had a total of 82 subjects with complete fNIRS still-face data available for future analysis. The above data processing was performed by NIRS brain AnalyzIR toolbox in MATLAB (Santosa et al., 2018).

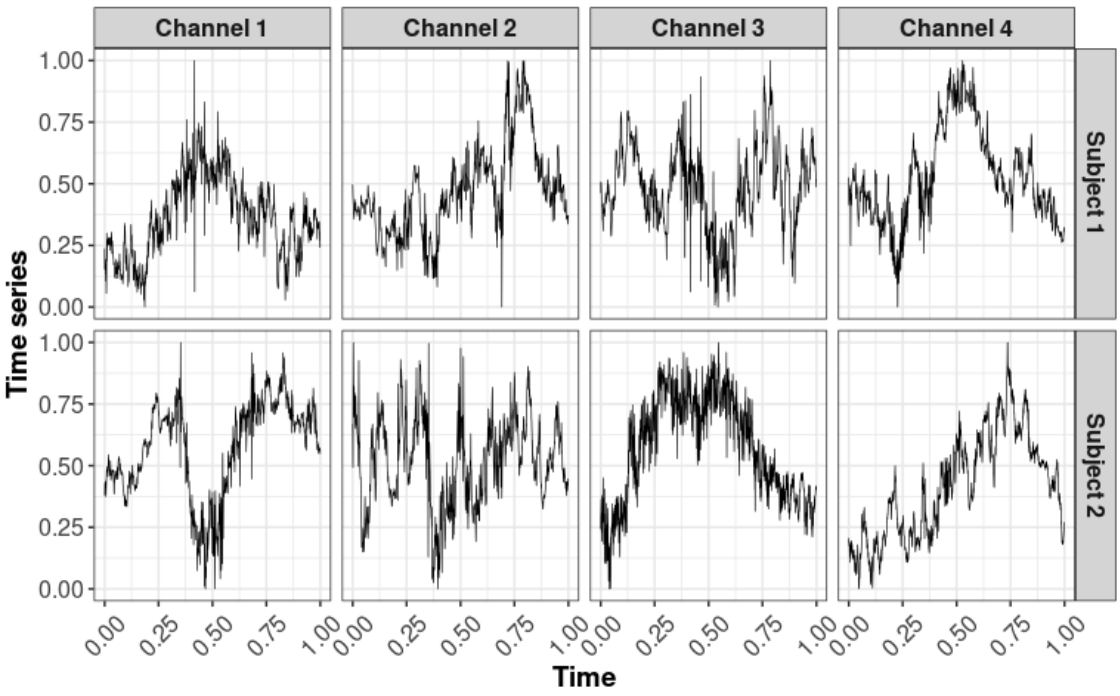
1.3.4.2 Outliers

For the data pre-processing, the modified Beer-Lambert law (Cope et al., 1988; Kocsis et al., 2006) was used to convert the measured light intensity to the relative concentration of oxy-hemoglobin (HbO_2) and deoxy-hemoglobin (Hb). To solve the issue of heavy-tailed noise from motion artifacts as detailed in Section 1.2.1.2, we implemented a motion correction method for fNIRS called Temporal Derivative Distribution Repair (TDDR) (Fishburn et al., 2019). TDDR is a novel motion correction method that implements robust regression to remove shifts from artifacts without using any user-supplied parameters. Before applying TDDR, we first used a PCA filter to remove systematic physiological noise, which was relatively narrow-band and quasi-stationary. This approach computes the temporal derivative of the measured signal, and then iteratively estimates robust observation weights, followed by applying the robust weights to the centered temporal derivative to produce a corrected derivative. To avoid the issue of variance inflation induced by high-frequency components, which affects the estimation of temporal derivative in TDDR, this approach splits the data into low and high-frequency parts with a low-pass filter. TDDR is applied to the low-frequency part and then added back to the high-frequency part (Fishburn et al., 2019). The TDDR motion correction method can also be done using the NIRS brain AnalyzIR toolbox in MATLAB (Santosa et al., 2018).

1.3.4.3 Data rescaling

After data pre-processing in MATLAB, we passed our processed data into R software for further processing. First, for each processed fNIRS time series, we took another step for

Figure 5. An example of processed fNIRS time series from two subjects and four channels.



dealing with outliers by removing any measurement points which were out of three standard deviations of the mean. By the end of this step, we removed outliers that were not removed by the TDDR in Section 1.3.4.2. Hereafter, interpolating splines were used to insert points (Catmull and Rom, 1974; Kochanek and Bartels, 1984). In general, interpolating splines is a widely-used type of interpolation which is preferred over polynomial interpolation because the interpolation error is small even when using low-degree polynomials for the splines (Hall and Meyer, 1976). By using interpolation, we are able to obtain processed fNIRS time series of equal lengths which are evenly spaced in each phase (interact, still-face and recovery), each channel, and each subject. Finally, to avoid the effect of extremely large or small mean values across different fNIRS time series, we rescaled time series measurements and sampling time to be between 0 and 1.

In conclusion, our final processed fNIRS data included a total of 82 subjects, each with processed time series of both oxy and deoxy-hemoglobin from twelve channels. For each fNIRS time series, we had a total of 1500 measurement points and each phase consisted of

500 points. All measurements and sampling times were between 0 and 1 after the rescaling step. Figure 5 shows an example of processed time series from two subjects and four selected channels. We will only focus on oxy-hemoglobin (HbO or HbO_2) in the future since we find that deoxy-hemoglobin is usually more erratic with high variability than oxy-hemoglobin.

1.4 Overview of the dissertation

My dissertation has four chapters. Chapter 1 provides an introduction to fNIRS brain-imaging data and still-face study, which serves as background information and motivation for Chapters 2-4.

In Chapter 2, I propose a group-based approach to clustering and estimating trajectories of univariate time series via a Bayesian mixture of smoothing splines. The approach assumes each time series is a mixture of multiple dynamic components using a spline model with different mixing weights for each component. Time-independent baseline covariates are assumed to be associated with the mixture components and are incorporated via the mixing weights using the mixture of experts model. This approach is formulated in a fully Bayesian framework using reversible jump Markov chain Monte Carlo (RJMCMC) and a data-based relabeling algorithm is adopted to solve the label switching issue. The superior performance of the approach in terms of subgroup detection and estimation is demonstrated through both simulation studies and applications to the analysis of fNIRS data.

In Chapter 3, I extend my proposed approach in Chapter 2 to multivariate time series. The study consists of multivariate time series observed on a collection of individuals, each with a multi-dimensional time series. Gibbs sampling is used as the sampling algorithm and the number of components is selected based on the deviance information criterion (DIC). The Equivalence Classes Representatives (ECR) algorithm is adopted to solve the label switching issue. The superior performance of the approach in terms of subgroup detection and estimation is demonstrated by simulation studies and applications to the analysis fNIRS data.

In Chapter 4, I propose a horseshoe prior-based generalized lasso for interpretable scalar

on function regression. The approach is able to penalize regression coefficients with selected orders of differences by specifying appropriate prior structures. The horseshoe prior is able to control both the global and local shrinkage levels of each coefficient simultaneously. The proposed method is demonstrated to have superior performance in terms of signal detection and prediction accuracy through simulation studies, and is applied to the analysis of fNIRS data.

2.0 Covariate-Guided Bayesian Mixture of Spline Experts for the Analysis of Univariate Time Series

2.1 Introduction

Time series are realizations of random processes, and obtaining estimated time series trajectories may provide insights into many practical problems. In many brain-imaging techniques, such as fNIRS introduced in Section 1.1, processed data are always nonstationary time series with non-constant mean and high variability across time, which poses many statistical challenges in inference and estimation. In addition, as in the case of fNIRS, it is critical to find an appropriate method to analyze a collection of time series that are from different subjects. However, time series are usually heterogeneous with high variability and different patterns across subjects. Thus, a methodological approach is proposed to deal with issues of nonstationary time series and heterogeneity across subjects.

Time series clustering is used to tackle heterogeneity across subjects by clustering similar time series together using various models or distance-based algorithms. Clustering of time series has been used in diverse scientific areas with the purpose of discovering patterns or trajectories, which leads to uncovering valuable information from complex and massive datasets (Liao, 2005; Aghabozorgi et al., 2015). The process of time series clustering partitions the whole collection of data into different groups such that homogeneous time series are grouped together based on a certain similarity measure. Challenges in time-series clustering include computational issues due to its large datasets, high-dimensionality with slow processing time (Lin et al., 2003; Keogh and Pazzani, 2000; Zhang et al., 2006) and identifying proper similarity measures (Wang et al., 2004; Lin et al., 2004).

In the review paper of Aghabozorgi et al. 2015, most of the methods for time series clustering can be classified into three categories: whole time series clustering, sub-sequence clustering, and time point clustering. Whole time-series clustering is the clustering of a set of univariate time series (Keogh and Lin, 2005). It is often used to cluster different subjects, where each subject has an individual time series. Sub-sequence clustering refers to

clustering on a set of sub-sequences of a time series, or clustering of segments from a single long time series. This approach is often used to estimate a single nonstationary time series by dividing it into locally stationary segments (Keogh and Lin, 2005; Adak, 1998). Time point clustering is used to cluster time points based on a combination of their temporal proximity and the similarity of time point values (Gionis and Mannila, 2003; Ultsch and Mörchen, 2005; Mörchen et al., 2005). Our proposed method focuses on whole time series clustering.

In general, there are three categories for whole time series clustering, namely model-based, feature-based and shape-based. Details on feature-based and shape-based clustering can be found in Keogh and Lin 2005. For model-based clustering approaches, certain parametric or nonparametric models are established to obtain estimates of model parameters, then an appropriate clustering algorithm is chosen to assign individual time series into different groups (Liao, 2005; Vlachos et al., 2004; Mitsa, 2010).

Typical modeling approaches for time series data consist of polynomial models (Bagnall and Janacek, 2005), linear mixed models (Biernacki et al., 2000), autoregressive–moving-average (ARMA) (Corduas and Piccolo, 2008), Markov chain (Ramoni et al., 2000) and Hidden Markov models (Bicego et al., 2003; Hu et al., 2006). Lots of work have been done in modeling nonstationary time series by assuming local stationarity in each segment of the time series. Kitagawa and Akaike 1978 suggested that a nonstationary time series can be partitioned into small segments and each segment was assumed to be stationary. Wood et al. 2011 proposed a Bayesian mixture of autoregressive (AR) models for the analysis of possibly nonstationary time series. Mixing weights for each component were computed as a function of time, and a common but unknown lag was considered for the AR components. However, this approach does not address the issue of the estimation of structural breaks. Davis et al. 2006 considered a piecewise AR model for nonstationary time series, where the number of segments, the locations of structural breaks, and the orders of AR models were all unknown. An objective function was obtained and optimized to find the best combination of the number of segments, locations of breaks, and orders of AR models. Ombao et al. 2001 implemented a nonparametric model for the estimation of piecewise stationary time series by using a smooth localized complex exponential transform, while Lau and So 2008 proposed

an infinite Bayesian mixture of AR models with a Dirichlet process prior. Furthermore, another approach for fitting time series models is to allow model parameters to evolve over time. State-space models with a smoothness prior are often used for dynamical estimators by implementing a random walk for AR processes (Kitagawa and Gersch, 1996; West et al., 1999). Prado and Huerta 2002 assumed that the order of AR model also changed over time by fitting a discrete random walk, and Gerlach et al. 2000 proposed an efficient Bayesian MCMC approach with rapid convergence of the posterior distributions for estimating a mixture of state-space models.

Instead of the time domain, much work has been focused on the frequency domain. Approaches to the spectral estimation of a single nonstationary time series include estimates of evolutionary parameters over time and model-based time-frequency analysis using time-varying AR models (Kitagawa and Gersch, 1996; Dahlhaus, 1997; Yang et al., 2016), fitting smoothing spline models for spectrum (Ombao et al., 2001; Guo et al., 2003) and piecewise AR models (Adak, 1998; Davis et al., 2006). Tuft et al. 2021 proposed an approach to time–frequency analysis that decomposes the power spectrum into orthogonal layers and provides a parsimonious representation of the time-varying power spectrum. Notably, mixture of spline models are widely used to perform spectral analysis. Wood et al. 2002 presented a Bayesian method for spatially adaptive nonparametric regression where regression functions were modeled using a mixture of splines. Rosen et al. 2009 proposed a Bayesian mixture of smoothing splines with time-varying mixing weights to estimate the evolution of the log spectrum. Later, Rosen et al. 2012 extended previous work and proposed an adaptive spectral estimation for a single nonstationary time series by adaptively dividing time series into an unknown number of segments using reversible jump MCMC (Green, 1995; Richardson and Green, 1997).

For the spectral estimation of multiple time series, many papers use a covariate-dependent model to associate mixture components with time-independent covariates. Bertolacci et al. 2021 extended the work of Rosen et al. 2012 by using a covariate-dependent infinite mixture model employing the logistic stick-breaking process (Rigon and Durante, 2021). Wang et al. 2021 considered a sparsity-inducing Dirichlet hyperprior for high-dimensional covariates in a tree-based model, which provided sparsity in covariate estimation and variable selection.

Krafty et al. 2011 introduced a mixed effect spectral model with the consideration of covariates on the second order of power spectrum and in 2017 Krafty et al. proposed a method to connect power spectra to clinical outcomes by using a tensor-product spline model of outcome-dependent power spectra. Bruce et al. 2018 introduced a method for investigating the association between the time-varying power spectrum and covariates by adaptively partitioning grids of time and covariate into different blocks by penalized spline. Cadonna et al. 2019 developed an approach to estimate spectral densities by adopting a mixture of Gaussian models with frequency-dependent mixing weights with frequency-dependent parameters.

The mixture-of-experts model (Jacobs et al., 1991; Jordan and Jacobs, 1994) use a multinomial logistic model as weights of the components, which are referred as experts. This model has a direct application to clustering, with mixing weights depending on external covariates. Waterhouse et al. 1995 established a Bayesian framework to estimate the parameters in the mixture-of-experts model, which avoided the overfitting issue. Zens 2019 proposed a variable selection method based on the Bayesian mixture-of-experts framework using Gibbs sampling. Mixture-of-experts have also been used in time series modeling. Huerta et al. 2003 addressed the issue of time series model mixing and allowed the incorporation of covariates for model comparisons using the hierarchical mixture-of-experts. In addition, Frühwirth-Schnatter et al. 2012 and Fröhwirth-Schnatter and Kaufmann 2008 applied the mixture-of-expert framework to model-based clustering under a fully Bayesian framework. Tang and Qu 2016 proposed an unbiased estimating equation approach for a longitudinal mixture model with correlated responses, where the mixture-of-experts model was used to model the mixing weights. Comprehensive overviews of the mixture-of-experts model can be found in Masoudnia and Ebrahimpour 2014 and Yuksel et al. 2012.

In our first project, we propose a mixture of spline experts for model-based clustering of multiple univariate time series. Smoothing splines are used to fit the time series with flexibility and a low-rank approximation approach (Wood et al., 2002; Wahba, 1980) was adopted to obtain smoothing coefficients based on a small set of basis functions. The mixture-of-experts model is incorporated into the proposed model to allow for the inclusion of time-independent covariates. The Pólya-Gamma data augmentation strategy (Polson et al., 2013) is used to simplify the sampling of the logistic parameters in the proposed model.

The proposed approach is formulated in a fully Bayesian framework and sampling from the posterior distributions is done via a reversible jump Markov chain Monte Carlo (RJMCMC) (Richardson and Green, 1997; Green, 1995; Rosen et al., 2012). The rest of Chapter 2 is organized as follows. In Sections 2.2 and 2.3 we present the proposed model and priors for each parameter. Section 2.4 introduces the Bayesian sampling scheme of the proposed RJMCMC algorithm. In Section 2.5 we report simulation results under different settings and Section 2.6 illustrates our proposed method using the data from the PGS-ECHO fNIRS still-face study introduced in Chapter 1. Section 2.7 gives a conclusion of our first project as well as limitations and future works.

2.2 Covariate-guided Bayesian mixture of spline experts model

In this section, I will introduce our proposed covariate-guided Bayesian mixture of spline experts model. Smoothing spline priors are placed on the time series trajectories, and the mixture of experts framework is adopted where the mixing weights for each mixture component depend on time-independent covariates.

2.2.1 Mixture of splines model

Smoothing splines is a nonparameteric modeling approach to estimate unknown functions which allows the regularization of smoothness by a smoothing parameter (Green and Silverman, 1993; Hastie and Tibshirani, 2017). The smoothing parameter controls the trade-off between the model goodness of fit and the roughness of the underlying function. In the Bayesian setting, smoothing splines are incorporated into the model by placing a prior on the underlying functions of interest (Kimeldorf and Wahba, 1970; Wahba, 1978).

We propose a Bayesian mixture of smoothing splines model for model-based clustering of univariate time series from multiple subjects. For each subject $i, i = 1, \ldots, N$, $\boldsymbol{y}_i$ is a univariate time series. Under the mixture model framework, we assume that subject i belongs to a latent component g, $g = 1, \ldots, G$. To simplify the computation, we introduce

latent indicators z_{ig}, such that $z_{ig} = 1$ if the ith time series belongs to the gth component and $z_{ig} = 0$ otherwise. We also denote $t_j = t_1, \ldots, t_n$ as the time of the jth time point. Let $\boldsymbol{y}_i = \left[y_i(t_1), \ldots, y_i(t_j), \ldots, y_i(t_n)\right]'$ be the univariate time series of length n for the ith subject The model for time series $\boldsymbol{y}_i$, conditional on component g, can be written as

$$\{\boldsymbol{y}_i \mid z_{ig} = 1\} = \boldsymbol{X}\boldsymbol{\alpha}_g + \boldsymbol{W}\boldsymbol{\beta}_g + \boldsymbol{\epsilon}_i, \tag{1}$$

where $\boldsymbol{X} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ t_1 & t_2 & \ldots & t_n \end{pmatrix}'$ and $\boldsymbol{W}$ is a $n \times m$ matrix of smoothing spline basis functions. The $n \times 1$ vector $\boldsymbol{X}\boldsymbol{\alpha}_g$ represents the linear part of the time series and $\boldsymbol{W}\boldsymbol{\beta}_g$ is the nonlinear part. The 2×1 vector $\boldsymbol{\alpha}_g$ contains the intercept and slope for gth component. The $m \times 1$ vector $\boldsymbol{\beta}_g$ contains the coefficients of the basis functions and m is the number of basis functions. The $n \times 1$ vector of errors $\boldsymbol{\epsilon}_i$ is assumed to have a $N(\boldsymbol{0}, \sigma_g^2 \boldsymbol{I}_n)$ distribution which means that the errors are independent and have a common variance over time. The parameter σ_g^2 is the error variance for the gth component.

Based on the time series pre-processing in Section 1.3.4, we will assume that each univariate time series are observed at the same n time points. Thus, common design matrix $\boldsymbol{X}$ and $\boldsymbol{W}$ across subjects and components are used in model (1).

Section 2.3 provides a detailed explanation of the smoothing spline model and the low-rank approximation in the Bayesian setting. To simplify the notation, we define $\boldsymbol{S} = [\boldsymbol{X}\ \boldsymbol{W}]$ and $\boldsymbol{\theta}_g = (\boldsymbol{\alpha}_g', \boldsymbol{\beta}_g')'$. Model (1) can thus be rewritten as:

$$\{\boldsymbol{y}_i \mid z_{ig} = 1\} = \boldsymbol{S}\boldsymbol{\theta}_g + \boldsymbol{\epsilon}_i. \tag{2}$$

2.2.2 Model for mixing weights

As introduced in Section 2.1, the mixture-of-experts approach (Jacobs et al., 1991) is applied to formulate a covariate-guided model, where selected covariates are used to predict mixing weights via the so-called gating functions. As in Rosen et al. 2009, the mixing weights are expressed using the multinomial logits model so that

$$\pi_{ig}(\boldsymbol{V}_i) = \frac{\exp(\boldsymbol{V}_i^T \boldsymbol{\delta}_g + \zeta_{ig})}{\sum_{h=1}^{G} \exp(\boldsymbol{V}_i^T \boldsymbol{\delta}_h + \zeta_{ih})}, \tag{3}$$

where $\boldsymbol{V}_i = (1, V_{i1}, \ldots, V_{iP})'$, $\boldsymbol{\delta}_g = (\delta_{g0}, \delta_{g1}, \ldots, \delta_{gP})'$, and we let $\boldsymbol{\delta}_g = \mathbf{0}$ for identifiability. The vector $\boldsymbol{V}_i$ contains the covariate values for subject i, P is the number of covariates and $\boldsymbol{\delta}_g$ is a vector of logistic parameters. To enhance the model performance and inference of the mixing weights, a random term ζ_{ig} for each subject and component is added to the mixing weights.

2.2.3 Likelihood

For brevity, we denote $\boldsymbol{\Theta}_g$ as the aggregation of all parameters of interest for component g, and $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1', \ldots, \boldsymbol{\Theta}_G')'$ are the parameters of interest for all components. Let $f_g(\boldsymbol{y}_i \mid \boldsymbol{\Theta}_g)$ be the probability density function of the gth component for time series $\boldsymbol{y}_i$. Thus, the contribution to the likelihood function from the ith subject can be written as:

$$L(\boldsymbol{\Theta} \mid \boldsymbol{y}_i) = \sum_{g=1}^{G} \pi_{ig} f_g(\boldsymbol{y}_i \mid \boldsymbol{\Theta}_g). \tag{4}$$

To simplify the computation, we use latent indicators z_{ig} to augment the likelihood function. The posterior weight of z_{ig} can be expressed as

$$p(z_{ig} = 1 \mid \boldsymbol{y}_i) = \frac{\pi_{ig} f_g(\boldsymbol{y}_i \mid \boldsymbol{\Theta}_g)}{\sum_{h=1}^{G} \pi_{ih} f_h(\boldsymbol{y}_i \mid \boldsymbol{\Theta}_h)} \tag{5}$$

Denoting $\boldsymbol{y}$ as univariate time series for all subjects. Thus, the augmented likelihood function with the latent indicator for all subjects is

$$L(\boldsymbol{\Theta}, z_{ig} \mid \boldsymbol{y}) = \prod_{i=1}^{N} \prod_{g=1}^{G} \left[\pi_{ig}(\boldsymbol{V}_i) f_g(\boldsymbol{y}_i \mid \boldsymbol{\Theta}_g)\right]^{z_{ig}} \tag{6}$$

2.3 Priors and joint posterior distribution

In this section, I will introduce the priors used for each parameter in the proposed covariate-guided mixture of spline experts model.

2.3.1 Smoothing splines priors

The conditional expectation of a mixture component in model (4) is given by $E(\boldsymbol{y}_i \mid z_{ig} = 1) = \boldsymbol{X}\boldsymbol{\alpha}_g + \boldsymbol{W}\boldsymbol{\beta}_g$. We place a smoothing spline prior on $\boldsymbol{\beta}_g$ and let $\boldsymbol{\mathcal{H}}_g = \boldsymbol{W}\boldsymbol{\beta}_g$, where $\boldsymbol{\mathcal{H}}_g = \big[\mathcal{H}_g(t_1), \ldots, \mathcal{H}_g(t_n)\big]'$ is a zero-mean Gaussian process with variance covariance matrix $\tau_g^2\boldsymbol{\Phi}$ (Wahba, 1980; Wood et al., 2002), such that $\text{cov}\big[\mathcal{H}_g(t_r), \mathcal{H}_g(t_h)\big] = \tau_g^2\phi_{rh}$, τ_g^2 is a smoothing parameter for component, and the (r,h)th element of $\boldsymbol{\Phi}$ is given by $\phi_{rh} = \frac{1}{2}t_r^2(t_h - \frac{t_r}{3})$ for $t_r \leq t_h$. The matrix $\boldsymbol{\Phi}$ is common to all subjects since time series are observed at common time points.

As seen above, the matrix $\boldsymbol{\Phi}$ is $n \times n$, and to avoid the computational burden for large n, a low-rank approximation is often adopted. To facilitate this approximation, we obtain basis functions via the spectral decomposition of $\boldsymbol{\Phi}$, as has been proposed in Wood et al. (2002) and used in Rosen et al. (2009, 2012); Krafty et al. (2011). In particular, the matrix $\boldsymbol{W}$ consists of m basis functions evaluated at times $t_1, \ldots, t_n$, and $\boldsymbol{\beta}_g$ is an m-dimensional vector of basis function coefficients. These basis functions are obtained by applying the spectral decomposition to $\boldsymbol{\Phi}$ such that $\boldsymbol{\Phi} = \boldsymbol{Q}\boldsymbol{\Gamma}\boldsymbol{Q}^T$, where $\boldsymbol{Q}$ is the matrix of eigenvectors of $\boldsymbol{\Phi}$, and $\boldsymbol{\Gamma}$ is a diagonal matrix containing the eigenvalues of $\boldsymbol{\Phi}$. We then let the design matrix $\boldsymbol{W} = \boldsymbol{Q}\boldsymbol{\Gamma}^{1/2}$ and place a normal prior $N(0, \tau_g^2\boldsymbol{I}_n)$ on $\boldsymbol{\beta}_g$, which leads to $\boldsymbol{\mathcal{H}}_g$ or $\boldsymbol{W}\boldsymbol{\beta}_g \sim N(\boldsymbol{0}, \tau_g^2\boldsymbol{\Phi})$ as mentioned above.

By using the low-rank approximation, the number of columns of $\boldsymbol{W}$ is reduced from n to m ($m < n$), which greatly reduces the computational burden without sacrificing the model fit (Wahba, 1980; Wood, 2006). Eubank (1999) indicated that the eigenvalues in the diagonal matrix $\boldsymbol{\Gamma}$ decay rapidly as m increases. Thus, we can achieve a good approximation by selecting a relatively small number m of basis functions. We use $m = 20$ basis functions in the real-data application, which are able to explain more than 98% of the total variance

based on an empirical finding in Krafty et al. 2011.

As in (9), where linear and spline parts are put together, we thus assume the Gaussian prior $\boldsymbol{\theta}_g \sim N(\mathbf{0}, \boldsymbol{D}_g)$, where $\boldsymbol{D}_g = \text{diag}(\sigma_\alpha^2 \mathbf{1}_2,\ \tau_{\text{g}}^2 \mathbf{1}_{\text{m}})$ is the covariance matrix for $\boldsymbol{\theta}_g$. The parameter σ_α^2 is the hyperprior variance for $\boldsymbol{\alpha}_g$, and we fix $\sigma_\alpha^2 = 100$ reflecting a noninformative prior across all components. The parameter τ_g^2 is the smoothing parameter for the gth component, and $\mathbf{1}_m$ is the m-vector of ones.

2.3.2 Priors on the smoothing parameters

We assume the smoothing parameters τ_g^2 vary across components g. Although the most common choice for the prior on a variance parameter is the inverse gamma distribution, Gelman (2006) and Wand et al. (2011) suggested that a half-t prior on the standard deviation can reflect a lack of information on a scale parameter. The half-t is a family of heavy-tailed distributions and has a good shrinkage performance. It can be expressed as a scale mixture of inverse gamma random variables using a latent variable that follows an inverse gamma distribution (Wand et al., 2011). Thus, we assume a half-t distribution such that $\tau_g \sim t_{\nu_\tau}^+(0, A_\tau)$, where ν_τ is a degrees of freedom parameter, and A_τ is a scale parameter. We set $\nu_\tau = 3$ and $A_\tau = 10$ for all components and entries.

2.3.3 Priors on the error variances

Similar to the priors on the smoothing parameters in section 2.3.2, we assume that the error variance σ_g^2 varies across components and follows a half-t distribution, such that $\sigma_g \sim t_{\nu_\sigma}^+(0, A_\sigma)$, where ν_σ is the degree of freedom and A_σ is a scale parameter. We set $\nu_\sigma = 3$ and $A_\sigma = 10$ for all components.

2.3.4 Priors on the logistic parameters and the variances of random intercepts

This section provides details on the prior distributions placed on the parameters of the logistic weights (3). For ease of notation, we denote $\boldsymbol{\delta}_g^* = (\boldsymbol{\delta}_g^T, \boldsymbol{\zeta}_g^T)^T$, where $\boldsymbol{\zeta}_g = (\zeta_{1g}, \ldots, \zeta_{Ng})^T$, $g = 1, \ldots, G$. We let $\boldsymbol{V}_i^* = (\boldsymbol{V}_i', \boldsymbol{e}_i')'$ where $\boldsymbol{e}_i$ is a vector of all zeros except for a single 1 in the

ith position, and $\boldsymbol{V}^*$ is a matrix consisting of the rows $\boldsymbol{V}_i^{*T}$, $i = 1, \ldots, N$. Gaussian priors are placed on the logistic parameters, i.e., $\boldsymbol{\delta}_g^* \sim N(\mathbf{0}, \boldsymbol{B}_g)$, where $\boldsymbol{B}_g = \text{diag}(\sigma_{\delta g}^2 \mathbf{1}_{\text{P}+1},\ \kappa_{\zeta g}^2 \mathbf{1}_{\text{N}})$, and the priors on the random intercepts satisfy $\boldsymbol{\zeta}_g \sim N(\mathbf{0}, \kappa_{\zeta g}^2 \boldsymbol{I}_N)$. As for the hyperparameters, we assume $\sigma_{\delta g}^2 = 10$ for all components and covariates, and $\kappa_{\zeta g} \sim t_{\nu_\kappa}^+(0, A_\kappa)$, where $\nu_\kappa = 3$ and $A_\kappa = 10$ for all components.

To sample the logistic parameters, Polson et al. (2013) proposed a data augmentation scheme incorporating Pólya-Gamma latent variables, which facilitates Gibbs steps. Details on sampling the logistic parameters are provided in the Appendix A.1.

2.3.5 Joint posterior distribution

Based on the augmented likelihood function in (6) and the prior distributions in Section 2.3, we denote all parameters of component g by $\boldsymbol{\Theta}_g = (\boldsymbol{\theta}_g', \tau_g^2, \sigma_g^2, \boldsymbol{\delta}_g^{*\prime}, \kappa_{\zeta g}^2)'$ and $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1', \ldots, \boldsymbol{\Theta}_G')'$. The joint augmented posterior distribution of all parameters $\boldsymbol{\Theta}$ can be written as:

$$
\begin{aligned}
f(\boldsymbol{\Theta} \mid \boldsymbol{y}, \boldsymbol{S}, \boldsymbol{V}^*) \propto & \prod_{i=1}^{N} \prod_{g=1}^{G} \left[\pi_{ig} f_g(\boldsymbol{y}_i \mid \boldsymbol{\Theta}_g)\right]^{z_{ig}} \\
& \times \prod_{g=1}^{G} f_\theta(\boldsymbol{\theta}_g \mid \boldsymbol{D}_g) \times \prod_{g=1}^{G} f_\tau(\tau_g^2 \mid \nu_\tau, A_\tau) \\
& \times \prod_{g=1}^{G} f_\sigma(\sigma_g^2 \mid \nu_\sigma, A_\sigma) \times \prod_{g=1}^{G} f_\delta(\boldsymbol{\delta}_g^* \mid \boldsymbol{B}_g) \\
& \times \prod_{g=1}^{G} f_\kappa(\kappa_{\zeta g}^2 \mid \nu_\kappa, A_\kappa),
\end{aligned}
\tag{7}
$$

where f_θ, for example, denotes the prior probability density function on $\boldsymbol{\theta}_g$.

2.4 Sampling scheme

We proposed a novel RJMCMC algorithm, aimed at obtaining trans-dimensional moves by allowing the number of components to either increase or decrease by 1.

2.4.1 Proposed RJMCMC algorithm

Reversible jump MCMC (RJMCMC) was proposed by Green 1995 by allowing the Markov Chain sampler to jump between parameter subspaces of different dimensionality. Richardson and Green 1997 applied RJMCMC to the analysis of univariate normal mixtures using a hierarchical prior model. Many papers in time series and spectral analysis have utilized RJMCMC (Rosen et al., 2009, 2012; Bertolacci et al., 2021; Li and Krafty, 2019).

In this project, we use RJMCMC as the sampling algorithm, which is an advanced Metropolis-Hastings algorithm allowing sampling from varying dimensions. In our algorithm, each RJMCMC iteration ℓ contains two types of moves: between-model moves and within-model moves. Between-model moves involve a change in parameter dimensions by proposing to add or remove one component. Within-model moves do not involve a change in parameter dimensions and Gibbs sampling is used to sample from the posterior distribution of the parameters.

To aid in developing the RJMCMC, we let $\boldsymbol{\Theta}_{g,G} = (\boldsymbol{\theta}_g', \tau_g^2, \sigma_g^2, \boldsymbol{\delta}_g^{*\prime}, \kappa_{\zeta g}^2)'$ be the aggregation of all parameters for gth component and $\boldsymbol{\Theta}_G = (\boldsymbol{\Theta}_{1,G}', \ldots, \boldsymbol{\Theta}_{g,G}', \ldots, \boldsymbol{\Theta}_{G,G}')'$ as the aggregation of all parameters of all components. We denote the current state as $(G^c, \boldsymbol{\Theta}_{G^c}^c)$ and the proposed state as $(G^p, \boldsymbol{\Theta}_{G^p}^p)$, where the superscript c and p denote the current and proposed states, respectively.

The number of components g is first initialized, followed by initializing the other model parameters. Between-model moves consist of steps to either split a component into two or merge two components into one. The proposed moves are then accepted or rejected using a Metropolis-Hasting step. Within-model moves involve sampling each model parameter using Gibbs sampling without changing the number of components. The RJMCMC sampling scheme is described as follows, and a detailed sampling scheme is provided in Appendix A.1.

1. **Between-model moves**

   For the between-model moves, a new value of G is proposed, and conditional on this value, parameter values $\boldsymbol{\theta}_g$, τ_g^2, σ_g^2, $\boldsymbol{\delta}_g^*$ and $\kappa_{\zeta g}^2$ are proposed.

   a. The number of components is proposed to either increase by 1 (split) or decrease by 1 (combine) with equal probabilities. Thus, we have $G^p = G^c + 1$ or $G^p = G^c - 1$, if

$G^c \neq 1$ or $G^c \neq G_{max}$, where G_{max} is the maximum number of components we allow.

b. If a split step is proposed, then a candidate component for splitting is drawn randomly. A new vector of logistic parameters is generated for the new component. New smoothing parameters and error variances are drawn based on the current parameter values. Finally, conditional on the smoothing parameters and error variances, two new sets of model parameters are drawn. We accept the proposed move with the probability

$$\alpha = \min\left\{1, \frac{p(G^p, \boldsymbol{\Theta}^p_{G^p} \mid \boldsymbol{y}) \times q(G^c, \boldsymbol{\Theta}^c_{G^c} \mid G^p, \boldsymbol{\Theta}^p_{G^p})}{p(G^c, \boldsymbol{\Theta}^c_{G^c} \mid \boldsymbol{y}) \times q(G^p, \boldsymbol{\Theta}^p_{G^p} \mid G^c, \boldsymbol{\Theta}^c_{G^c})}\right\},$$

where $p(\cdot)$ denotes the target density of the proposed or current states and is the product of the joint likelihood function and the prior densities. The function $q(\cdot)$ denotes the proposed density conditional on the current or proposed states. If the proposed state is accepted, we move to the model with G^p components. If the proposed state is rejected, we stay at the current state with G^c components.

c. If a combine step is proposed, then two components are selected to be combined, with one component to which the smallest number of subjects are allocated. We begin a combine proposal by first removing one set of logistic parameters. Values of the combined smoothing parameters and error variances are then computed. Finally, conditional on the values of the combined smoothing parameters and error variances, the new set of model parameters of the combined component are drawn. The acceptance rate is the inverse of that from the split proposal.

2. **Within-model moves**

After completing between-model moves, given the new value of G^c, parameters corresponding to a model with G^c components can be updated using Gibbs sampling. Denote ℓ as the index of the Gibbs sampling iteration. For the $(\ell+1)$th iteration, we carry out the following steps:

a. Draw $\boldsymbol{\theta}_g^{(\ell+1)}$ from $(\boldsymbol{\theta}_g^{(\ell+1)} \mid \boldsymbol{y}, \boldsymbol{S}, \tau_g^{2(\ell)}, \sigma_g^{2(\ell)}) \sim N(\boldsymbol{u}_g, \sigma_g^2 \boldsymbol{\Lambda}_g)$, where $\boldsymbol{u}_g$ and $\boldsymbol{\Lambda}_g$ are posterior means and covariance matrix. Detailed formulas are displayed in Appendix A.1.

b. Draw $\sigma_g^{2(\ell+1)}$ from $(\sigma_g^{2(\ell+1)} \mid \boldsymbol{\epsilon}_{ig}^{(\ell+1)}, a_{\sigma_g}^{(\ell+1)}) \sim IG\Big(\frac{nN_g+\nu_\sigma}{2}, \frac{\sum_{i=1}^{N} z_{ig}\boldsymbol{\epsilon}_{ig}'\boldsymbol{\epsilon}_{ig}}{2} + \frac{\nu_\sigma}{a_{\sigma_g}}\Big)$, where N_g is the number of subjects in the gth component, $\boldsymbol{\epsilon}_{ig}$ is the error vector for the gth component and ith subject and a_{σ_g} is a latent variable related to the augmentation of the half-t distribution (Wand et al., 2011).

c. Draw $\tau_g^{2(\ell+1)}$ from $(\tau_g^{2(\ell+1)} \mid \boldsymbol{\beta}_g^{(\ell+1)}, a_{\tau_g}^{(\ell+1)}) \sim IG\Big(\frac{\nu_\tau+m}{2}, \frac{\boldsymbol{\beta}_g'\boldsymbol{\beta}_g}{2} + \frac{\nu_\tau}{a_{\tau_g}}\Big)$, where a_{τ_g} is a latent variable related to the augmentation of the half-t distribution.

d. Draw $\boldsymbol{\delta}_g^{*(\ell+1)}$ from $(\boldsymbol{\delta}_g^{*(\ell+1)} \mid \boldsymbol{V^*}, z_{ig}^{(\ell)}, \omega_{ig}^{(\ell+1)}, \kappa_{\zeta g}^{2(\ell)}) \sim N(\boldsymbol{M}_g, \boldsymbol{\Sigma}_g)$, where $\omega_{ig}^{(\ell+1)}$ is a Pólya-Gamma latent variable related to the augmentation of the Pólya-Gamma distribution (Polson et al., 2013). The terms $\boldsymbol{M}_g$ and $\boldsymbol{\Sigma}_g$ are posterior mean and covariance matrices. Detailed formulas are displayed in Appendix A.1.

e. Draw $\kappa_{\zeta g}^{2(\ell+1)}$ from $(\kappa_{\zeta g}^{2(\ell+1)} \mid \boldsymbol{\zeta}_g^{(\ell+1)}, a_{\kappa_g}^{(\ell+1)}) \sim IG\Big(\frac{\nu_\kappa}{2}, \frac{\boldsymbol{\zeta}_g'\boldsymbol{\zeta}_g}{2} + \frac{\nu_\kappa+N}{a_{\kappa g}}\Big)$, where a_{κ_g} is a latent variable related to the augmentation of the half-t distribution.

f. Mixing weight $\pi_{ig}^{(\ell+1)}$ can be obtained by computing $p(\pi_{ig}^{(\ell+1)} \mid \boldsymbol{V}^*, \boldsymbol{\delta}_g^{*(\ell+1)}, z_{ig}^{(\ell)})$ following (3).

g. Draw $z_{ig}^{(\ell+1)} \sim p(z_{ig}^{(\ell+1)} = 1 \mid \boldsymbol{y}, \boldsymbol{S}, \boldsymbol{\theta}_g^{*(\ell+1)}, \sigma_g^{2(\ell+1)}, \pi_{ig}^{(\ell+1)})$ following (5).

Details of conditional posterior distributions are given in Appendix A.1.

2.4.2 Label switching

The label switching issue is a well-known issue in mixture models. Although the likelihood is invariant under the permutations of labeling, it becomes a problem when we aim at estimating trajectories of one component (Rossell and Steel, 2019). Consider a random sample from a population of two normal components. If the means of the two components are well separated, labeling by the posterior mean will be equivalent to the population labeling. However, if the level of separation reduces and the two posterior distributions overlap, label switching issues will occur (Jasra et al., 2005). The issue of label switching will mix inferences of posteriors from different components and impede the correct inferences of posterior distributions of a single component.

Richardson and Green 1997 mentioned the issue of label switching in the context of RJMCMC in the univariate mixture case. Label switching can be tackled by choosing to

order on means, variances, weights, or some combination of all three parameters. Richardson and Green suggested to post-process the MCMC runs according to different choices of labels.

However, in more complex models such as our time series model, ordering approaches do not work well because it is rarely true that the selected ordering constraint is able to separate symmetric posterior modes (Jasra et al., 2005; Fúquene et al., 2019), especially in RJMCMC (Spezia, 2009). Many solutions have been proposed to solve or reduce the issue of label switching. Stephens 2000 was the first who proposed a solution to deal with label switching. A Kullback-Leibler divergence between an averaged matrix of classification probabilities in each MCMC iteration is minimized using a permutation-based method. Marin et al. 2005, 2007 proposed the pivotal reordering algorithm, which is a data-driven approach that selects the permutation that minimizes the Euclidean distance between the pivot and the set of permuted parameter vectors in each MCMC iteration. Papastamoulis and Iliopoulos 2010, 2013 proposed different versions of Equivalence Classes Representatives (ECR) algorithms based on the assumption that equivalent allocation vectors are mutually exclusive from the label switching solution. Different ECR algorithms differ in the choice of equivalent allocation vectors. In addition, Sperrin et al. 2010 presented a probabilistic relabeling algorithm based on an EM-type algorithm and Rodriguez and Walker 2014 proposed a data-based labeling method that is based on a k-mean type loss function between the cluster pivots and the observed data.

For our proposed method, we choose to use a data-based relabeling algorithm to solve the issue of label switching by post-processing the RJMCMC iterations. This algorithm is a permutation-based method based on the latent indicators $z_i^{(t)}$ for each subject i at each iteration $t = 1, \ldots, m$. First, cluster centers m_{kr} and dispersion parameters s_{kr} are estimated for each component $k = 1, \ldots, K$ and each dimension $r = 1, \ldots, d$. Then the optimal permutations are determined based on the optimization of a k-means type loss function between the cluster pivots and the observed data x_{ir}. The data-based relabeling algorithm is given below.

1. Obtain estimates m_{kr} and s_{kr} for $k = 1, \ldots, K$ and $r = 1, \ldots, d$.

2. For each iteration $t = 1, \ldots, m$, find a permutation τ that minimizes

$$\sum_{k=1}^{K}\sum_{\ell=1}^{K} I(z_i^{(t)} = \tau_l) \sum_{i:\tau z_i^{(t)}=\ell} \sum_{r=1}^{d} (\frac{x_{ir} - m_{kr}}{s_{kr}})^2.$$

To implement the above methods in R, Papastamoulis 2015 has created an R package, `label.switching`, which includes all of the above methods and provides user-friendly functions to implement these relabeling algorithms.

2.5 Simulation studies

We have conducted simulation studies and compared our method to different existing methods.

2.5.1 Comparing performance of the proposed method to other methods

We conducted simulation studies according to six model settings, denoted by M1 to M6. All model settings generate a collection of univariate time series with two components. We generated $s = 100$ replicates, each consisting of a collection of $N = 150$ univariate time series of length $n = 300$. Models based on regression splines used $m = 10$ basis functions and $P = 4$ covariates (including the intercept). Different model parameters such as spline coefficients, linear coefficients, error variances, and smoothing parameters were assigned values for each component g and for each model setting.

To form different model settings, we assume two true models, which are evaluated by three methods. M1, M2 and M3 generate a two-component mixture model where each mixture follows a cubic function of time. The true cubic models $f_g(t)$ for the two mixture components are

$$f_1(t) = 3 - 2t + 1.5t^2 - 0.1t^3$$

and

$$f_2(t) = 10 + t - 0.5t^2 - 0.05t^3,$$

where the error variances are 2 and 4 respectively. M1, M2 and M3 then fit different models using our proposed model, an `R` package `lcmm` and `TRAJ` in `SAS`. M4, M5 and M6 generate a 2-component mixture model where each mixture follows a regression spline model. M4, M5 and M6 then fit different models using the three methods listed above. Parameters used for the cubic and the regression splines model are presented in Table 2.

Proust-Lima and Liquet 2011; Proust-Lima et al. 2015 developed an `R` package called `lcmm`, which is an abbreviation for latent class mixed model. It can be used to fit various extensions of mixed models including latent class mixed models, joint latent class mixed models, mixed models for curvilinear outcomes or mixed models for multivariate longitudinal outcomes using maximum likelihood estimation (MLE). A numerical method called the modified Marquardt algorithm is used to obtain the MLEs with strict convergence criteria based on the parameter values and likelihood stability, as well as the negativity of the second derivatives. In our simulation studies, we apply the latent class random intercept model for Gaussian longitudinal outcomes. A finite number of latent classes needs to be specified. In addition, covariates can be incorporated into the class membership model via multinomial logistic regression.

`TRAJ` performs a group-based trajectory modeling for longitudinal data (Jones et al., 2001). It is based on a polynomial model of different orders. `TRAJ` assumes that conditional on the group membership, longitudinal outcomes are independently distributed, although longitudinal outcomes are not conditionally independent at the population level. The procedure obtains MLE using a numerical quasi-Newton iterative algorithm and allows the inclusion of covariates via a class membership model. As described before, the amount of smoothing in our approach is controlled by the τ_g^2.

The data for each replicate of the regression splines case are generated as follows:

1. Generate fixed and evenly-spaced time points with a length of 300 from 0 to 5
2. Generate 10 sets of basis functions as a function of time
3. Set values for the parameters τ_g^2, σ_g^2,$\boldsymbol{\alpha}_g$ and generate $\boldsymbol{\beta}_g \sim N(\mathbf{0}, \tau_g^2 \boldsymbol{I}_m)$ for each component g. The values of intercepts for two components are 1 and 5, and the values of slopes for two components are -2 and 1. Error variances are set to be 2 and 4 for each component, as well as 0.8 and 1 for the smoothing parameters

Table 1. Simulation results for the logistic parameters in the 2-component mixture model: values in each cell are in the format of RMSE (bias, variance).

Method	Model setting	δ_0	δ_1	δ_2	δ_3
Proposed	M1	0.90 (0.08,0.81)	0.51(-0.18,0.23)	0.28(0.11,0.07)	0.35(-0.02,0.12)
lcmm	M2	1.15(0.49,1.11)	0.67(-0.34,0.34)	0.34(0.12,0.10)	0.36(0.06,0.13)
PROC TRAJ	M3	1.28(0.22,1.64)	0.74(-0.20,0.52)	0.35(0.09,0.12)	0.34(0.02,0.11)
Proposed	M4	1.12 (0.41,1.12)	0.54(-0.25,0.24)	0.25(0.07,0.06)	0.33(-0.01,0.11)
lcmm	M5	1.41(0.50,1.76)	0.67(-0.25,0.38)	0.26(0.05,0.06)	0.33(-0.01,0.11)
PROC TRAJ	M6	1.53(0.52,1.84)	0.78(-0.23,0.56)	0.28(0.09,0.07)	0.34(0.02,0.11)

4. Generate values of four covariates (including the intercept) for each univariate time series, which follow normal distributions with different means and variances.
5. Set the logistic parameters $\boldsymbol{\delta}_g$, with the values of 5, -3, 1 and 0.1 for each covariate (including the intercept). These values are also used in the cubic model in model settings M1 to M3.
6. Plug the covariate values and the true logistic parameters into the multinomial logits to obtain the mixing weights
7. Sample latent indicators z_{ig} for each univariate time series and each component, based on the weights from step 6.
8. For each component, simulate each univariate time series $\boldsymbol{y}_i$ according to model (1)

Table 1 shows the results of the logistic parameters in the 2-component mixture model in terms of RMSE (bias, variance) for the six model settings. Model settings M1 to M6 are introduced. Our proposed method outperforms the other two methods for both the cubic and regression splines (M1 and M4), especially for the intercept δ_0 and the first coefficient of the covariate δ_1. Both bias and variances are smaller than for `lcmm` and `TRAJ` methods. Our covariate-guided model can be viewed as a regularization model with a penalty since we assume the logistic parameters follow a normal distribution, which corresponds to ridge regression in the frequentist setting. The shrinkage parameter or the penalty is able to result in more accurate estimates for some extreme cases, such as perfect separation and unbalanced

designs, where logistic parameter estimates are hugely inflated in ordinary logistic regression. The small RMSEs of the proposed method demonstrate the stability of our proposed model and its capability for accurate clustering compared to the other two methods, in which no penalty is added to the class membership model.

We also investigate the performance of the estimated trajectories for each component by calculating the averaged root square error (ARSE) of each component

$$\text{ARSE} = \sqrt{\frac{1}{nK}\sum_{j=1}^{n}\sum_{k=1}^{K}\left(\hat{y}_j - y_j\right)^2},$$

where $\hat{y}_j$ is the estimated value of the true y_j for the jth time point, $j = 1, \ldots, n$. Table 2 provides the ARSE, averaged bias (A-bias) and V-bias for each component, where V-bias is computed by calculating the sample variance of the bias over time points.

For model settings M1, M2 and M3 in Table 2, unsurprisingly, `lcmm` and `TRAJ` outperform our proposed model since they use the true model, while our proposed model fits a penalized splines to the truth of a cubic model. However, our proposed method still shows a relatively good performance with a comparable bias to the other two methods. For model settings M4, M5 and M6 in Table 2, our proposed model outperforms the other two methods, especially compared to `TRAJ`. Notably, M4, M5 and M6 use a regression spline model as the true model. Thus, our proposed method demonstrates good performance when the model is correctly specified. `lcmm` is able to fit a linear mixed model using spline basis functions as input covariates, thus leading to an accurate estimation. However, since `TRAJ` only allows fitting polynomial models, it performs the worst among the three methods because of several poor fits among all replicates. In general, our proposed method is able to give precise trajectory estimates for both underlying models. An example of estimated trajectories with the true trajectory from one replicate for M1 and M4 model settings is presented in Appendix A.2.

Table 2. Simulation results of estimated trajectories in the 2-component mixture model: values in each cell are in the format of ARSE×100 (SD×100).

Method	Model setting	Statistics	First component	Second component
Proposed	M1	ARSE	3.2 (0.5)	4.3 (0.9)
		A-Bias	0.05 (0.8)	-0.04 (1.4)
		V-bias	0.1 (0.03)	0.2 (0.04)
lcmm	M2	ARSE	1.7 (0.6)	2.6 (0.9)
		A-Bias	0.05 (0.8)	-0.04 (1.4)
		V-bias	0.03 (0.02)	0.06 (0.04)
TRAJ	M3	ARSE	1.8 (0.6)	2.8 (1.0)
		A-Bias	0.05 (0.8)	-0.04 (1.4)
		V-bias	0.03 (0.02)	0.05 (0.04)
Proposed	M4	ARSE	2.9 (0.6)	3.9 (0.9)
		A-Bias	0.04 (0.8)	-0.02 (1.4)
		V-bias	0.1 (0.03)	0.1 (0.06)
lcmm	M5	ARSE	3.0 (0.7)	4.3 (0.8)
		A-Bias	0.02 (0.8)	-0.02 (1.3)
		V-bias	0.1 (0.04)	0.2 (0.07)
TRAJ	M6	ARSE	26.8 (6.9)	28.9 (7.2)
		A-Bias	1.6 (8.1)	-1.5 (8.6)
		V-bias	7.2 (0.38)	8.4 (0.42)

2.6 Real-data application results

We apply our proposed method to the analysis of the fNIRS still-face study. Introduction to fNIRS data and the motivating study are given in Chapter 1. Five covariates are included

in our covariate-guided model, including Infant Behavior Questionnaire-Revised negative emotionality (IBQ-NE) score, gestational age (in Days), infant age (in Months), head circumference (in cm) and sex. All continuous covariates are scaled to follow the standard normal distribution.

Infant Behavior Questionnaire-Revised (IBQ-R) is a widely used parent report of measure developed to assess dimensions of temperament along multiple scales (Gartstein and Rothbart, 2003). The IBQ-NE construct combines data from the following subscales: Sadness, Distress to Limitations, Fear, and Falling Reactivity/Rate of Recovery from Distress (Gartstein and Rothbart, 2003; Gartstein et al., 2010). We assume that these covariates are associated with the mixing weights of each component and the trajectory pattern of that component. The data pre-processing steps in Section 1.3.4 resulted in a sample of 82 subjects with complete covariates, each with a univariate time series from each channel. Each univariate time series consists of a grid of 1500 time points of which 500 measurements are in the still-face period. Here we present results the RJMCMC results after post-processing using the data-based relabeling algorithm Papastamoulis 2015 introduced in Section 2.4.2.

Table 3 lists the estimated posterior probabilities for each number of components, as well as the acceptance rates for the twelve channels. Two-component models have the largest posterior probabilities for all channels. Three-component models have the second-largest probabilities for almost all channels. In addition, five-component and six-component models are less appealing with very small probabilities. The acceptance rates for the twelve channels range from 8.9% to 11.3%, which are reasonable for RJMCMC. A very low acceptance rate prevents the algorithm from exploring the entire parameter space and requires a large number of iterations to achieve convergence. A very high acceptance rate would indicate poor model fitting without stability.

We present results from two selected channels S1D1 and S5D4. Results of the trajectory estimates for the other three channels S1D3, S6D4 ad S7D4 are given in Appendix A.3. Since two-component models are the most appealing, we look at the estimated trajectories based on the 2-component model. Estimated trajectories with pointwise 95% credible intervals for the 2-component model are displayed in Figures 6 and 7. Components in the right panel of Figures 6 and 7 are the reference component for the covariate-guided model analysis.

Table 3. Estimated posterior Probabilities of the number of components, along with the acceptance rate for each channel.

Channel	P(G=1)	P(G=2)	P(G=3)	P(G=4)	P(G=5)	P(G=6)	Acceptance rate
S1D1	0.0782	0.5000	0.2977	0.1024	0.0198	0.0018	0.0937
S1D2	0.1260	0.4469	0.3191	0.0959	0.0119	0.0002	0.1128
S1D3	0.0686	0.5238	0.3088	0.0833	0.0146	0.0009	0.0943
S2D2	0.1230	0.4608	0.2966	0.0968	0.0202	0.0026	0.1005
S3D2	0.1440	0.5560	0.2370	0.0555	0.0067	0.0008	0.0904
S4D2	0.1253	0.5130	0.2846	0.0655	0.0106	0.0010	0.0904
S5D1	0.0696	0.5338	0.3104	0.0748	0.0110	0.0004	0.0894
S5D3	0.2379	0.4718	0.2234	0.0582	0.0083	0.0004	0.0904
S5D4	0.0700	0.4446	0.3404	0.1217	0.0218	0.0014	0.0995
S6D4	0.1286	0.5031	0.2694	0.0815	0.0157	0.0017	0.1027
S7D4	0.1143	0.4508	0.2988	0.1129	0.0216	0.0015	0.1019
S8D4	0.2072	0.4642	0.2367	0.0740	0.0156	0.0022	0.1076

We are interested in the brain activation trajectories for the still-face period (S), while the interact (I) period serves as the reference level. Both the S1D1 and S5D4 channels show clear trajectory patterns. One component has a decreasing trend during the still-face period, while the other component has an increasing trend during the still-face period in both channels.

Table 4 displays the logistic coefficient estimates for channels S1D1 and S5D4, where the number of components $G = 2$ as well as pooled results across all components (2-6). Pooled results refer to pooling results of logistic parameter estimates from iterations that have 2 to 6 components. Since all iterations included at least two components, where one component shows increasing brain activity whereas the other has decreasing brain activity in the still-face period, we are able to pool the iterations of the logistic parameter estimates for these two components together and find pooled posterior means as well as 95% credible intervals. The results based on the pooled analysis are similar to the results conditional on two components, where the directions of all coefficients are the same. Though all credible

Figure 6. Estimated trajectories with pointwise 95% credible intervals for the two-component model for the S1D1 channel. **I**: Interact **S**: Still-face **R**: Recovery.

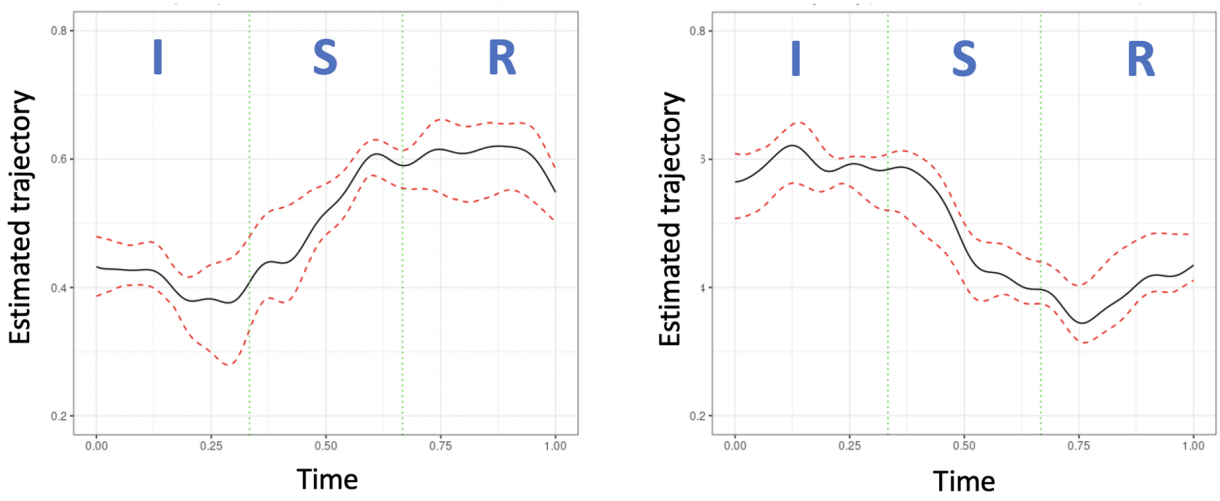
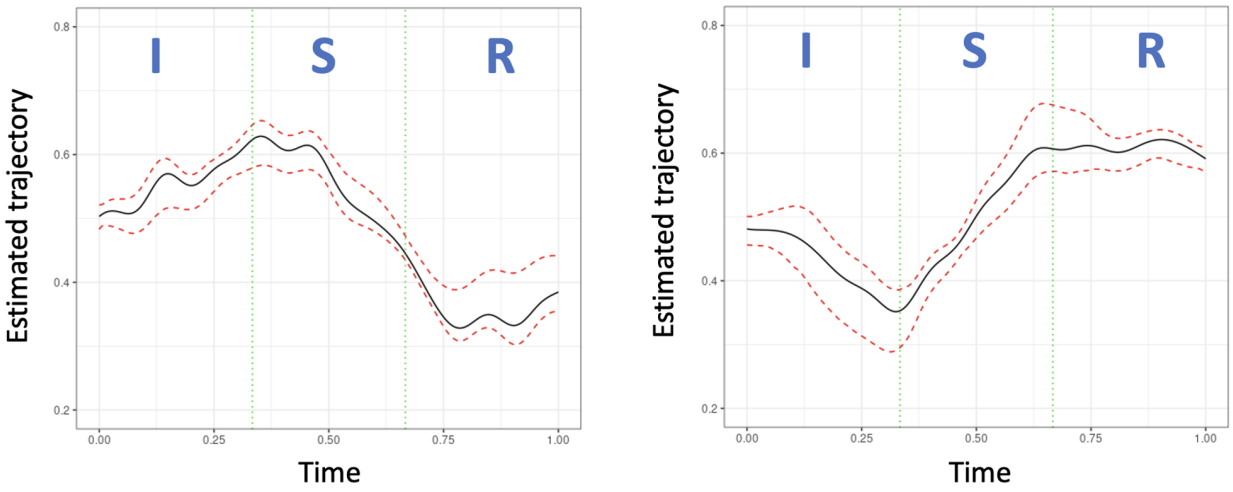


Figure 7. Estimated trajectories with pointwise 95% credible intervals for the two-component model for the S5D4 channel **I**: Interact **S**: Still-face **R**: Recovery.



intervals include zero, the negative posterior mean estimate of the IBQ-NE score from the S1D1 channel and positive posterior mean estimate of the IBQ-NE score from the S5D4 channel could still indicate that high IBQ-NE score is associated with a decreasing brain activation for the still-face period.

Table 4. Logistic coefficient estimates for channels S1D1 and S5D4.

Covariates	S1D1 G=2			S5D4 G=2		
	Mean	Lower 95% CI	Upper 95% CI	Mean	Lower 95% CI	Upper 95% CI
Intercept	-0.083	-4.890	4.432	0.669	-3.989	5.102
IBQ-NE	-1.523	-5.028	1.877	1.489	-2.352	5.033
Infant age	2.212	-1.954	5.698	2.133	-1.674	5.896
Gestational days	-0.069	-3.855	3.520	0.954	-2.910	4.335
Head Circumference	0.785	-2.651	4.688	0.542	-3.358	4.206
Sex	0.142	-4.038	4.354	-0.175	-4.231	4.165
Covariates	S1D1 pooled			S5D4 pooled		
	Mean	Lower 95% CI	Upper 95% CI	Mean	Lower 95% CI	Upper 95% CI
Intercept	-0.041	-4.727	4.449	0.729	-3.902	5.289
IBQ-NE	-1.794	-5.597	2.245	0.952	-2.822	4.614
Infant age	1.809	-2.238	5.618	1.693	-2.284	5.655
Gestational days	-0.064	-3.846	3.527	0.97	-2.838	4.752
Head Circumference	0.592	-3.007	4.369	0.563	-3.227	4.338
Sex	0.171	-3.965	4.424	-0.184	-4.114	4.285

2.7 Discussion

Our proposed covariate-guided Bayesian mixture of spline experts model aims to perform model-based clustering of univariate time series from multiple subjects. The proposed method is a Bayesian mixture of splines, where covariates are incorporated into the mixing weight of each component. The performance of our proposed method is illustrated in simulation studies in Section 2.5. Results from the simulation studies demonstrate that the Bayesian penalized spline model is flexible enough to be able to provide a good fit to a parametric cubic model. Simulation studies also compare our proposed method to two existing methods and results show that our proposed method outperforms these methods in terms of both trajectory estimates and logistic parameter estimates, especially when the true trajectory is wiggly.

We apply our proposed method to an fNIRS still-face study, aiming to discover trajectory

patterns for a single channel and to see how selected covariates are associated with the trajectory patterns of each component. Based on results from several channels, we conclude that clear patterns emerge in the components making up the trajectory estimates within the still-face period. From the logistic coefficients estimates in Table 4, we are able to conclude that a higher IBQ-NE score is related to a lack of response in the still-face period, with a decreasing trajectory of oxy-hemoglobin. The potential explanation for this association is that infants with a high IBQ-NE are not able to or attempt to regulate their emotions, which indicates less blood flowing to their brains and thus results in a decreasing trajectory of responses. Though the 95% credible interval of infant age does include zero, we can still reach a conclusion that young infants tend not to have the control to attempt to regulate emotions, which leads to the decreasing trend of brain activity level as seen in the second component of S1D1 and the first component of S5D4 in Figures 6 and 7.

Our proposed method has some limitations. First, our proposed RJMCMC algorithm still suffers from the issue of label switching. The credible intervals are still wide for some channels, indicating the possible existence of label switching even after performing the post-processing steps. Second, our proposed method focuses on the clustering of univariate time series. However, in the area of brain-imaging, time series from different channels, voxels, or electrodes are usually multi-dimensional. Thus, it is natural to extend our proposed method to multivariate time series, which will be more meaningful in real-data applications of brain-imaging data. Our next project in Chapter 3 extends our proposed model to the clustering of multivariate time series from multiple subjects.

3.0 Covariate-guided Bayesian mixture of spline experts for the analysis of multivariate time series

3.1 Introduction

With the rapid development of modern technologies in diverse scientific areas, the analysis of multivariate time series data becomes more and more important, especially in the brain-imaging area. With the fast computing technologies such as cloud computing, many statistical approaches are able to be implemented by considering the heterogeneity and dimensionality of multivariate time series. Multivariate time series is a common data structure in brain-imaging data, such as fNIRS introduced in Section 1.1, where each subject has a multi-dimensional time series of oxy-hemoglobin (HbO) measurements in multiple channels. Multivariate time series have different definitions in different settings. In our project, it refers to a collection of individuals, each with a multi-dimensional time series. Multivariate time series are often heterogeneous across subjects and even for different dimensions or variates within a subject, hence causing many statistical challenges in inference and estimation. Thus, we extended our previous work in Chapter 2 to the multivariate case with the purpose of clustering multivariate time series under the Bayesian framework.

Several authors have focused on proposing different clustering algorithms for multivariate time series. Johnson et al. 2014 and McLachlan 2005 have applied the discriminant and cluster analysis to the case of multivariate time series. Kakizawa et al. 1998 used Kullback-Leibler discrimination information as the minimum discrimination criteria for the clustering of multivariate Gaussian time series. Parzen 1990 proposed the Chernoff information measures and Zhang and Taniguchi 1994, 1995 have shown its robustness in clustering of multivariate non-Gaussian time series. Singhal and Seborg 2005 proposed a new method for multivariate time series clustering based on two similarity factors, where one factor is based on the principal component within the dimension and another one is based on the Mahalanobis distance between multivariate time series. Wang et al. 2007 used a modified K-mean clustering algorithm for the clustering of multivariate time series based on univariate

structures. A variety of papers have established different model-based clustering methods for the clustering of multivariate time series, such as multivariate AR models (Kalpakis et al., 2001; Xiong and Yeung, 2002) and Hidden Markov Models (Li et al., 2001; Wang et al., 2002). Other feature-based clustering approaches with dimension reduction include principle component analysis (Li, 2019; Rani and Sikka, 2012; Ye et al., 2004) and independent component analysis (Verdoolaege and Rosseel, 2010; Guo et al., 2008). A comprehensive review of methods of time series clustering can be found in Liao 2005.

Many works have been performed for a single multivariate time series in both the time and the frequency domains. Dahlhaus 2000 proposed a multivariate locally stationary process with time-varying parameters using a generalization of Whittle likelihood. Krafty and Collinge 2013 presented a novel approach for the smoothness and estimation of multivariate power spectrum using a penalized multivariate Whittle likelihood. Guo and Dai 2006 proposed a nonparametric smoothing method for the estimation of a time-varying spectrum that is assumed to be smooth over both time and frequency. A smoothing splines ANOVA was used to smooth Cholesky components of the power spectrum for both time and frequency. Other contributions of nonparametric models of multivariate spectral analysis include the multivariate smooth localized complex exponential (SLEX) model (Ombao et al., 2005) and stationary wavelet models (Sanderson et al., 2010; Park et al., 2014). Bayesian approaches have been widely used for the spectral analysis of multivariate time series. Zhang 2016 extended the Bayesian adaptive spectral analysis from Rosen et al. to the multivariate nonstationary setting. Rosen and Stoffer 2007 proposed a Bayesian method for multivariate time series that is able to fit the smoothing splines model to each component separately by using the Cholesky decomposition of the spectral matrix. Li and Krafty 2019 introduced an approach to the adaptive spectral analysis of multivariate time-varying power spectrum. This approach can be formulated in the Bayesian framework with the advantage of approximating both abrupt and slowly varying changes in spectral matrices. Recently, Li et al. 2021 developed an approach to analyze the association between covariates and multivariate time series across multiple subjects. This method was fully Bayesian and assumed that the number of groups as well as the covariate partition defining groups are random.

Different statistical and computational methods of multivariate correlated time series

have been applied to the brain-imaging areas. In the area of functional data analysis, multivariate functional PCA (Chiou et al., 2014; Happ and Greven, 2018) have been used to analyze functional processes. Zhang et al. 2021 used an interpretable functional principal component analysis for the analysis of multilevel multivariate functional data with the application to the electroencephalography (EEG) data, where EEG measures brain activities from different locations and with different frequency bands. Multivariate time series are often high-dimensional and have a correlated structure in brain-imaging data such as EEG. Hector and Song 2021 proposed a divide-and-conquer algorithm for the estimation of regression parameters that allows a fully distributed computation at each data source with a pairwise composite likelihood. Later Hector and Song performed a joint integrative analysis based on a distributed quadratic inference function. In addition, Baladandayuthapani et al. 2008 proposed a Bayesian approach to analyze hierarchical spatially correlated functional data with multiple nested hierarchy layers and spatial correlations.

Many works have been focused on the latent class analysis of multivariate longitudinal data. Jones et al. 2001 and Nagin et al. 2018 developed a group-based multivariate trajectory modeling approach with a quasi-Newton procedure and aimed to identify trajectory estimates based on multivariate longitudinal data. This approach fits polynomial models with different orders for the estimation of trajectories and it is implemented in `SAS` software through a procedure called `TRAJ`. Magrini 2022 presented another group-based multivariate trajectory modeling approach using the EM algorithm. Proust-Lima et al. 2015, 2017 proposed a latent class mixed model to estimate trajectories of multivariate markers by using different marker-specific link functions from different families. This approach can be extended to a joint analysis with the time-to-event data.

In our project, we extend our proposed mixture of spline experts model to the case of multivariate time series, which refer to a collection of multi-dimensional time series across subjects. Models with different parameters are fit separately for each component of univariate time series across subjects. Smoothing splines models were used to fit the time series model with flexibility and a low-rank approximation approach (Wood et al., 2002; Wahba, 1980) was adopted to obtain smoothing coefficients based on a small set of basis functions. Mixture-of-experts model was incorporated to allow for the inclusion of time-independent

covariates. The Pólya-Gamma data augmentation strategy (Polson et al., 2013) was used to obtain logistic parameters and mixing weights for each component. The proposed approach was formulated in a fully Bayesian framework and sampled from the posterior distributions via Gibbs sampling. The number of components was selected using an adjusted deviance information criteria (DIC) based on the posterior variance (Celeux et al., 2006; Gelman et al., 1995). The rest of Chapter 3 is organized as follows. In Section 3.2 and 3.3 we present the proposed model and priors for each parameter. Section 3.4 introduces the Bayesian sampling scheme based on Gibbs sampling. In Section 3.5 we report simulation results under different settings and Section 3.6 illustrates our proposed method with the real-data application of the PGS-ECHO fNIRS still-face study, which is given a detailed introduction in Chapter 1. Section 3.7 gives a conclusion of our first project as well as limitations and future works.

3.2 Multivariate mixture of spline experts model

In this section, we provide a detailed description of the proposed covariate-guided Bayesian mixture of spline experts model. The proposed model consists of spline components whose mixing weights depend on covariates.

3.2.1 Mixture of splines model

We propose a tensor-product mixture of splines model for multivariate time series. For each subject $i = 1, \ldots, N$, let $\boldsymbol{y}_i = (\boldsymbol{y}_{i1}', \ldots, \boldsymbol{y}_{ik}', \ldots, \boldsymbol{y}_{iK}')'$ be the nK-vector corresponding to the K-dimensional time series for $k = 1, \ldots, K$, where $\boldsymbol{y}_{ik} = \big[y_{ik}(t_1), \ldots, y_{ik}(t_j), \ldots, y_{ik}(t_n)\big]'$ is the kth entry of the time series of length n for $j = 1, \ldots, n$, and $\boldsymbol{\epsilon}_i = (\boldsymbol{\epsilon}_{i1}', \ldots, \boldsymbol{\epsilon}_{iK}')'$ is the nK-vector of errors. Following the model representation of Krafty et al. (2017), the tensor-product model for the K-dimensional multivariate time series, conditional on component g, $g = 1, \ldots, G$, can be written as:

$$\{\boldsymbol{y}_i \mid z_{ig} = 1\} = (\boldsymbol{I}_K \otimes \boldsymbol{X})\boldsymbol{\alpha}_g + (\boldsymbol{I}_K \otimes \boldsymbol{W})\boldsymbol{\beta}_g + \boldsymbol{\epsilon}_i, \tag{8}$$

where $\{z_{ig}\}_{g=1}^{G}$ are latent indicators as described in Section 3.2.3, $\boldsymbol{\alpha}_g = (\boldsymbol{\alpha}'_{g1}, \ldots, \boldsymbol{\alpha}'_{gK})'$ is a $2K$-vector of intercepts and slopes, $\boldsymbol{\beta}_g = (\boldsymbol{\beta}'_{g1}, \ldots, \boldsymbol{\beta}'_{gK})'$ is a mK-vector of basis function coefficients as described in Section 3.3.1, $\boldsymbol{I}_K$ is a $K \times K$ identity matrix and $\otimes$ denotes a tensor product. The matrix $\boldsymbol{X}$ is given by $\boldsymbol{X} = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ t_1 & t_2 & \ldots & t_n \end{pmatrix}'$ and the m columns of the matrix $\boldsymbol{W}$ are smoothing splines basis functions as described in Section 3.3.1. We assume the error vector $\boldsymbol{\epsilon}_i$ follows a MVN$(\boldsymbol{0}, \boldsymbol{\Psi}_g \otimes \boldsymbol{U})$ distribution, where $\boldsymbol{U} = \boldsymbol{I}_n$ is the $n \times n$ identity matrix, and $\boldsymbol{\Psi}_g = \text{diag}(\boldsymbol{\sigma}_g^2)$ is a $K \times K$ diagonal matrix with the error variances $\boldsymbol{\sigma}_g^2 = (\sigma_{g1}^2, \ldots, \sigma_{gK}^2)'$. We assume each subject has a common grid of time points across all K entries, such that $\boldsymbol{X}$ and $\boldsymbol{W}$ are common to all subjects, although our proposed method can be generalized to the case where subjects are observed at different grids of time points. In addition, we assume $\boldsymbol{E}(\boldsymbol{y}_{ik}, \boldsymbol{y}_{ih}) = \boldsymbol{0}_{n \times n}$ for $k \neq h$.

To simplify notation, we let $\boldsymbol{S} = [\boldsymbol{X}\ \boldsymbol{W}]$ and $\boldsymbol{\theta}_g = (\boldsymbol{\alpha}'_{g1}, \boldsymbol{\beta}'_{g1}, \ldots, \boldsymbol{\alpha}'_{gK}, \boldsymbol{\beta}'_{gK})'$. The model (8) can then be rewritten as:

$$\{\boldsymbol{y}_i \mid z_{ig} = 1\} = (\boldsymbol{I}_K \otimes \boldsymbol{S})\boldsymbol{\theta}_g + \boldsymbol{\epsilon}_i. \tag{9}$$

3.2.2 Model for mixing weights

As in 2.2.2, the mixture-of-experts model (Jacobs et al., 1991) is also applied to form a covariate-guided structure for the clustering of multivariate time series, where the mixing weights are multinomial logits that are functions of selected covariates. As in Sun et al. (2007), the mixing weights are expressed as

$$\pi_{ig}(\boldsymbol{V}_i) = \frac{\exp(\boldsymbol{V}'_i\boldsymbol{\delta}_g + \zeta_{ig})}{\sum_{h=1}^{G} \exp(\boldsymbol{V}'_i\boldsymbol{\delta}_h + \zeta_{ih})}, \tag{10}$$

where $\boldsymbol{V}_i = (1, V_{i1}, \cdots, V_{iP})'$ is a vector of length $(P+1)$ containing values of P covariates for subject i, and $\boldsymbol{\delta}_g = (\delta_{g0}, \delta_{g1}, \cdots, \delta_{gP})'$ is the corresponding coefficient vector. For identifiability, we set $\boldsymbol{\delta}_G = \boldsymbol{0}$. Equation (10) differs slightly from the weights in the traditional mixture of experts model in that it includes a random term ζ_{ig} for each subject. This term accounts for unmeasured factors beyond the observed covariates and enhances model performance and inference of the mixing weights.

3.2.3 Augmented likelihood

To account for heterogeneity across subjects, we assume that the kth entry of the multivariate time series, $\boldsymbol{y}_{ik}$, comes from a mixture model with G components, i.e.,

$$\boldsymbol{y}_{ik} \sim \sum_{g=1}^{G} \pi_{ig} f_{gk}(\boldsymbol{y}_{ik} \mid \boldsymbol{\mu}_{gk}, \sigma_{gk}^2 \boldsymbol{I}_n), \tag{11}$$

where $f_{gk}(\boldsymbol{y}_{ik} \mid \boldsymbol{\mu}_{gk}, \sigma_{gk}^2 \boldsymbol{I}_n)$ is the probability density function of the multivariate normal distribution with mean vector $\boldsymbol{\mu}_{gk} = \boldsymbol{X}\boldsymbol{\alpha}_{gk} + \boldsymbol{W}\boldsymbol{\beta}_{gk}$ and covariance matrix $\sigma_{gk}^2 \boldsymbol{I}_n$ for the gth component and the kth entry. The π_{ig} are mixing weights that depend on covariates as described in Section 3.2.2.

As is common in mixture models, augmenting the likelihood with latent variables indicating the component from which a time series originates simplifies the computation greatly (Dempster et al., 1977). In particular, let $z_{ig} = 1$ if the ith multivariate time series belongs to the gth component and $z_{ig} = 0$, otherwise. Let $\boldsymbol{y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N)'$ be all observed multivariate time series and $\boldsymbol{\Theta}_{gk}$ be the aggregation of all parameters for component g and entry k. The parameter vector for all components and all entries are then denoted by $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_{11}', \ldots, \boldsymbol{\Theta}_{GK}')'$. The augmented likelihood of all N multivariate time series is given by

$$L(\boldsymbol{\Theta} \mid \boldsymbol{y}, Z) = \prod_{i=1}^{N} \prod_{g=1}^{G} \left[\pi_{ig} \prod_{k=1}^{K} f_{gk}(\boldsymbol{y}_{ik} \mid \boldsymbol{\Theta}_{gk}) \right]^{z_{ig}}, \tag{12}$$

where $Z = \{z_{ig}\}$ is the matrix of all indicators, and $f_{gk}(\boldsymbol{y}_{ik} \mid \boldsymbol{\Theta}_{gk})$ is the probability density function as appeared in the (11). From Bayes' rule, the distribution of the latent indicators z_{ig} is given by

$$p(z_{ig} = 1 \mid \boldsymbol{y}, \boldsymbol{S}, \boldsymbol{\Theta}, \pi_{ig}) = \frac{\pi_{ig} \prod_{k=1}^{K} f_{gk}(\boldsymbol{y}_{ik} \mid \boldsymbol{\Theta}_{gk})}{\sum_{h=1}^{G} \pi_{ih} \prod_{k=1}^{K} f_{hk}(\boldsymbol{y}_{ik} \mid \boldsymbol{\Theta}_{hk})}. \tag{13}$$

3.3 Priors

3.3.1 Smoothing splines prior

Following the smoothing splines prior in the univariate case and the low-rank approximation method introduced in 2.3.1, we assume the following diagonal Gaussian priors in the multivariate case

$$\boldsymbol{\theta}_g \sim N(\mathbf{0}, \boldsymbol{D}_g),$$

where $\boldsymbol{D}_g = \text{diag}(\sigma^2_{\alpha 1}\mathbf{1}_2,\ \tau^2_{\text{g1}}\mathbf{1}_{\text{m}},\ \ldots\ , \sigma^2_{\alpha \text{K}}\mathbf{1}_2,\ \tau^2_{\text{gK}}\mathbf{1}_{\text{m}})$ is the covariance matrix for $\boldsymbol{\theta}_g$. The vector $(\sigma^2_{\alpha 1}, \ldots, \sigma^2_{\alpha K})'$ contains fixed prior variances for the regression coefficients $\boldsymbol{\alpha}_{gk}$, common to all components and entries. In particular, we fix the common prior variance $\sigma^2_\alpha = 100$. The vector $\boldsymbol{\tau}^2_g = (\tau^2_{g1}, \ldots, \tau^2_{gK})'$ contains the smoothing parameters for the gth mixture component and $\mathbf{1}_m$ is an m-vector of ones. We assume independence between the regression coefficients $\boldsymbol{\alpha}_{gk}$ and the basis function coefficients $\boldsymbol{\beta}_{gk}$.

3.3.2 Priors on the smoothing parameters

We assume the smoothing parameter $\boldsymbol{\tau}^2_g = (\tau^2_{g1}, \ldots, \tau^2_{gK})'$ varies from components and entries. For the prior of the smoothing parameter, we assume it follows a half-t distribution as in Section 2.3.2 such that $\tau_{gk} \sim t^+_{\nu_\tau}(0, A_\tau)$, where ν_τ is the degree of freedom and A_τ is the scale parameter. We set $\nu_\tau = 3$ and $A_\tau = 10$ for all components and entries.

3.3.3 Priors on the error variances

Similar as priors for smoothing parameters in section 3.3.2, we assume the error variance $\boldsymbol{\sigma}^2_g = (\sigma^2_{g1}, \ldots, \sigma^2_{gK})'$ varies from components and entries. For the prior of the error variance, we assume it follows a half-t distribution such that $\sigma_{gk} \sim t^+_{\nu_\sigma}(0, A_\sigma)$, where ν_σ is the degree of freedom and A_σ is the scale parameter. We set $\nu_\sigma = 3$ and $A_\sigma = 10$ for all components and entries.

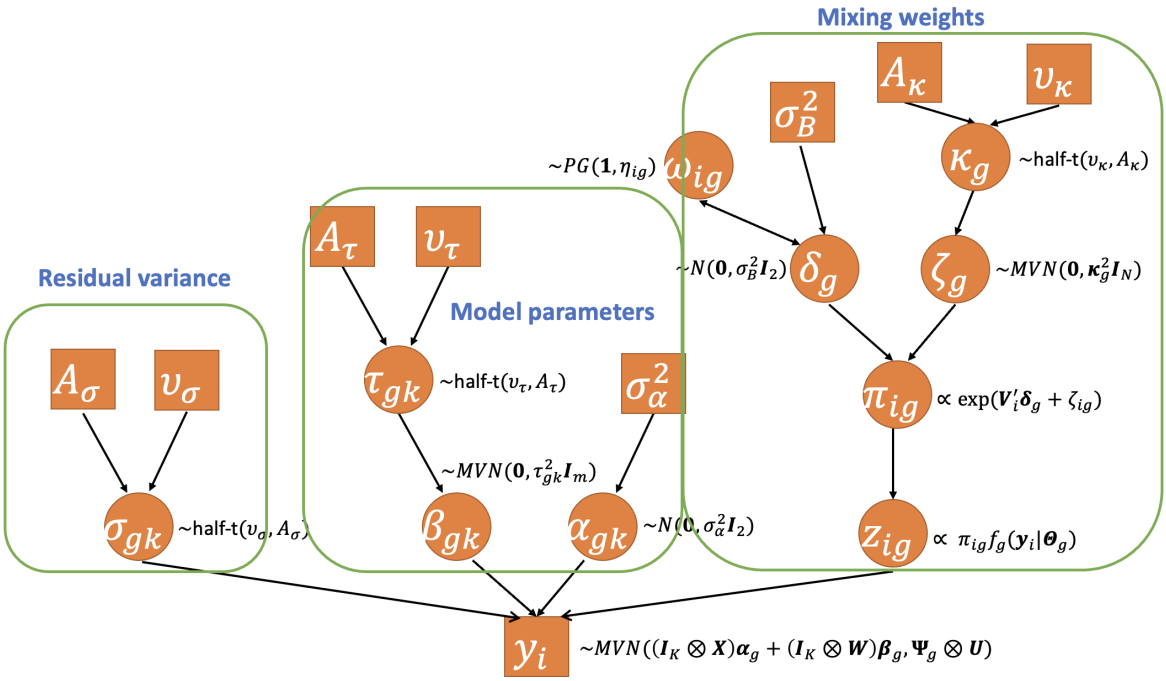
3.3.4 Priors on the logistic parameters and the variances of random intercepts

Priors for logistic parameters and variances of random intercepts are the same as the univariate case in Section 2.3.4. For the simplification of notations, we denote $\boldsymbol{\delta}_g^* = (\boldsymbol{\delta}_g^T, \boldsymbol{\zeta}_g^T)^T$, where $\boldsymbol{\zeta}_g = (\zeta_{1g}, \ldots, \zeta_{Ng})^T$. We also denote $\boldsymbol{V}_i^*$ as the covariates with a subvector indicating the random intercept for subject i and $\boldsymbol{V}^*$ is a matrix with each subject as a row. We assume the priors of logistic parameters $\boldsymbol{\delta}_g^* \sim N(\mathbf{0}, \boldsymbol{B}_g)$, where $\boldsymbol{B}_g = \text{diag}(\sigma^2_{\zeta \text{g}} \mathbf{1}_{\text{P}+1},\ \kappa^2_{\zeta \text{g}} \mathbf{1}_{\text{N}})$ and the priors of random intercept $\boldsymbol{\zeta}_g \sim N(\mathbf{0}, \kappa^2_{\zeta g} \boldsymbol{I}_N)$. For hyperparameters, we give a weakly informative prior with $\sigma^2_{\zeta g} = 10$ by assuming independence across all components and covariates. We also give a half-t hyperprior distribution for κ_g with $\kappa_g \sim t^+_{\nu_\kappa}(0, A_\kappa)$. We set $\nu_\kappa = 3$ and $A_\kappa = 10$ for all components.

For the estimates of logistic parameters using the Bayesian method, Polson et al. 2013 proposed a data augmentation strategy with the inclusion of a Pólya-Gamma latent variable that follows the Pólya-Gamma distribution. This data augmentation strategy has been shown to outperform other known data augmentation strategies with computational efficiency. Details about the implementation of Pólya-Gamma augmentation strategy are displayed in the sampling scheme in Appendix B.1.

For the purpose of explanation, Figure 8 shows a directed acyclic graph (DAG) for the Bayesian hierarchical structure of the proposed Bayesian mixture of spline experts model in the multivariate time series. Priors of error variances, model parameters, and mixing weights are in different branches.

Figure 8. Directed acyclic graph (DAG) for the Bayesian hierarchical structure.



3.3.5 Joint posterior distribution

Based on the augmented likelihood function in (12) and prior distributions in Section 3.3, the joint posterior distribution of $\boldsymbol{\Theta}$ can be written as:

$$
\begin{aligned}
f(\boldsymbol{\Theta} \mid \boldsymbol{y}, \boldsymbol{S}, \boldsymbol{V}^*) \propto & \prod_{i=1}^{N} \prod_{g=1}^{G} \{\pi_{ig} \prod_{j=1}^{n} f_g(\boldsymbol{y}_{it_j} \mid \boldsymbol{\Theta}_g)\}^{z_{ig}} \\
& \times \prod_{g=1}^{G} f_\theta(\boldsymbol{\theta}_g \mid \boldsymbol{\tau}_g^2, \sigma_\alpha^2) \times \prod_{g=1}^{G} f_\tau(\boldsymbol{\tau}_g^2 \mid \nu_\tau, A_\tau) \\
& \times \prod_{g=1}^{G} f_\sigma(\boldsymbol{\sigma}_g^2 \mid \nu_\sigma, A_\sigma) \times \prod_{g=1}^{G} f_\delta(\boldsymbol{\delta}_g^* \mid \sigma_{\zeta g}^2, \kappa_{\zeta g}^2) \\
& \times \prod_{g=1}^{G} f_\kappa(\kappa_{\zeta g}^2 \mid \nu_\kappa, A_\kappa),
\end{aligned}
\tag{14}
$$

where f_θ, for example, denotes the prior probability density function for the model parameters.

3.4 Sampling scheme

This section outlines the Gibbs steps for sampling from the conditional posterior distributions of all the model parameters. More details are given in Appendix B.1.

3.4.1 Gibbs sampling steps

Letting ℓ denote the current Gibbs sampling iteration, parameter values at the $(\ell+1)$th iteration are drawn according to the following steps.

1. Draw $\boldsymbol{\theta}_{gk}^{(\ell+1)}$ from $(\boldsymbol{\theta}_{gk}^{(\ell+1)} \mid \boldsymbol{y}, \boldsymbol{S}, \tau_{gk}^{2(\ell)}, \sigma_{gk}^{2(\ell)}) \sim N(\boldsymbol{u}_{gk}, \sigma_{gk}^2 \boldsymbol{\Lambda}_{gk})$, where $\boldsymbol{u}_{gk}$ and $\boldsymbol{\Lambda}_{gk}$ are mean vectors and covariance matrices.
2. Draw $\sigma_{gk}^{2(\ell+1)}$ from $(\sigma_{gk}^{2(\ell+1)} \mid \boldsymbol{\epsilon}_{igk}^{(\ell+1)}, a_{\sigma_{gk}}^{(\ell+1)}) \sim IG\Big((nN_g^{(\ell)} + \nu_\sigma)/2, \sum_{i=1}^{N} z_{ig} \boldsymbol{\epsilon}_{igk}' \boldsymbol{\epsilon}_{igk}/2 + \nu_\sigma / a_{\sigma_{gk}}\Big)$, where $N_g^{(\ell)}$ is the current number of subjects in the gth component, $\boldsymbol{\epsilon}_{igk}$ is the error vector for the gth component, the ith subject and the kth entry, and $a_{\sigma_{gk}}$ is a latent variable in the IG scale mixture underlying the half-t distribution.
3. Draw $\tau_{gk}^{2(\ell+1)}$ from $(\tau_{gk}^{2(\ell+1)} \mid \boldsymbol{\beta}_{gk}^{(\ell+1)}, a_{\tau_{gk}}^{(\ell+1)}) \sim IG\Big((\nu_\tau + m)/2, \boldsymbol{\beta}_{gk}' \boldsymbol{\beta}_{gk}/2 + \nu_\tau / a_{\tau_{gk}}\Big)$, where $a_{\tau_{gk}}$ is a latent variable as in 2.
4. Draw $\boldsymbol{\delta}_g^{*(\ell+1)}$ from $(\boldsymbol{\delta}_g^{*(\ell+1)} \mid \boldsymbol{V^*}, z_{ig}^{(\ell)}, \omega_{ig}^{(\ell+1)}, \kappa_{\zeta g}^{2(\ell)}) \sim N(\boldsymbol{M}_g, \boldsymbol{\Sigma}_g)$, where $\omega_{ig}^{(\ell+1)}$ is a Pólya-Gamma latent variable in the augmentation described in Section 3.3.4.
5. Draw $\kappa_{\zeta g}^{2(\ell+1)}$ from $(\kappa_{\zeta g}^{2(\ell+1)} \mid \boldsymbol{\zeta}_g^{(\ell+1)}, a_{\kappa_g}^{(\ell+1)}) \sim IG\Big(\nu_\kappa/2, \boldsymbol{\zeta}_g' \boldsymbol{\zeta}_g/2 + (\nu_\kappa + N)/a_{\kappa_g}\Big)$, where a_{κ_g} is a latent variable as in 2 and 3.
6. The mixing weights $\pi_{ig}^{(\ell+1)}$ are obtained by computing $p(\pi_{ig}^{(\ell+1)} \mid \boldsymbol{V}^*, \boldsymbol{\delta}_g^{*(\ell+1)}, z_{ig}^{(\ell)})$ from Equation (10).
7. Draw $z_{ig}^{(\ell+1)} \sim p(z_{ig}^{(\ell+1)} = 1 \mid \boldsymbol{y}, \boldsymbol{S}, \boldsymbol{\theta}_{gk}^{(\ell+1)}, \sigma_{gk}^{2(\ell+1)}, \pi_{ig}^{(\ell+1)})$ according to Equation (13).

3.4.2 Label switching

The issue of label switching still persists for our proposed Gibbs sampling algorithm. Section 2.4.2 has listed a set of solutions for solving the label switching issue. For the case of multivariate time series with Gibbs sampling, we use Equivalence Classes Representatives

(ECR) algorithm version 1 (Papastamoulis and Iliopoulos, 2010), which uses the simulated allocation variables as the equivalent allocation vectors.

ECR algorithm is based on the equivalent allocation vectors, which are mutually exclusive by simply permuting its labels. First, it partitions the set of allocation vectors $z_i^{(t)}$ into equivalence classes for each subject $i, i = 1, \ldots, n$ at each iteration $t = 1, \ldots, m$ and selects one representative from each class. Then the optimal permutation is determined by the one that reorders the allocations and is identical to the representative of its class. ECR algorithm version 1 uses only the vector of latent indicators as input and selects one representative at random. The ECR algorithm version 1 is given below.

1. Choose m initial permutations $\tau^{(t)}$ for each iteration $t = 1, \ldots, m$ (set to identity for each iteration).
2. Update $z_i^* = \text{mode}\{\tau z_i^{(t)}; t = 1, \ldots, m\}$ for each subject $i = 1, \ldots, n$.
3. For each iteration $t = 1, \ldots, m$, find the optimal permutation $\tau^{(t)}$ that maximizes $\sum_{i=1}^{n} I(\tau z_i^{(t)} = z_i^*)$.
4. If there is an improvement of $\sum_{t=1}^{m} \sum_{i=1}^{n} I(\tau z_i^{(t)} = z_i^*)$, go back to step 2, finish otherwise.

One advantage of ECR algorithm version 1 is this algorithm only needs allocation vectors for each MCMC iteration and it is computationally efficient compared to other methods listed in Section 2.4.2. More details about the ECR algorithm can be found in Papastamoulis and Iliopoulos 2010.

Papastamoulis 2015 developed an `R` package called `label.switching`, which includes the user-friendly R function to implement the ECR algorithm. We post-process our proposed Gibbs sampling results using the ECR algorithm to solve the label switching issue before making inferences.

3.4.3 Select number of components

In the second project, we did not apply the RJMCMC for the clustering of multivariate time series. Instead, we fit the proposed model and update model parameters using Gibbs sampling. Thus, the best model based on the number of components needs to be selected based on certain model selection criteria.

Akaike information criterion (AIC) and Bayesian information criterion (BIC) are the two most widely-used model selection criteria for the frequentist way. However, Spiegelhalter et al. 2014 argued that AIC is not asymptotically consistent since it is not seeking to select the true model. Gelman et al. 1995 stated that BIC is not intended to assess the model performance and it is completely possible for a complex model to have a good prediction performance while having a relatively high BIC due to a large penalty function. Hence, it is necessary to seek other model selection criteria which may have a good performance under the Bayesian framework.

Spiegelhalter et al. 2002 suggested the use of deviance information criterion (DIC) for a Bayesian criterion of model selection by a measure of an effective number of parameters. DIC under the Bayesian framework is defined as

$$\begin{aligned} \mathrm{DIC}_1 &= \overline{D(\theta)} + p_D \\ &= \boldsymbol{E}_{\theta|y}\big[-4 \log p(\boldsymbol{y} \mid \boldsymbol{\theta})\big] + 2 \log p(\boldsymbol{y} \mid \bar{\boldsymbol{\theta}}), \end{aligned}$$

where $\overline{D(\theta)}$ is the posterior mean of deviance, p_D is the proposed effective number of parameters or the penalty function. $\bar{\boldsymbol{\theta}}$ is the posterior mean.

Many works have been done for the improvements or the extension of DIC to other models. Celeux et al. 2006 extended the use of DIC to different areas including missing data models, random effect models and mixture models with multiple versions of DICs. Kim 2021 provided theoretical properties of DICs proposed by Celeux et al. 2006. Watanabe and Opper 2010 proposed a fully Bayesian approach for evaluating the model performance called Watanabe-Akaike information criterion (WAIC), with the desirable property of averaging over the posterior distribution rather than conditioning on a point estimate. Considering the complexity of our proposed model, we adopt the DIC to assess the performance of a set of proposed models varying by the number of components.

In a review paper by Spiegelhalter et al., many criticisms about the DIC in Spiegelhalter et al. have been discussed. Firstly, p_D is not invariant to reparameterization and can be negative if the posterior of θ is very skewed. Thus, $\bar{\theta}$ does not give an accurate estimate of θ. Secondly, it suffers from the lack of consistency if the label switching issue persists due

to the overlapping of posterior distributions in mixtures.

Gelman et al. 2003 introduced an alternative measure of an effective number of parameters using the variance log predictive density over MCMC iterations. DIC can be expressed as

$$\mathrm{DIC}_{2=} - 2\log p(\boldsymbol{y} \mid \bar{\boldsymbol{\theta}}) + 4V_{\ell=1}^{L}[\log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{\ell})],$$

where $V_{\ell=1}^{L}\big[\log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{\ell})\big] = \frac{1}{L-1}\sum_{\ell=1}^{L}\big[\log p(\boldsymbol{y} \mid \boldsymbol{\theta}^{\ell}) - \overline{\log p(\boldsymbol{y} \mid \boldsymbol{\theta})}\big]^2$ and $\overline{\log p(\boldsymbol{y} \mid \boldsymbol{\theta})}$ is the averaged log predictive density over all MCMC iterations. This DIC is remarkably robust and more accurate than the original DIC in Spiegelhalter et al.. Moreover, this DIC has the advantage of always being positive and not affected by reparameterizations (Gelman et al., 2003; Spiegelhalter et al., 2014).

For the evaluation of models varying from the different numbers of components in our project, we adopted DIC_2 as the model selection criterion by extending it into the framework of the mixtures of the time series model.

3.5 Simulation studies

We have conducted two sets of simulation studies. Simulation I aims to evaluate the performance of the proposed penalized splines model under different true models. Simulation II aims to evaluate the performance of our proposed method by comparing it to other existing methods.

3.5.1 Simulation I: Evaluate the performance of the proposed method under different true models

We conducted simulation I based on two scenarios. Scenario 1 generates a collection of bivariate time series with four components. Scenario 2 generates a collection of trivariate time series with two components. For each scenario, we generated $s = 100$ replicates, each with a collection of $N = 150$ multivariate time series of length $n = 70$ for each variate.

Two scenarios both used $m = 10$ number of basis functions for the proposed model and with $P = 4$ number of covariates (including the intercept). Different model parameters such as basis function coefficients, linear coefficients, error variances, and smoothing parameters were given for each component g and each variate k for each scenario.

For each scenario, we generated four model settings, which were listed as M1, M2, M3 and M4. M1 and M2 generate a mixture model where each mixture follows a cubic model on time. M1 fits a Bayesian cubic model with noninformative priors and a covariate-guided structure using generated covariates on generated data, while our proposed Bayesian mixture of spline experts model is fitted for M2. M3 and M4 generate a mixture model where each mixture follows a regression splines model on time. M3 fits a Bayesian regression splines model with noninformative priors on spline coefficients with no smoothing parameter, while M4 utilizes our proposed method, which introduces the regularization of roughness by adding the smoothing parameter. In general, our proposed method implements a penalized splines model by adding a hyperprior distribution of the prior variance of model coefficients. The prior variance of model coefficients is assumed to be random and thus can control the smoothness of the model as compared to the ridge regression in the frequentist way.

From above, M1 and M2 have the same underlying true model but are evaluated by either the true or proposed penalized splines model. So as M3 and M4. Data generation steps of each replicate for the mixture regression splines model are shown below:

1. Generate fixed and evenly-spaced time points with a length of 70 from 0 to 5
2. Generate 10 sets of basis functions based on time
3. Set true values of parameters τ_{gk}^2, σ_{gk}^2,$\boldsymbol{\alpha}_{gk}$ and generate $\boldsymbol{\beta}_{gk} \sim N(\mathbf{0}, \tau_{gk}^2 \boldsymbol{I}_m)$ for each component g and entry k
4. Generate values of four covariates (include the intercept) for each multivariate time series
5. Set true logistic parameters $\boldsymbol{\delta}_g$
6. Compute mixing weights based on covariates and true logistic parameters
7. Sample latent indicators z_{ig} for each multivariate time series and each component
8. For each component, simulate each multivariate time series $\boldsymbol{y}_i(t)$ as

$$\{\boldsymbol{y}_i(t) \mid z_{ig} = 1\} = \boldsymbol{\alpha}_{0g} + \boldsymbol{\alpha}_{1g} t + \boldsymbol{\beta}_{1g} W_1(t) + \ldots + \boldsymbol{\beta}_{mg} W_m(t) + \boldsymbol{\epsilon}_{igt},$$

where $\boldsymbol{\alpha}_{0g}$ and $\boldsymbol{\alpha}_{1g}$ are multivariate intercept and slope for each component g, $\boldsymbol{\beta}_g = (\boldsymbol{\beta}'_{1g}, \ldots, \boldsymbol{\beta}'_{mg})'$ is a vector of multivariate spline coefficients for component g, $\{W_1(t), \ldots, W_m(t)\}$ are m basis functions, and $\boldsymbol{\epsilon}_{gt}$ are independent zero-mean multivariate Gaussian random variable as

$$\boldsymbol{\epsilon}_{gt} \sim \text{MVN}\big(\mathbf{0}, \text{diag}(\sigma^2_{g1}, \ldots, \sigma^2_{gK})\big).$$

Table 5 shows the results of logistic parameters for the four-component bivariate scenario in terms of RMSE (bias, variance) for simulation I. Component 4 is always used as the reference component and different components are listed as C1, C2, C3 and C4 in the table. δ_0,δ_1,δ_2 and δ_3 are the intercept and three covariates. Comparing the results of M1 vs M2, M3 vs M4 for which true models are the same, we found that RMSE, bias and variance for each covariate and each comparison are comparable without any large differences. Those findings are reasonable since our proposed penalized splines model does not affect estimates of logistic parameters if all multivariate time series are assigned to the correct cluster. Those similar RMSE of logistic parameters demonstrate the stability of our proposed model and the capability of accurate clustering classification.

We also investigate the performance of estimated trajectories for each component by calculating the averaged root square error (ARSE) of each component

$$\text{ARSE} = \sqrt{\frac{1}{nK}\sum_{j=1}^{n}\sum_{k=1}^{K}\Big(\hat{y}_{kj} - y_{kj}\Big)^2},$$

where $\hat{y}_{kj}$ is the estimated value of the true y_{kj} for the kth entry and the jth time point. All estimated values are posterior means. In addition to ARSE, we also report the averaged bias (A-bias) and the variance of the bias (V-bias), where

$$\text{A-bias} = \frac{1}{nK}\sum_{j=1}^{n}\sum_{k=1}^{K}\Big(\hat{y}_{kj} - y_{kj}\Big),$$

and V-bias is computed by calculating the sample variance of the bias over entries and time points. Boxplots of ARSE, A-bias, and V-bias for each component are given in Figures 9 and 10.

In terms of M1 vs M2 in Figure 9, unsurprisingly, the Bayesian cubic model outperforms

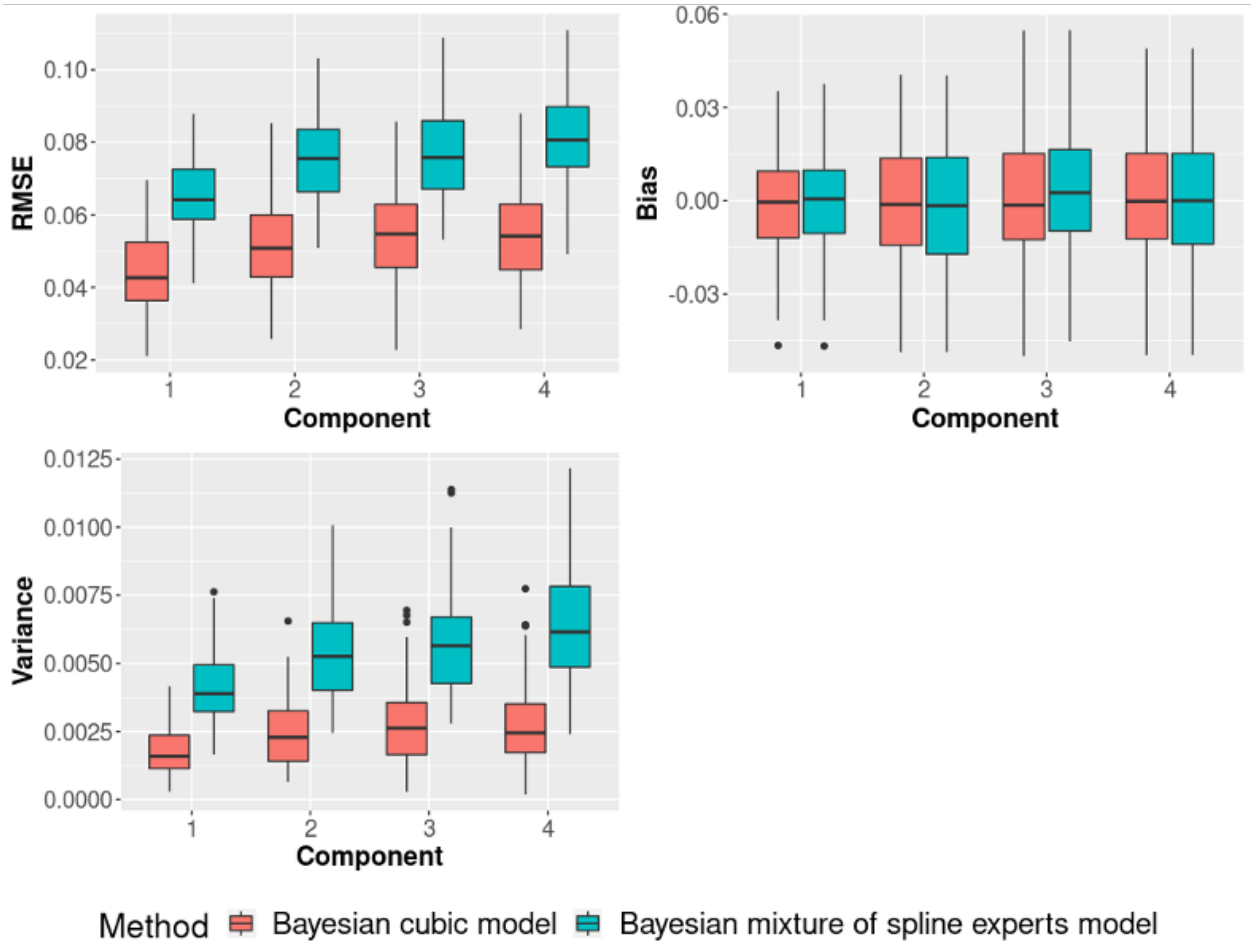
Table 5. Results of logistic parameters for the four-component bivariate scenario in Simulation I: values in each cell are in the format of RMSE (bias, variance).

Model setting	Comparison	δ_0	δ_1	δ_2	δ_3
M1	C1 vs C4	0.80 (0.01, 0.65)	0.52 (-0.06,0.27)	0.29 (-0.01, 0.09)	0.41 (-0.05, 0.17)
	C2 vs C4	0.99 (0.03, 0.98)	0.49 (0.13, 0.22)	0.49 (-0.20, 0.20)	0.37 (-0.03, 0.14)
	C3 vs C4	0.81 (-0.22, 0.61)	0.40 (0.07, 0.15)	0.27 (0.004,0.07)	0.30 (-0.03, 0.09)
M2	C1 vs C4	0.75 (0.02, 0.57)	0.57 (-0.05,0.32)	0.30 (-0.01, 0.09)	0.42 (-0.04, 0.18)
	C2 vs C4	0.96 (0.06, 0.93)	0.44 (0.12, 0.18)	0.45 (-0.20, 0.17)	0.38 (-0.06, 0.14)
	C3 vs C4	0.79 (-0.23, 0.57)	0.40 (0.05, 0.16)	0.27 (0.02, 0.07)	0.29 (-0.03, 0.08)
M3	C1 vs C4	0.76 (-0.09, 0.57)	0.44 (-0.05,0.20)	0.30 (0.05,0.09)	0.34 (-0.06,0.11)
	C2 vs C4	1.48 (0.29, 2.13)	0.74 (0.01,0.56)	0.55 (-0.10,0.30)	0.34 (0.02,0.11)
	C3 vs C4	1.02 (-0.30, 0.97)	0.54 (0.11,0.28)	0.34 (-0.01,0.12)	0.31 (-0.06, 0.09)
M4	C1 vs C4	0.77 (-0.14,0.58)	0.47 (-0.01,0.23)	0.31 (0.04,0.10)	0.35 (-0.05,0.12)
	C2 vs C4	1.47 (0.23,2.14)	0.75 (0.05,0.57)	0.57 (-0.13,0.31)	0.37 (0.03,0.14)
	C3 vs C4	0.99 (-0.28,0.91)	0.56 (0.14,0.30)	0.35 (-0.04,0.12)	0.32 (-0.05,0.10)

our proposed model since it uses the true model. However, our proposed method still shows a relatively good performance with small ARSE values and a similar bias level compared to the Bayesian cubic model. In terms of M3 vs M4 in Figure 10, our proposed model outperforms the Bayesian regression splines model by adding the regularization using smoothing parameters. Adding smoothing parameters will lower ARSE by decreasing the variance estimates. This is demonstrated with boxplots in 10 in that our proposed Bayesian mixture of spline experts model exhibits smaller ARSE in terms of variance compared to the Bayesian regression splines model.

In conclusion, Simulation I demonstrates the flexible use of our proposed Bayesian penalized splines model even if the true model follows a certain parametric model. Additional simulation results of the two-component trivariate model are placed in Appendix B.2.

Figure 9. Boxplots of RMSE, bias and variance of trajectory estimates for model setting M1 vs M2 in Simulation I: four-component bivariate model.

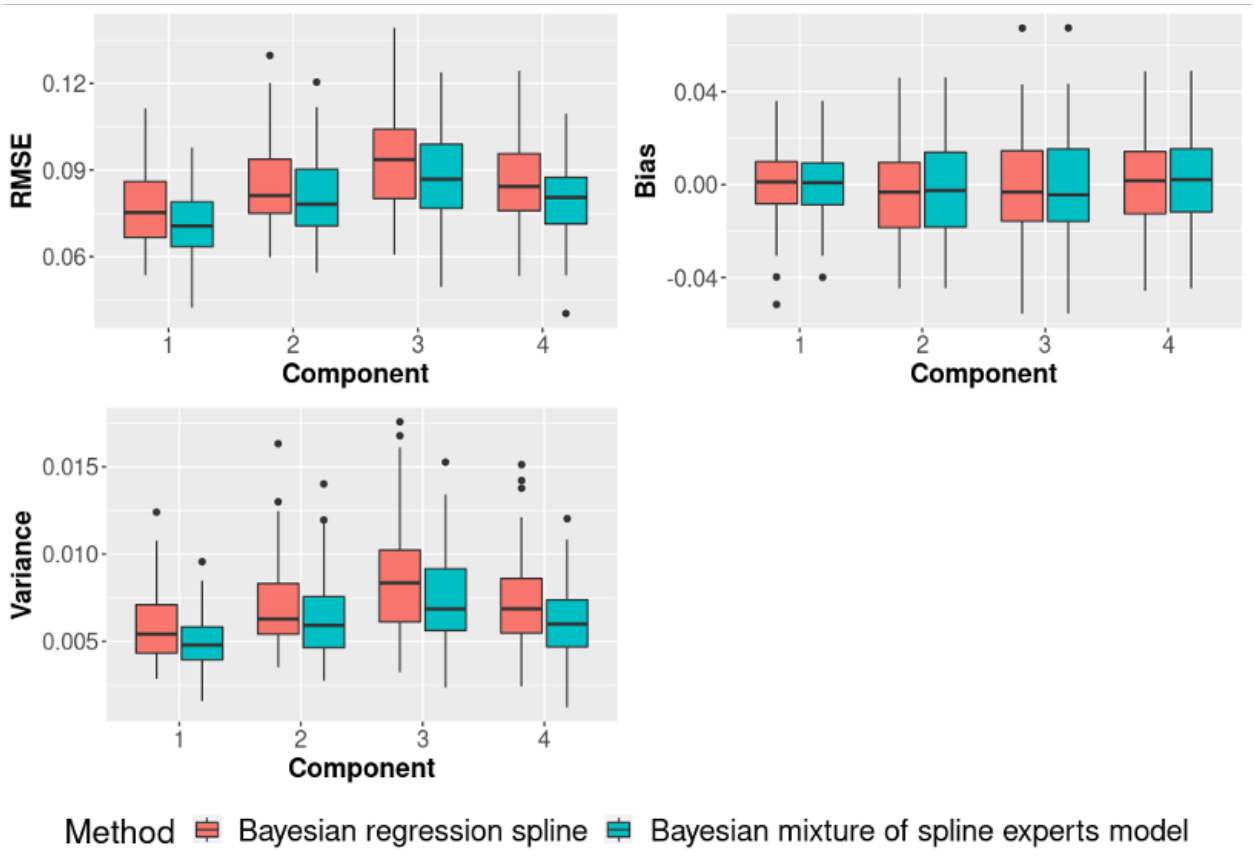


3.5.2 Simulation II: Comparing the performance of the proposed model to other existing methods

In simulation II, we compare the performance of our proposed model to other existing methods including `TRAJ` procedure in `SAS` and `gbmt R` package.

`TRAJ` performs a group-based trajectory modeling based on longitudinal data (Jones et al., 2001). Nagin et al. 2018 extended this method to the multivariate trajectory modeling based on a polynomial model with different orders. `TRAJ` assumes conditional on the group membership, longitudinal outcomes are independently distributed. But longitudinal outcomes are not conditionally independent at the population level. It obtains MLE using

Figure 10. Boxplots of RMSE, bias and variance of trajectory estimates for model setting M3 vs M4 in Simulation I: four-component bivariate model.



a numerical quasi-Newton procedure with iterations and it allows the inclusion of covariates via a class membership model.

`gbmt` is an abbreviation for group-based multivariate trajectory and it is developed by Magrini 2022. This method also uses a polynomial model to construct group trajectories and EM algorithm is used to update model parameters. However, this method does not allow to use covariates to predict mixing weights.

To demonstrate the performance of the proposed method, we conduct simulation studies by generating data sets from the proposed model under two scenarios: a two-component mixture ($G = 2$) of trivariate time series ($K = 3$) and a four-component mixture ($G = 4$) of bivariate time series ($K = 2$). We simulate 100 replicates in each simulation setting with $N = 150$ time series of length $n = 50$. A total of $20,000$ Gibbs sampling iterations are

Table 6. Root mean square errors (RMSEs) of each logistic parameter for the four-component bivariate model from 100 replicates of 150 four-component bivariate time series of length 50. RMSEs of the proposed method were compared to `TRAJ` procedure in `SAS`. Parameters δ_0, δ_1, δ_2 and δ_3 are intercept, first, second and third logistic parameters, respectively. The fourth component was used as the reference component. The true values of logistic parameters are $5, -3.5, 1, 0.1$ (first component), $-4, 2.5, -2, -0.2$ (second component), $3, -2, 0.8, 0.2$ (third component). C1, C2, C3 and C4 denote first, second, third and fourth components, respectively.

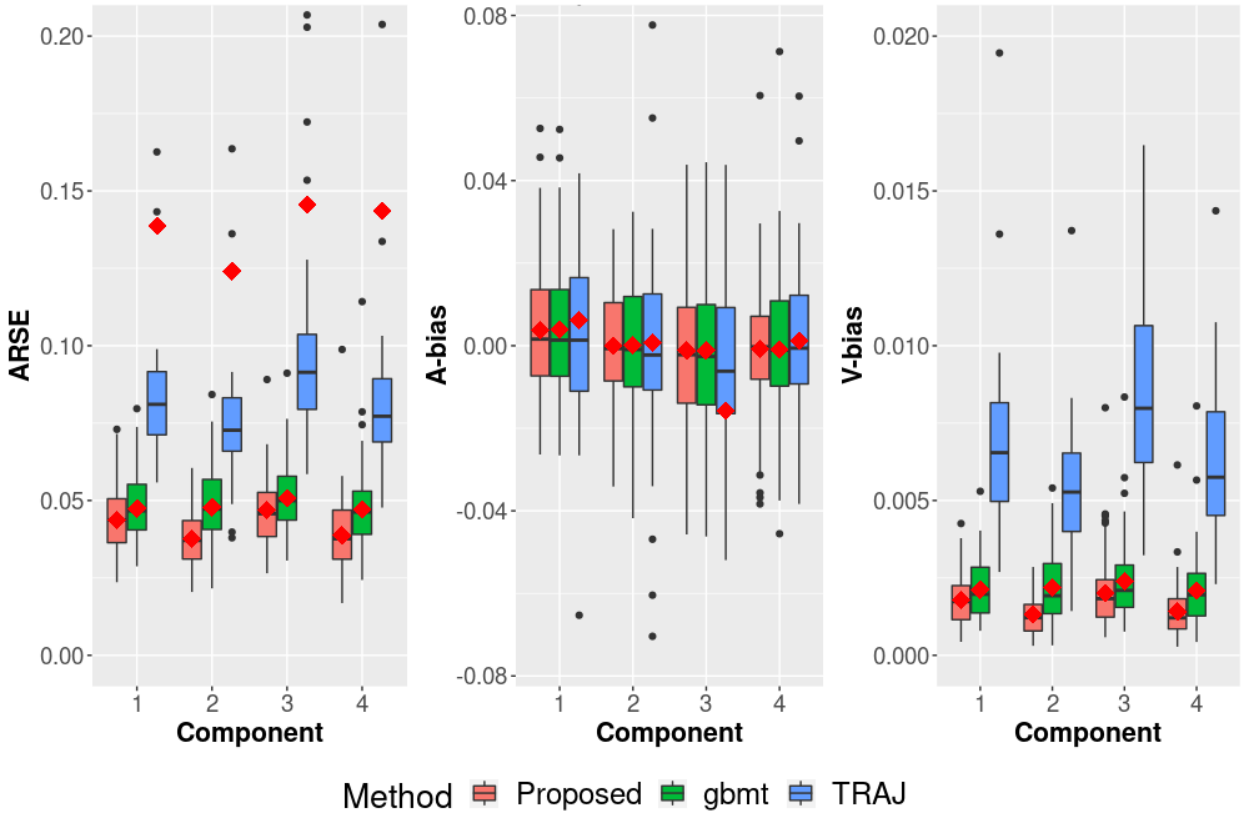
n	N	Method	Comparison	δ_0	δ_1	δ_2	δ_3
50	150	Proposed	C1 vs C4	0.81	0.53	0.30	0.39
			C2 vs C4	1.11	0.46	0.42	0.36
			C3 vs C4	0.89	0.42	0.28	0.34
		TRAJ	C1 vs C4	1.20	0.74	0.35	0.41
			C2 vs C4	3.81	2.27	1.33	0.49
			C3 vs C4	2.07	1.33	0.76	0.32

run with a burn-in of $4,000$. Data generation steps are similar to simulation I. Table 6 and Figure 11 display simulation results of logistic parameters and trajectory estimates for the 4-component bivariate scenario.

For the performance of estimates of logistic parameters, our proposed method outperforms `TRAJ` for almost all comparisons and covariates. The smaller RMSE comes from both bias and variances. This demonstrates the importance of adding the shrinkage parameter or penalty in the multinomial logistic model. `TRAJ` fits a multinomial logistic model without the penalty term, which may lead to biased estimates and inflated coefficients in some extreme cases such as the unbalanced design and perfect separation. Our proposed method adds the penalty by specifying a prior variance for logistic parameters, thus resulting in more accurate estimates.

For the performance of trajectory estimates, Figure 11 compares our proposed method

Figure 11. Boxplots of the averaged root square error (ARSE), the averaged bias (A-bias) and the variance of bias (V-bias) of estimated trajectories for each component from 100 replicates of 150 four-component bivariate time series of length 50. Estimates of the proposed method were compared to `R` package `gbmt` and `TRAJ` procedure in `SAS`. The diamond markers denote the means of each estimate. All boxplots are zoomed in for better visualization.



with `gbmt` and `TRAJ`. The proposed method is able to outperform both `gbmt` and `TRAJ` in terms of ARSE. The smaller ARSE comes from the variance estimates. This is within our expectation in that our proposed method uses a penalized splines model for each component, which aims to reduce ARSE in terms of reducing variance. The red diamond marks in Figure 11 gives the mean for each component and each method. Notably, `TRAJ` has a larger mean of variance across replicates, which could be an indication that `TRAJ` has several poor trajectory estimates. Our proposed method is stable with no obvious outliers in boxplots.

More simulation results of different numbers of time series N and time series length n

under two scenarios (two-component trivariate and four-component bivariate) are displayed in Appendix B.2.

3.6 Real-data application results

We apply our proposed method to the analysis of the fNIRS still-face study introduced in Chapter 1. Six covariates are included in our covariate-guided model, including Infant Behavior Questionnaire-Revised negative emotionality (IBQ-NE) score, Infant Behavior Questionnaire-Revised effortful control (IBQ-EC) score, gestational age (in Days), infant age (in Months), head circumference (in cm) and sex. All continuous covariates are centered and scaled to follow the standard normal distribution. We set the number of basis functions $m = 20$ and run a total of $30,000$ Gibbs iterations with a burn-in period of $6,000$. The same values of hyperparameters are used as in the simulation studies.

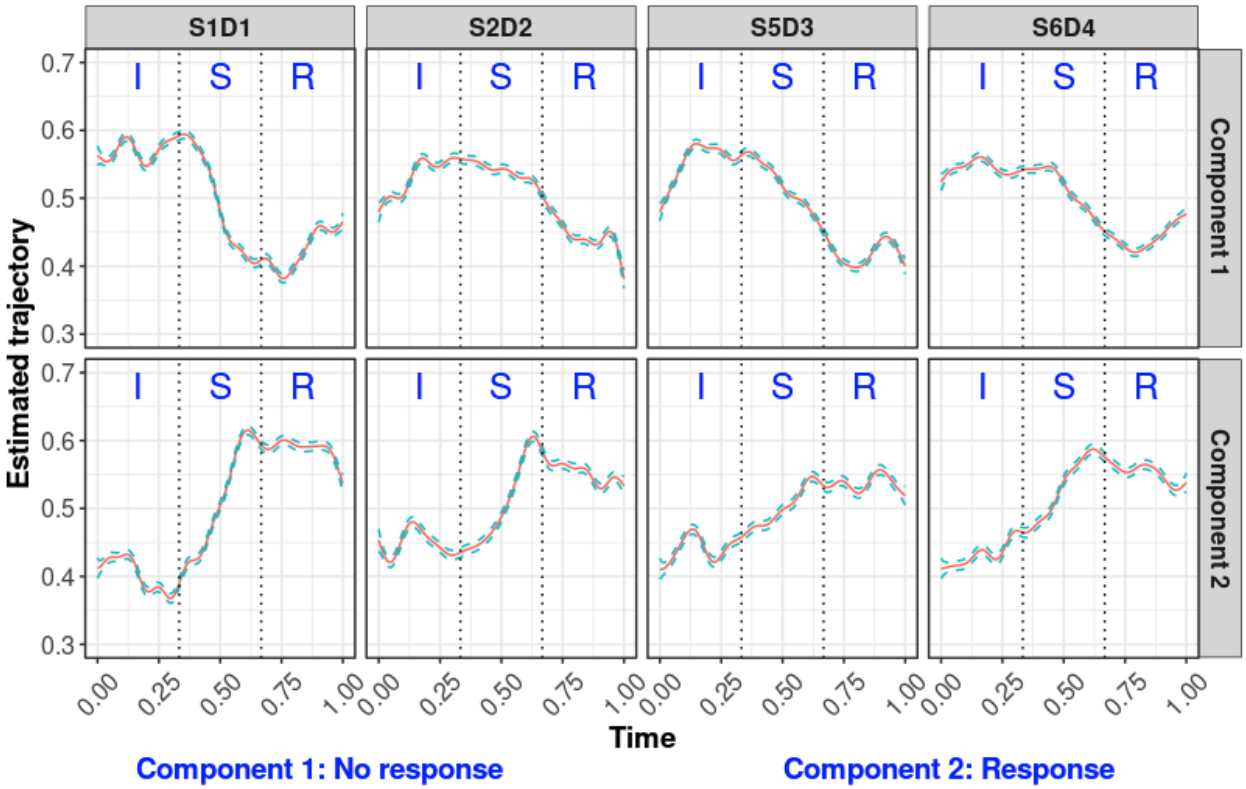
Infant Behavior Questionnaire-Revised (IBQ-R) is a widely-used parent-report measure developed to assess dimensions of temperament along multiple scales (Gartstein and Rothbart, 2003). The IBQ-NE construct combines data from the following subscales: Sadness, Distress to Limitations, Fear, and Falling Reactivity/Rate of Recovery from Distress. IBQ-EC refers to the ability to inhibit a dominant response to perform a subdominant one and has been shown to be protective against a myriad of difficulties (Gartstein et al., 2013). We assume that those covariates are associated with the mixing weights of each component and the trajectory pattern of that component. Through data pre-processing steps in Section 1.3.4, we had a sample size of 79 subjects, each with multivariate time series from twelve channels. Each univariate time series consists of a grid of 1500 time points and 500 measurements for each still-face period (Interact, still-face and recovery). Here we presented a set of four-channel results and all-channel results.

3.6.1 Four-channel analysis

We presented the results of the four-channel analysis by selecting four channels from four detector. Four selected channels are S1D1, S2D2, S5D3 and S6D4. S1D1 and S5D3 are in the center prefrontal region, while S2D2 and S6D4 are in the left and right prefrontal regions, respectively. We fitted our proposed model with the number of components varying from 2 to 6. Based on values of DIC_2 in Section 3.4.3, the 2-component model was selected as the best model for this four-channel analysis.

Figure 12 gives estimated trajectories of the 2-component model with four selected channels. We are interested in brain activation signals in the still-face period while the interact period is used as the reference level. For component 1, a decreasing trajectory is observed for the still-face period in all four channels. In contrast, an increasing trend is observed for the still-face period in all four channels for component 2. We define component 1 as the no response component and component 2 as the response component based on trajectory patterns in the still-face period. Figure 13 shows logistic coefficient estimates for all covariates in the 2-component model. Component 2 is used as the reference component in the covariate-guide model. IBQ-NE score is a significant covariate and 95% pointwise credible interval does not include zero. Positive coefficients of IBQ-NE indicate that a higher IBQ-NE score is associated with component 1, which has decreased brain activation levels in the still-face period for all four selected channels. Though other logistic coefficients have 95% credible intervals including zero, the negative posterior mean estimate of the IBQ-EC score could still be evidence indicating that a high IBQ-EC score is associated with an increasing brain activation as shown for component 2. These conclusions are consistent with findings in Gartstein et al. (2013) that IBQ-NE is negatively associated with IBQ-EC. Enlow et al. (2016) reported a negative association between activity level and IBQ-NE among infants whose families encourage a high level of activity. Furthermore, a negative posterior mean of infant age may suggest that younger infant tends to have a decreasing brain activation level in the still-face period. Results of another four-channel analysis are given in Appendix B.3.

Figure 12. Estimated trajectories of the two-component model with four selected channels. **I**: Interact **S**: Still-face **R**: Recovery.



3.6.2 All-channel analysis

We conducted an all-channel analysis by fitting the proposed model for all twelve channels. Based on values of DIC_2 in Section 3.4.3, the three-component model was selected as the best model for this all-channel analysis.

Figure 14, 15 and 16 provide estimated trajectories of the three-component model with all channels. Component 1 is considered as the no response component since decreasing trajectory patterns were found in this component for all channels. Component 3 is considered the response component since increasing brain activities were discovered for this component for most of the channels. In addition, we have an additional component which we call the mixture response component since it involves different trajectory trends in the still-face period. For example, in Figure 15, channel S1D1 and S3D2 have increased brain activity in

the still-face period, while for channel S6D4 and S8D4 we observe a more wiggly trend.

Logistic coefficient estimates for the all-channel analysis are shown in Figure 17. Component 3 is used as the reference component. 95% pointwise credible intervals of all covariates contain zero, which indicates that the association between covariates and components is not strong. However, the positive posterior mean of IBQ-NE of no response component still suggested that a higher IBQ-NE score is related to the no response component with decreasing brain activity in the still-face period, which achieves the same conclusion as for the previous four-channel analysis and in Enlow et al. (2016). In contrast to the four-channel analysis, the positive posterior mean estimate of the mixture component indicates that the older infant is related to the mixture response component, which exhibits a discordant trajectory pattern among twelve channels.

Heatmaps with averaged first derivatives among certain grids of time for estimated trajectories are an effective visualization method to uncover different trajectory patterns for different components. Heatmaps of the four-channel and all-channel analyses are placed in Appendix B.3.

Figure 13. Logistic coefficient estimates and 95% pointwise credible intervals of the two-component model.

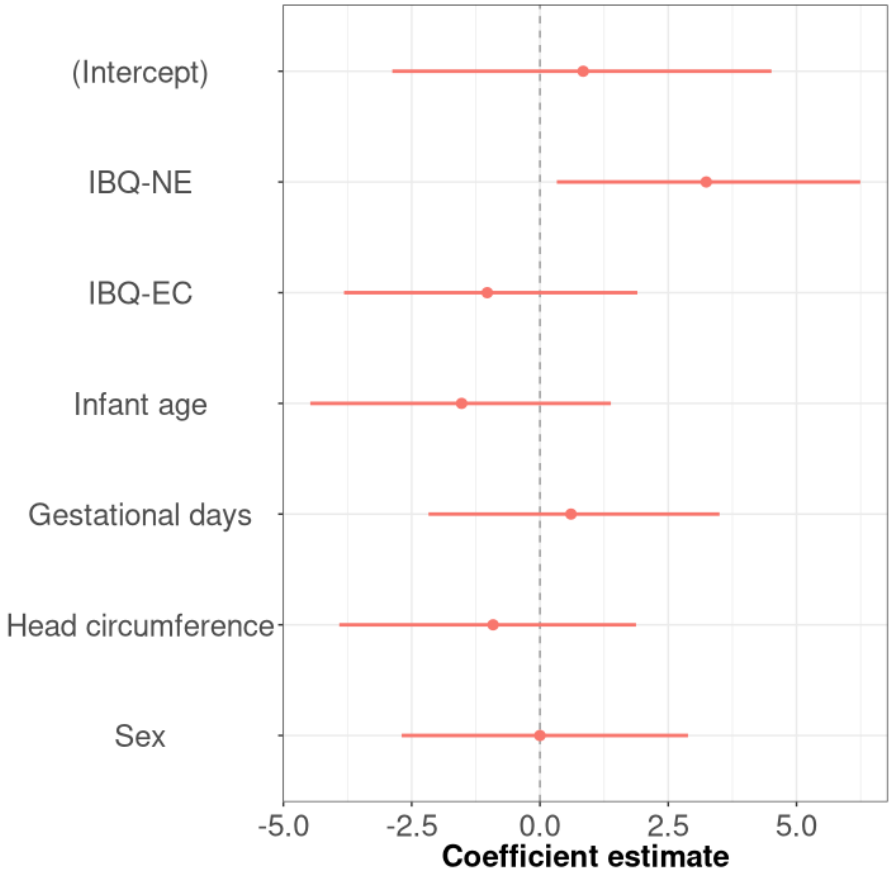


Figure 14. Estimated trajectories of component 1 for a three-component model with all channels. **I**: Interact **S**: Still-face **R**: Recovery.

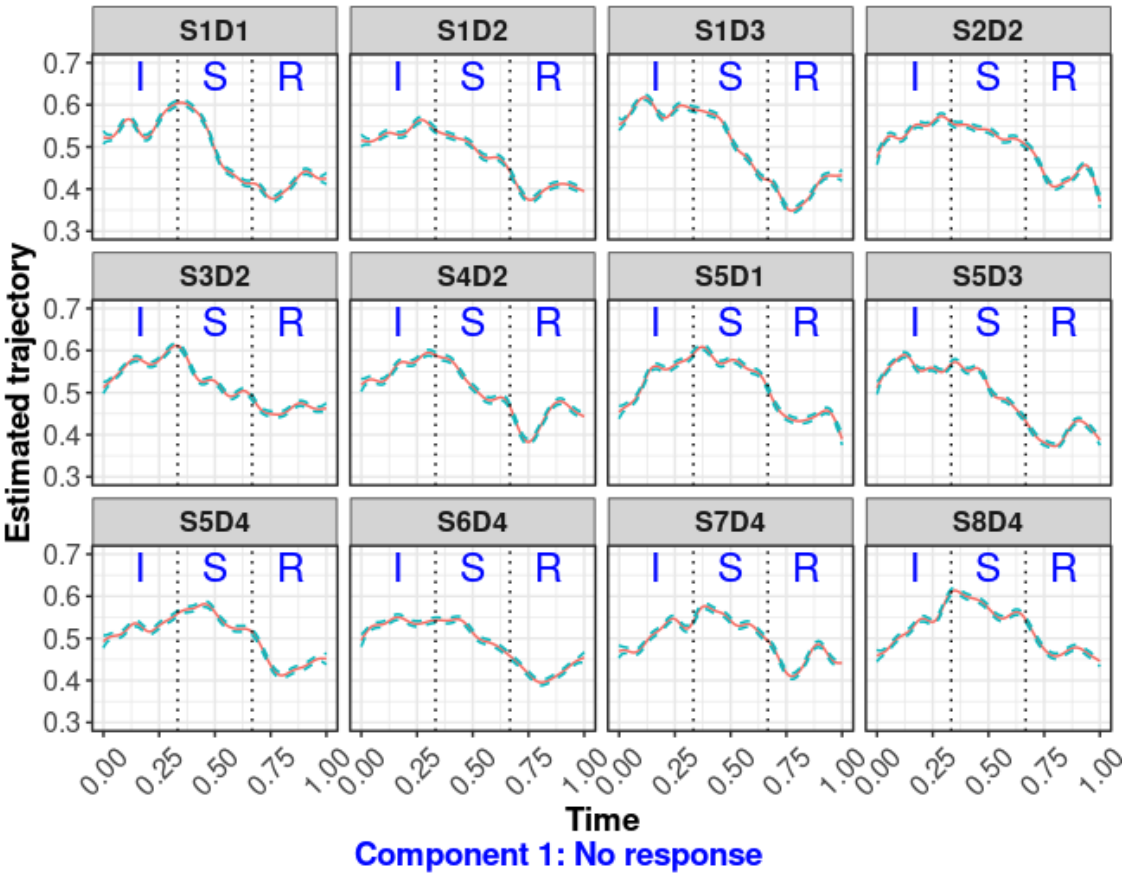


Figure 15. Estimated trajectories of component 2 for a three-component model with all channels. **I**: Interact **S**: Still-face **R**: Recovery.

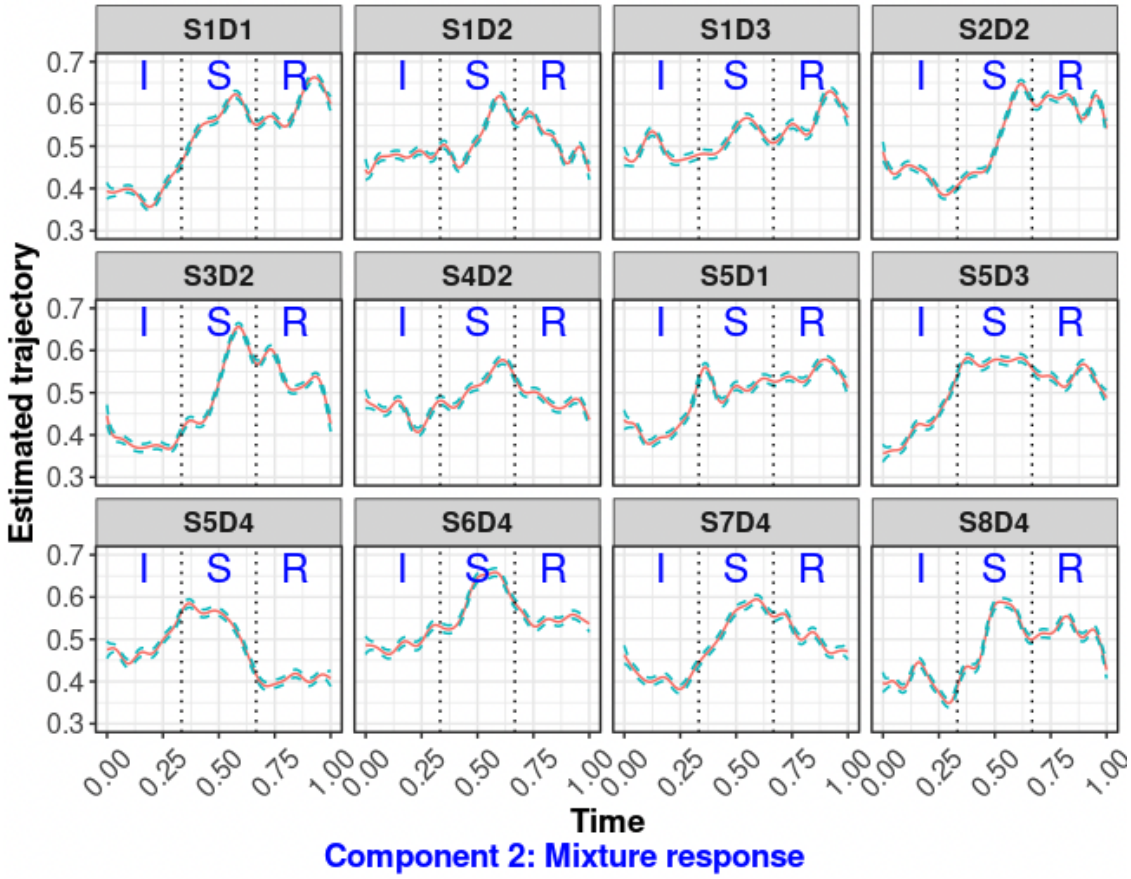


Figure 16. Estimated trajectories of component 3 for a three-component model with all channels. **I**: Interact **S**: Still-face **R**: Recovery.

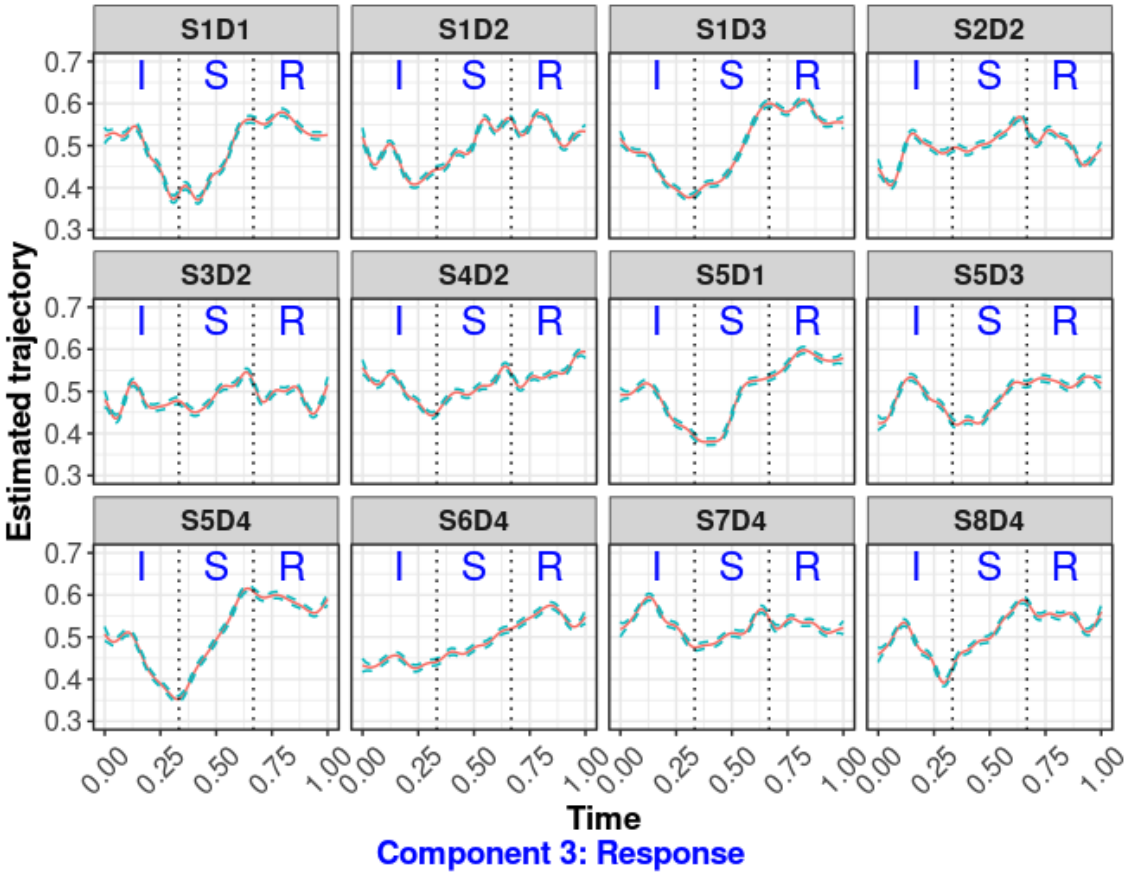
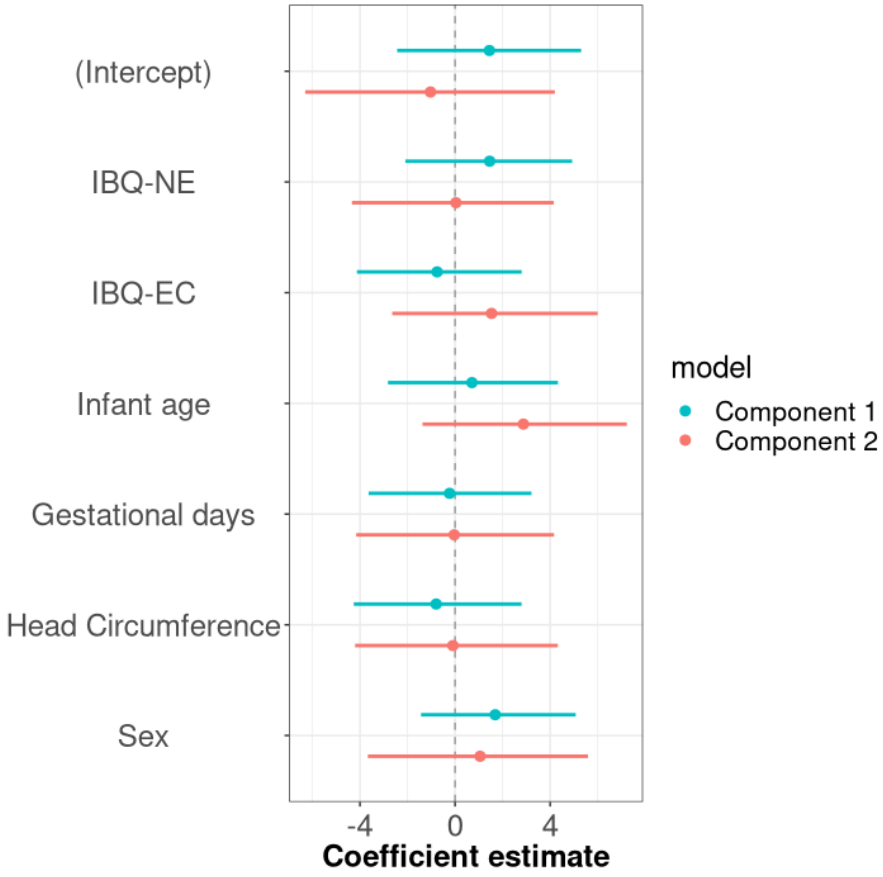


Figure 17. Logistic coefficient estimates and 95% pointwise credible interval of the three-component model.



3.7 Discussion

Our proposed covariate-guided Bayesian mixture of spline experts model aims to perform a model-based clustering of multivariate time series from multiple subjects. The proposed method implements a Bayesian penalized splines approach to fit the model of each mixture component, while covariates are incorporated to compute mixing weights. The performance of our proposed method is illustrated in two simulation studies in Section 3.5. Simulation I demonstrates that the Bayesian penalized splines model we use is flexible enough to be able to have a good fit for a parametric model. In the real-world setting, the true model is often unknown and non-parametric. A non-parametric splines model with the regularization of smoothness is one of the good fits for different types of data. Simulation II compares our proposed method to two existing methods and results have shown that our proposed method outperforms other methods in terms of both trajectory estimates and logistic parameter estimates by introducing penalty terms.

We apply our proposed method to an fNIRS still-face study, aiming to discover trajectory patterns of different components and how selected covariates are associated with patterns of components. Through results from four-channel and all-channel analyses, we conclude that clear patterns of response and no response component can be discovered by both analyses based on estimated trajectories within the still-face period. To the best of our knowledge, this is the first still-face study using the fNIRS technique with the purpose of finding trajectory components and covariates that are associated with a certain component. To be more specific in estimated trajectories, we also observe an increasing trend in the recovery period of component 1, which can be seen in Figure 12 and 14 for some channels. This may be an indication of a 'delayed' effect from the still-face period. Infants belonging to component 1 tend to have delayed brain activation from the still-face period, which are reflected in the later recovery period. In terms of covariates in multinomial logits, we find that a high IBQ-NE score is associated with the decreasing brain activation level for the all-channel analysis and the four-channel analysis including channels S1D1, S2D2, S5D3 and S6D4. We are able to reach the same conclusion as the univariate case that a high IBQ-NE score is associated with a lack of response in the still-face period, which could be due to infant's

failure to regulate emotions. In addition, young infants are not able to attempt to regulate their emotions compared to old infants, which indicate that they might have decreasing brain activity or response in the still-face period. Lastly, a high IBQ-EC score is related to the increasing brain activity in that infants with a high IBQ-EC score are able to regulate their emotions and behavior, which might result in responses or increasing blood flowing in the still-face period due to these regulations.

Our proposed method has some limitations. First, the proposed method assumes independence between different time series entries. However, there exist different ways of spatial correlations among different fNIRS channels. A multilevel multivariate model can be a choice to incorporate correlations within a level. Secondly, our proposed method will not work well in the case that there is large heterogeneity within entries. Our method assumes components for all entries and if large amount of differences are present within entries, our proposed method will fail to give a parsimonious number of components. Lastly, our proposed method uses DIC to select the best number of components since RJMCMC introduces severe label switching issues and components are indistinguishable for models with more components. However, there should be no best model and model selection based on DIC is not always the best approach. Model averaging approaches such as Bayesian model averaging and stacking could be considered later.

4.0 Bayesian generalized LASSO on selected orders of differences via the horseshoe prior with application to functional regression

4.1 Introduction

With the emergence of modern technologies to record data, functional data analysis has become an increasingly important analytic tool in many areas such as health and economy. Functional regression is one of the most popular tools in the area of functional data analysis, where the response, the predictors, or both the response and predictors are functional. Scalar-on-functional regression refers to functional regression with a scalar response y_i for each subject i and functional predictors over a period of time. Sometimes we refer $\beta(t)$ as the coefficient function. Many books and review papers have been focusing on functional data analysis and introducing topics on scalar-on-functional regression (Ramsay and Silverman, 2013, 2002; Ramsay et al., 2009; Reiss et al., 2017).

In general, there are two common frequentist approaches to obtain the estimation of the coefficient function $\beta(t)$. The first approach uses basis functions to obtain a smooth estimate of the coefficient function. Some popular choices of basis functions include B-splines (Frank and Friedman, 1993), Fourier (Marx and Eilers, 1999), eigenfunctions (Rice and Silverman, 1991) and wavelet (Ogden, 1997). The second approach introduces regularization to the coefficient function $\beta(t)$ by adding a penalty term, which is able to produce a shrink estimate of the coefficient function. Some popular choices of this penalization approach include cubic B-splines (Ramsay and Silverman, 2013) and penalized B-splines (Eilers and Marx, 1996). However, both basis function and penalized approaches give complicated coefficient function estimates with wiggles and are difficult to interpret in practice. Thus, out of consideration for interpretation and simplicity, a statistical approach that is able to give interpretable trends over some time regions and shrink other parts of $\beta(t)$ to zero appears to be more desirable.

LASSO (Tibshirani, 1996) is one of the most popular regularization methods that is able to shrink the magnitude of coefficients by introducing an ℓ_1 penalty term. In addition,

LASSO can perform the variable selection by shrinking coefficients towards zero, where other regularization methods, such as ridge regression (Hoerl and Kennard, 1970) and best subset selection (Hocking and Leslie, 1967), are not able to perform regularization and variable selection simultaneously. However, LASSO only penalizes regression coefficients themselves without considering spatial or temporal correlations between these coefficients, which cannot be ignored in the case of functional regression. Tibshirani et al. (2005) proposed the fused lasso, which includes both the lasso penalty and an additional penalty of first-order differences. Fused lasso takes spatial and temporal structure into consideration and it is widely used in time series and functional data analysis. Tibshirani et al. (2005) also extended the fused lasso to the generalized fused lasso, where differences of neighboring features, not necessarily the adjacent ones, are penalized. However, fused lasso only gives the constraint on the first-order differences. In functional regression, it is common to assume sparsity not only for the coefficients (zeroth-order differences) and the first-order differences but also for the higher-order differences. By selecting appropriate orders of differences, one can produce a set of highly-flexible and interpretable coefficient curves.

Several approaches have been proposed to address the constraints of multiple selected orders of differences. James et al. (2009) proposed an approach called FLiRTI, which performs variable selection on multiple derivatives of the coefficient function $\beta(t)$ using LASSO or Dantzig selector (Candes and Tao, 2007). FLiRTI transforms coefficients themselves to the differences of coefficients with selected orders based on differential operators and adds constraints to each selected derivative. This approach is proven to be theoretically desirable and can produce accurate and interpretable estimates. Another approach to tackle multiple selected orders of differences is the generalized lasso proposed by Tibshirani and Taylor (2011). Generalized lasso utilizes a common penalty matrix for the coefficient vector $\boldsymbol{\beta}$, which is able to create various constraints including lasso, the first-order difference (fused lasso), the first-order difference of selected neighbors (generalized fused lasso) and multiple orders of differences. Hence, generalized lasso has the advantage of being able to induce multiple orders of differences and is a good choice in functional regression. Despite the computational simplicity of these penalization methods, the quantification of estimation uncertainty has remained challenging.

Bayesian modeling approaches have become more and more popular in recent decades since they give prior beliefs to coefficients and are able to consider uncertainties for coefficient estimates. In general, there are two main types of Bayesian modeling approaches: discrete mixtures and shrinkage priors. The first approach, also known as spike-and-slab model (Mitchell and Beauchamp, 1988; George and McCulloch, 1993; Ishwaran and Rao, 2005), is widely used in Bayesian variable selection. Priors of coefficients follow a two-point mixture distribution with a degenerate distribution at zero as the spike part and a flat distribution as the slab part. Spike-and-slab priors can be applied to the group-level variable selection. Xu and Ghosh (2015) proposed the sparse group lasso with the spike-and-slab priors to perform both group selection and within-group variable selection. The second approach utilizes continuous priors centered at zero, which induce shrinkage for coefficient estimates. Tibshirani (1996) emphasized that lasso can be interpreted as the linear regression with Laplace priors on coefficients. Park and Casella (2008) proposed Bayesian lasso by expressing the Laplace distribution as a scale mixture of normals (Andrews and Mallows, 1974) and using Gibbs sampling to draw model parameters. To consider spatial and temporal structures, Casella et al. (2010) proposed the Bayesian fused lasso with the Laplace prior and a hierarchical representation using Gibbs sampling. Zhang et al. (2014) proposed the hierarchical structured variable selection method, which enabled the group selection with spike-and-slab priors and incorporated the Bayesian fused lasso for within-group selection based on the Laplace prior.

Both discrete mixtures and shrinkage approaches have their own issues. Computational issues related to the marginal distributions of latent indicators exist for the discrete mixtures. Laplace prior tends to overshrink coefficients compared to other priors with heavy-tailed distributions. In addition to using the Laplace prior for shrinkage, many other shrinkage priors are implemented into Bayesian hierarchical models and are demonstrated to have outstanding performances. Tipping (2001) used the Student-t prior with inverse-gamma mixing as the shrinkage prior. Griffin and Brown (2005) and Shimamura et al. (2019) used the normal-exponential-gamma (NEG) prior for Bayesian variable selection. Carvalho et al. (2010) proposed the horseshoe prior and it shares features of both discrete mixtures and Laplace priors. The horseshoe prior includes a global shrinkage parameter for all regression coefficients to control the global shrinkage level, as well as local shrinkage parameters for each regression

coefficient to obtain parameter-specific shrinkage. In addition, the horseshoe prior is proven to have an infinite spike at zero and a Cauchy-like tail, which results in weak shrinkage on non-zero coefficients and strong shrinkage on exact-zero coefficients. Thus, horseshoe prior is a popular choice in recent years for Bayesian modeling and variable selection. Kakikawa et al. (2022) proposed a Bayesian fused lasso modeling with the Laplace prior on regression coefficients and the horseshoe prior on the difference of successive regression coefficients.

In this project, we propose a Bayesian generalized LASSO on selected orders of differences via the horseshoe prior, which imposes multiple constraints with selected orders of differences and apply the proposed method to functional data with the purpose of obtaining interpretable coefficient estimates. The rest of Chapter 4 is organized as follows. In Section 4.2 we present the proposed model and priors for each parameter. Section 4.3 introduces the Bayesian sampling scheme based on Gibbs sampling. In Section 4.4 we report simulation results under different settings and Section 4.5 illustrates our proposed method with the real-data application of the PGS-ECHO fNIRS still-face study as described in Chapter 1. Section 4.6 gives a discussion of the last project as well as future works.

4.2 Model and priors

4.2.1 Scalar-on-functional regression with a simple grid basis

As in the FLiRTI approach proposed by James et al. (2009), the scalar-on-functional regression model can be expressed as

$$Y_i = \beta_0 + \int X_i(t)\beta(t)dt + \epsilon_i, \tag{15}$$

where Y_i is the response for ith subject, $i = 1, \ldots, n$, β_0 is the intercept, $\beta(t)$ is the coefficient function with $t \in [0, 1]$ and errors ϵ_i are iid with mean zero and a constant variance σ^2. Let $\boldsymbol{Q}(t) = [q_1(t), \ldots, q_m(t), \ldots, q_p(t)]'$ be a p-dimensional simple grid basis, where $q_m(t) = 1$ if $\frac{m-1}{p} < t \leq \frac{m}{p}$ and 0 otherwise. Thus, the coefficient function β_t can be expressed as

$$\beta(t) = \boldsymbol{Q}(t)\boldsymbol{\beta}, \tag{16}$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of basis function coefficients associated with the simple grid basis $\boldsymbol{Q}(t)$. Based on (15) and (16), the scalar-on-functional regression model with a simple grid basis is given as

$$Y_i = \beta_0 + \boldsymbol{X}_i'\boldsymbol{\beta} + \epsilon_i, \tag{17}$$

where $\boldsymbol{X}_i = \int X_i(t)\boldsymbol{Q}(t)dt$ is the vector of functional observations for ith subject since the simple grid basis $\boldsymbol{Q}(t)$ is just a $p \times p$ identity matrix. By centering each functional observation, we can get rid of the intercept and (17) can be expressed as the familiar linear model in matrix form

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{18}$$

where $\boldsymbol{Y}$ is a vector of responses with length n, $\boldsymbol{\epsilon}$ is an error vector with length n, $\boldsymbol{X}$ is a $n \times p$ matrix where the jth column is the jth covariate for $j = 1, \ldots, p$, and $\boldsymbol{\beta}$ is a vector of corresponding coefficients with length p for the design matrix $\boldsymbol{X}$.

By using the simple grid basis, we convert the scalar-on-functional regression to the conventional regression model. We use the simple grid basis and (18) for the rest of the paper. Notably, our proposed method can be extended to other complicated basis such as B-splines, Fourier and wavelet by specifying the corresponding basis function $\boldsymbol{Q}(t)$.

4.2.2 Bayesian linear regression with the horseshoe prior

Based on the model (18), Makalic and Schmidt (2015) proposed the Bayesian linear regression with horseshoe prior using hierarchical model formulation. This method corresponds to adding the constraint only to the coefficients themselves, which we also refer to as the zeroth-order difference. The Bayesian hierarchical models can be expressed as

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 &\sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n), \\
\beta_j|\omega_j^2, \tau^2, \sigma^2 &\sim N(0, \omega_j^2\tau^2\sigma^2), \\
\sigma^2 &\sim \sigma^{-2}d\sigma^2, \\
\omega_j &\sim C^+(A), \\
\tau &\sim C^+(A),
\end{aligned} \tag{19}$$

where τ^2 is the global shrinkage parameter and ω_j^2 are the local shrinkage parameters for the horseshoe prior, $C^+(A)$ is a half-Cauchy distribution with the scale parameter $A, A > 0$. Local shrinkage parameters ω_j^2 adjust the level of local shrinkage for regression coefficient β_j, while τ^2 determines the degree of global shrinkage for all regression coefficients and σ^2 controls the scale of regression coefficients. With two types of shrinkage parameters, the horseshoe prior is able to control the overall shrinkage level, as well as shrinkage levels for each coefficient. The half-Cauchy distribution of shrinkage parameters allows strong coefficients to remain large due to its property of a heavy tail.

To deal with the hierarchy of the half-Cauchy distribution, Wand et al. (2011) introduced a latent variable for the hierarchical expression of the half-Cauchy distribution based on the inverse-gamma distribution. For example, if $\tau \sim C^+(A)$, then we have

$$\tau^2|a \sim IG(\frac{1}{2}, \frac{1}{a}), a \sim IG(\frac{1}{2}, \frac{1}{A^2}),$$

where a is a latent variable, then we have $a|\tau^2 \sim IG(1, \frac{1}{\tau^2} + \frac{1}{A^2})$. Hence, the hierarchical model (19) can be rewritten as:

$$\begin{aligned}
\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 &\sim N(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n), \\
\beta_j|\omega_j^2, \tau^2, \sigma^2 &\sim N(0, \omega_j^2 \tau^2 \sigma^2), \\
\sigma^2 &\sim \sigma^{-2} d\sigma^2, \\
\omega_j^2|\zeta_j &\sim IG(\frac{1}{2}, \frac{1}{\zeta_j}), \\
\tau^2|\eta &\sim IG(\frac{1}{2}, \frac{1}{\eta}), \\
\zeta_1, \ldots, \zeta_p, \eta &\sim IG(\frac{1}{2}, \frac{1}{A^2}),
\end{aligned} \tag{20}$$

where $(\zeta_1, \ldots, \zeta_p, \eta)$ are latent variables associated with local and global shrinkage parameters $(\omega_1^2, \ldots, \omega_p^2, \tau^2)$.

4.2.3 Bayesian generalized lasso with the horseshoe prior

Tibshirani and Taylor (2011) introduced generalized lasso, which is able to solve the lasso-type optimization problem with multiple penalties on coefficients. The optimization

problem can be expressed as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{D}\boldsymbol{\beta}\|_1, \tag{21}$$

where $\lambda > 0$ is a tuning parameter for the lasso-type penalty term, $\boldsymbol{D}$ includes all selected finite difference operators. To be more specific, (21) can be rewritten as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \sum_{k\in E} \lambda_k\|\boldsymbol{D}^{(k)}\boldsymbol{\beta}\|_1, \tag{22}$$

where k is a non-negative integer denoting kth-order difference, E is a set with selected orders of differences, λ_k is the tuning parameter for the kth-order difference, and $\boldsymbol{D}^{(k)}$ is the kth-order difference operator. For example, zeroth-order difference operator $\boldsymbol{D}^{(0)} = \boldsymbol{I}_p$, where $\boldsymbol{I}_p$ is a $p \times p$ identity matrix. The first-order difference operator $\boldsymbol{D}^{(1)}$ is a $(p-1) \times p$ matrix with

$$\boldsymbol{D}^{(1)} = \begin{pmatrix} -1 & 1 & 0 & \dots & 0 & 0 \\ 0 & -1 & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & -1 & 1 \end{pmatrix},$$

and the second-order difference operator $\boldsymbol{D}^{(2)}$ is a $(p-2) \times p$ matrix with

$$\boldsymbol{D}^{(2)} = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & -2 & 1 \end{pmatrix}.$$

For the Bayesian generalized lasso with the horseshoe prior, we place horseshoe priors

on each selected order of difference. We propose the following hierarchical models

$$
\begin{aligned}
\boldsymbol{y}|\boldsymbol{X},\boldsymbol{\beta},\sigma^2 &\sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n),\\
\boldsymbol{D}^{(k)}\boldsymbol{\beta} &\sim N(\boldsymbol{0}_{p-k},\sigma^2\boldsymbol{B}_k),\ \boldsymbol{B}_k=\mathrm{diag}(\tau_k^2\omega_{1,k}^2,\ldots,\tau_k^2\omega_{p-k,k}^2), \qquad k\in E,\\
\sigma^2 &\sim \mathrm{IG}(b_1,b_2),\\
\tau_k^2|\eta_k &\sim \mathrm{IG}(\frac{1}{2},\frac{1}{\eta_k}), \qquad k\in E,\\
\omega_{j,k}^2|\zeta_{j,k} &\sim \mathrm{IG}(\frac{1}{2},\frac{1}{\zeta_{j,k}}), \qquad j=1,\ldots,p-k,\ k\in E,\\
\eta_k,\zeta_{j,k} &\sim \mathrm{IG}(\frac{1}{2},\frac{1}{A^2}), \qquad j=1,\ldots,p-k,\ k\in E,
\end{aligned}
\tag{23}
$$

where $\boldsymbol{D}^{(k)}$ is the kth-order difference operator, $\boldsymbol{0}_{p-k}$ is a vector of 0 with length $p-k$, τ_k^2 is the global shrinkage parameter for the kth-order difference, $\omega_{j,k}^2$ is the local shrinkage parameter for the jth covariate and the kth-order difference, η_k and $\zeta_{j,k}$ are corresponding latent variables for τ_k^2 and $\omega_{j,k}^2$. For hyperparameters, we set $b_1=b_2=0.01$ and $A=10$. All priors are non-informative priors by using the above hierarchical structures and selecting appropriate hyperprior values.

The priors in (23) can be represented using a scale mixture of normals (Andrews and Mallows, 1974), by integrating out all latent variables $\{\tau_k^2,\omega_{j,k}^2,\eta_k,\zeta_{j,k}^2\}$ for $k\in E$ and $j=1,\ldots,p-k$. Hence, the priors in (23) can be expressed as

$$
\begin{aligned}
f(\boldsymbol{\beta}|\sigma^2)\propto &\int\cdots\int\prod_{k\in E}\frac{1}{\sqrt{(2\pi\sigma^2)^{p-k}|\boldsymbol{B}_k|}}\exp\Big\{-\frac{1}{2\sigma^2}\sum_{k\in E}(\boldsymbol{D}^{(k)}\boldsymbol{\beta})'\boldsymbol{B}_k^{-1}(\boldsymbol{D}^{(k)}\boldsymbol{\beta})\Big\}\\
&\times\prod_{k\in E}f(\tau_k^2|\eta_k)\times\prod_{k\in E}f(\eta_k)\times\prod_{k\in E}\prod_{j=1}^{p-k}f(\omega_{j,k}^2|\zeta_{j,k})\times\prod_{k\in E}\prod_{j=1}^{p-k}f(\zeta_{j,k})\\
&d(\prod_{k\in E}\tau_k^2)\ d(\prod_{k\in E}\eta_k)\ d(\prod_{k\in E}\prod_{j=1}^{p-k}\omega_{j,k}^2)\ d(\prod_{k\in E}\prod_{j=1}^{p-k}\zeta_{j,k}),
\end{aligned}
\tag{24}
$$

where $f(\cdot|\cdot)$ denotes probability density function for the corresponding parameters listed in (23). Let $S=\exp\Big\{-\frac{1}{2\sigma^2}\sum_{k\in E}(\boldsymbol{D}^{(k)}\boldsymbol{\beta})'\boldsymbol{B}_k^{-1}(\boldsymbol{D}^{(k)}\boldsymbol{\beta})\Big\}$, then S can be rewritten as

$$
\begin{aligned}
S&=\exp\Big\{-\frac{1}{2\sigma^2}\sum_{k\in E}(\boldsymbol{D}^{(k)}\boldsymbol{\beta})'\boldsymbol{B}_k^{-1}(\boldsymbol{D}^{(k)}\boldsymbol{\beta})\Big\}\\
&=\exp\Big\{-\frac{1}{2\sigma^2}\boldsymbol{\beta}'(\sum_{k\in E}\boldsymbol{D}^{(k)\prime}\boldsymbol{B}_k^{-1}\boldsymbol{D}^{(k)})\boldsymbol{\beta}\Big\}.
\end{aligned}
\tag{25}
$$

Let $\boldsymbol{\Sigma}_E^{-1} = \sum_{k\in E} \boldsymbol{D}^{(k)\prime}\boldsymbol{B}_k^{-1}\boldsymbol{D}^{(k)}$, we can conclude that $\boldsymbol{\beta}$ follows a multivariate normal distribution with mean vector $\mathbf{0}_p$ and variance covariance matrix $\sigma^2\boldsymbol{\Sigma}_E$.

For example, if $E = \{0, 1\}$, which indicates that we choose to penalize zeroth and first-order differences. Therefore, we have $\Sigma_{\{0,1\}}^{-1} = \boldsymbol{D}^{(0)\prime}\boldsymbol{B}_0^{-1}\boldsymbol{D}^{(0)} + \boldsymbol{D}^{(1)\prime}\boldsymbol{B}_1^{-1}\boldsymbol{D}^{(1)}$. Through matrix calculation, we can express $\boldsymbol{\Sigma}_{\{0,1\}}^{-1}$ as

$$\boldsymbol{\Sigma}_{\{0,1\}}^{-1} = \begin{pmatrix} \frac{1}{\tau_0^2\omega_{1,0}^2} + \frac{1}{\tau_1^2\omega_{1,1}^2} & -\frac{1}{\tau_1^2\omega_{1,1}^2} & 0 & \dots & 0 \\ -\frac{1}{\tau_1^2\omega_{1,1}^2} & \frac{1}{\tau_0^2\omega_{2,0}^2} + \frac{1}{\tau_1^2\omega_{1,1}^2} + \frac{1}{\tau_1^2\omega_{2,1}^2} & -\frac{1}{\tau_1^2\omega_{2,1}^2} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & -\frac{1}{\tau_1^2\omega_{p-1,1}^2} \\ 0 & 0 & 0 & \dots & \frac{1}{\tau_0^2\omega_{p,0}^2} + \frac{1}{\tau_1^2\omega_{p-1,1}^2} \end{pmatrix}.$$

Though our proposed method shares some similarities with the HSVS method in Zhang et al. (2014) and the HORSES method in Kakikawa et al. (2022), which focus on Bayesian fused lasso modeling, it is able to extend the fused lasso to a more common case called generalized lasso, with the advantage of adding any types of penalties on coefficients. In addition, the implementation of horseshoe priors on all types of penalties has superiority over Laplace and spike-and-slab priors because of simultaneous controls of both global and local shrinkage levels.

4.3 Sampling scheme

With the model in (18) and prior structures in (23), we are able to derive the conditional posterior distribution of all parameters. Let $\boldsymbol{\omega}_k^2 = (\omega_{1,k}^2, \dots, \omega_{p-k,k}^2)'$ be a vector of length $(p-k)$ of the local shrinkage parameters for the kth-order difference, $\boldsymbol{\Omega}_k = \text{diag}(\omega_{1,k}^2, \dots, \omega_{p-k,k}^2)$ be a $(p-k)\times(p-k)$ diagonal matrix with each value of $\boldsymbol{\omega}_k^2$ as the diagonal element, and $\boldsymbol{D}_j^{(k)}$ be the jth row of the kth difference operator $\boldsymbol{D}^{(k)}$, for $j = 1, \dots, p-k$, $k \in E$. The full conditional posterior distributions are listed as follows:

$$
\begin{aligned}
\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{X},\{\boldsymbol{D}^{(k)},\tau_k^2,\boldsymbol{\omega}_k^2\},\sigma^2 &\sim N\Big((\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}_E^{-1})^{-1}\boldsymbol{X}'\boldsymbol{y},\sigma^2(\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}_E^{-1})^{-1}\Big),\\
\sigma^2|\boldsymbol{y},\boldsymbol{X},\{\boldsymbol{D}^{(k)},\tau_k^2,\boldsymbol{\omega}_k^2\},\boldsymbol{\beta} &\sim \mathrm{IG}\Big(\frac{2b_1+n+\sum_{k\in E}(p-k)}{2},\frac{2b_2+\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2+\boldsymbol{\beta}'\boldsymbol{\Sigma}_E^{-1}\boldsymbol{\beta}}{2}\Big),\\
\tau_k^2|\boldsymbol{\beta},\sigma^2,\boldsymbol{\omega}_k^2,\boldsymbol{D}^{(k)},\eta_k &\sim \mathrm{IG}\Big(\frac{p-k+1}{2},\frac{1}{2\sigma^2}(\boldsymbol{D}^{(k)}\boldsymbol{\beta})'\boldsymbol{\Omega}_k^{-1}(\boldsymbol{D}^{(k)}\boldsymbol{\beta})+\frac{1}{\eta_k}\Big),\\
\omega_{j,k}^2|\boldsymbol{\beta},\sigma^2,\tau_k^2,\boldsymbol{D}_j^{(k)},\zeta_{j,k} &\sim \mathrm{IG}\Big(1,\frac{1}{2\tau_k^2\sigma^2}(\boldsymbol{D}_j^{(k)}\boldsymbol{\beta})'(\boldsymbol{D}_j^{(k)}\boldsymbol{\beta})+\frac{1}{\zeta_{j,k}}\Big),\\
\zeta_{j,k}|\omega_{j,k}^2 &\sim \mathrm{IG}\Big(1,\frac{1}{\omega_{j,k}^2}+\frac{1}{A^2}\Big),\\
\eta_k|\tau_k^2 &\sim \mathrm{IG}\Big(1,\frac{1}{\tau_k^2}+\frac{1}{A^2}\Big).
\end{aligned}
\tag{26}
$$

Therefore, we can sample each set of parameters iteratively based on the above posterior distributions using Gibbs sampling.

4.4 Simulation studies

We conduct simulation studies to evaluate the performance of our proposed method. Simulated data is generated from the true model (18), where $\boldsymbol{\beta}$ is the true coefficient vector of length p, and the error vector $\boldsymbol{\epsilon}$ is iid and follows a distribution of $N(\mathbf{0}_n,\sigma^2\boldsymbol{I}_n)$. We also assume that all covariates are independent and generated from the standard normal distribution.

We consider nine different cases with three different numbers of covariates p and three prior structures with different selected orders of differences E. Nine cases are listed as follows

- Number of covariates $p=40$

  Case 1: $\boldsymbol{\beta}_{C1,nz}=(\mathbf{2}_5',\mathbf{1}_5')'$ with $E=\{0,1\}$,

  Case 2: $\boldsymbol{\beta}_{C2,nz}=(-2.5,-2,-1.5,-1,-0.5,3,2.5,2,1.5,1)'$ with $E=\{0,2\}$,

  Case 3: $\boldsymbol{\beta}_{C3,nz}=(-2.5,-2,-1.5,-1,-0.5,\mathbf{1}_5')'$ with $E=\{0,1,2\}$,

- Number of covariates $p=80$

  Case 4: $\boldsymbol{\beta}_{C4,nz}=(\mathbf{-3}_5',\mathbf{1}_5',\mathbf{-1.5}_5',\mathbf{2}_5')'$ with $E=\{0,1\}$,

Case 5: $\boldsymbol{\beta}_{C5,nz} = (\boldsymbol{\beta}'_{C2,nz}, 1, 1.5, 2, 2.5, 3, -0.5, -1, -1.5, -2, -2.5)'$ with $E = \{0, 2\}$,

Case 6: $\boldsymbol{\beta}_{C6,nz} = (-2.5, -2, -1.5, -1, -0.5, \mathbf{2}'_{\mathbf{5}}, 3, 2.5, 2, 1.5, 1, -\mathbf{2}'_{\mathbf{5}})'$ with $E = \{0, 1, 2\}$,

- Number of covariates $p = 120$

Case 7: $\boldsymbol{\beta}_{C7,nz} = (\boldsymbol{\beta}'_{C4,nz}, -\mathbf{1.5}'_{\mathbf{5}}, \mathbf{2}_{\mathbf{5}})'$ with $E = \{0, 1\}$,

Case 8: $\boldsymbol{\beta}_{C8,nz} = (\boldsymbol{\beta}'_{C5,nz}, -2, -1, 0, 1, 2, 2, 1, 0, -1, -2)'$ with $E = \{0, 2\}$,

Case 9: $\boldsymbol{\beta}_{C9,nz} = (-0.5, -1, -1.5, -2, -2.5, -\mathbf{1}'_{\mathbf{5}}, -3, -2, -1, 0, 1, \mathbf{2}'_{\mathbf{5}}, 2, 1, 0, -1, -2, -\mathbf{2}'_{\mathbf{5}})'$ with $E = \{0, 1, 2\}$,

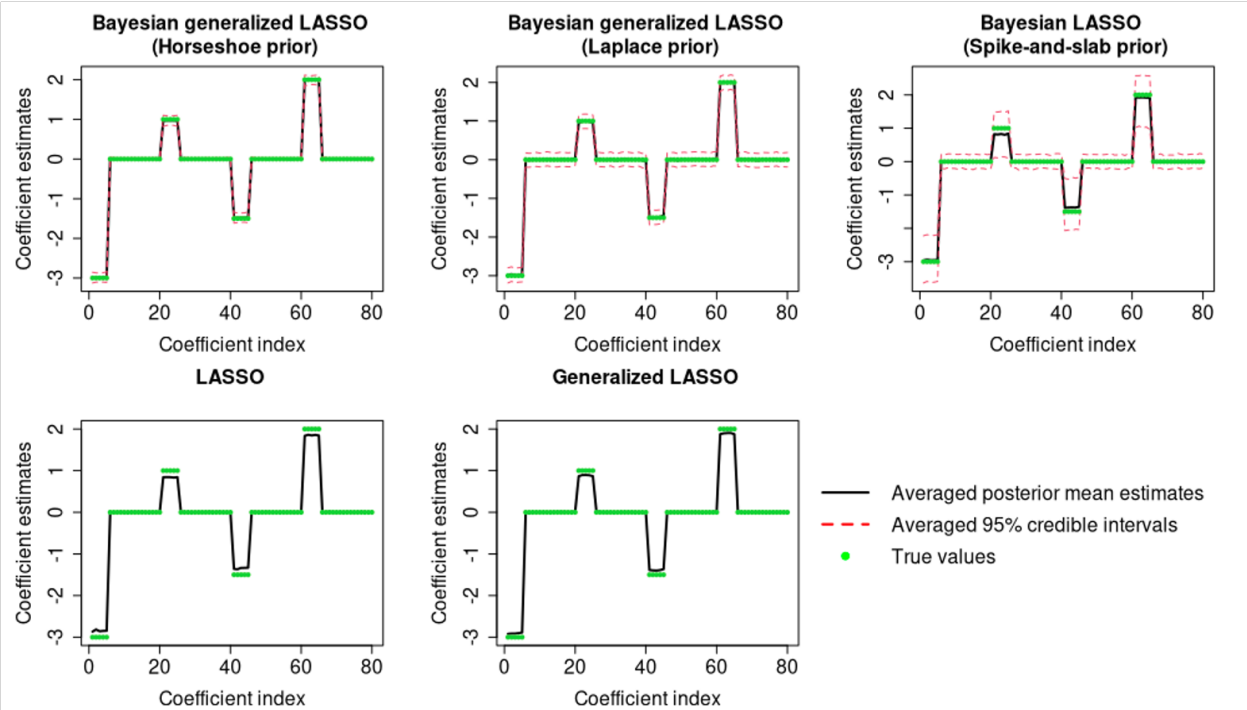
where for example, $\boldsymbol{\beta}_{C1,nz}$ is a vector of non-zero coefficients for Case 1 and $\mathbf{2}'_{\mathbf{5}}$ is a vector of 2 with lengh 5. Other than non-zero coefficients, all other coefficients are zero for all cases. In addition, we assume the error variance $\sigma^2 = 1$ for all nine cases. For each case, we simulate $r = 100$ replicates, each with $n_1 = 100$ for training the model and $n_2 = 100$ for testing the model performance and accuracy.

We compare our proposed method with four existing methods using both Bayesian and frequentist approaches in each above case. The first two methods are Bayesian with different prior structures. The first one is the Bayesian generalized lasso with Laplace priors on selected orders on differences. The second one puts a spike-and-slab prior on regression coefficients themselves (only zeroth-order difference) (Ishwaran and Rao, 2005; Xu and Ghosh, 2015). The third one is the so-called LASSO (Tibshirani, 1996) and the last one is the generalized lasso with selected orders of differences from (Tibshirani and Taylor, 2011). Notably, for the proposed method, the first and the last compared methods, we are able to give constraints on selected orders on differences E. However, for the second and third compared methods, we are only able to evaluate the simulated data with a constraint on the coefficients themselves. In addition, for two frequentist methods, we select the best tuning parameters using 5-fold cross-validation.

To evaluate the accuracy between true and estimated coefficients, we compute mean square error (MSE)

$$\text{MSE} = \frac{1}{100}\sum_{r=1}^{100}(\hat{\boldsymbol{\beta}}^{(r)}_{mean} - \boldsymbol{\beta})'(\hat{\boldsymbol{\beta}}^{(r)}_{mean} - \boldsymbol{\beta}), \tag{27}$$

Figure 18. Coefficient function plots of five methods for $p = 80$ and selected orders of differences $\mathrm{E} = \{0, 1\}$ (Case 4).



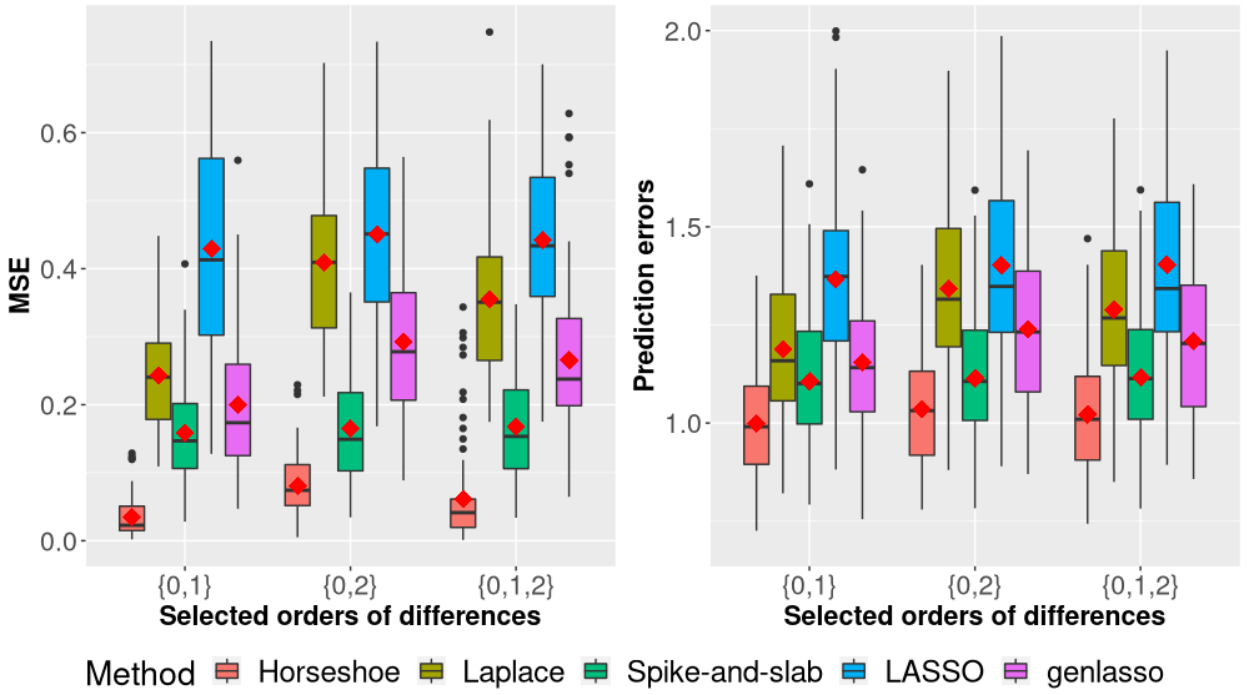
where $\hat{\boldsymbol{\beta}}^{(r)}_{mean}$ is a p-dimensional vector of the posterior mean estimates of regression coefficients for the rth replicate. In addition, we also assess the prediction accuracy on testing data by computing the prediction errors (PE)

$$\mathrm{PE} = \frac{1}{100}\sum_{r=1}^{100}\frac{(\boldsymbol{Y}^{(r)} - \boldsymbol{X}^{(r)}\hat{\boldsymbol{\beta}}^{(r)}_{mean})'(\boldsymbol{Y}^{(r)} - \boldsymbol{X}^{(r)}\hat{\boldsymbol{\beta}}^{(r)}_{mean})}{n_2}, \tag{28}$$

where $\boldsymbol{Y}^{(r)}$ is rth response vector of testing data with length $n_2 = 100$ and $\boldsymbol{X}^{(r)}$ is rth design matrix containing covariates of subjects in testing data.

Figure 18 shows coefficient function plots of five methods (proposed plus four compared methods) using Case 4 as an example. Our proposed method, Bayesian generalized lasso with the horseshoe prior, achieves accurate coefficient estimates with narrow 95% credible intervals. Bayesian generalized lasso with the Laplace prior has good coefficient estimates but with slightly wider 95% credible intervals. Bayesian lasso with the spike-and-slab prior tends to shrink non-zero coefficients with the widest 95% credible intervals. Two frequentist

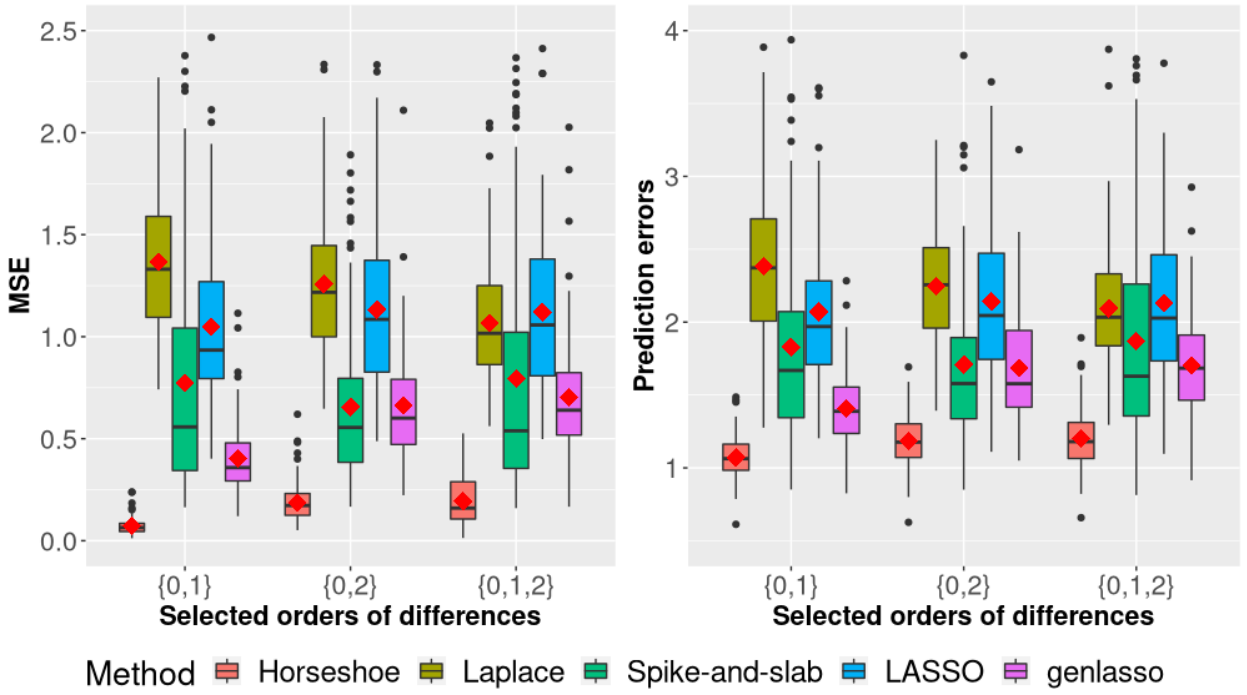
Figure 19. Boxplots of MSE and prediction errors for five methods with $p = 40$ (Case 1,2 and 3). Diamond markers denote the means of ease case and method.



methods, LASSO, and generalized lasso, also perform some levels of shrinkage on non-zero coefficients. Coefficient function plots of other simulation cases are given in supplemental materials. More coefficient function plots of other simulation cases are given in Appendix C.1.

Simulation results of MSE and prediction errors for each case and each method are shown in Figure 19, 20 and 21. Each simulation case corresponds to each figure with one set of selected orders of differences E. For a better visualization, we plot MSE and prediction errors for $p = 120$ under the log scale. For all nine cases, our proposed method has the lowest MSE and prediction errors compared to the other four methods. In Figure 19 with $p = 40$ (Case 1, 2 and 3), the spike-and-slab model shows a good performance in terms of both MSE and PE, only inferior to the proposed method with the horseshoe prior. However, it has the worst performance for $p = 120$ (Case 7,8,9), where the number of covariates is larger than the number of observations. The model with the Laplace prior also tends to

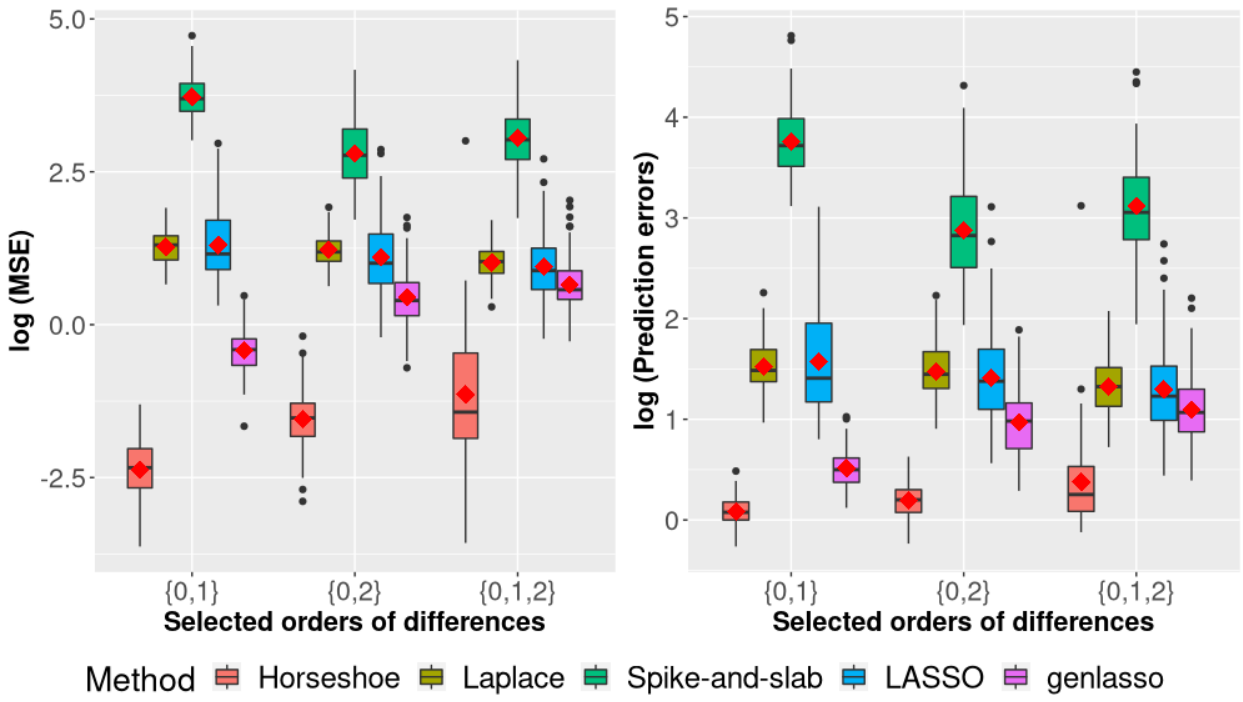
Figure 20. Boxplots of MSE and prediction errors for five methods with $p = 80$ (Case 4,5 and 6). Diamond markers denote the means of ease case and method.



have larger MSE and PE with the increase of p, but not much as the spike-and-slab prior. Between the two frequentist methods, generalized lasso always performs better than LASSO since it is able to add desired orders of differences into the penalty matrix, while LASSO only penalizes coefficients themselves (zeroth-order difference).

In conclusion, our proposed method, Bayesian generalized lasso with the horseshoe prior, can achieve a better model fit and prediction accuracy compared to other existing methods. Good performance is demonstrated under multiple simulation settings. Comparing our proposed method to other Bayesian models with Laplace and spike-and-slab prior, our method is able to produce coefficient estimates with less shrinkage and stable results with narrow 95% credible intervals. Tables of simulation results of all cases and methods can be found in Appendix C.1.

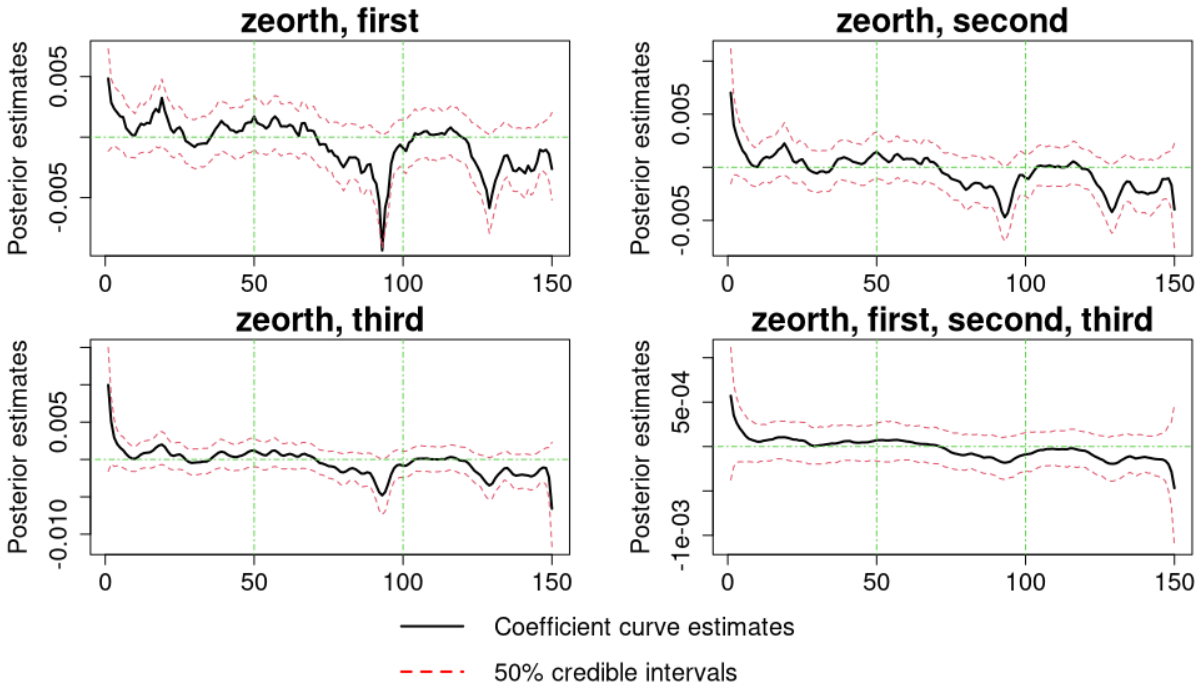
Figure 21. Boxplots of MSE and prediction errors for five methods with $p = 120$ (Case 7,8 and 9). Values are presented using the log scale. Diamond markers denote the means of ease case and method.



4.5 Real-data application results

We apply our proposed method to the scalar-on-functional regression using fNIRS still-face data introduced in Chapter 1. Measurements of concentration of oxygen Hemoglobin (HbO) are treated as functional predictors to predict Infant Behavior Questionnaire-Revised negative emotionality (IBQ-NE) or Infant Behavior Questionnaire-Revised effortful control (IBQ-EC) score. To save computational time, we take an average of every ten measurements so that the number of functional predictors equals 150 per subject. In addition, all functional predictors are centered and rescaled to follow the standard normal distribution. We run a total of 10,000 Gibbs iterations with a burn-in period of 2,000 and use hyperparameters as provided in Section 4.2.3. We ran multiple analyses with measurements from each channel being treated as functional predictors and only presented results of coefficient curve estimates

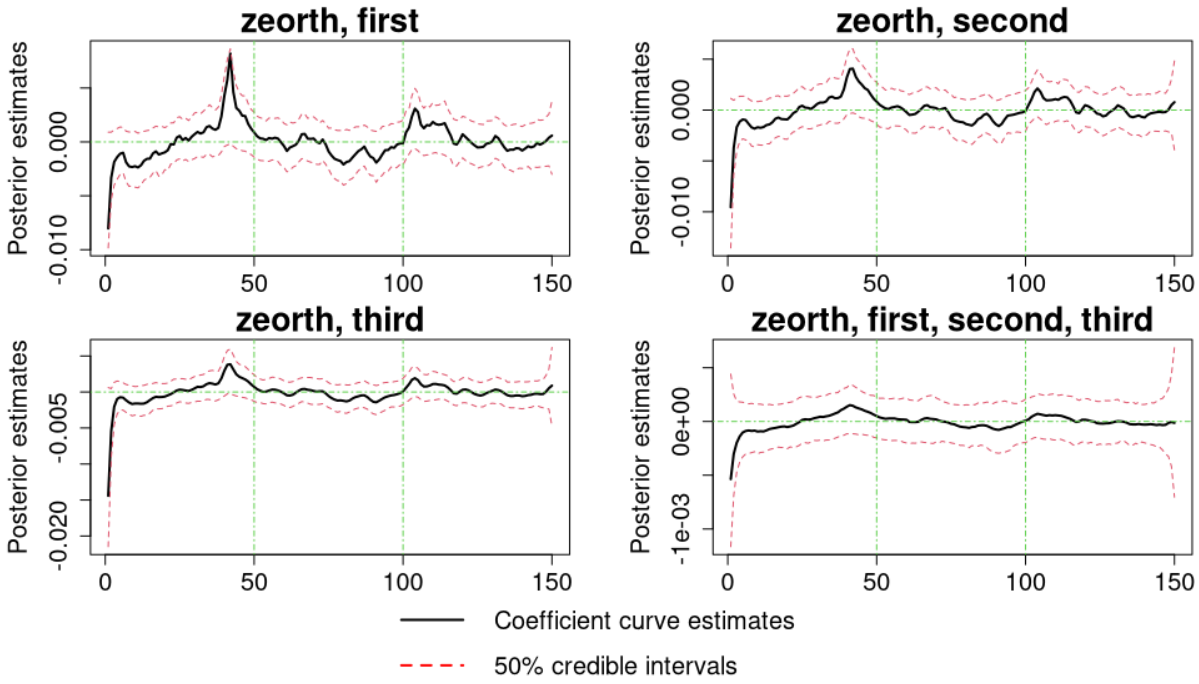
Figure 22. Coefficient plots of selected orders of differences for the model with IBQ-NE as the outcome and measurements of channel S1D1 as functional predictors. The horizontal green dashed line is the line of zero and two vertical green dashed lines are separations of three FFSF phases (Interact, still-face, recovery).



from channel S1D1.

Figure 22 gives coefficient curve estimates of selected orders of differences for the model with IBQ-NE as the outcome, along with 50% credible intervals. Results from four sets of selected orders of differences are shown: zeroth and first-order differences, zeroth and second-order differences, zeroth and third-order differences, and zeroth, first, second, and third-order differences. From the coefficient plot with zeroth and first-order differences, we can observe the positive coefficient estimates and a decreasing trend of HbO at the beginning of the interact period (baseline), followed with sudden drops to the negative level at the end of the still-face period and the middle of the recovery period. In general, the IBQ-NE score is related to the increasing brain activity at the interact period, but the relationship is reversed and the IBQ-NE score is related to the decreasing brain activity at the still-face and recovery periods. Other three coefficient plots in Figure 22 also shows the above trends, but with

Figure 23. Coefficient plots of selected orders of differences for the model with IBQ-EC as the outcome and measurements of channel S1D1 as functional predictors. The horizontal green dashed line is the line of zero and two vertical green dashed lines are separations of three FFSF phases (Interact, still-face, recovery).



weaker coefficient estimates as the constraints of high-order differences are incorporated in the prior structures. The hierarchical model with four selected orders of differences (zeroth, first, second and third) tends to shrinkage coefficients more towards zero.

Figure 23 gives coefficient curve estimates of four selected orders of differences for the model with IBQ-EC as the outcome, along with 50% credible intervals. From the coefficient plot with zeroth and first-order differences, coefficients are negative at the beginning and are close to zero later. We observe a large positive signal at the end of the interact period and a small positive signal at the beginning of the recovery period. As described in the above model with the outcome IBQ-NE, the hierarchical model with a more complicated prior structure (more constraints of high-order differences) gives more shrinkage towards zero.

4.6 Discussion

In this Chapter, we propose a Bayesian version of generalized lasso, with the purpose of giving constraints on certain orders of differences. The proposed method is able to correctly find the trend (constant, linear, quadratic, etc) of coefficients by manipulating prior structures of coefficients. In addition, the well-defined Bayesian hierarchical model structures give more flexibility to choose the appropriate orders of differences and the desired trend filtering. We also use the horseshoe prior instead of the traditional Laplace prior, which gives both global and local shrinkage priors for regression coefficients and is able to control the global shrinkage level and local shrinkage level of each coefficient simultaneously. We conduct simulations with different settings and compare our proposed method to two Bayesian methods using Laplace and spike-and-slab priors, as well as two frequentist methods: LASSO and generalized lasso. Simulation results show the superiority of the proposed method over other four methods in terms of smaller MSE, prediction errors, and stable coefficient estimates with the tightest 95% credible intervals in all simulation settings. We also apply our proposed method to the fNIRS still-face study and aim to obtain coefficient curve estimates by using IBQ-NE or IBQ-EC scores as the outcome and functional measurements in a selected channel as predictors.

However, for the real-data application, we did not find any strong signals between the outcome and functional predictors using the fNIRS still-face data. From Section 4.5, we did not find any large coefficient estimates as desired in the simulation studies. All coefficient estimates seem to be around zero, which show a small contribution to the prediction of the outcome. In addition, we only plot the 50% credible intervals and the 95% credible intervals are very wide. Overall, fitting models have a large prediction error as well as a large error variance, which also confirm the poor fit of the data. Later we will apply our proposed method to another dataset with more promising results.

Our future works include the following two directions. First, we can incorporate the feature of group selection into our proposed model. Currently, our proposed method is designed to fit the model with functional predictors only from one source. However, for functional imaging data, there are always more than one source. Thus, it is meaningful to extend

our proposed method with a group selection feature, which is able to select the appropriate group that has large effects on prediction and penalizes other groups to exact zero. Second, as emphasized in Section 4.2.1, we use a simple grid basis to convert the scalar-on-functional regression to the conventional regression model. In reality, more complicated basis such as polynomial, B-splines and Fourier basis are more common in functional regression analysis. We can try our proposed method based on any of these common basis functions.

Appendix A Chapter 2

A.1 Detailed RJMCMC sampling scheme

We use the reversible jump MCMC (RJMCMC) as the Bayesian sampling algorithm, with the advantage of jumping over the number of components through MCMC iterations. RJMCMC is an advanced Metropolis-Hastings algorithm and it can allow posterior distribution sampled from varying dimensions. In our study, each RJMCMC iteration ℓ contains two types of moves: between-model moves and within-model moves. Between-model moves involve the change of parameter dimensions by proposing to add or remove one component. RJMCMC can be used to fulfill this type of move. Within-model moves do not involve the change of parameter dimensions and Gibbs sampling is used to draw different parameters.

For ease if notations, we define $\boldsymbol{\Theta}_{g,G} = (\boldsymbol{\theta}'_g, \tau_g^2, \sigma_g^2, \boldsymbol{\delta}_g^{*\prime}, \kappa_{\zeta g}^2)'$ as the aggregation of all parameters for gth of a total of G components and $\boldsymbol{\Theta}_G = (\boldsymbol{\Theta}'_{1,G}, \ldots, \boldsymbol{\Theta}'_{g,G}, \ldots, \boldsymbol{\Theta}'_{G,G})'$ as the aggregation of all parameters from all components. We also denote the current state as $(G^c, \boldsymbol{\Theta}^c_{G^c})$ and the proposed state as $(G^p, \boldsymbol{\Theta}^p_{G^p})$, where the superscript c and p are the current and proposed state, respectively. The proposed RJMCMC sampling algorithm is detailed below:

1. **Between-model moves**

   We propose to move from $(G^c, \boldsymbol{\Theta}^c_{G^c})$ to $(G^p, \boldsymbol{\Theta}^p_{G^p})$ by drawing $(G^p, \boldsymbol{\Theta}^p_{G^p})$ from a proposed density $q(G^p, \boldsymbol{\Theta}^p_{G^p} \mid G^c, \boldsymbol{\Theta}^c_{G^c})$ and accepting it with the acceptance rate

   $$\alpha = \min\left\{1, \frac{p(G^p, \boldsymbol{\Theta}^p_{G^p} \mid \boldsymbol{y}) \times q(G^c, \boldsymbol{\Theta}^c_{G^c} \mid G^p, \boldsymbol{\Theta}^p_{G^p})}{p(G^c, \boldsymbol{\Theta}^c_{G^c} \mid \boldsymbol{y}) \times q(G^p, \boldsymbol{\Theta}^p_{G^p} \mid G^c, \boldsymbol{\Theta}^c_{G^c})}\right\}, \tag{29}$$

   where $p(\cdot)$ denotes the target density and is the product of the joint likelihood function times prior densities. The proposed density function $q(\cdot)$ varies from different moving types and proposed distributions. In our case, the proposed density of the proposed state

$q(G^p, \boldsymbol{\Theta}^p_{G^p} \mid G^c, \boldsymbol{\Theta}^c_{G^c})$, given the current state is:

$$\begin{aligned} & q(G^p, \boldsymbol{\Theta}^p_{G^p} \mid G^c, \boldsymbol{\Theta}^c_{G^c}) = q(G^p \mid G^c) \times q(\boldsymbol{\Theta}^p_{G^p} \mid G^p, G^c, \boldsymbol{\Theta}^c_{G^c}) \\ & = q(G^p \mid G^c) \times q(\boldsymbol{\tau}^{2p}_{G^p} \mid G^p, G^c, \boldsymbol{\tau}^{2c}_{G^c}) \times q(\boldsymbol{\sigma}^{2p}_{G^p} \mid G^p, G^c, \boldsymbol{\sigma}^{2c}_{G^c}) \\ & \times q(\boldsymbol{\theta}^p_{G^p} \mid G^p, G^c, \boldsymbol{\tau}^{2p}_{G^p}, \boldsymbol{\sigma}^{2p}_{G^p}) \times q(\boldsymbol{\delta}^{*p}_{G^p} \mid G^p, G^c, \boldsymbol{\delta}^c_{G^c}, \boldsymbol{\kappa}^{2p}_{\zeta G^p}) \\ & \times q(\boldsymbol{\kappa}^{2p}_{\zeta G^p} \mid G^p, G^c, \boldsymbol{\kappa}^{2c}_{\zeta G^c}), \end{aligned} \tag{30}$$

where $\{\boldsymbol{\tau}^{2p}_{G^p}, \boldsymbol{\sigma}^{2p}_{G^p}, \boldsymbol{\theta}^p_{G^p}, \boldsymbol{\delta}^{*p}_{G^p}, \boldsymbol{\kappa}^{2p}_{\zeta G^p}\}$ are the corresponding parameters of all components at the proposed state. From (30), we can first draw G^p and then follow by drawing $\boldsymbol{\tau}^{2p}_{G^p}$, $\boldsymbol{\sigma}^{2p}_{G^p}$, $\boldsymbol{\theta}^p_{G^p}$, $\boldsymbol{\kappa}^{2p}_{\zeta G^p}$ and $\boldsymbol{\delta}^{*p}_{G^p}$.

The first step of the proposed RJMCMC algorithm is to decide whether to split one component into two or combine two components into one. For the following part, we will call the two steps split and combine. For the first step, G^p is proposed from $q(G^p \mid G^c)$ and $q(G^p = G^c - 1) = q(G^p = G^c + 1) = 0.5$, which assumes equal probability of split or combine. Notably, split proposal will be accepted with probability 1 if $G^c = 1$ and combine proposal will be accepted with probability 1 if G^c reaches a predefined maximum number of components, saying $G_{\max}$.

a. **Split proposal**

Suppose split proposal is selected and thus we have $G^p = G^c + 1$. The detailed split proposal is listed below.

i. **A component r is randomly selected to split**

ii. **Sampling new variance of the random intercept**

Denoting $\boldsymbol{\kappa}^{2p}_{\zeta G^p} = (\kappa^{2c}_{\zeta,1}, \ldots, \kappa^{2c}_{\zeta,r}, \kappa^{2p}_{\zeta,r+1}, \kappa^{2c}_{\zeta,r+1}, \ldots, \kappa^{2c}_{\zeta,G^c-1})'$ as the new vector of variances of random intercepts, where $\kappa^{2p}_{\zeta,r+1}$ is the new variance of the random intercept at the proposed state. We propose to generate $\kappa^{2p}_{\zeta,r+1}$ by drawing a common random variable u_3 from a Uniform distribution such that $u_3 \sim U(0,1)$ and let

$$\kappa^{2p}_{\zeta,r+1} = \kappa^{2c}_{\zeta,r} \times \frac{u_3}{1-u_3}.$$

iii. **Sampling new logistic parameters**

Denoting $\boldsymbol{\delta}^{*p}_{G^p} = (\boldsymbol{\delta}^{*c\prime}_1, \ldots, \boldsymbol{\delta}^{*c\prime}_r, \boldsymbol{\delta}^{*p\prime}_{r+1}, \boldsymbol{\delta}^{*c\prime}_{r+1}, \ldots, \boldsymbol{\delta}^{*c\prime}_{G^c-1})'$ as the new vector of logis-

tic parameters. $\boldsymbol{\delta}_{r+1}^{*p}$ is the vector of logistic parameters of the new proposed component and can be drawn by:

A. Computing the Mahalanobis distance between each subject which is classified to component r and the rth estimated component trajectory given $\boldsymbol{\theta}_{G^c}^{c}$.

B. Performing an initiated two-component clustering of those subjects using the K-means clustering algorithm. The split proposal will be rejected if none or only one subject belongs to the selected component r for the current state.

C. Sampling $\boldsymbol{\delta}_{r+1}^{*p}$ with a Polya-Gamma augmentation strategy using Gibbs sampling, which is detailed in the within-model moves later.

iv. **Computing mixing weights**

Based on logistic parameter $\boldsymbol{\delta}_{r}^{*c}$ for the rth component and the newly-generated logistic parameter $\boldsymbol{\delta}_{r+1}^{*p}$, we need to recompute mixing weights $\boldsymbol{\pi}_{r}^{*p}$ and $\boldsymbol{\pi}_{r+1}^{p}$, where $\boldsymbol{\pi}_{r}^{p}$ denotes mixing weights of all subjects for rth component at the proposed state and $\boldsymbol{\pi}_{r+1}^{p}$ represents mixing weights of all subjects for $(r+1)$th component at the proposed state. To keep mixing weights of other components unchanged, $\boldsymbol{\pi}_{r}^{p}$ and $\boldsymbol{\pi}_{r+1}^{p}$ need to be rescaled such that $\boldsymbol{\pi}_{r}^{p} + \boldsymbol{\pi}_{r+1}^{p} = \boldsymbol{\pi}_{r}^{c}$, to ensure that mixing weights for each subject sum up to 1.

v. **Sampling new smoothing parameters**

Under the split proposal, let $\boldsymbol{\tau}_{G^p}^{2p} = (\tau_1^{2c}, \ldots, \tau_{r-1}^{2c}, \tau_r^{2p}, \tau_{r+1}^{2p}, \tau_{r+1}^{2c}, \ldots \tau_{G^c}^{2c})'$ be the new vector of smoothing parameters, where τ_r^{2p} and τ_{r+1}^{2p} are values of new smoothing parameters at the proposed state splitting from τ_r^{2c}. We propose to split τ_r^{2c} by drawing a common random variable u_1 across dimensions from a Uniform distribution such that $u_1 \sim U(0,1)$ and let

$$\tau_r^{2p} = \tau_r^{2c} \times \frac{u_1}{1-u_1},$$

$$\tau_{r+1}^{2p} = \tau_r^{2c} \times \frac{1-u_1}{u_1}.$$

vi. **Sampling new error variances**

Let $\boldsymbol{\sigma}_{G^p}^{2p} = (\sigma_1^{2c}, \ldots, \sigma_{r-1}^{2c}, \sigma_r^{2p}, \sigma_{r+1}^{2p}, \sigma_{r+1}^{2c}, \ldots \sigma_{G^c}^{2c})'$ be the new vector of error variances, where σ_r^{2p} and σ_{r+1}^{2p} are values of new error variances at the proposed state

splitting from σ_r^{2c}. We propose to split σ_r^{2c} by drawing a common random variable u_2 across dimensions from a Uniform distribution such that $u_2 \sim U(0,1)$ and let

$$\sigma_r^{2p} = \sigma_r^{2c} \times \frac{u_2}{1-u_2},$$
$$\sigma_{r+1}^{2p} = \sigma_r^{2c} \times \frac{1-u_2}{u_2}.$$

vii. **Sampling new model coefficients**

Let $\boldsymbol{\theta}_{G^p}^p = (\boldsymbol{\theta}_1^{c\prime}, \ldots, \boldsymbol{\theta}_{r-1}^{c\prime}, \boldsymbol{\theta}_r^{p\prime}, \boldsymbol{\theta}_{r+1}^{p\prime}, \boldsymbol{\theta}_{r+1}^{c\prime}, \ldots \boldsymbol{\theta}_{G^c}^{c\prime})'$ be the new vector of model coefficients, which includes regression and smoothing spline coefficients. Thus, $\boldsymbol{\theta}_r^{p\prime}$ and $\boldsymbol{\theta}_{r+1}^{p\prime}$ can be drawn from the conditional posterior distribution based on $\tau_{G^p}^{2p}$ and $\sigma_{G^p}^{2p}$ in a Gibbs manner, which is detailed in the within-model moves later.

viii. **Subject reallocation**

After splitting a component into two by sampling all parameters $\{\boldsymbol{\tau}_{G^p}^{2p}, \boldsymbol{\sigma}_{G^p}^{2p}, \boldsymbol{\theta}_{G^p}^p, \boldsymbol{\delta}_{G^p}^{*p}, \boldsymbol{\kappa}_{\zeta G^p}^{2p}\}$ at the proposed state, we need to reallocate subjects that belong to component r to two new components by computing the distribution of latent indicator z_{ir} and $z_{i,r+1}$ with

$$f(z_{ir} = 1 \mid \boldsymbol{y}_i, \boldsymbol{\Theta}_r^p) = \frac{\pi_{ir} f_r(\boldsymbol{y}_i \mid \boldsymbol{\Theta}_r^p)}{\sum_{h=r}^{r+1} \pi_{ih} f_h(\boldsymbol{y}_i \mid \boldsymbol{\Theta}_h^p)}, \tag{31}$$

thus z_{ir} and $z_{i,r+1}$ can be drawn from the multinomial distribution.

ix. **Computing the acceptance rate**

The acceptance rate for the split proposal is $\alpha = \min\{1, A\}$, where

$$\begin{aligned} A = & \frac{p(\boldsymbol{y} \mid G^p, \boldsymbol{\Theta}_{G^p}^p)\, p(\boldsymbol{\Theta}_{G^p}^p \mid G^p)\, p(G^p)}{p(\boldsymbol{y} \mid G^c, \boldsymbol{\Theta}_{G^c}^c)\, p(\boldsymbol{\Theta}_{G^c}^c \mid G^c)\, p(G^c)} \\ & \times \frac{q(G^c \mid G^p)\, q(\boldsymbol{\theta}_r^c)}{q(G^p \mid G^c)\, q(Alloc)\, q(u_1)\, q(u_2)\, q(u_3)\, q(\boldsymbol{\theta}_r^p)\, q(\boldsymbol{\theta}_{r+1}^p)\, q(\boldsymbol{\delta}_{r+1}^{*p})} \times |J|, \end{aligned} \tag{32}$$

where $q(Alloc)$ is the probability that this reallocation is made, $q(u_1) = q(u_2) = q(u_3) = 1$ is the probability density function of $U(0,1)$ and the jacobian J is:

$$|J| = \left| \frac{\partial(\tau_r^{2p}, \tau_{r+1}^{2p}, \sigma_r^{2p}, \sigma_{r+1}^{2p}, \kappa_{\zeta,r+1}^{2p})}{\partial(\tau_r^{2c}, \sigma_r^{2c}, u_1, u_2, u_3)} \right|.$$

b. **Combine proposal**

Suppose combine proposal is selected and thus we have $G^p = G^c - 1$. The combine proposal is the reversed step of split proposal since detailed balance condition is required to hold for the proposed RJMCMC algorithm. The combine proposal is listed below.

i. **Selecting two components to combine**

Under the combine proposal, we need to first select two components to combine. One component is the component that is allocated with the fewest number of subjects and another component is selected based on the weights of the number of subjects that belong to each component.

ii. **Removing variance of the random intercept**

Since a new variance of the random intercept is generated via the split proposal, we need to remove one variance of the random intercept in order to reduce one component. Thus, one variance of the random intercept is randomly selected to be removed from two selected components.

iii. **Removing logistic parameters**

Since a new set of logistic parameters is generated via the split proposal, we need to remove one set of logistic parameters in order to reduce one component. Thus, the set of logistic parameters which belong to the same component removed by the variance of the random intercept is randomly selected to be removed from two selected components.

iv. **Computing mixing weights**

Since two components are selected to combine, the mixing weights of two components are added together to form the new mixing weights of the combined component.

v. **Computing combined smoothing parameters**

Let $\boldsymbol{\tau}^{2p}_{G^p} = (\tau^{2c}_1, \ldots, \tau^{2c}_{r-1}, \tau^{2p}_r, \tau^{2c}_{r+2}, \ldots \tau^{2c}_{G^c})'$ be the new vector of smoothing parameters, where τ^{2p}_r is the value of new smoothing parameters at the proposed state, combining from τ^{2c}_r and τ^{2c}_{r+1}. τ^{2p}_r is computed by reversing the step described in sampling new smoothing parameters in the split proposal and thus we

have:

$$\tau_r^{2p} = \sqrt{\tau_r^{2c}\,\tau_{r+1}^{2c}}.$$

vi. **Computing combined error variances**

Let $\boldsymbol{\sigma}_{G^p}^{2p} = (\sigma_1^{2c}, \ldots, \sigma_{r-1}^{2c}, \sigma_r^{2p}, \sigma_{r+2}^{2c}, \ldots \sigma_{G^c}^{2c})'$ be the new vector of error variances, where σ_r^{2p} is the value of new error variances at the proposed state, combining from σ_r^{2c} and σ_{r+1}^{2c}. Similar to the combine of smoothing parameters, σ_r^{2p} is computed by reversing the step described in sampling new error variances in the split proposal and thus we have:

$$\sigma_r^{2p} = \sqrt{\sigma_r^{2c}\,\sigma_{r+1}^{2c}}.$$

vii. **Sampling combined model coefficients**

Let $\boldsymbol{\theta}_{G^p}^{p} = (\boldsymbol{\theta}_1^{c\prime}, \ldots, \boldsymbol{\theta}_{r-1}^{c\prime}, \boldsymbol{\theta}_r^{p\prime}, \boldsymbol{\theta}_{r+2}^{c\prime}, \ldots \boldsymbol{\theta}_{G^c}^{c\prime})'$ be the new vector of model coefficients. Thus, $\boldsymbol{\theta}_r^p$ can be drawn from the conditional posterior distribution based on τ_r^{2p} and σ_r^{2p} in a Gibbs manner, which is detailed in within-model moves later.

viii. **Computing the acceptance rate**

Since the detailed balanced condition holds for the proposed RJMCMC algorithm, which indicates that the split and combine proposal can be reversible. Thus, the acceptance rate for the combine proposal is $\alpha = \min\{1, A^{-1}\}$, where A can be obtained using (32).

2. **Within-model moves**

Within-model moves sample $\boldsymbol{\Theta}_G = (\boldsymbol{\Theta}_{1,G}', \ldots, \boldsymbol{\Theta}_{g,G}', \ldots, \boldsymbol{\Theta}_{G,G}')'$ without the change of parameter dimensions or the number of components. For ease of notation, we suppress the subscript G in within-model moves. As notations in 2.2.1, we denote $\boldsymbol{\Theta}_g = (\boldsymbol{\theta}_g', \tau_g^2, \sigma_g^2, \boldsymbol{\delta}_g^{*\prime}, \kappa_{\zeta g}^2)'$ as parameters for gth component, detailed within-model moves sampling scheme is shown as below.

a. **Sampling model coefficients**

For each component g, based on the augmented likelihood and priors on $\boldsymbol{\theta}_g =$

$(\boldsymbol{\alpha}_g', \boldsymbol{\beta}_g')'$, the conditional posterior distribution of $(\boldsymbol{\theta}_g \mid \boldsymbol{y}, \boldsymbol{S}, \tau_g^2, \sigma_g^2)$ is:

$$
\begin{aligned}
p(\boldsymbol{\theta}_g \mid \boldsymbol{y}, \boldsymbol{S}, \tau_g^2, \sigma_g^2) &\propto p(\boldsymbol{y} \mid \boldsymbol{S}, \boldsymbol{\theta}_g, \sigma_g^2) \cdot p(\boldsymbol{\theta}_g \mid \tau_g^2) \\
&\propto \prod_{i=1}^{N} \Big\{ (\sigma_g^2)^{-n/2} \exp\Big[-\frac{1}{2\sigma_g^2} (\boldsymbol{y}_i - \boldsymbol{S}\boldsymbol{\theta}_g)'(\boldsymbol{y}_i - \boldsymbol{S}\boldsymbol{\theta}_g) \Big] \Big\}^{z_{ig}} \\
&\times |\boldsymbol{D}_g|^{-1/2} \exp\Big(-\frac{1}{2} \boldsymbol{\theta}_g' \boldsymbol{D}_g^{-1} \boldsymbol{\theta}_g \Big) \\
&\propto \exp\Big\{ -\frac{1}{2\sigma_g^2} \Big[\sum_{i=1}^{N} z_{ig} (\boldsymbol{y}_i - \boldsymbol{S}\boldsymbol{\theta}_g)'(\boldsymbol{y}_i - \boldsymbol{S}\boldsymbol{\theta}_g) + \boldsymbol{\theta}_g' \sigma_g^2 \boldsymbol{D}_g^{-1} \boldsymbol{\theta}_g \Big] \Big\} \\
&\propto \exp\Big[-\frac{1}{2\sigma_g^2} (\boldsymbol{\theta}_g - \boldsymbol{\mu}_g)' (\boldsymbol{\Lambda}_g)^{-1} (\boldsymbol{\theta}_g - \boldsymbol{\mu}_g) \Big] \\
&\sim N(\boldsymbol{\mu}_g, \sigma_g^2 \boldsymbol{\Lambda}_g),
\end{aligned}
$$

where $\boldsymbol{\Lambda}_g = (N_g \boldsymbol{S}'\boldsymbol{S} + \sigma_g^2 \boldsymbol{D}_g^{-1})^{-1}$ and $\boldsymbol{\mu}_g = \boldsymbol{\Lambda}_g \sum_{i=1}^{N} z_{ig} \boldsymbol{S}' \boldsymbol{y}_i$, N_g is the number of time series that belongs to the gth component, $\boldsymbol{D}_g = \mathrm{diag}(\sigma_\alpha^2 \mathbf{1}_2, \tau_g^2 \mathbf{1}_m)$ is the prior covariance matrix for $\boldsymbol{\theta}_g$. Hence, for each component g, we can draw $(\boldsymbol{\theta}_g \mid \sigma_g^2, \tau_g^2, \boldsymbol{y}, \boldsymbol{S}) \sim N(\boldsymbol{\mu}_g, \sigma_g^2 \boldsymbol{\Lambda}_g)$.

b. **Sampling error variances**

We expand the half-t prior by augmenting the posterior of error variances with a latent variable a_{σ_g}, using the hierarchical structure

$$
\sigma_g^2 \mid a_{\sigma_g} \sim IG\Big(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma}{a_{\sigma_g}}\Big), a_{\sigma_g} \sim IG\Big(\frac{1}{2}, \frac{1}{A_\sigma^2}\Big),
$$

so that the full conditional distributions are

$$
p(a_{\sigma_g} \mid \sigma_g^2) \propto p(\sigma_g^2 \mid a_{\sigma_g}) p(a_{\sigma_g}) \propto \exp\Big[-\frac{1}{a_{\sigma_g}} \Big(\frac{\nu_\sigma}{\sigma_g^2} + \frac{1}{A_\sigma^2} \Big) \Big] \cdot (a_{\sigma_g})^{-(\frac{1}{2}+1+\frac{\nu_\sigma}{2})},
$$

which can be sampled from $IG\Big(\frac{\nu_\sigma+1}{2}, \frac{\nu_\sigma}{\sigma_g^2} + \frac{1}{A_\sigma^2}\Big)$. Denoting $\boldsymbol{\epsilon}_{ig}$ as the error of time series $\boldsymbol{y}_i$ for the component g and $\boldsymbol{\epsilon}_{ig} = \boldsymbol{y}_i - \boldsymbol{S}\boldsymbol{\theta}_g$, where $\boldsymbol{\epsilon}_{ig} \sim N(\mathbf{0}, \sigma_g^2 \boldsymbol{I}_n)$. Thus we

have

$$p(\sigma_g^2 \mid \boldsymbol{\epsilon}_{ig}, \gamma_g) \propto p(\boldsymbol{\epsilon}_{ig} \mid \sigma_g^2) p(a_{\sigma_g} \mid \sigma_g^2) p(\sigma_g^2)$$
$$\propto \prod_{i=1}^{N} \Big[(\sigma_g^2)^{-\frac{n}{2}} \exp\Big(-\frac{1}{2\sigma_g^2} \boldsymbol{\epsilon}_{ig}' \boldsymbol{\epsilon}_{ig} \Big) \Big]^{z_{ig}} \times (\sigma_g^2)^{-(\frac{\nu_\sigma}{2}+1)} \exp\Big(-\frac{\nu_\sigma}{\sigma_g^2 a_{\sigma_g}} \Big)$$
$$\propto (\sigma_g^2)^{-(\frac{n}{2}N_g + \frac{\nu_\sigma}{2}+1)} \cdot \exp\Big[-\frac{1}{\sigma_g^2} \big(\frac{\sum_{i=1}^{N} z_{ig} \boldsymbol{\epsilon}_{ig}' \boldsymbol{\epsilon}_{ig}}{2} + \frac{\nu_\sigma}{a_{\sigma_g}} \big) \Big],$$

which can be sampled from $IG\Big(\frac{nN_g+\nu_\sigma}{2}, \frac{\sum_{i=1}^{N} z_{ig} \boldsymbol{\epsilon}_{ig}' \boldsymbol{\epsilon}_{ig}}{2} + \frac{\nu_\sigma}{a_{\sigma_g}}\Big)$. The sampling scheme proceeds by first sampling $(a_{\sigma_g} \mid \sigma_g^2)$ than $(\sigma_g^2 \mid \boldsymbol{\epsilon}_{ig}, a_{\sigma_g})$.

c. **Sampling smoothing parameters**

Following the same procedure of sampling error variances, we can sample smoothing parameter τ_g^2 by introducing a latent variable a_{τ_g}. We first draw $(a_{\tau_g} \mid \tau_g^2) \sim IG\Big(\frac{\nu_\tau+1}{2}, \frac{\nu_\tau}{\tau_g^2} + \frac{1}{A_\tau^2}\Big)$, then we have

$$p(\tau_g^2 \mid \boldsymbol{\beta}_g, a_{\tau_g}) \propto p(\boldsymbol{\beta}_g \mid \tau_g^2) p(a_{\tau_g} \mid \tau_g^2) p(\tau_g^2)$$
$$\propto (\tau_g^2)^{-\frac{m+\nu_\tau}{2}} \cdot \exp\Big[-\frac{1}{\tau_g^2} \big(\frac{\nu_\tau}{a_{\tau_g}} + \frac{\boldsymbol{\beta}_g' \boldsymbol{\beta}_g}{2} \big) \Big],$$

which can be sampled from $IG\Big(\frac{\nu_\tau+m}{2}, \frac{\boldsymbol{\beta}_g' \boldsymbol{\beta}_g}{2} + \frac{\nu_\tau}{a_{\tau_g}}\Big)$. The sampling scheme proceeds by first sampling $(a_{\tau_g} \mid \tau_g^2)$ then $(\tau_g^2 \mid \boldsymbol{\beta}_g, a_{\tau_g})$.

d. **Sampling logistic parameters**

Based on (3) and Section 2.3, the conditional posterior distribution of $(\boldsymbol{\delta}_g^* \mid z_{ig}, \boldsymbol{V}^*)$ is

$$p(\boldsymbol{\delta}_g^* \mid z_{ig}, \boldsymbol{V}^*) \propto p(z_{ig} = 1 \mid \boldsymbol{\delta}_g^*, \boldsymbol{V}^*) p(\boldsymbol{\delta}_g^*)$$
$$= \prod_{i=1}^{N} \Big[\frac{\exp(\boldsymbol{V}_i^{*\prime} \boldsymbol{\delta}_g^*)}{\sum_{h=1}^{G} \exp(\boldsymbol{V}_i^{*\prime} \boldsymbol{\delta}_h^*)} \Big]^{z_{ig}} \cdot p(\boldsymbol{\delta}_g^*).$$

To sample from the posterior distribution of $p(\boldsymbol{\delta}_g^* \mid z_{ig}, \boldsymbol{V}^*)$, we use a data augmentation strategy from Polson et al. 2013 by introducing a latent Póyla-Gamma variable ω_{ig}, which comes from the Pólya-Gamma distribution. Thus, we have

$$p(\boldsymbol{\delta}_g^* \mid z_{ig}, \omega_{ig}, \boldsymbol{V}^*) \propto p(z_{ig} = 1 \mid \omega_{ig}, \boldsymbol{\delta}_g^*, \boldsymbol{V}^*) p(\boldsymbol{\delta}_g^*)$$
$$\propto \exp\Big(-\frac{\omega_{ig} \eta_{ig}^2}{2} \Big) p(\omega_{ig} \mid 1, 0) |\boldsymbol{B}_g|^{-P/2} \exp\Big(-\frac{1}{2} \boldsymbol{\delta}_g^{*\prime} \boldsymbol{B}_g^{-1} \boldsymbol{\delta}_g^* \Big),$$

where $\eta_{ig} = \boldsymbol{V}_i^{*\prime}\boldsymbol{\delta}_g^* - C_{ig}$ and $C_{ig} = log\sum_{h\neq j}\exp(\boldsymbol{V}_i^{*\prime}\boldsymbol{\delta}_h^*)$, $p(\omega_{ig} \mid 1, 0)$ is the Pólya-gamma distribution $PG(b, c)$ with $b = 1$ and $c = 0$, $\boldsymbol{B}_g$ is the prior covariance matrix from section 2.3 and $\boldsymbol{B}_g = \text{diag}(\sigma_{\zeta g}^2\boldsymbol{I}_{P+1}, \kappa_{\zeta g}^2\boldsymbol{I}_N)$. Using the conjugate prior $\boldsymbol{\delta}_g^* \sim N(\boldsymbol{0}, \boldsymbol{B}_g)$, we have

$$\omega_{ig} \mid \boldsymbol{\delta}_g^*, \boldsymbol{V}^* \sim PG(1, \eta_{ig}),$$

$$\boldsymbol{\delta}_g^* \mid z_{ig}, \omega_{ig}, \boldsymbol{V}^* \sim N(\boldsymbol{M}_g, \boldsymbol{\Sigma}_g),$$

where posterior variance $\boldsymbol{\Sigma}_g = (\boldsymbol{V}^{*\prime}\boldsymbol{\Omega}_g\boldsymbol{V}^* + \boldsymbol{B}_g^{-1})^{-1}$ and posterior mean $\boldsymbol{M}_g = \boldsymbol{\Sigma}_g\Big[\boldsymbol{V}^{*\prime}(\boldsymbol{\Omega}_g\boldsymbol{C}_g + \boldsymbol{\xi}_g)\Big]$. Here $\boldsymbol{\Omega}_g = \text{diag}(\omega_{1g}, \ldots, \omega_{Ng})$, $\boldsymbol{C}_g = (C_{1j}, \ldots, C_{Ng})'$, and $\boldsymbol{\xi}_g = (\xi_{1g}, \ldots, \xi_{Ng})'$, where $\xi_{ig} = z_{ig} - \frac{1}{2}$. Thus, $\boldsymbol{\delta}_g^*$ can be sampled by first drawing $(\omega_{ig} \mid \boldsymbol{\delta}_g^*, \boldsymbol{V}^*)$ and then drawing $(\boldsymbol{\delta}_g^* \mid z_{ig}, \omega_{ig}, \boldsymbol{V}^*)$.

e. **Sampling variances of the random intercept**

Following the same procedure for sampling τ_g^2, we first draw $(a_{\kappa_g} \mid \kappa_{\zeta g}^2) \sim IG\Big(\frac{\nu_\kappa+1}{2}, \frac{\nu_\kappa}{\kappa_{\zeta g}^2} + \frac{1}{A_\kappa^2}\Big)$, then we have

$$\begin{aligned} p(\kappa_{\zeta g}^2 \mid \boldsymbol{\zeta}_g, a_{\kappa_g}) &\propto p(\boldsymbol{\zeta}_g \mid \kappa_{\zeta g}^2)p(a_{\kappa_g} \mid \kappa_{\zeta g}^2)p(\kappa_{\zeta g}^2) \\ &\propto (\kappa_{\zeta g}^2)^{-\frac{N+\nu_\kappa}{2}+1} \cdot \exp\Big[-\frac{1}{\kappa_{\zeta g}^2}(\frac{\nu_\kappa}{a_{\kappa_g}} + \frac{\boldsymbol{\zeta}_g^T\boldsymbol{\zeta}_g}{2})\Big], \end{aligned}$$

which can be sampled from $IG\Big(\frac{\nu_\kappa+N}{2}, \frac{\boldsymbol{\zeta}_g^T\boldsymbol{\zeta}_g}{2} + \frac{\nu_\kappa}{a_{\kappa_g}}\Big)$. The sampling scheme proceeds by first sampling $(a_{\kappa_g} \mid \kappa_{\zeta g}^2)$ than $(\kappa_{\zeta g}^2 \mid \boldsymbol{\zeta}_g, a_{\kappa_g})$.

f. **Computing mixing weights**

After drawing $\boldsymbol{\delta}_g^*$ from its posterior distribution, we can directly compute mixing weights π_{ig} from (3) with known design matrix $\boldsymbol{V}^*$.

g. **Sampling latent indicators**

After obtaining all estimated parameters and computing mixing weights, the final step is to allocate subjects to different components by sampling the latent indicator z_{ig}. The latent indicator z_{ig} has the following distribution as in (5):

$$f(z_{ig} = 1 \mid \boldsymbol{\Theta}, \boldsymbol{y}_i) = \frac{\pi_{ig}f_g(\boldsymbol{y}_i \mid \boldsymbol{\Theta}_g)}{\sum_{h=1}^G \boldsymbol{\pi}_{ih}f_h(\boldsymbol{y}_i \mid \boldsymbol{\Theta}_h)},$$

where the latent indicator z_{ig} can be drawn from the multinomial distribution.

A.2 Additional simulation results

Figure A.2.1. An example of estimated trajectories with true trajectories from one replicate of M1.

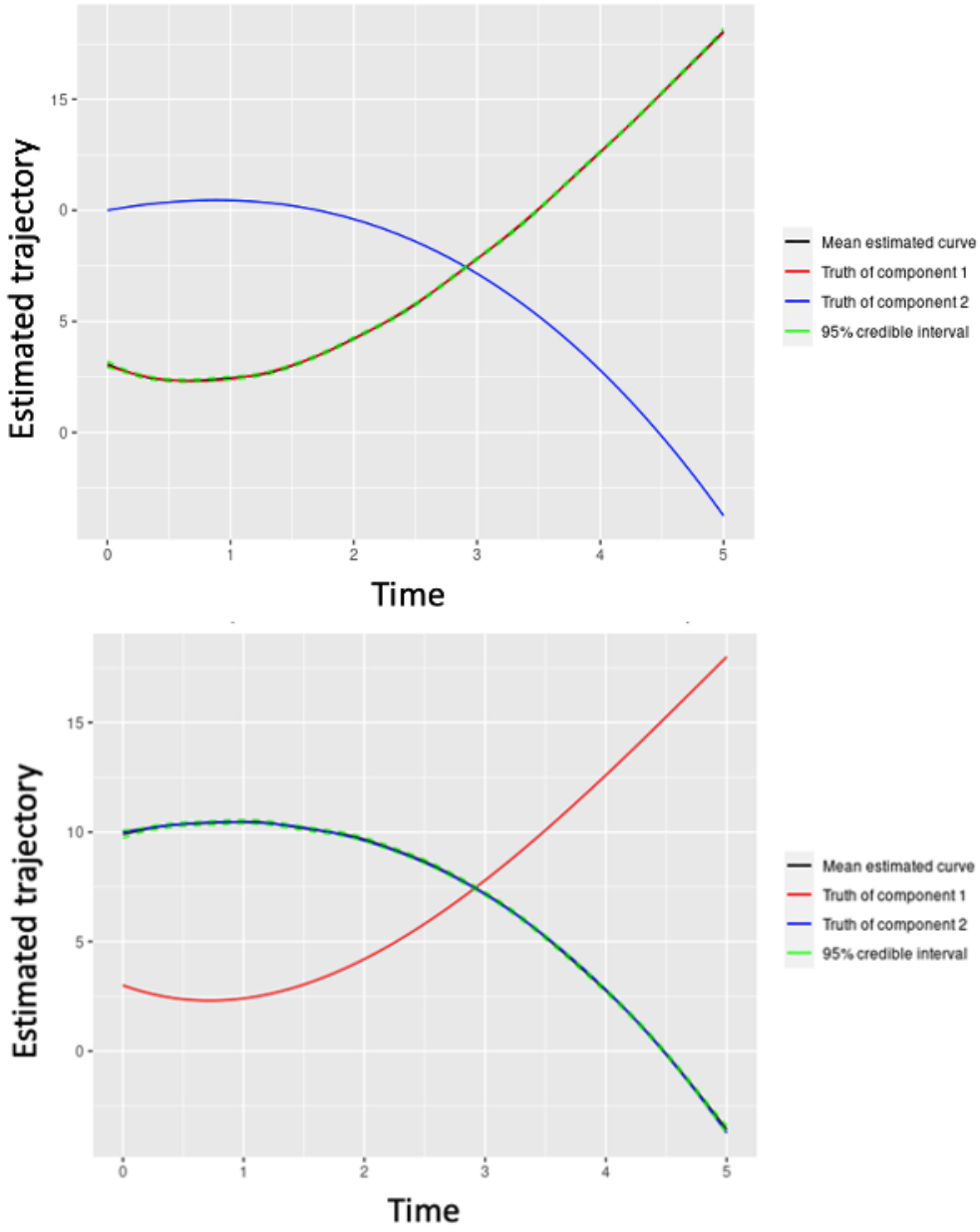
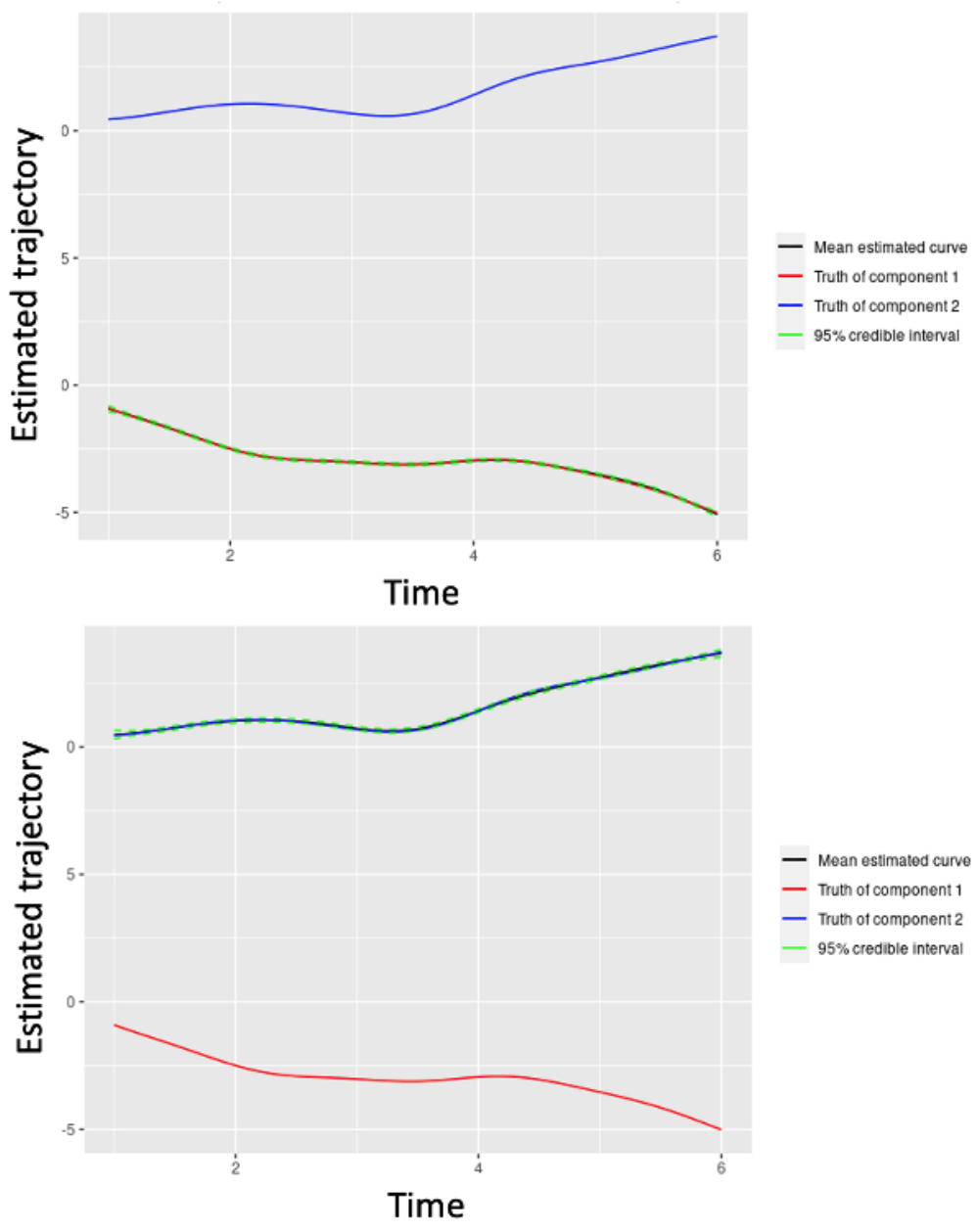


Figure A.2.2. An example of estimated trajectories with true trajectories from one replicate of M4.



A.3 Additional real-data results

Figure A.3.3. Estimated trajectories with 95% pointwise credible intervals for the two-component model of channel S1D3. **I**: Interact **S**: Still-face **R**: Recovery.

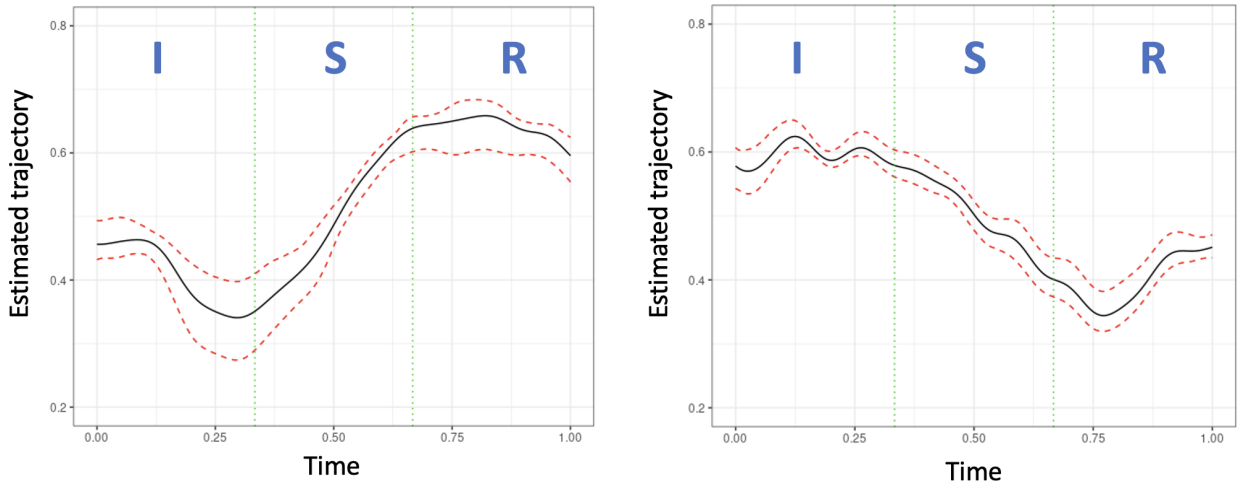


Figure A.3.4. Estimated trajectories with pointwise 95% credible intervals for the two-component model of channel S5D1. **I**: Interact **S**: Still-face **R**: Recovery.

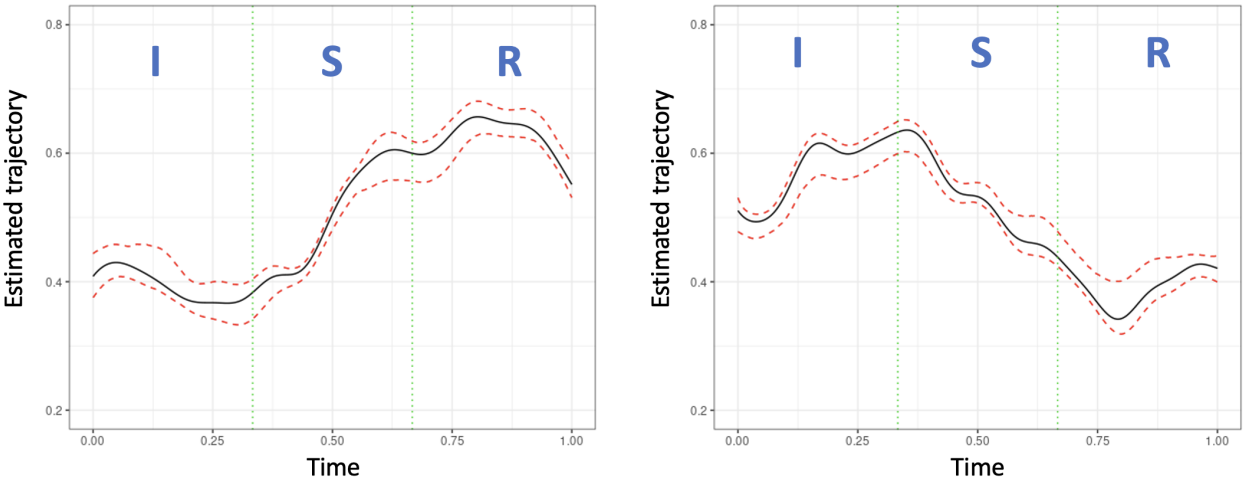


Figure A.3.5. Estimated trajectories with 95% pointwise credible intervals for the two-component model of channel S6D4. **I**: Interact **S**: Still-face **R**: Recovery.

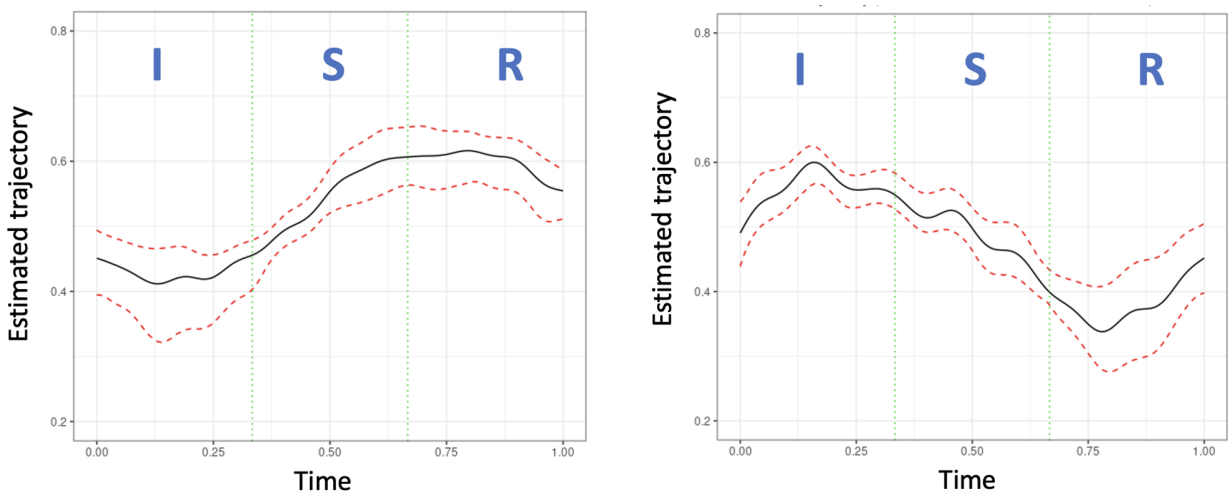
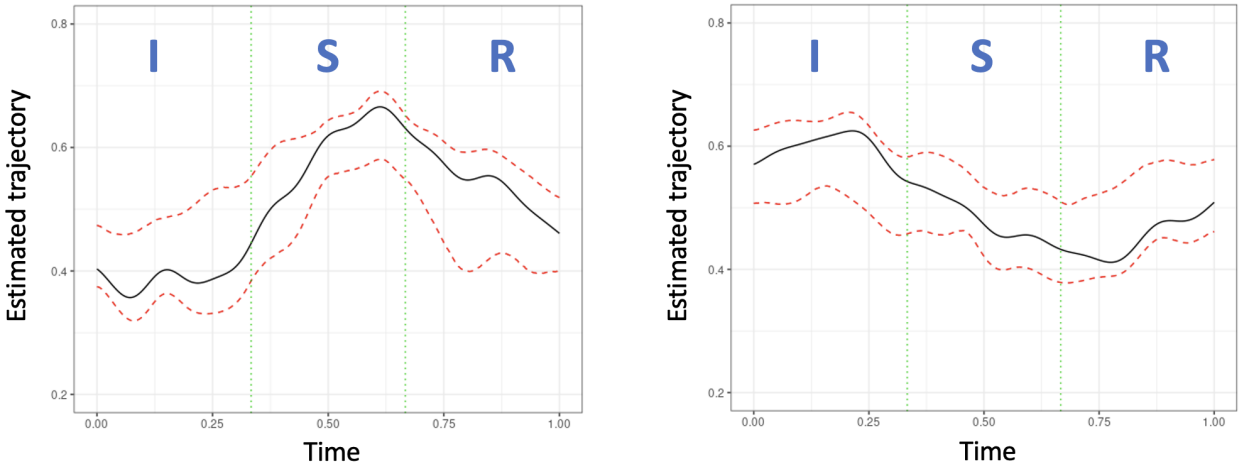


Figure A.3.6. Estimated trajectories with pointwise 95% credible intervals for the two-component model of channel S7D4. **I**: Interact **S**: Still-face **R**: Recovery.



Appendix B Chapter 3

B.1 Detailed Gibbs sampling scheme

Sampling $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1', \ldots, \boldsymbol{\Theta}_g', \ldots, \boldsymbol{\Theta}_G')'$ does not require the change of parameter dimensions or the number of components. We can sample parameters from each kth dimension separately from their conditional posterior distribution using Gibbs sampling. Denoting $\boldsymbol{\Theta}_{gk} = (\boldsymbol{\theta}_{gk}', \tau_{gk}^2, \sigma_{gk}^2, \boldsymbol{\delta}_g^{*\prime}, \kappa_{\zeta g}^2)'$ as parameters for gth component and kth dimension, detailed Gibbs sampling scheme is shown as below.

1. **Sampling model coefficients**

   For each component g and time series dimension k, based on the augmented likelihood and priors on $\boldsymbol{\theta}_{gk} = (\boldsymbol{\alpha}_{gk}', \boldsymbol{\beta}_{gk}')'$, the conditional posterior distribution of $\boldsymbol{\theta}_{gk} \mid \boldsymbol{y}, \boldsymbol{S}, \tau_{gk}^2, \sigma_{gk}^2$ is:

$$
\begin{aligned}
p(\boldsymbol{\theta}_{gk} \mid \boldsymbol{y}, \boldsymbol{S}, \tau_{gk}^2, \sigma_{gk}^2) &\propto p(\boldsymbol{y} \mid \boldsymbol{S}, \boldsymbol{\theta}_{gk}, \sigma_{gk}^2) \cdot p(\boldsymbol{\theta}_{gk} \mid \tau_{gk}^2) \\
&\propto \prod_{i=1}^{N} [(\sigma_{gk}^2)^{-n/2} \exp\{-\frac{1}{2\sigma_{gk}^2}(\boldsymbol{y}_{ik} - \boldsymbol{S}\boldsymbol{\theta}_{gk})'(\boldsymbol{y}_{ik} - \boldsymbol{S}\boldsymbol{\theta}_{gk})\}]^{z_{ig}} \\
&\times |\boldsymbol{D}_{gk}|^{-1/2} \exp\{-\frac{1}{2}\boldsymbol{\theta}_{gk}' \boldsymbol{D}_{gk}^{-1} \boldsymbol{\theta}_{gk}\} \\
&\propto \exp\{-\frac{1}{2\sigma_{gk}^2}[\sum_{i=1}^{N} z_{ig}(\boldsymbol{y}_{ik} - \boldsymbol{S}\boldsymbol{\theta}_{gk})'(\boldsymbol{y}_{ik} - \boldsymbol{S}\boldsymbol{\theta}_{gk}) + \boldsymbol{\theta}_{gk}' \sigma_{gk}^2 \boldsymbol{D}_{gk}^{-1} \boldsymbol{\theta}_{gk}]\} \\
&\propto \exp\{-\frac{1}{2\sigma_{gk}^2}(\boldsymbol{\theta}_{gk} - \boldsymbol{\mu}_{gk})'(\boldsymbol{\Lambda}_{gk})^{-1}(\boldsymbol{\theta}_{gk} - \boldsymbol{\mu}_{gk})\} \\
&\sim N(\boldsymbol{\mu}_{gk}, \sigma_{gk}^2 \boldsymbol{\Lambda}_{gk}),
\end{aligned}
$$

   where $\boldsymbol{\Lambda}_{gk} = (N_g \boldsymbol{S}'\boldsymbol{S} + \sigma_{gk}^2 \boldsymbol{D}_{gk}^{-1})^{-1}$ and $\boldsymbol{\mu}_{gk} = \boldsymbol{\Lambda}_{gk} \sum_{i=1}^{N} z_{ig} \boldsymbol{S}' \boldsymbol{y}_{ik}$, N_g is the number of time series that belongs to the gth component, $\boldsymbol{D}_{gk} = \text{diag}(\sigma_\alpha^2 \mathbf{1}_2, \tau_{gk}^2 \mathbf{1}_m)$ is the prior covariance matrix for $\boldsymbol{\theta}_{gk}$. Hence, for each component g and entry k, we can draw $\boldsymbol{\theta}_{gk} \mid \sigma_{gk}^2, \tau_{gk}^2, \boldsymbol{y}, \boldsymbol{S} \sim N(\boldsymbol{\mu}_{gk}, \sigma_{gk}^2 \boldsymbol{\Lambda}_{gk})$.

2. **Sampling error variances**

   We follow Wand et al.(2012), expanding the half-t prior in 3.3.3 by augmenting the

posterior of error variances with a latent variable $a_{\sigma_{gk}}$, using the hierarchical structure

$$\sigma^2_{gk} \mid a_{\sigma_{gk}} \sim IG(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma}{a_{\sigma_{gk}}}), a_{\sigma_{gk}} \sim IG(\frac{1}{2}, \frac{1}{A^2_\sigma}),$$

so that the full conditional distributions are

$$p(a_{\sigma_{gk}} \mid \sigma^2_{gk}) \propto p(\sigma^2_{gk} \mid a_{\sigma_{gk}}) p(a_{\sigma_{gk}}) \propto \exp\Big(- \frac{1}{a_{\sigma_{gk}}} (\frac{\nu_\sigma}{\sigma^2_{gk}} + \frac{1}{A^2_\sigma}) \Big) \cdot (a_{\sigma_{gk}})^{-(\frac{1}{2}+1+\frac{\nu_\sigma}{2})},$$

which can be sampled from $IG(\frac{\nu_\sigma+1}{2}, \frac{\nu_\sigma}{\sigma^2_{gk}} + \frac{1}{A^2_\sigma})$. Denoting $\boldsymbol{\epsilon}_{igk}$ as the error of time series $\boldsymbol{y}_{ik}$ for the component g and $\boldsymbol{\epsilon}_{igk} = \boldsymbol{y}_{ik} - \boldsymbol{S}\boldsymbol{\theta}_{gk}$, where $\boldsymbol{\epsilon}_{igk} \sim N(\boldsymbol{0}, \sigma^2_{gk}\boldsymbol{I}_n)$. Thus we have

$$\begin{aligned} p(\sigma^2_{gk} \mid \boldsymbol{\epsilon}_{igk}, a_{\sigma_{gk}}) &\propto p(\boldsymbol{\epsilon}_{igk} \mid \sigma^2_{gk}) p(a_{\sigma_{gk}} \mid \sigma^2_{gk}) p(\sigma^2_{gk}) \\ &\propto \prod_{i=1}^{N} [(\sigma^2_{gk})^{-\frac{n}{2}} \exp(-\frac{1}{2\sigma^2_{gk}} \boldsymbol{\epsilon}'_{igk}\boldsymbol{\epsilon}_{igk})]^{z_{ig}} \times (\sigma^2_{gk})^{-(\frac{\nu_\sigma}{2}+1)} \exp(-\frac{\nu_\sigma}{\sigma^2_{gk} a_{\sigma_{gk}}}) \\ &\propto (\sigma^2_{gk})^{-(\frac{n}{2}N_g + \frac{\nu_\sigma}{2}+1)} \cdot \exp\Big(- \frac{1}{\sigma^2_{gk}} (\frac{\sum_{i=1}^{N} z_{ig} \boldsymbol{\epsilon}'_{igk}\boldsymbol{\epsilon}_{igk}}{2} + \frac{\nu_\sigma}{a_{\sigma_{gk}}}) \Big), \end{aligned}$$

which can be sampled from $IG(\frac{nN_g+\nu_\sigma}{2}, \frac{\sum_{i=1}^{N} z_{ig}\boldsymbol{\epsilon}'_{igk}\boldsymbol{\epsilon}_{igk}}{2} + \frac{\nu_\sigma}{a_{\sigma_{gk}}})$. The sampling scheme proceeds by first sampling $a_{\sigma_{gk}} \mid \sigma^2_{gk}$ than $\sigma^2_{gk} \mid \boldsymbol{\epsilon}_{igk}, a_{\sigma_{gk}}$.

3. **Sampling smoothing parameters**

   Following the same procedure in sampling error variance, we can sample smoothing parameter τ^2_{gk} by introducing a latent variable $a_{\tau_{gk}}$. We first draw $a_{\tau_{gk}} \mid \tau^2_{gk} \sim IG(\frac{\nu_\tau+1}{2}, \frac{\nu_\tau}{\tau^2_{gk}} + \frac{1}{A^2_\tau})$, then we have

$$\begin{aligned} p(\tau^2_{gk} \mid \boldsymbol{\beta}_{gk}, a_{\tau_{gk}}) &\propto p(\boldsymbol{\beta}_{gk} \mid \tau^2_{gk}) p(a_{\tau_{gk}} \mid \tau^2_{gk}) p(\tau^2_{gk}) \\ &\propto (\tau^2_{gk})^{-\frac{m+\nu_\tau}{2}} \cdot \exp\Big(- \frac{1}{\tau^2_{gk}} (\frac{\nu_\tau}{a_{\tau_{gk}}} + \frac{\boldsymbol{\beta}'_{gk}\boldsymbol{\beta}_{gk}}{2}) \Big), \end{aligned}$$

   which can be sampled from $IG(\frac{\nu_\tau+m}{2}, \frac{\boldsymbol{\beta}'_{gk}\boldsymbol{\beta}_{gk}}{2} + \frac{\nu_\tau}{a_{\tau_{gk}}})$. The sampling scheme proceeds by first sampling $a_{\tau_{gk}} \mid \tau^2_{gk}$ then $\tau^2_{gk} \mid \boldsymbol{\beta}_{gk}, a_{\tau_{gk}}$.

4. **Sampling logistic parameters**

   Sampling logistic parameters utilize the same method as shown in the sampling logistic parameters of within-model moves of RJMCMC in Appendix A.1.

5. **Sampling variances of the random intercept**

Sampling variances of the random intercept adopt the same method as shown in the sampling variances of the random intercept in RJMCMC within-model moves in Appendix A.1.

6. **Computing mixing weights**

   After drawing $\boldsymbol{\delta}_g$ from its posterior distribution, we can directly compute mixing weights π_{ig} for each component from (10) with known design matrix V^*.

7. **Sampling latent indicators**

   After obtaining all estimated parameters and computing mixing weights, the final step is to allocate subjects to different components by sampling the latent indicator z_{ig}. The latent indicator z_{ig} has the following distribution as in (13):

$$f(z_{ig} = 1 \mid \boldsymbol{\Theta}_g, \boldsymbol{S}, \boldsymbol{y}_i) = \frac{\pi_{ig} \prod_{j=1}^{n} f_g(\boldsymbol{y}_{it_j} \mid \boldsymbol{\Theta}_g)}{\sum_{h=1}^{G} \boldsymbol{\pi}_{ih} \prod_{j=1}^{n} f_h(\boldsymbol{y}_{it_j} \mid \boldsymbol{\Theta}_h)}, \tag{33}$$

   which can be drawn directly from the multinomial distribution.

B.2 Additional simulation results

Table B.2.1. Results of logistic parameters for the two-component trivariate scenario in Simulation I: values in each cell are in the format of RMSE (bias, variance).

Model Setting	Comparison	δ_0	δ_1	δ_2	δ_3
M1	C1 vs C2	0.90 (0.10,0.81)	0.42 (-0.09,0.17)	0.23 (0.03,0.05)	0.31 (-0.01,0.10)
M2	C1 vs C2	0.89 (0.09,0.80)	0.41(-0.09,0.16)	0.23 (0.03,0.05)	0.31 (-0.01,0.10)
M3	C1 vs C2	0.86 (0.07,0.73)	0.50(-0.13,0.23)	0.29 (0.08,0.08)	0.31 (-0.05,0.10)
M4	C1 vs C2	0.86 (0.07,0.74)	0.50 (-0.13,0.23)	0.29 (0.08,0.08)	0.31 (-0.05,0.10)

Figure B.2.7. Boxplots of RMSE, bias and variance of trajectory estimates for model setting M1 vs. M2 in Simulation I: two-component trivariate model.

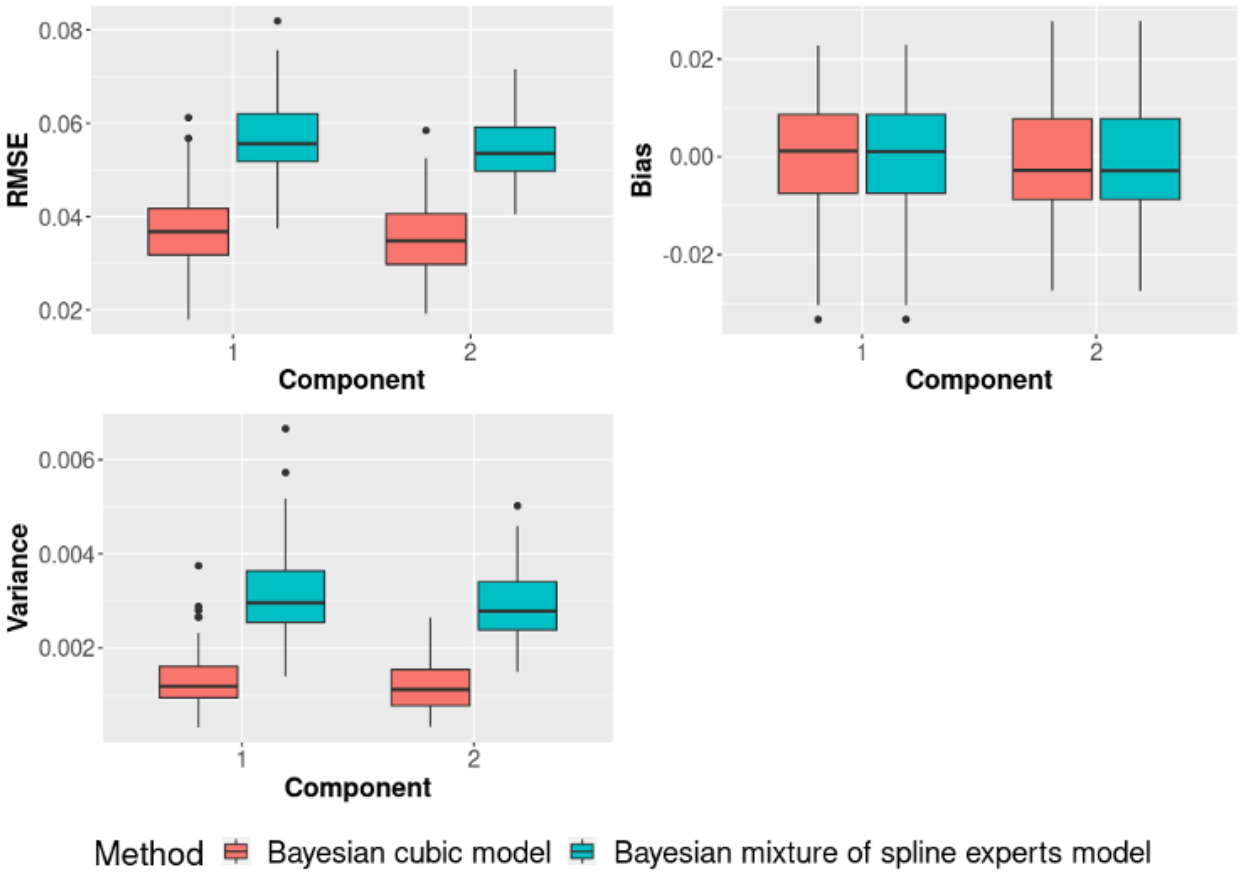


Figure B.2.8. Boxplots of RMSE, bias and variance of trajectory estimates for model setting M3 vs. M4 in Simulation I: two-component trivariate model.

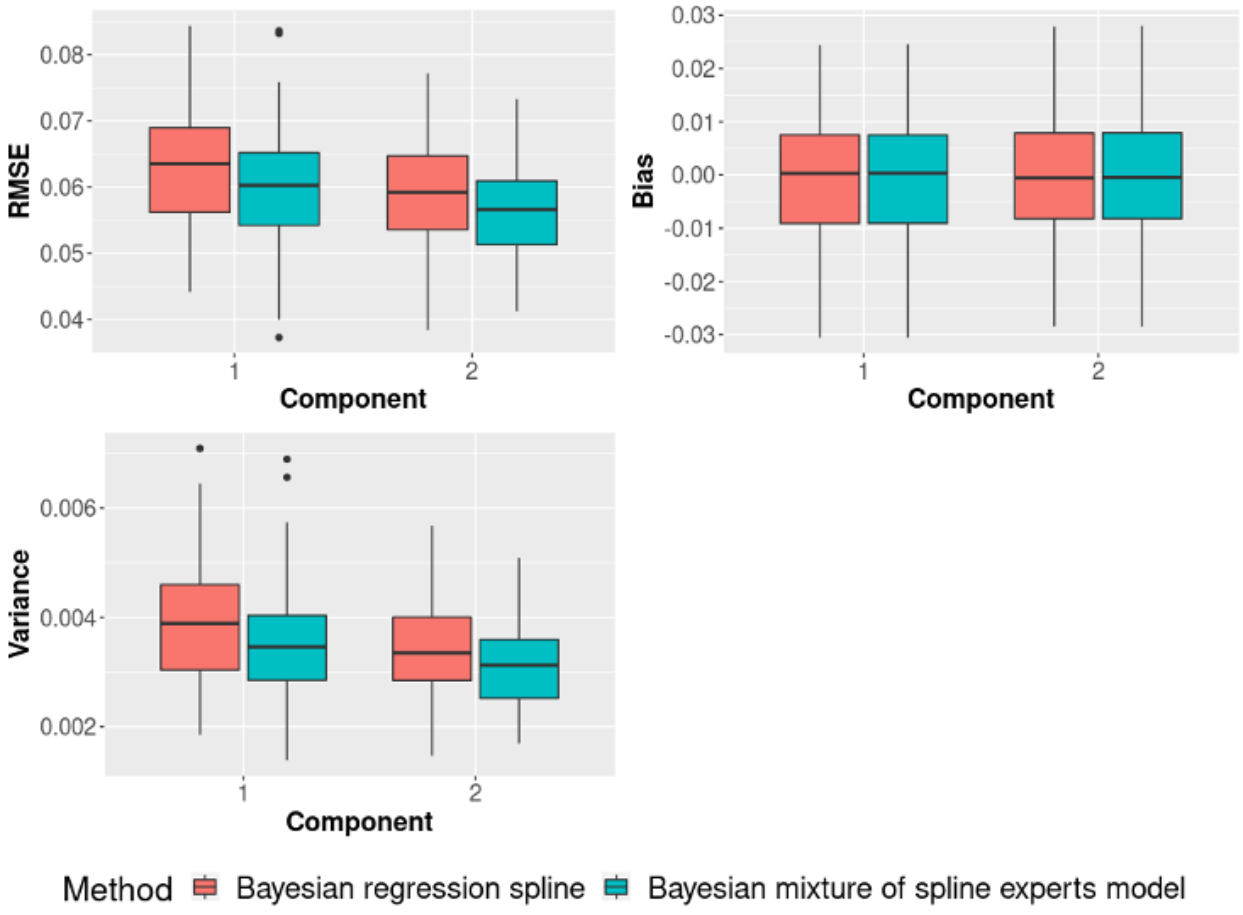


Table B.2.2. Root mean square errors (RMSEs) of each logistic parameter for the two-component trivariate model from 100 replicates of N two-component trivariate time series of length n. RMSEs of the proposed method were compared to `TRAJ` procedure in `SAS`. Parameters δ_0, δ_1, δ_2 and δ_3 are intercept, first, second and third logistic parameters, respectively. The true values of logistic parameters are $5, -3.5, 1, 0.1$, respectively.

n	N	Method	δ_0	δ_1	δ_2	δ_3
50	150	Proposed	0.89	0.52	0.29	0.32
		TRAJ	1.57	0.87	0.36	0.34
70	150	Proposed	0.86	0.50	0.29	0.31
		TRAJ	1.55	0.86	0.36	0.34
50	250	Proposed	0.77	0.40	0.22	0.23
		TRAJ	0.96	0.50	0.23	0.24
70	250	Proposed	0.77	0.41	0.22	0.23
		TRAJ	0.97	0.51	0.24	0.24

Figure B.2.9. Boxplots of the averaged root square error (ARSE), the averaged bias (A-bias) and the variance of bias (V-bias) of estimated trajectories for each component from 100 replicates of 150 two-component trivariate time series of length 50. Estimates of the proposed method were compared to `R` package `gbmt` and `TRAJ` procedure in `SAS`. The diamond markers denote the means of each estimate.

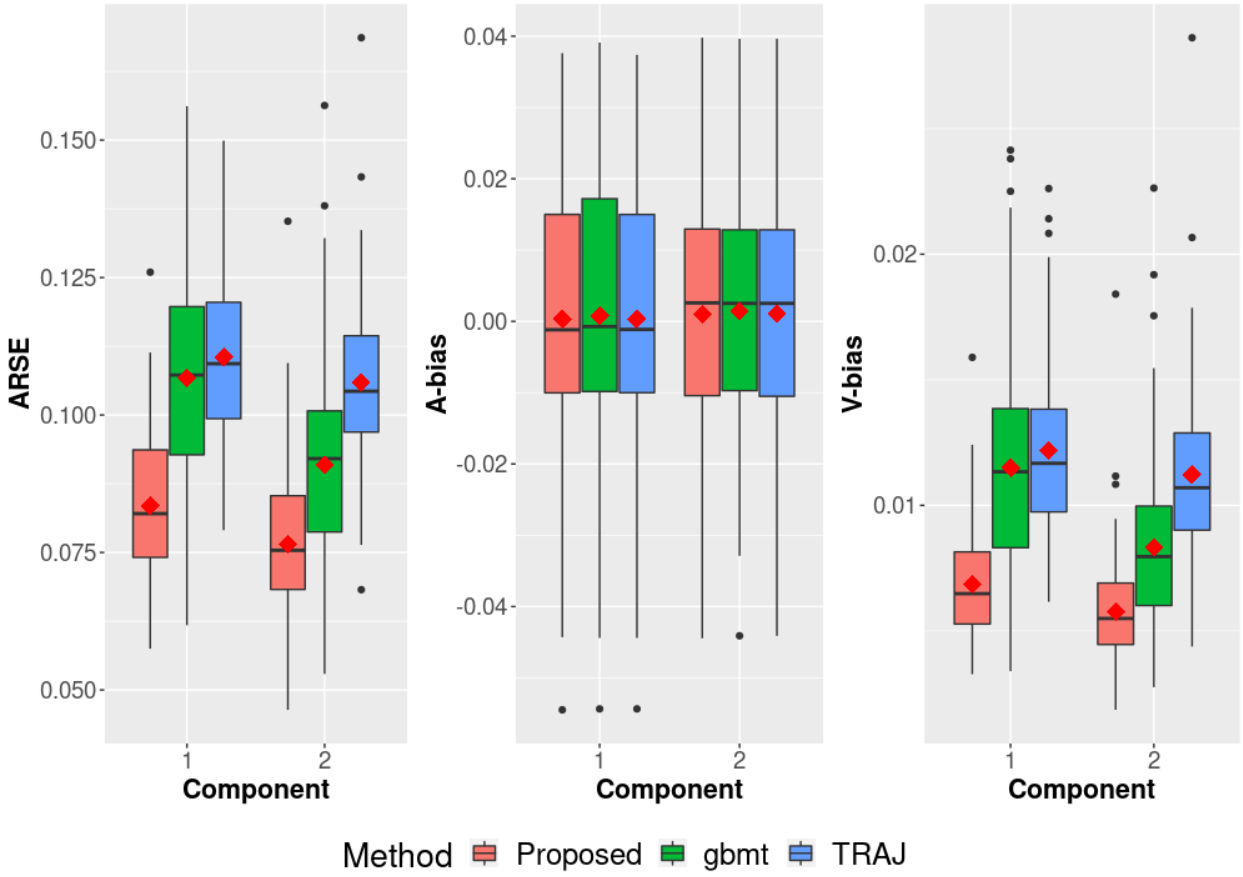


Table B.2.3. Mean (standard deviation) of the averaged root square error (ARSE), the averaged bias (A-bias) and the variance of bias (V-bias) of estimated trajectories for each component from 100 replicates of N two-component trivariate time series of length n. Estimates of the proposed method were compared to `R` package `gbmt` and `TRAJ` procedure in `SAS`. C1 and C2 denote the first and the second components. Means were calculated by averaging over estimates of 100 replicates. Standard deviations are Monte Carlo standard deviations from estimates of 100 replicates. Each value was reported $\times 10^2$.

n	N	Method	ARSE C1	A-bias C1	V-bias C1	ARSE C2	A-bias C2	V-bias C2
50	150	Proposed	8.35	0.03	0.68	7.65	0.10	0.58
			(1.26)	(1.83)	(0.22)	(1.38)	(1.76)	(0.22)
		gbmt	10.67	0.03	1.15	9.08	0.10	0.83
			(1.91)	(1.83)	(0.42)	(1.72)	(1.76)	(0.33)
		TRAJ	11.06	0.03	1.22	10.59	0.10	1.12
			(1.48)	(1.83)	(0.33)	(1.52)	(1.76)	(0.34)
70	150	Proposed	7.16	0.24	0.51	6.53	-0.11	0.42
			(1.04)	(1.38)	(0.16)	(1.07)	(1.42)	(0.14)
		gbmt	9.91	0.24	1.01	8.19	-0.11	0.68
			(1.96)	(1.38)	(0.40)	(1.65)	(1.42)	(0.29)
		TRAJ	9.34	0.24	0.87	8.95	-0.11	0.80
			(1.10)	(1.38)	(0.22)	(1.13)	(1.42)	(0.20)
50	250	Proposed	6.81	0.07	0.46	6.22	0.02	0.38
			(1.02)	(1.33)	(0.14)	(0.94)	(1.31)	(0.12)
		gbmt	9.79	0.07	0.98	8.00	0.02	0.65
			(1.91)	(1.33)	(0.40)	(1.53)	(1.31)	(0.26)
		TRAJ	8.70	0.07	0.76	8.20	0.02	0.67
			(1.18)	(1.33)	(0.21)	(1.03)	(1.31)	(0.17)
70	250	Proposed	5.65	0.08	0.32	5.27	-0.06	0.27
			(0.84)	(1.00)	(0.10)	(0.82)	(1.42)	(0.09)
		gbmt	9.15	0.08	0.87	7.43	-0.06	0.56
			(1.96)	(1.00)	(0.38)	(1.60)	(1.42)	(0.26)
		TRAJ	7.18	0.08	0.52	6.80	-0.06	0.45
			(0.94)	(1.00)	(0.14)	(0.77)	(1.42)	(0.10)

Table B.2.4. Mean (standard deviation) of the averaged root square error (ARSE), the averaged bias (A-bias) and the variance of bias (V-bias) of estimated trajectories for each component from 100 replicates of 150 four-component bivariate time series of length 50. Estimates of the proposed method were compared to `R` package `gbmt` and `TRAJ` procedure in `SAS`. C1, C2, C3 and C4 denote first, second, third and fourth component, respectively. Means were calculated by averaging over estimates of 100 replicates. Standard deviations are Monte Carlo standard deviations from estimates of 100 replicates. Each value was reported $\times 10^2$.

n	N	Method	ARSE C1	A-bias C1	V-bias C1	ARSE C2	A-bias C2	V-bias C2
50	150	Proposed	4.38	0.38	0.18	3.76	-0.01	0.13
			(1.04)	(1.59)	(0.08)	(0.87)	(1.37)	(0.06)
		gbmt	4.75	0.38	0.21	4.79	0.01	0.22
			(1.04)	(1.59)	(0.09)	(1.17)	(1.65)	(0.11)
		TRAJ	13.87	0.62	3.39	12.41	0.08	2.41
			(15.59)	(9.91)	(9.24)	(13.42)	(9.74)	(5.37)
n	N	Method	ARSE C3	A-bias C3	V-bias C3	ARSE C4	A-bias C4	V-bias C4
50	150	Proposed	4.69	-0.11	0.20	3.88	-0.09	0.14
			(1.14)	(1.83)	(0.12)	(1.15)	(1.56)	(0.08)
		gbmt	5.08	-0.12	0.24	4.70	-0.09	0.21
			(1.12)	(1.83)	(0.12)	(1.32)	(1.78)	(0.12)
		TRAJ	14.55	-1.58	3.24	14.36	0.12	3.99
			(14.82)	(10.31)	(7.35)	(17.01)	(9.92)	(10.70)

Table B.2.5. Mean (standard deviation) of the averaged root square error (ARSE), the averaged bias (A-bias) and the variance of bias (V-bias) of estimated trajectories for each component from 100 replicates of 150 four-component bivariate time series of length 70. Estimates of the proposed method were compared to `R` package `gbmt` and `TRAJ` procedure in `SAS`. C1, C2, C3 and C4 denote first, second, third and fourth component, respectively. Means were calculated by averaging over estimates of 100 replicates. Standard deviations are Monte Carlo standard deviations from estimates of 100 replicates. Each value was reported $\times 10^2$.

n	N	Method	ARSE C1	A-bias C1	V-bias C1	ARSE C2	A-bias C2	V-bias C2
70	150	Proposed	3.82	0.44	0.14	3.22	-0.07	0.10
			(0.95)	(1.30)	(0.07)	(0.85)	(0.95)	(0.06)
		gbmt	4.05	0.44	0.16	4.11	-0.08	0.17
			(0.97)	(1.30)	(0.08)	(1.12)	(1.15)	(0.10)
		TRAJ	13.51	-0.30	3.86	10.00	-0.25	1.64
			(17.04)	(9.34)	(10.73)	(10.11)	(6.21)	(4.02)
n	N	Method	ARSE C3	A-bias C3	V-bias C3	ARSE C4	A-bias C4	V-bias C4
70	150	Proposed	4.12	-0.29	0.15	3.52	0.24	0.12
			(0.90)	(1.69)	(0.06)	(0.85)	(1.22)	(0.06)
		gbmt	4.38	-0.29	0.17	4.13	0.27	0.16
			(1.01)	(1.69)	(0.07)	(0.99)	(1.40)	(0.09)
		TRAJ	13.03	0.30	3.86	11.77	0.39	2.57
			(17.04)	(10.49)	(10.73)	(13.78)	(8.49)	(6.45)

Table B.2.6. Root mean square errors (RMSEs) of each logistic parameter for the four-component bivariate model from 100 replicates of 150 four-component bivariate time series of length 70. RMSEs of the proposed method were compared to `TRAJ` procedure in `SAS`. Parameters δ_0, δ_1, δ_2 and δ_3 are intercept, first, second and third logistic parameters, respectively. The fourth component was used as the reference component. The true values of logistic parameters are $5, -3.5, 1, 0.1$ (first component), $-4, 2.5, -2, -0.2$ (second component), $3, -2, 0.8, 0.2$ (third component). C1, C2, C3 and C4 denote first, second, third and fourth component, respectively.

n	N	Method	Comparison	δ_0	δ_1	δ_2	δ_3
70	150	Proposed	C1 vs C4	0.81	0.51	0.29	0.41
			C2 vs C4	1.42	0.73	0.58	0.36
			C3 vs C4	1.05	0.58	0.37	0.31
		TRAJ	C1 vs C4	1.13	0.66	0.31	0.45
			C2 vs C4	3.12	1.66	0.99	0.55
			C3 vs C4	1.15	0.74	0.48	0.35

Table B.2.7. Mean (standard deviation) of the averaged root square error (ARSE), the averaged bias (A-bias) and the variance of bias (V-bias) of estimated trajectories for each component from 100 replicates of 250 four-component bivariate time series of length 50. Estimates of the proposed method were compared to `R` package `gbmt` and `TRAJ` procedure in `SAS`. C1, C2, C3 and C4 denote first, second, third and fourth component, respectively. Means were calculated by averaging over estimates of 100 replicates. Standard deviations are Monte Carlo standard deviations from estimates of 100 replicates. Each value was reported $\times 10^2$.

n	N	Method	ARSE C1	A-bias C1	V-bias C1	ARSE C2	A-bias C2	V-bias C2
50	250	Proposed	3.42	0.18	0.11	2.86	-0.14	0.08
			(0.78)	(1.19)	(0.05)	(0.61)	(0.98)	(0.04)
		gbmt	3.57	0.18	0.12	3.68	-0.15	0.13
			(0.85)	(1.19)	(0.06)	(0.79)	(1.20)	(0.06)
		TRAJ	11.66	0.53	2.50	8.90	1.47	1.05
			(15.03)	(10.63)	(6.30)	(9.38)	(7.81)	(2.16)
n	N	Method	ARSE C3	A-bias C3	V-bias C3	ARSE C4	A-bias C4	V-bias C4
50	250	Proposed	3.93	-0.06	0.14	3.28	-0.10	0.10
			(0.92)	(1.48)	(0.07)	(0.76)	(1.19)	(0.05)
		gbmt	4.16	-0.06	0.16	3.83	-0.13	0.14
			(0.95)	(1.49)	(0.07)	(0.83)	(1.36)	(0.06)
		TRAJ	10.80	0.49	1.83	10.17	-0.17	1.84
			(9.92)	(5.78)	(3.70)	(12.54)	(8.83)	(5.20)

Table B.2.8. Root mean square errors (RMSEs) of each logistic parameter for the four-component bivariate model from 100 replicates of 250 four-component bivariate time series of length 50. RMSEs of the proposed method were compared to `TRAJ` procedure in `SAS`. Parameters δ_0, δ_1, δ_2 and δ_3 are intercept, first, second and third logistic parameters, respectively. The fourth component was used as the reference component. The true values of logistic parameters are $5, -3.5, 1, 0.1$ (first component), $-4, 2.5, -2, -0.2$ (second component), $3, -2, 0.8, 0.2$ (third component). C1, C2, C3 and C4 denote first, second, third and fourth component, respectively.

n	N	Method	Comparison	δ_0	δ_1	δ_2	δ_3
50	250	Proposed	C1 vs C4	0.63	0.41	0.26	0.29
			C2 vs C4	1.00	0.46	0.40	0.27
			C3 vs C4	0.63	0.33	0.23	0.24
		TRAJ	C1 vs C4	0.91	0.56	0.30	0.28
			C2 vs C4	1.40	0.86	0.61	0.35
			C3 vs C4	2.24	1.40	0.85	0.27

Table B.2.9. Mean (standard deviation) of the averaged root square error (ARSE), the averaged bias (A-bias) and the variance of bias (V-bias) of estimated trajectories for each component from 100 replicates of 250 four-component bivariate time series of length 70. Estimates of the proposed method were compared to `R` package `gbmt` and `TRAJ` procedure in `SAS`. C1, C2, C3 and C4 denote first, second, third and fourth component, respectively. Means were calculated by averaging over estimates of 100 replicates. Standard deviations are Monte Carlo standard deviations from estimates of 100 replicates. Each value was reported $\times 10^2$.

n	N	Method	ARSE C1	A-bias C1	V-bias C1	ARSE C2	A-bias C2	V-bias C2
70	250	Proposed	2.94	-0.04	0.08	2.61	-0.01	0.06
			(0.60)	(1.06)	(0.04)	(0.57)	(0.87)	(0.03)
		gbmt	3.10	-0.04	0.09	3.18	0.01	0.10
			(0.63)	(1.06)	(0.04)	(0.70)	(1.05)	(0.05)
		TRAJ	13.52	-1.58	3.71	11.51	-0.19	2.54
			(17.70)	(11.09)	(8.80)	(14.85)	(9.98)	(7.10)
n	N	Method	ARSE C3	A-bias C3	V-bias C3	ARSE C4	A-bias C4	V-bias C4
70	250	Proposed	3.30	-0.02	0.10	2.85	-0.07	0.08
			(0.76)	(1.21)	(0.05)	(0.73)	(0.97)	(0.04)
		gbmt	3.51	-0.01	0.12	3.26	-0.09	0.10
			(0.79)	(1.21)	(0.06)	(0.80)	(1.06)	(0.05)
		TRAJ	13.07	1.48	2.90	10.68	0.65	2.16
			(15.21)	(10.52)	(7.11)	(12.93)	(8.11)	(5.66)

Table B.2.10. Root mean square errors (RMSEs) of each logistic parameter for the four-component bivariate model from 100 replicates of 250 four-component bivariate time series of length 70. RMSEs of the proposed method were compared to `TRAJ` procedure in `SAS`. Parameters δ_0, δ_1, δ_2 and δ_3 are intercept, first, second and third logistic parameters, respectively. The fourth component was used as the reference component. The true values of logistic parameters are $5, -3.5, 1, 0.1$ (first component), $-4, 2.5, -2, -0.2$ (second component), $3, -2, 0.8, 0.2$ (third component). C1, C2, C3 and C4 denote first, second, third and fourth component, respectively.

n	N	Method	Comparison	δ_0	δ_1	δ_2	δ_3
70	250	Proposed	C1 vs C4	0.64	0.40	0.26	0.28
			C2 vs C4	0.92	0.42	0.41	0.28
			C3 vs C4	0.63	0.31	0.23	0.23
		TRAJ	C1 vs C4	0.82	0.50	0.27	0.28
			C2 vs C4	1.47	0.86	0.61	0.36
			C3 vs C4	1.60	0.96	0.57	0.25

B.3 Additional real-data results

Figure B.3.10. Estimated trajectories of the three-component model with four selected channels. **I**: Interact **S**: Still-face **R**: Recovery. Red curves are posterior mean and two green dashed curves are 95% pointwise credible intervals.

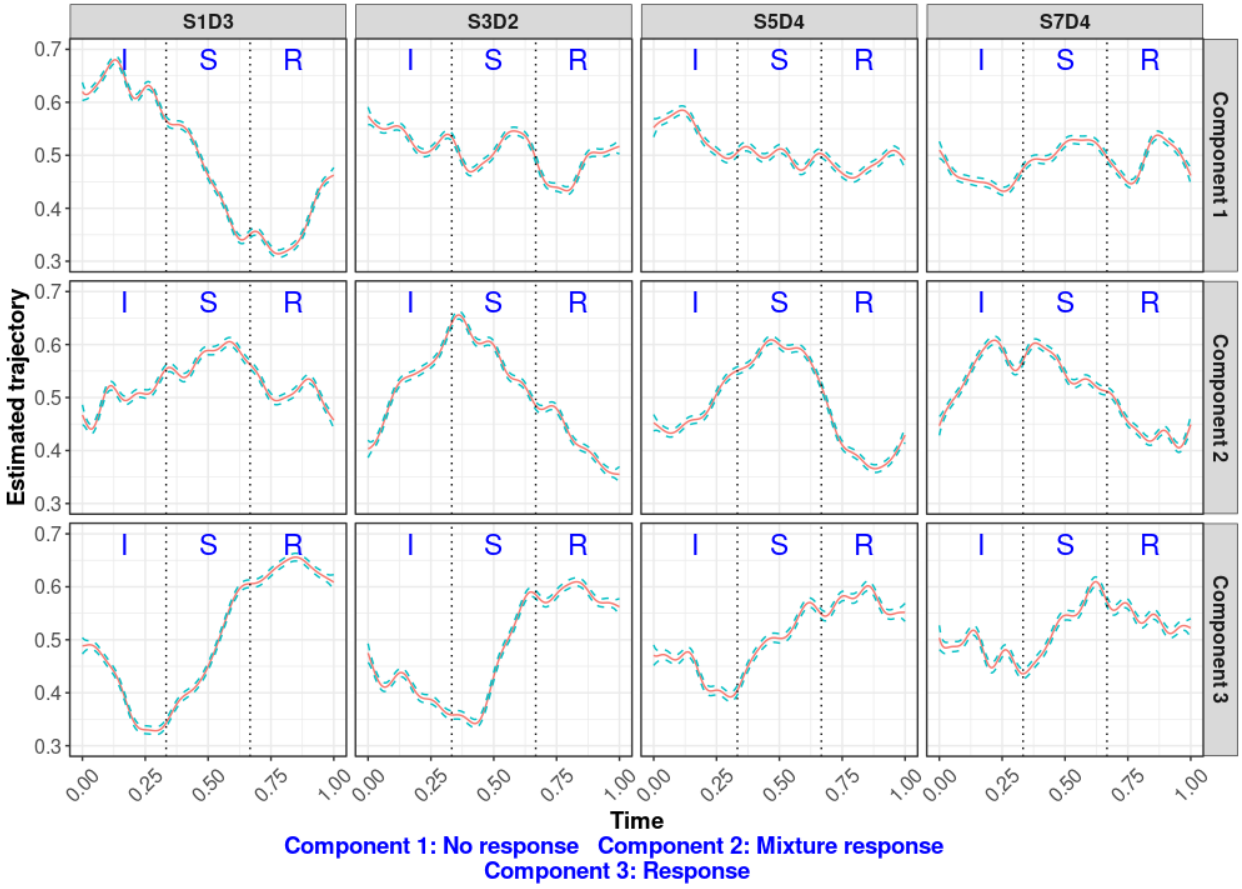


Figure B.3.11. Logistic coefficient estimates and 95% credible intervals for each covariate of the three-component model.

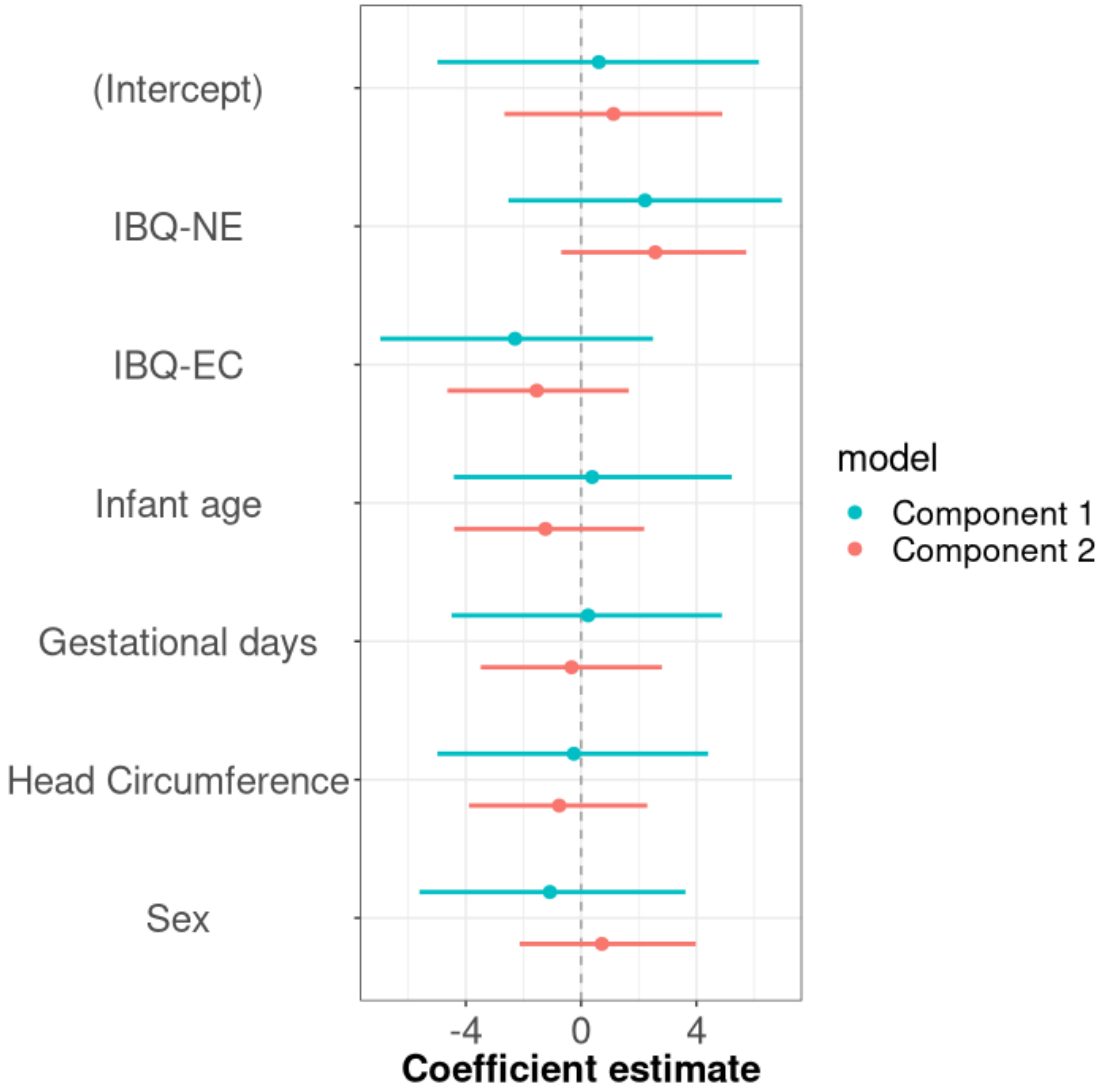


Figure B.3.12. Heatmap of averaged first derivatives of estimated trajectories for combinations of all components and selected channels.

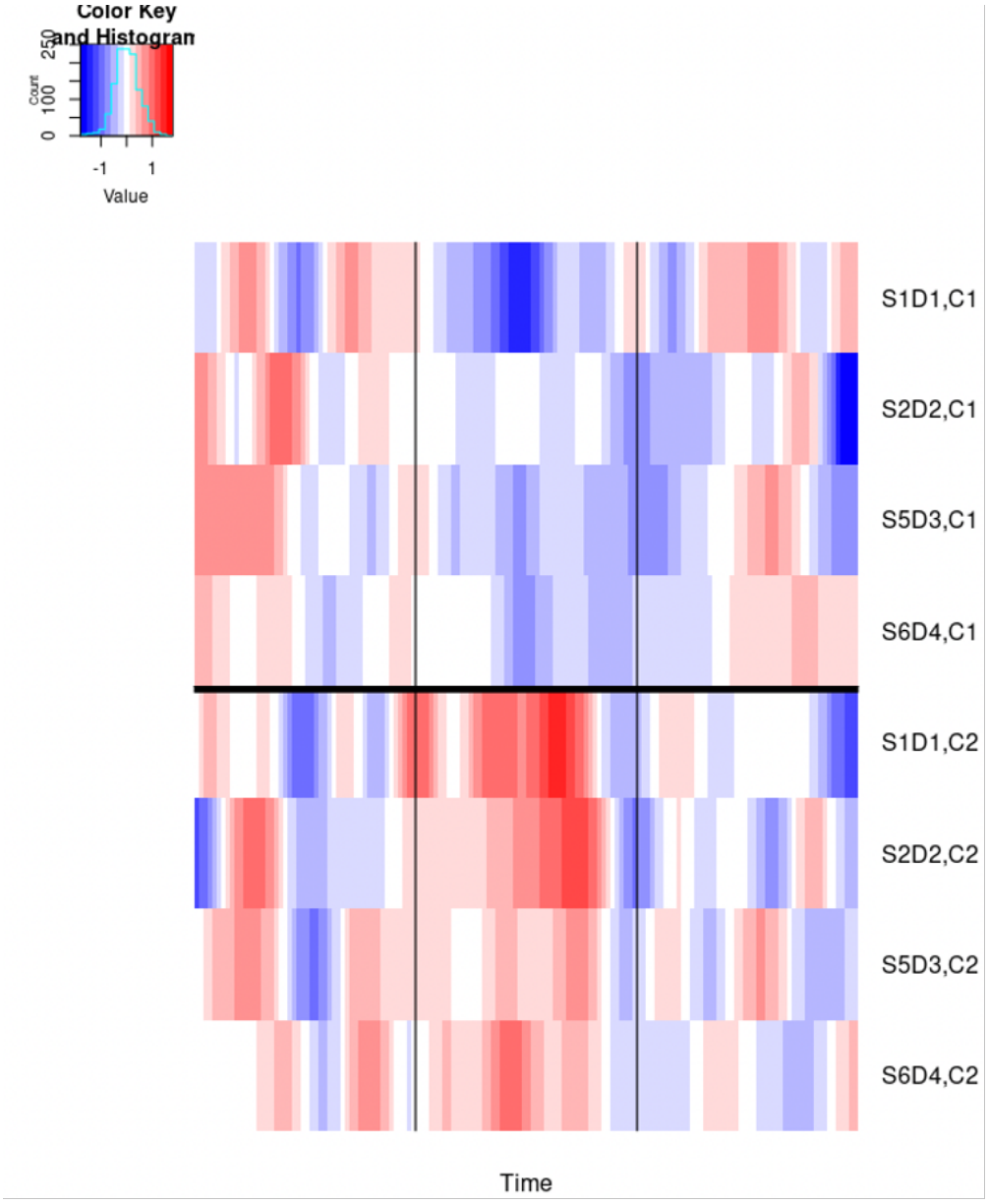
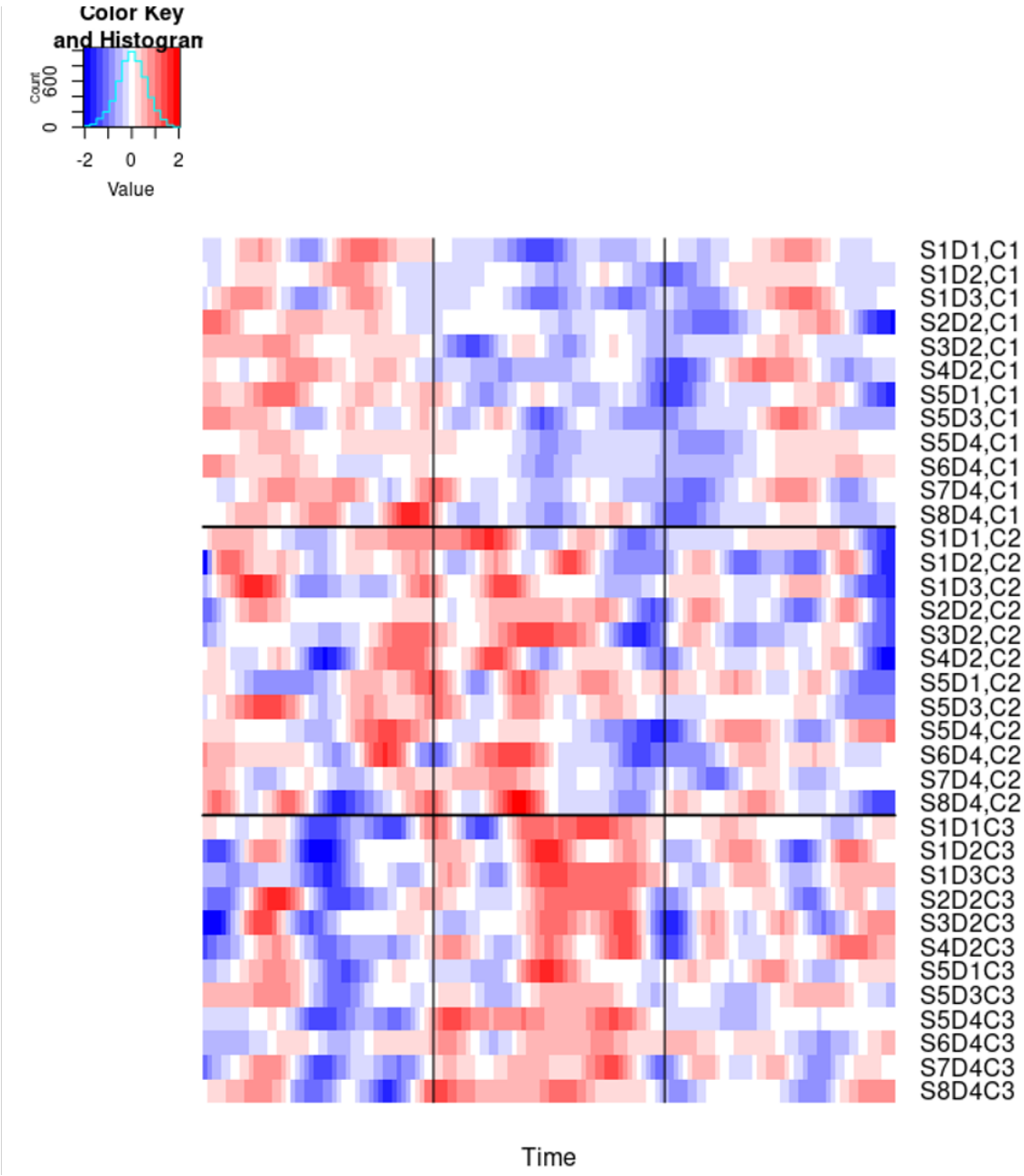


Figure B.3.13. Heatmap of averaged first derivatives of estimated trajectories for combinations of all components and all channels.



Appendix C Chapter 4

C.1 Additional simulation results

Figure C.1.14. Coefficient function plots of five methods for $p = 40$ and selected orders of differences $\mathrm{E} = \{0, 1\}$ (Case 1).

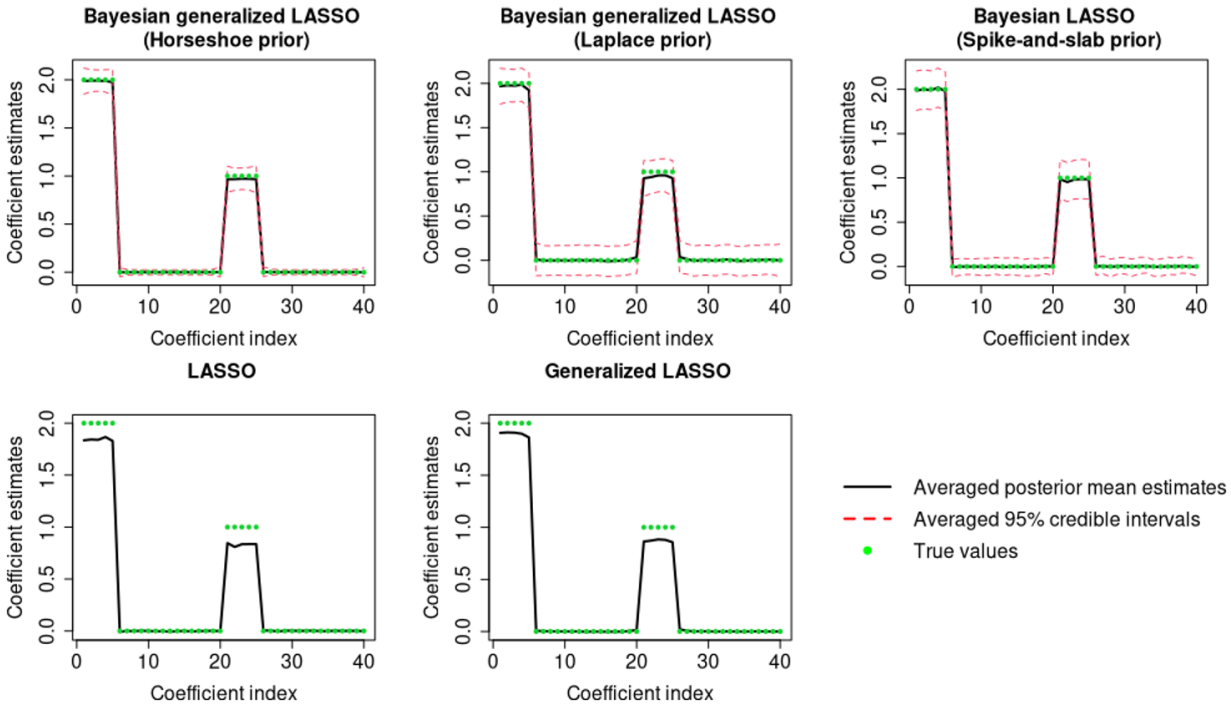


Figure C.1.15. Coefficient function plots of five methods for $p = 40$ and selected orders of differences $\mathrm{E} = \{0, 2\}$ (Case 2).

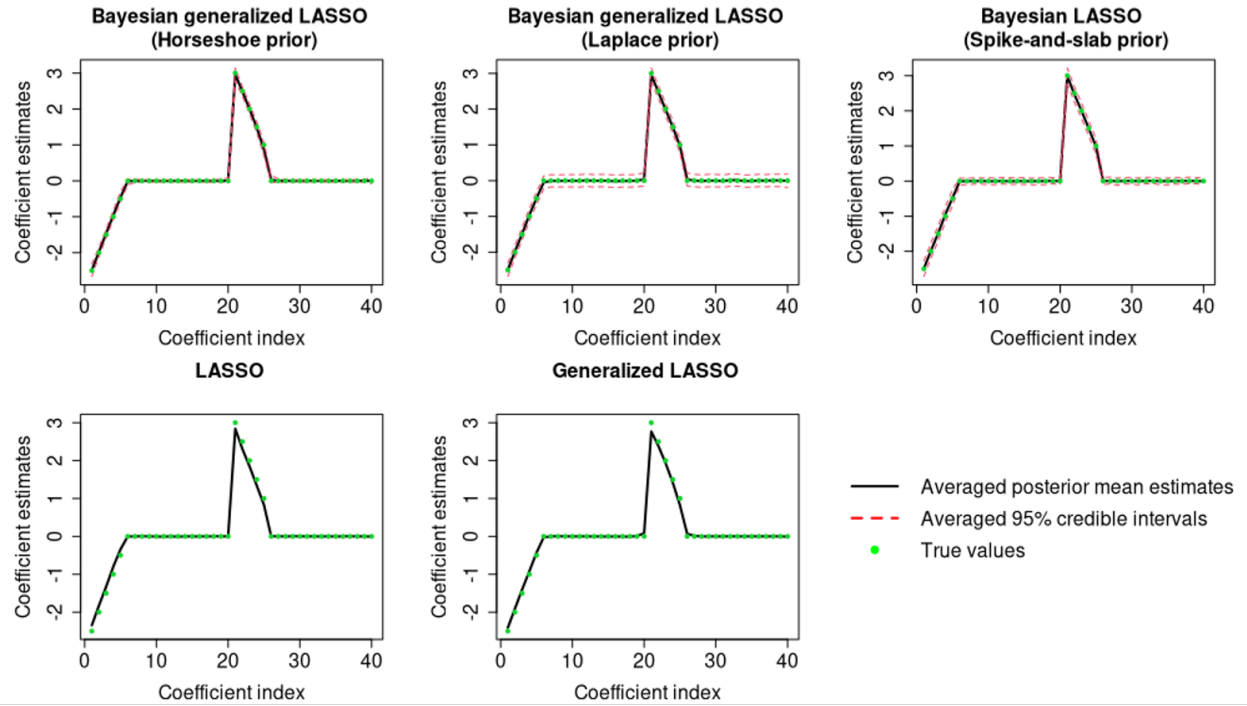


Figure C.1.16. Coefficient function plots of five methods for $p = 40$ and selected orders of differences $\mathrm{E} = \{0, 1, 2\}$ (Case 3).

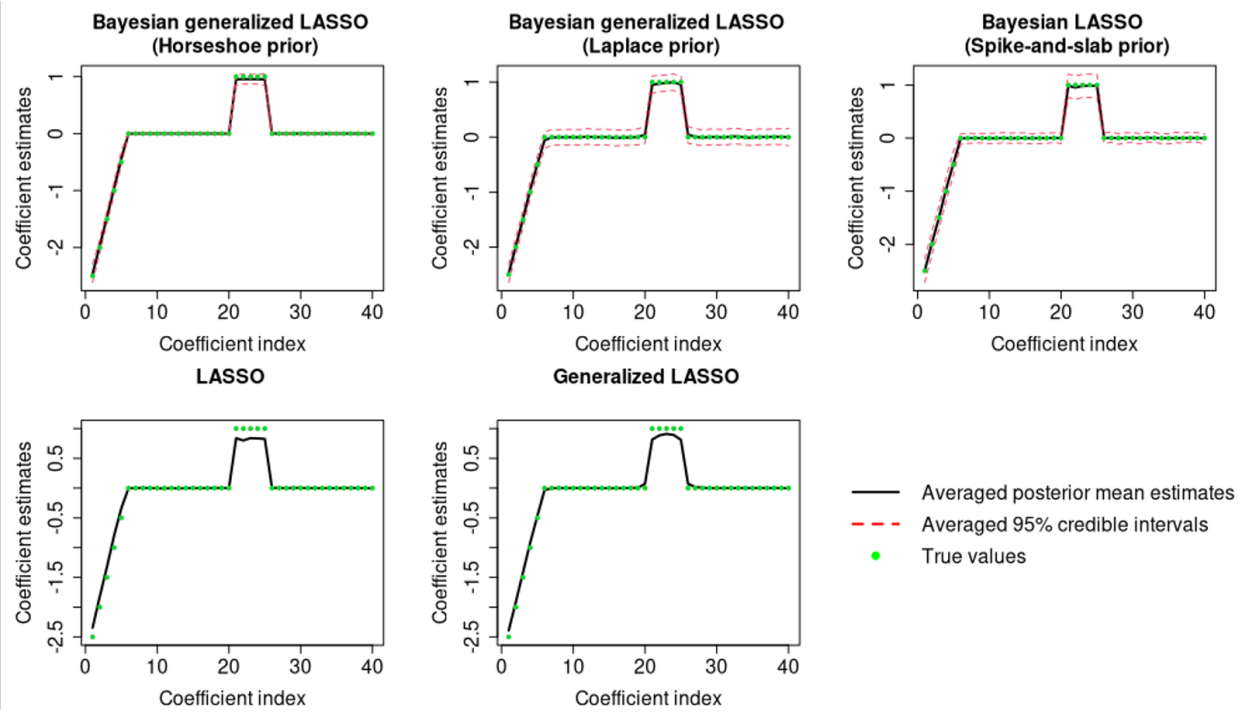


Figure C.1.17. Coefficient function plots of five methods for $p = 80$ and selected orders of differences $\mathrm{E} = \{0, 2\}$ (Case 5).

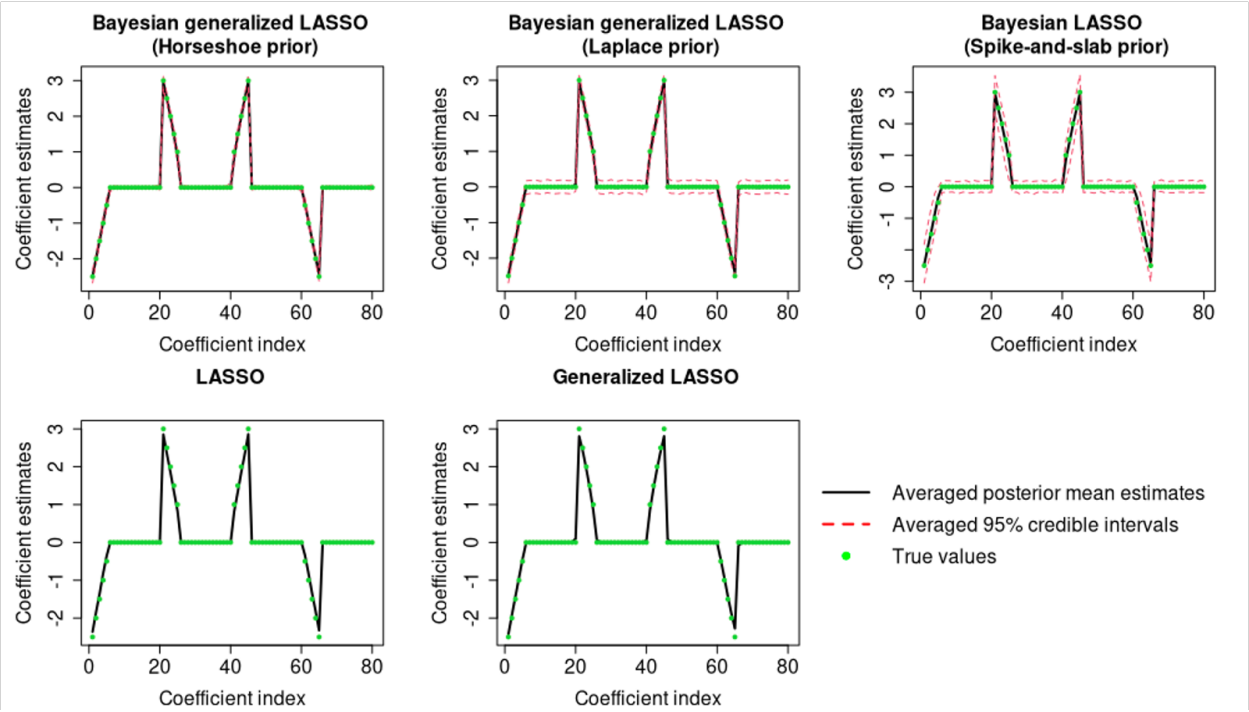


Figure C.1.18. Coefficient function plots of five methods for $p = 80$ and selected orders of differences $\mathrm{E} = \{0, 1, 2\}$ (Case 6).

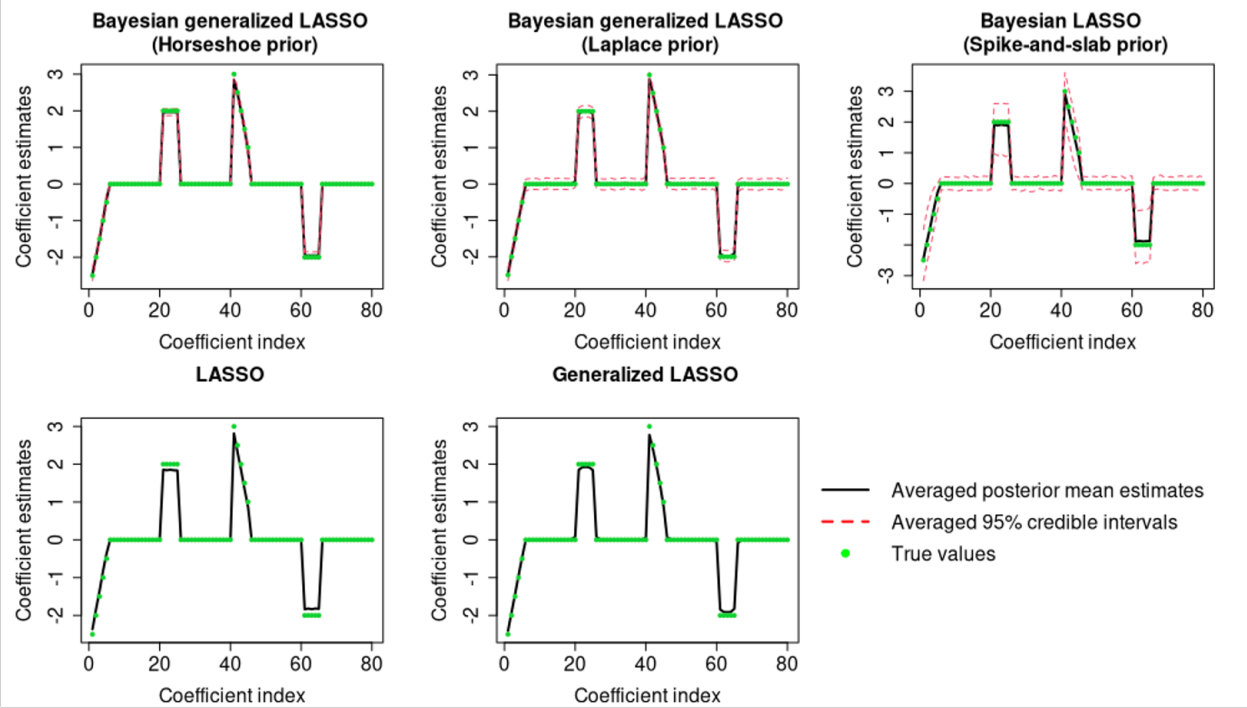


Figure C.1.19. Coefficient function plots of five methods for $p = 120$ and selected orders of differences $\mathrm{E} = \{0, 1\}$ (Case 7).

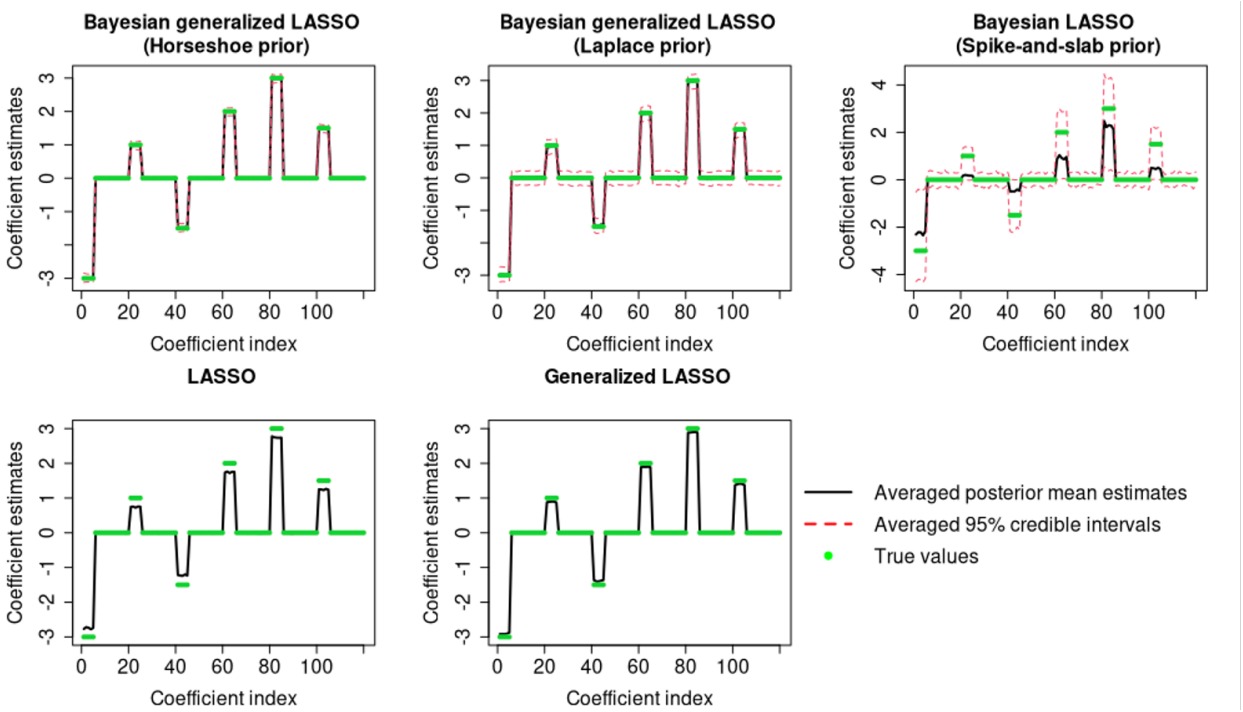


Figure C.1.20. Coefficient function plots of five methods for $p = 120$ and selected orders of differences $\mathrm{E} = \{0, 2\}$ (Case 8).

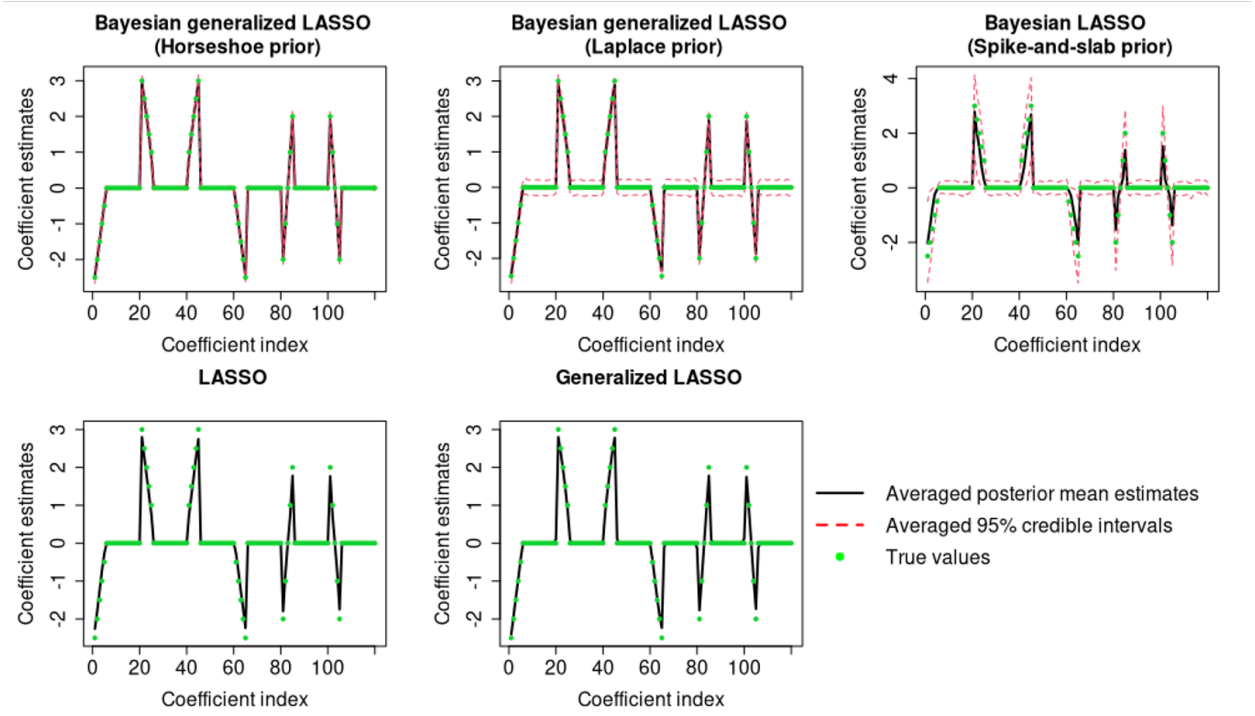


Figure C.1.21. Coefficient function plots of five methods for $p = 120$ and selected orders of differences $\mathrm{E} = \{0, 1, 2\}$ (Case 9).

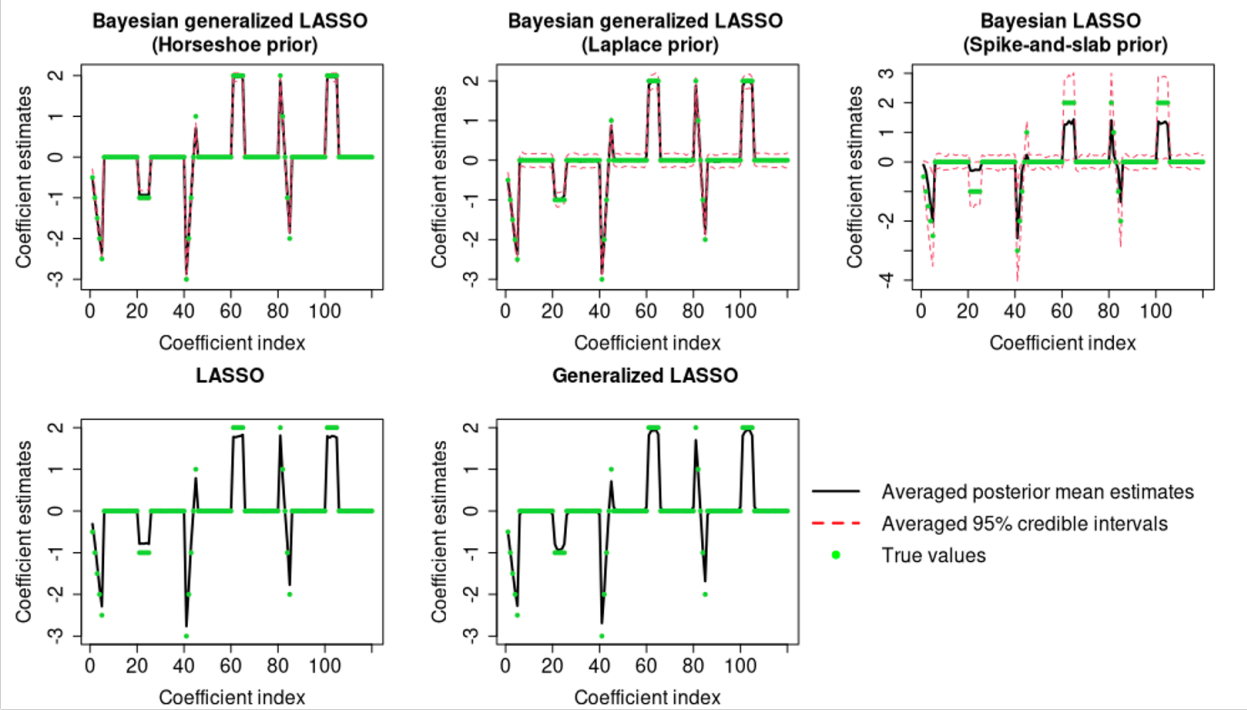


Table C.1.11. Mean (standard deviation) of prediction errors for five methods and nine simulation cases with different numbers of covariates and orders of differences (i.e. $0, 1$ indicates zeroth and first-order differences are selected).

Type	p=40			p=80			p=120		
	0,1	0,2	0,1,2	0,1	0,2	0,1,2	0,1	0,2	0,1,2
Bayesian generalized LASSO (Horseshoe prior)	0.99 (0.14)	1.03 (0.14)	1.02 (0.15)	1.07 (0.15)	1.19 (0.19)	1.20 (0.21)	1.10 (0.15)	1.23 (0.20)	1.72 (2.20)
Bayesian generalized LASSO (Laplace prior)	1.19 (0.18)	1.35 (0.24)	1.30 (0.22)	2.38 (0.52)	2.28 (0.49)	2.10 (0.44)	4.72 (1.21)	4.53 (1.21)	3.90 (1.11)
Bayesian LASSO (Spike-and-slab prior)	1.11 (0.17)	1.11 (0.17)	1.12 (0.17)	2.02 (1.06)	1.78 (0.71)	2.04 (1.08)	45.43 (17.84)	20.39 (11.89)	25.51 (13.92)
LASSO	1.39 (0.27)	1.44 (0.29)	1.44 (0.30)	2.10 (0.57)	2.17 (0.56)	2.15 (0.57)	5.64 (3.72)	4.66 (2.91)	4.08 (2.26)
Generalized LASSO	1.15 (0.19)	1.24 (0.20)	1.21 (0.19)	1.24 (0.20)	1.69 (0.37)	1.71 (0.35)	1.70 (0.33)	2.79 (0.99)	3.17 (1.24)

Table C.1.12. Mean (standard deviation) of mean square errors for five methods and nine simulation cases with different numbers of covariates and orders of differences (i.e. $0, 1$ indicates zeroth and first-order differences are selected).

Type	p=40			p=80			p=120		
	0,1	0,2	0,1,2	0,1	0,2	0,1,2	0,1	0,2	0,1,2
Bayesian generalized LASSO (Horseshoe prior)	0.03 (0.03)	0.08 (0.05)	0.06 (0.07)	0.07 (0.04)	0.19 (0.10)	0.19 (0.12)	0.10 (0.04)	0.24 (0.12)	0.70 (2.04)
Bayesian generalized LASSO (Laplace prior)	0.24 (0.08)	0.42 (0.13)	0.36 (0.11)	1.37 (0.35)	1.26 (0.35)	1.07 (0.29)	3.70 (0.97)	3.53 (1.01)	2.88 (0.81)
Bayesian LASSO (Spike-and-slab prior)	0.16 (0.07)	0.17 (0.07)	0.17 (0.07)	1.04 (1.11)	0.79 (0.74)	1.07 (1.18)	44.03 (16.57)	19.03 (11.04)	23.87 (12.88)
LASSO	0.45 (0.18)	0.50 (0.21)	0.51 (0.21)	1.07 (0.43)	1.16 (0.47)	1.15 (0.48)	4.54 (3.43)	3.61 (2.69)	3.02 (2.07)
Generalized LASSO	0.20 (0.10)	0.29 (0.12)	0.27 (0.11)	0.29 (0.12)	0.66 (0.28)	0.70 (0.30)	0.69 (0.23)	1.77 (0.97)	2.14 (1.19)

Bibliography

Aarabi, A. and Huppert, T. J. (2016). Characterization of the relative contributions from systemic physiological noise to whole-brain resting-state functional near-infrared spectroscopy data using single-channel independent component analysis. *Neurophotonics*, 3(2):025004.

Aasted, C. M., Yücel, M. A., Cooper, R. J., Dubb, J., Tsuzuki, D., Becerra, L., Petkov, M. P., Borsook, D., Dan, I., and Boas, D. A. (2015). Anatomical guidance for functional near-infrared spectroscopy: Atlasviewer tutorial. *Neurophotonics*, 2(2):020801.

Adak, S. (1998). Time-dependent spectral analysis of nonstationary time series. *Journal of the American Statistical Association*, 93(444):1488–1501.

Adamson, L. B. and Frick, J. E. (2003). The still face: A history of a shared experimental paradigm. *Infancy*, 4(4):451–473.

Aghabozorgi, S., Shirkhorshidi, A. S., and Wah, T. Y. (2015). Time-series clustering–a decade review. *Information Systems*, 53:16–38.

Ainsworth, M. D. S. and Bell, S. M. (1972). Mother-infant interaction and the development of competence.

Ainsworth, M. D. S., Blehar, M. C., Waters, E., and Wall, S. N. (2015). *Patterns of attachment: A psychological study of the strange situation*. Psychology Press.

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):99–102.

Ashby, F. G. (2019). *Statistical analysis of fMRI data*. MIT press.

Bagnall, A. and Janacek, G. (2005). Clustering time series with clipped data. *Machine learning*, 58(2):151–178.

Baker, W. B., Parthasarathy, A. B., Busch, D. R., Mesquita, R. C., Greenberg, J. H., and Yodh, A. (2014). Modified beer-lambert law for blood flow. *Biomedical optics express*, 5(11):4053–4075.

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

Barker, J. W., Aarabi, A., and Huppert, T. J. (2013). Autoregressive model based algorithm for correcting motion and serially correlated errors in fnirs. *Biomedical optics express*, 4(8):1366–1379.

Belsky, J., Hsieh, K.-H., and Crnic, K. (1998). Mothering, fathering, and infant negativity as antecedents of boys' externalizing problems and inhibition at age 3 years: Differential susceptibility to rearing experience? *Development and psychopathology*, 10(2):301–319.

Bertolacci, M., Rosen, O., Cripps, E., and Cripps, S. (2021). Adaptspec-x: Covariate dependent spectral modeling of multiple nonstationary time series. *Journal of Computational and Graphical Statistics*, (just-accepted):1–40.

Bicego, M., Murino, V., and Figueiredo, M. A. (2003). Similarity-based clustering of sequences using hidden markov models. In *International Workshop on Machine Learning and Data Mining in Pattern Recognition*, pages 86–95. Springer.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7):719–725.

Bigelow, A. E., MacLean, K., Proctor, J., Myatt, T., Gillis, R., and Power, M. (2010). Maternal sensitivity throughout infancy: Continuity and relation to attachment security. *Infant behavior and Development*, 33(1):50–60.

Boas, D. A., Dale, A. M., and Franceschini, M. A. (2004). Diffuse optical imaging of brain activation: approaches to optimizing image sensitivity, resolution, and accuracy. *Neuroimage*, 23:S275–S288.

Brodmann, K. (1909). *Vergleichende Lokalisationslehre der Grosshirnrinde in ihren Prinzipien dargestellt auf Grund des Zellenbaues*. Barth.

Bruce, S. A., Hall, M. H., Buysse, D. J., and Krafty, R. T. (2018). Conditional adaptive bayesian spectral analysis of nonstationary biomedical time series. *Biometrics*, 74(1):260–269.

Bullmore, E., Brammer, M., Williams, S. C., Rabe-Hesketh, S., Janot, N., David, A., Mellers, J., Howard, R., and Sham, P. (1996). Statistical methods of estimation and inference for functional mr image analysis. *Magnetic Resonance in Medicine*, 35(2):261–277.

Cadonna, A., Kottas, A., and Prado, R. (2019). Bayesian spectral modeling for multiple time series. *Journal of the American Statistical Association*.

Candes, E. and Tao, T. (2007). The dantzig selector: Statistical estimation when p is much larger than n. *The annals of Statistics*, 35(6):2313–2351.

Carter, A. S., Mayes, L. C., and Pajer, K. A. (1990). The role of dyadic affect in play and infant sex in predicting infant response to the still-face situation. *Child Development*, 61(3):764–773.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Casella, G., Ghosh, M., Gill, J., and Kyung, M. (2010). Penalized regression, standard errors, and bayesian lassos. *Bayesian analysis*, 5(2):369–411.

Catmull, E. and Rom, R. (1974). A class of local interpolating splines. In *Computer aided geometric design*, pages 317–326. Elsevier.

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). Deviance information criteria for missing data models. *Bayesian analysis*, 1(4):651–673.

Chiarelli, A. M., Maclin, E. L., Fabiani, M., and Gratton, G. (2015). A kurtosis-based wavelet algorithm for motion artifact correction of fnirs data. *NeuroImage*, 112:128–137.

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

Cooper, R., Selb, J., Gagnon, L., Phillip, D., Schytz, H. W., Iversen, H. K., Ashina, M., and Boas, D. A. (2012). A systematic comparison of motion artifact correction techniques for functional near-infrared spectroscopy. *Frontiers in neuroscience*, 6:147.

Cope, M., Delpy, D., Reynolds, E., Wray, S., Wyatt, J., and Zee, P. (1988). Methods of quantitating cerebral near infrared spectroscopy data. In *Oxygen Transport to Tissue X*, pages 183–189. Springer.

Corduas, M. and Piccolo, D. (2008). Time series clustering and classification by the autoregressive metric. *Computational statistics & data analysis*, 52(4):1860–1872.

Cui, X., Bray, S., and Reiss, A. L. (2010). Functional near infrared spectroscopy (nirs) signal improvement based on negative correlation between oxygenated and deoxygenated hemoglobin dynamics. *Neuroimage*, 49(4):3039–3046.

Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *The annals of Statistics*, 25(1):1–37.

Dahlhaus, R. (2000). A likelihood approximation for locally stationary processes. *The Annals of Statistics*, 28(6):1762–1794.

Davis, R. A., Lee, T. C. M., and Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, 101(473):223–239.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Edwards, R. C. and Hans, S. L. (2015). Infant risk factors associated with internalizing, externalizing, and co-occurring behavior problems in young children. *Developmental Psychology*, 51(4):489.

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

Enlow, M. B., White, M. T., Hails, K., Cabrera, I., and Wright, R. J. (2016). The infant behavior questionnaire-revised: Factor structure in a culturally and sociodemographically diverse sample in the united states. *Infant Behavior and Development*, 43:24–35.

Eubank, R. L. (1999). *Nonparametric regression and spline smoothing*. CRC press.

Feldman, R., Greenbaum, C. W., and Yirmiya, N. (1999). Mother–infant affect synchrony as an antecedent of the emergence of self-control. *Developmental psychology*, 35(1):223.

Ferrari, M. and Quaresima, V. (2012). A brief review on the history of human functional near-infrared spectroscopy (fnirs) development and fields of application. *Neuroimage*, 63(2):921–935.

Fishburn, F. A., Ludlum, R. S., Vaidya, C. J., and Medvedev, A. V. (2019). Temporal derivative distribution repair (tddr): a motion correction method for fnirs. *Neuroimage*, 184:171–179.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Friston, K. J., Holmes, A. P., Poline, J., Grasby, P., Williams, S., Frackowiak, R. S., and Turner, R. (1995). Analysis of fmri time-series revisited. *Neuroimage*, 2(1):45–53.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J.-P., Frith, C. D., and Frackowiak, R. S. (1994). Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210.

Friston, K. J., Penny, W., Phillips, C., Kiebel, S., Hinton, G., and Ashburner, J. (2002). Classical and bayesian inference in neuroimaging: theory. *NeuroImage*, 16(2):465–483.

Fröhwirth-Schnatter, S. and Kaufmann, S. (2008). Model-based clustering of multiple time series. *Journal of Business & Economic Statistics*, 26(1):78–89.

Frühwirth-Schnatter, S., Pamminger, C., Weber, A., and Winter-Ebmer, R. (2012). Labor market entry and earnings dynamics: Bayesian inference using mixtures-of-experts markov chain clustering. *Journal of Applied Econometrics*, 27(7):1116–1137.

Fúquene, J., Steel, M., and Rossell, D. (2019). On choosing mixture components via non-local priors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(5):809–837.

Garey, L. J. (1999). *Brodmann's' localisation in the cerebral cortex'*. World Scientific.

Gartstein, M. A., Bridgett, D. J., Rothbart, M. K., Robertson, C., Iddins, E., Ramsay, K., and Schlect, S. (2010). A latent growth examination of fear development in infancy: contributions of maternal depression and the risk for toddler anxiety. *Developmental psychology*, 46(3):651.

Gartstein, M. A., Bridgett, D. J., Young, B. N., Panksepp, J., and Power, T. (2013). Origins of effortful control: Infant and parent contributions. *Infancy*, 18(2):149–183.

Gartstein, M. A. and Rothbart, M. K. (2003). Studying infant temperament via the revised infant behavior questionnaire. *Infant behavior and development*, 26(1):64–86.

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

Gelman, A., Carlin, J., Stern, H., Rubin, D., et al. (2003). Bayesian data analysis.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman and Hall/CRC.

George, E. I. and McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889.

Gerlach, R., Carter, C., and Kohn, R. (2000). Efficient bayesian inference for dynamic mixture models. *Journal of the American Statistical Association*, 95(451):819–828.

Gillman, M. W. and Blaisdell, C. J. (2018). Environmental influences on child health outcomes, a research program of the nih. *Current opinion in pediatrics*, 30(2):260.

Gionis, A. and Mannila, H. (2003). Finding recurrent sources in sequences. In *Proceedings of the seventh annual international conference on Research in computational molecular biology*, pages 123–130.

Green, P. (1995). Reversible jump mcmc computation and bayesian model determination. biomeirika, 82: 711-732 hastings, wk 1970. monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57:97–109.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Crc Press.

Griffin, J. and Brown, P. (2005). Alternative prior distributions for variable selection with very many more variables than observations. Technical report, Technical report, University of Warwick.

Guo, C., Jia, H., and Zhang, N. (2008). Time series clustering based on ica for stock data analysis. In *2008 4th international conference on wireless communications, networking and mobile computing*, pages 1–4. IEEE.

Guo, W. and Dai, M. (2006). Multivariate time-dependent spectral analysis using cholesky decomposition. *Statistica Sinica*, pages 825–845.

Guo, W., Dai, M., Ombao, H. C., and Von Sachs, R. (2003). Smoothing spline anova for time-dependent spectral analysis. *Journal of the American Statistical Association*, 98(463):643–652.

Hall, C. A. and Meyer, W. W. (1976). Optimal error bounds for cubic spline interpolation. *Journal of Approximation Theory*, 16(2):105–122.

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

Hastie, T. J. and Tibshirani, R. J. (2017). *Generalized additive models*. Routledge.

Hector, E. C. and Song, P. X.-K. (2020). Joint integrative analysis of multiple data sources with correlated vector outcomes. *arXiv preprint arXiv:2011.14996*.

Hector, E. C. and Song, P. X.-K. (2021). A distributed and integrated method of moments for high-dimensional correlated data analysis. *Journal of the American Statistical Association*, 116(534):805–818.

Hipwell, A. E., Tung, I., Northrup, J., and Keenan, K. (2019). Transgenerational associations between maternal childhood stress exposure and profiles of infant emotional reactivity. *Development and psychopathology*, 31(3):887–898.

Hocking, R. R. and Leslie, R. (1967). Selection of the best subset in regression analysis. *Technometrics*, 9(4):531–540.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Holland, P. W. and Welsch, R. E. (1977). Robust regression using iteratively reweighted least-squares. *Communications in Statistics-theory and Methods*, 6(9):813–827.

Hu, J., Ray, B., and Han, L. (2006). An interweaved hmm/dtw approach to robust time series clustering. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 145–148. IEEE.

Huerta, G., Jiang, W., and Tanner, M. A. (2003). Time series modeling via hierarchical mixtures. *Statistica Sinica*, pages 1097–1118.

Huppert, T. J. (2016). Commentary on the statistical properties of noise and its implication on general linear models in functional near-infrared spectroscopy. *Neurophotonics*, 3(1):010401.

Huppert, T. J., Diamond, S. G., Franceschini, M. A., and Boas, D. A. (2009). Homer: a review of time-series analysis methods for near-infrared spectroscopy of the brain. *Applied optics*, 48(10):D280–D298.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, 33(2):730–773.

Izzetoglu, M., Devaraj, A., Bunce, S., and Onaral, B. (2005). Motion artifact cancellation in nir spectroscopy using wiener filtering. *IEEE Transactions on Biomedical Engineering*, 52(5):934–938.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression that's interpretable. *The Annals of Statistics*, 37(5A):2083–2108.

Jang, K.-E., Tak, S., Jung, J., Jang, J., Jeong, Y., and Ye, Y. C. (2009). Wavelet minimum description length detrending for near-infrared spectroscopy. *Journal of biomedical optics*, 14(3):034004.

Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain monte carlo methods and the label switching problem in bayesian mixture modeling. *Statistical Science*, 20(1):50–67.

Jobsis, F. F. (1977). Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters. *Science*, 198(4323):1264–1267.

Johnson, R. A., Wichern, D. W., et al. (2014). *Applied multivariate statistical analysis*, volume 6. Pearson London, UK:.

Jones, B. L., Nagin, D. S., and Roeder, K. (2001). A sas procedure based on mixture models for estimating developmental trajectories. *Sociological methods & research*, 29(3):374–393.

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the em algorithm. *Neural computation*, 6(2):181–214.

Kakikawa, Y., Shimamura, K., and Kawano, S. (2022). Bayesian fused lasso modeling via horseshoe prior. *arXiv preprint arXiv:2201.08053*.

Kakizawa, Y., Shumway, R. H., and Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *Journal of the American Statistical Association*, 93(441):328–340.

Kalpakis, K., Gada, D., and Puttagunta, V. (2001). Distance measures for effective clustering of arima time-series. In *Proceedings 2001 IEEE international conference on data mining*, pages 273–280. IEEE.

Karim, H., Fuhrman, S. I., Furman, J. M., and Huppert, T. J. (2013a). Neuroimaging to detect cortical projection of vestibular response to caloric stimulation in young and older adults using functional near-infrared spectroscopy (fnirs). *Neuroimage*, 76:1–10.

Karim, H., Fuhrman, S. I., Sparto, P., Furman, J., and Huppert, T. (2013b). Functional brain imaging of multi-sensory vestibular processing during computerized dynamic posturography using near-infrared spectroscopy. *Neuroimage*, 74:318–325.

Karim, H., Schmidt, B., Dart, D., Beluk, N., and Huppert, T. (2012). Functional near-infrared spectroscopy (fnirs) of brain function during active balancing using a video game system. *Gait & posture*, 35(3):367–372.

Keenan, K., Hipwell, A., Chung, T., Stepp, S., Stouthamer-Loeber, M., Loeber, R., and McTigue, K. (2010). The pittsburgh girls study: overview and initial findings. *Journal of Clinical Child & Adolescent Psychology*, 39(4):506–521.

Keogh, E. and Lin, J. (2005). Clustering of time-series subsequences is meaningless: implications for previous and future research. *Knowledge and information systems*, 8(2):154–177.

Keogh, E. J. and Pazzani, M. J. (2000). A simple dimensionality reduction technique for fast similarity search in large time series databases. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 122–133. Springer.

Kim, C. (2021). Deviance information criteria for mixtures of distributions. *Communications in Statistics-Simulation and Computation*, 50(10):2935–2948.

Kim, S., Fonagy, P., Allen, J., Martinez, S., Iyengar, U., and Strathearn, L. (2014). Mothers who are securely attached in pregnancy show more attuned infant mirroring 7 months postpartum. *Infant Behavior and Development*, 37(4):491–504.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502.

Kitagawa, G. and Akaike, H. (1978). A procedure for the modeling of non-stationary time series. *Annals of the Institute of Statistical Mathematics*, 30(2):351–363.

Kitagawa, G. and Gersch, W. (1996). *Smoothness priors analysis of time series*, volume 116. Springer Science & Business Media.

Kochanek, D. H. and Bartels, R. H. (1984). Interpolating splines with local tension, continuity, and bias control. In *Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, pages 33–41.

Kocsis, L., Herman, P., and Eke, A. (2006). The modified beer–lambert law revisited. *Physics in Medicine & Biology*, 51(5):N91.

Krafty, R. T. and Collinge, W. O. (2013). Penalized multivariate whittle likelihood for power spectrum estimation. *Biometrika*, 100(2):447–458.

Krafty, R. T., Hall, M., and Guo, W. (2011). Functional mixed effects spectral analysis. *Biometrika*, 98(3):583–598.

Krafty, R. T., Rosen, O., Stoffer, D. S., Buysse, D. J., and Hall, M. H. (2017). Conditional spectral analysis of replicated multiple time series with application to nocturnal physiology. *Journal of the American Statistical Association*, 112(520):1405–1416.

Lau, J. W. and So, M. K. (2008). Bayesian mixture of autoregressive models. *Computational Statistics & Data Analysis*, 53(1):38–60.

Levendosky, A. A., Leahy, K. L., Bogat, G. A., Davidson, W. S., and Von Eye, A. (2006). Domestic violence, maternal parenting, maternal mental health, and infant externalizing behavior. *Journal of Family Psychology*, 20(4):544.

Li, C., Biswas, G., Dale, M., and Dale, P. (2001). Building models of ecological dynamics using hmm based temporal data clustering—a preliminary study. In *International Symposium on Intelligent Data Analysis*, pages 53–62. Springer.

Li, H. (2019). Multivariate time series clustering based on common principal component analysis. *Neurocomputing*, 349:239–247.

Li, R., Potter, T., Huang, W., and Zhang, Y. (2017). Enhancing performance of a hybrid eeg-fnirs system using channel selection and early temporal features. *Frontiers in human neuroscience*, 11:462.

Li, Z., Bruce, S. A., Wutzke, C. J., and Long, Y. (2021). Conditional adaptive bayesian spectral analysis of replicated multivariate time series. *Statistics in Medicine*, 40(8):1989–2005.

Li, Z. and Krafty, R. T. (2019). Adaptive bayesian time–frequency analysis of multivariate time series. *Journal of the American Statistical Association*, 114(525):453–465.

Liao, T. W. (2005). Clustering of time series data—a survey. *Pattern recognition*, 38(11):1857–1874.

Lin, J., Keogh, E., and Truppel, W. (2003). Clustering of streaming time series is meaningless. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 56–65.

Lin, J., Vlachos, M., Keogh, E., and Gunopulos, D. (2004). Iterative incremental clustering of time series. In *International Conference on Extending Database Technology*, pages 106–122. Springer.

Magrini, A. (2022). Assessment of agricultural sustainability in european union countries: a group-based multivariate trajectory approach. *AStA Advances in Statistical Analysis*, pages 1–31.

Makalic, E. and Schmidt, D. F. (2015). A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182.

Marin, J.-M., Mengersen, K., and Robert, C. P. (2005). Bayesian modelling and inference on mixtures of distributions. *Handbook of statistics*, 25:459–507.

Marin, J.-M., Robert, C. P., et al. (2007). *Bayesian core: a practical approach to computational Bayesian statistics*, volume 268. Springer.

Marx, B. D. and Eilers, P. H. (1999). Generalized linear regression on sampled signals and curves: a p-spline approach. *Technometrics*, 41(1):1–13.

Masoudnia, S. and Ebrahimpour, R. (2014). Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293.

McLachlan, G. J. (2005). *Discriminant analysis and statistical pattern recognition*. John Wiley & Sons.

Mesman, J., Linting, M., Joosen, K. J., Bakermans-Kranenburg, M. J., and Van IJzendoorn, M. H. (2013). Robust patterns and individual variations: Stability and predictors of infant behavior in the still-face paradigm. *Infant Behavior and Development*, 36(4):587–598.

Mesman, J., van IJzendoorn, M. H., and Bakermans-Kranenburg, M. J. (2009). The many faces of the still-face paradigm: A review and meta-analysis. *Developmental review*, 29(2):120–162.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Mitsa, T. (2010). *Temporal data mining*. Chapman and Hall/CRC.

Miyai, I., Tanabe, H. C., Sase, I., Eda, H., Oda, I., Konishi, I., Tsunazawa, Y., Suzuki, T., Yanagida, T., and Kubota, K. (2001). Cortical mapping of gait in humans: a near-infrared spectroscopic topography study. *Neuroimage*, 14(5):1186–1192.

Mörchen, F., Ultsch, A., and Hoos, O. (2005). Extracting interpretable muscle activation patterns with time series knowledge mining. *International Journal of Knowledge-based and Intelligent Engineering Systems*, 9(3):197–208.

Nagin, D. S., Jones, B. L., Passos, V. L., and Tremblay, R. E. (2018). Group-based multi-trajectory modeling. *Statistical methods in medical research*, 27(7):2015–2023.

Naseer, N. and Hong, K.-S. (2015). fnirs-based brain-computer interfaces: a review. *Frontiers in human neuroscience*, 9:3.

Northrup, J. B., Ridley, J., Foley, K., Moses-Kolko, E. L., Keenan, K., and Hipwell, A. E. (2019). A transactional model of infant still-face response and maternal behavior during the first year. *Infancy*, 24(5):787–806.

Ogden, R. T. (1997). *Essential wavelets for statistical applications and data analysis*. Springer.

Ombao, H., Von Sachs, R., and Guo, W. (2005). Slex analysis of multivariate nonstationary time series. *Journal of the American Statistical Association*, 100(470):519–531.

Ombao, H. C., Raz, J. A., von Sachs, R., and Malow, B. A. (2001). Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association*, 96(454):543–560.

Papastamoulis, P. (2015). label. switching: An r package for dealing with the label switching problem in mcmc outputs. *arXiv preprint arXiv:1503.02271*.

Papastamoulis, P. and Iliopoulos, G. (2010). An artificial allocations based solution to the label switching problem in bayesian analysis of mixtures of distributions. *Journal of Computational and Graphical Statistics*, 19(2):313–331.

Papastamoulis, P. and Iliopoulos, G. (2013). On the convergence rate of random permutation sampler and ecr algorithm in missing data models. *Methodology and Computing in Applied Probability*, 15(2):293–304.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Park, T., Eckley, I. A., and Ombao, H. C. (2014). Estimating time-evolving partial coherence between signals via multivariate locally stationary wavelet processes. *IEEE Transactions on Signal Processing*, 62(20):5240–5250.

Parzen, E. (1990). Time series, statistics, and information.

Plichta, M. M., Heinzel, S., Ehlis, A.-C., Pauli, P., and Fallgatter, A. J. (2007). Model-based analysis of rapid event-related functional near-infrared spectroscopy (nirs) data: a parametric validation study. *Neuroimage*, 35(2):625–634.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, 108(504):1339–1349.

Prado, R. and Huerta, G. (2002). Time-varying autoregressions with model order uncertainty. *Journal of Time Series Analysis*, 23(5):599–618.

Proust-Lima, C. and Liquet, B. (2011). Lcmm: an r package for estimation of latent class mixed models and joint latent class models. In *The R User Conference, useR! 2011 August 16-18 2011 University of Warwick, Coventry, UK*, page 66. Citeseer.

Proust-Lima, C., Philipps, V., Diakite, A., and Liquet, B. (2017). lcmm: Extended mixed models using latent classes and latent processes. *R package version*, 1(7).

Proust-Lima, C., Philipps, V., and Liquet, B. (2015). Estimation of extended mixed models using latent classes and latent processes: the r package lcmm. *arXiv preprint arXiv:1503.00890*.

Purdon, P. L. and Weisskoff, R. M. (1998). Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fmri. *Human brain mapping*, 6(4):239–249.

Ramoni, M., Sebastiani, P., and Cohen, P. (2000). Multivariate clustering by dynamics. In *AAAI/IAAI*, pages 633–638.

Ramsay, J., Hooker, G., and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*. Use R! Springer New York.

Ramsay, J. and Silverman, B. (2002). *Applied Functional Data Analysis: Methods and Case Studies*. Springer Series in Statistics. Springer New York.

Ramsay, J. and Silverman, B. (2013). *Functional Data Analysis*. Springer Series in Statistics. Springer New York.

Rani, S. and Sikka, G. (2012). Recent techniques of clustering of time series data: a survey. *International Journal of Computer Applications*, 52(15).

Reiss, P. T., Goldsmith, J., Shang, H. L., and Ogden, R. T. (2017). Methods for scalar-on-function regression. *International Statistical Review*, 85(2):228–249.

Rice, J. A. and Silverman, B. W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society: Series B (Methodological)*, 53(1):233–243.

Richardson, S. and Green, P. J. (1997). On bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: series B (statistical methodology)*, 59(4):731–792.

Rigon, T. and Durante, D. (2021). Tractable bayesian density regression via logit stick-breaking priors. *Journal of Statistical Planning and Inference*, 211:131–142.

Robertson, F. C., Douglas, T. S., and Meintjes, E. M. (2010). Motion artifact removal for functional near infrared spectroscopy: a comparison of methods. *IEEE Transactions on Biomedical Engineering*, 57(6):1377–1387.

Rodriguez, C. E. and Walker, S. G. (2014). Label switching in bayesian mixture models: Deterministic relabeling strategies. *Journal of Computational and Graphical Statistics*, 23(1):25–45.

Rosen, O. and Stoffer, D. S. (2007). Automatic estimation of multivariate spectra via smoothing splines. *Biometrika*, 94(2):335–345.

Rosen, O., Stoffer, D. S., and Wood, S. (2009). Local spectral analysis via a bayesian mixture of smoothing splines. *Journal of the American Statistical Association*, 104(485):249–262.

Rosen, O., Wood, S., and Stoffer, D. S. (2012). Adaptspec: Adaptive spectral estimation for nonstationary time series. *Journal of the American Statistical Association*, 107(500):1575–1589.

Rossell, D. and Steel, M. F. (2019). Continuous mixtures with skewness and heavy tails. In *Handbook of Mixture Analysis*, pages 219–237. Chapman and Hall/CRC.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *The annals of statistics*, pages 1346–1370.

Sanderson, J., Fryzlewicz, P., and Jones, M. (2010). Estimating linear dependence between nonstationary time series using the locally stationary wavelet model. *Biometrika*, 97(2):435–446.

Santosa, H., Zhai, X., Fishburn, F., and Huppert, T. (2018). The nirs brain analyzir toolbox. *Algorithms*, 11(5):73.

Schroeter, M. L., Bücheler, M. M., Müller, K., Uludağ, K., Obrig, H., Lohmann, G., Tittgemeyer, M., Villringer, A., and von Cramon, D. Y. (2004). Towards a standard analysis for functional near-infrared imaging. *NeuroImage*, 21(1):283–290.

Shimamura, K., Ueki, M., Kawano, S., and Konishi, S. (2019). Bayesian generalized fused lasso modeling via neg distribution. *Communications in Statistics-Theory and Methods*, 48(16):4132–4153.

Sidell, B. D. and O'Brien, K. M. (2006). When bad things happen to good fish: the loss of hemoglobin and myoglobin expression in antarctic icefishes. *Journal of Experimental Biology*, 209(10):1791–1802.

Singhal, A. and Seborg, D. E. (2005). Clustering multivariate time-series data. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 19(8):427–438.

Sperrin, M., Jaki, T., and Wit, E. (2010). Probabilistic relabelling strategies for the label switching problem in bayesian mixture models. *Statistics and Computing*, 20(3):357–366.

Spezia, L. (2009). Reversible jump and the label switching problem in hidden markov models. *Journal of Statistical Planning and Inference*, 139(7):2305–2315.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the royal statistical society: Series b (statistical methodology)*, 64(4):583–639.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):485–493.

Sravish, A. V., Tronick, E., Hollenstein, T., and Beeghly, M. (2013). Dyadic flexibility during the face-to-face still-face paradigm: A dynamic systems analysis of its temporal organization. *Infant Behavior and Development*, 36(3):432–437.

Stephens, M. (2000). Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809.

Sun, Z., Rosen, O., and Sampson, A. R. (2007). Multivariate bernoulli mixture models with application to postmortem tissue studies in schizophrenia. *Biometrics*, 63(3):901–909.

Suzuki, M., Miyai, I., Ono, T., and Kubota, K. (2008). Activities in the frontal cortex and gait performance are modulated by preparation. an fnirs study. *Neuroimage*, 39(2):600–607.

Tang, X. and Qu, A. (2016). Mixture modeling for longitudinal data. *Journal of Computational and Graphical Statistics*, 25(4):1117–1137.

Tarabulsy, G. M., Provost, M. A., Deslandes, J., St-Laurent, D., Moss, E., Lemelin, J.-P., Bernier, A., and Dassylva, J.-F. (2003). Individual differences in infant still-face response at 6 months. *Infant behavior and development*, 26(3):421–438.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The annals of statistics*, 39(3):1335–1371.

Tipping, M. E. (2001). Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244.

Tronick, E., Als, H., Adamson, L., Wise, S., and Brazelton, T. B. (1978). The infant's response to entrapment between contradictory messages in face-to-face interaction. *Journal of the American Academy of Child psychiatry*, 17(1):1–13.

Tuft, M., Hall, M. H., and Krafty, R. T. (2021). Spectra in low-rank localized layers (spelll) for interpretable time–frequency analysis. *Biometrics*.

Ultsch, A. and Mörchen, F. (2005). *ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM*, volume 46. Univ.

Verdoolaege, G. and Rosseel, Y. (2010). Activation detection in event-related fmri through clustering ofwavelet distributions. In *2010 IEEE International Conference on Image Processing*, pages 4393–4396. IEEE.

Villringer, A. and Chance, B. (1997). Non-invasive optical spectroscopy and imaging of human brain function. *Trends in neurosciences*, 20(10):435–442.

Villringer, A., Planck, J., Hock, C., Schleinkofer, L., and Dirnagl, U. (1993). Near infrared spectroscopy (nirs): a new tool to study hemodynamic changes during activation of brain function in human adults. *Neuroscience letters*, 154(1-2):101–104.

Vlachos, M., Gunopulos, D., and Das, G. (2004). Indexing time-series under conditions of noise. In *Data mining in time series databases*, pages 67–100. World Scientific.

von Economo, C. F. and Koskinas, G. N. (1925). *Die cytoarchitektonik der hirnrinde des erwachsenen menschen*. J. Springer.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 40(3):364–372.

Wahba, G. (1980). Automatic smoothing of the log periodogram. *Journal of the American Statistical Association*, 75(369):122–132.

Wand, M. P., Ormerod, J. T., Padoan, S. A., and Frühwirth, R. (2011). Mean field variational bayes for elaborate distributions. *Bayesian Analysis*, 6(4):847–900.

Wang, L., Mehrabi, M. G., and Kannatey-Asibu Jr, E. (2002). Hidden markov model-based tool wear monitoring in turning. *J. Manuf. Sci. Eng.*, 124(3):651–658.

Wang, X., Smith, K. A., Hyndman, R., and Alahakoon, D. (2004). A scalable method for time series clustering. *Unrefereed research papers*, 1.

Wang, X., Wirth, A., and Wang, L. (2007). Structure-based statistical features and multivariate time series clustering. In *Seventh IEEE international conference on data mining (ICDM 2007)*, pages 351–360. IEEE.

Wang, Y., Li, Z., and Bruce, S. A. (2021). Adaptive bayesian sum of trees model for covariate dependent spectral analysis. *arXiv preprint arXiv:2109.14677*.

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

Waterhouse, S., MacKay, D., and Robinson, A. (1995). Bayesian methods for mixtures of experts. *Advances in neural information processing systems*, 8.

Wells, J. C. and Hung, T. T. (1990). Longman pronunciation dictionary. *RELC Journal*, 21(2):95–97.

West, M., Prado, R., and Krystal, A. D. (1999). Evaluation and comparison of eeg traces: Latent structure in nonstationary time series. *Journal of the American Statistical Association*, 94(446):375–387.

Wood, S., Rosen, O., and Kohn, R. (2011). Bayesian mixtures of autoregressive models. *Journal of Computational and Graphical Statistics*, 20(1):174–195.

Wood, S. A., Jiang, W., and Tanner, M. (2002). Bayesian mixture of splines for spatially adaptive nonparametric regression. *Biometrika*, 89(3):513–528.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. chapman and hall/CRC.

Worsley, K. J. and Friston, K. J. (1995). Analysis of fmri time-series revisited—again. *Neuroimage*, 2(3):173–181.

Wyatt, J. S., Delpy, D. T., Cope, M., Wray, S., and Reynolds, E. (1986). Quantification of cerebral oxygenation and haemodynamics in sick newborn infants by near infrared spectrophotometry. *The Lancet*, 328(8515):1063–1066.

Xiong, Y. and Yeung, D.-Y. (2002). Mixtures of arma models for model-based time series clustering. In *2002 IEEE International Conference on Data Mining, 2002. Proceedings.*, pages 717–720. IEEE.

Xu, X. and Ghosh, M. (2015). Bayesian variable selection and estimation for group lasso. *Bayesian Analysis*, 10(4):909–936.

Yang, W.-H., Holan, S. H., and Wikle, C. K. (2016). Bayesian lattice filters for time-varying autoregression and time–frequency analysis. *Bayesian Analysis*, 11(4):977–1003.

Ye, J., Janardan, R., and Li, Q. (2004). Gpca: An efficient dimension reduction scheme for image compression and retrieval. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 354–363.

Ye, J. C., Tak, S., Jang, K. E., Jung, J., and Jang, J. (2009). Nirs-spm: statistical parametric mapping for near-infrared spectroscopy. *Neuroimage*, 44(2):428–447.

Yuksel, S. E., Wilson, J. N., and Gader, P. D. (2012). Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193.

Zens, G. (2019). Bayesian shrinkage in mixture-of-experts models: identifying robust determinants of class membership. *Advances in Data Analysis and Classification*, 13(4):1019–1051.

Zhang, G. and Taniguchi, M. (1994). Discriminant analysis for stationary vector time series. *Journal of Time Series Analysis*, 15(1):117–126.

Zhang, G. and Taniguchi, M. (1995). Nonparametric approach for discriminant analysis in time series. *Journaltitle of Nonparametric Statistics*, 5(1):91–101.

Zhang, H., Ho, T. B., Zhang, Y., and Lin, M.-S. (2006). Unsupervised feature extraction for time series clustering using orthogonal wavelet transform. *Informatica*, 30(3).

Zhang, J., Siegle, G. J., Sun, T., D'andrea, W., and Krafty, R. T. (2021). Interpretable principal component analysis for multilevel multivariate functional data. *Biostatistics*.

Zhang, L., Baladandayuthapani, V., Mallick, B. K., Manyam, G. C., Thompson, P. A., Bondy, M. L., and Do, K.-A. (2014). Bayesian hierarchical structured variable selection methods with application to molecular inversion probe studies in breast cancer. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 63(4):595–620.

Zhang, S. (2016). Adaptive spectral estimation for nonstationary multivariate time series. *Computational Statistics & Data Analysis*, 103:330–349.