Title
Mixed Membership Models with Applications to Neuroimaging

Author
Marco, Nicholas Daya

Publication Date
2023

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Mixed Membership Models with Applications to Neuroimaging

A dissertation submitted in partial satisfaction

of the requirements for the degree

Doctor of Philosophy in Biostatistics

by

Nicholas Daya Marco

2023

ABSTRACT OF THE DISSERTATION

Mixed Membership Models with Applications to Neuroimaging

by

Nicholas Daya Marco
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2023
Professor Donatello Telesca, Chair

Mixed membership models, or partial membership models, are a flexible unsupervised learning method that allows each observation to belong to multiple clusters. In this dissertation, we propose a Bayesian mixed membership model for multivariate Gaussian data in Chapter 2 and functional data in Chapter 3. Compared to previous work on mixed membership models, our proposal allows for increased modeling flexibility, with the benefit of a directly interpretable mean and covariance structure. Our work is primarily motivated by studies in functional brain imaging through electroencephalography (EEG) of children with autism spectrum disorder (ASD). In this context, our work formalizes the clinical notion of "spectrum" in terms of feature membership probabilities. In Chapter 4, we extend the functional mixed membership model proposed in Chapter 3 to include covariate dependence. Using age as our covariate, we revisit the ASD study to illustrate the effect age has on alpha oscillations of developing children. The dissertation concludes with a discussion on possible extensions of the mixed membership framework in Chapter 5.

The dissertation of Nicholas Daya Marco is approved.

Joanne B. Weidhaas

Damla Şentürk

Michele Guindani

Donatello Telesca, Committee Chair

University of California, Los Angeles

2023

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

VITA

2014 - 2017 B.S., Mathematics: Option in Statistics, California State University, Long Beach

PUBLICATIONS

Marco, N., Şentürk, D., Jeste, S., DiStefano, C., Dickinson, A., & Telesca, D. (2023+). Covariate Adjusted Functional Mixed Membership Models. (in preparation)

Shamshoian, J., **Marco, N.**, Şentürk, D., & Telesca, D. (2023+). Bayesian Covariance Regression in Functional Data Analysis with Applications to Functional Brain Imaging. (in preparation)

Marco, N., Şentürk, D., Jeste, S., DiStefano, C., Dickinson, A., & Telesca, D. (2022). Flexible Regularized Estimation in High-Dimensional Mixed Membership Models. *arXiv preprint arXiv*:2212.06906.

Marco, N., Şentürk, D., Jeste, S., DiStefano, C., Dickinson, A., & Telesca, D. (2022). Functional Mixed Membership Models. *arXiv preprint arXiv*:2206.12084.

Gunatilaka, A. B., **Marco, N.**, Read, G. H., Sweeney, M., Regan, G., Tsang, C., ... & Weidhaas, J. B. (2022, May). Viral burden and clearance in asymptomatic COVID-19 Patients. *In Open forum infectious diseases* (Vol. 9, No. 5, p. ofac126). US: Oxford University Press.

Kishan, A. U., **Marco, N.**, Schulz-Jaavall, M. B., Steinberg, M. L., Tran, P. T., Juarez, J. E., ... & Weidhaas, J. B. (2022). Germline variants disrupting microRNAs predict long-term genitourinary toxicity after prostate cancer radiation. *Radiotherapy and Oncology*, 167, 226-232.

Weidhaas, J., **Marco, N.**, Scheffler, A. W., Kalbasi, A., Wilenius, K., Rietdorf, E., ... & Telesca, D. (2022). Germline biomarkers predict toxicity to anti-PD1/PDL1 checkpoint therapy. *Journal for immunotherapy of cancer*, 10(2).

PRESENTATIONS

Functional Mixed Membership Models, **invited colloquium talk**, Mathematics Colloquium, California State University, Long Beach, CA, USA, March 2023.

Functional Partial Membership Models, **joint invited talk with Donatello Telesca**, O'Bayes 2022: Objective Bayes Methodology Conference, Santa Cruz, CA, USA, September 2022.

Functional Partial Membership Models, **poster**, O'Bayes 2022: Objective Bayes Methodology Conference, Santa Cruz, CA, USA, September 2022.

Bayesian Functional Partial Membership Models, **contributed presentation**, Joint Statistical Meetings, Washington D.C., USA, August 2022.

Bayesian Functional Partial Membership Models, **poster**, 2022 ISBA World Meeting, Montreal, Canada, June 2022

CHAPTER 1

Introduction

Cluster analysis often aims to identify homogeneous subgroups of statistical units within a data-set [Hennig et al., 2015]. A typical underlying assumption of both heuristic and model-based procedures posits the existence of a finite number of sub-populations, from which each sample is extracted with some probability, akin to the idea of *u*ncertain membership. A natural extension of this framework allows each observation to belong to multiple clusters simultaneously; leading to the concept of mixed membership models, or partial membership models [Blei et al., 2003, Erosheva et al., 2004], akin to the idea of *p*artial membership. This dissertation aims to present a interpretable and flexible mixed membership framework for multivariate data, as well as functional data.

The idea that cluster analysis should allow observations to belong to more than one cluster originated in a paper on *fuzzy sets* by Zadeth [1965], gibing rise to a subgroup of cluster analysis called *fuzzy clustering* [Ruspini et al., 2019]. Historically, fuzzy clustering has referred to cost-based algorithms, meaning probabilistic uncertainty on the memberships to each cluster is unobtainable. In this dissertation, we will use the term *mixed membership models* to refer to a probabilistic representation of fuzzy clustering, leading to the idea of *mixed membership*. One of the earliest uses of mixed membership models was to model individuals into sub-populations using genotype data [Pritchard et al., 2000]. The added flexibility of a mixed membership model allowed them to study admixed individuals, or individuals that have parents that belong to two separate sub-populations. The use of mixed membership models in the field of genetics, sometimes referred to as *admixture models* in

the genetics literature, became a popular model for reconstructing ancestries from genotype data [Tang et al., 2005, Alexander et al., 2009]. Mixed membership models also became popular in topic modeling with the advent of the latent Dirichlet allocation (LDA) model [Blei et al., 2003], specifically within topic modeling of text corpora. While the first works on mixed membership models were largely application or domain specific, Heller et al. [2008] introduced a fully probabilistic mixed membership framework for data that is assumed to have come from the exponential family of distributions (Heller et al. [2008] referred to the model as a partial membership model). While the Gaussian case is covered in the framework described by Heller et al. [2008], their framework had two main drawbacks: flexibility and interpretability.

Beyond clustering, the ability to model an observation's membership on a spectrum is particularly advantageous when we consider applications to biomedical data. For example, in diagnostic settings, the severity of symptoms may vary from person to person. It is therefore important to ask which symptomatic features characterize the sample, and whether subjects exhibit one or more of these features. Within this general context, our work is motivated by functional brain imaging studies of neurodevelopmental disorders, such as autism spectrum disorder (ASD). In this setting, typical patterns of neuronal activity are examined through the use of brain imaging technologies (e.g. electroencephalography (EEG) or functional magnetic resonance imaging (fMRI)), and result in observations which are naturally characterized as random functions on a specific evaluation domain. We are primarily interested in the discovery of distinct features characterizing the sample and in the explanation of how these features determine subject-level heterogeneity. As our primary motivating case study is a neurodevelopmental study, we expect patterns of alpha oscillations to change as children age. Therefore, we develop a covariate adjusted mixed membership model in Chapter 4, which allows us to study how the mean and covariance structure of the mixed membership model changes as children age.

The remainder of the dissertation is organized as follows. Section 1.1 discusses the general

framework of the proposed mixed membership models and shows how it differs from traditional finite mixture models. Section 1.2 compares the general framework of the proposed covariate adjusted mixed membership models to well established covariate-dependent clustering techniques such as mixture of regressions and mixture of experts models. Chapters 2 and 3 contain an in-depth discussion of multivariate Gaussian mixed membership models and functional mixed membership models, respectively. Section 1.2 discusses the general framework of covariate adjusted mixed membership models, and how they compare to mixture of regressions and mixture of experts models. Chapter 4 contains an in-depth discussion on covariate adjusted functional mixed membership models, along with an example of how they can be utilized to analyze neurodevelopmental data. Lastly, Chapter 5 contains a discussion on possible extensions to the models proposed in this dissertation.

1.1 Overview of Mixed Membership Models

Finite mixture models are a well-studied class of models that provide a flexible and formal framework for model-based clustering [Melnykov and Maitra, 2010, McLachlan and Basford, 1988, McLachlan and Peel, 2004]. In this section, we will show that our proposed model can also be thought of as an extension of a Gaussian finite mixture model. For this section, we will assume that the number of features or clusters, K, are known *a-priori*. While the number of features are known *a-priori* for the technical developments of the paper, Sections 2.2.2 and 3.3.2 contain an in-depth discussion on the use of information criteria to aid in choosing the number of features.

We will start by letting $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be the observed noise-free data, where $\mathbf{x}_i \in \mathbb{R}^P$. Under the Gaussian finite mixture model framework, we would typically assume that $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are independent and identically distributed from a distribution such that

$$p\left(\mathbf{x}_i | \rho_{(1:K)}, \boldsymbol{\nu}_{(1:K)}, \mathbf{C}_{(1:K)}\right) = \sum_{k=1}^{K} \rho_k \mathcal{N}\left(\mathbf{x}_i | \boldsymbol{\nu}_k, \mathbf{C}_k\right),$$

where $\boldsymbol{\nu}_k$ is the mean of the k^{th} cluster, $\mathbf{C}_k$ is the covariance of the k^{th} cluster, and ρ_k is the mixing proportion for the k^{th} cluster. The mixing proportion, ρ_k, can be thought of as the proportion of observations that belong to the k^{th} cluster. In the finite mixture model framework, we have the constraint that $\sum_{k=1}^{K} \rho_k = 1$. In a Gaussian finite mixture model framework, we assume that each observation comes from exactly one of the mixing components or clusters, with a corresponding mean and covariance of $\boldsymbol{\nu}^{(k)}$ and $\mathbf{C}^{(k)}$, respectively. By introducing the latent variables $\boldsymbol{\pi}_i = [\pi_{i1}, \ldots, \pi_{iK}]$ ($\pi_{ik} \in \{0, 1\}$ and $\sum_{k=1}^{K} \pi_{ik} = 1$), we can equivalently express our model as

$$p\left(\mathbf{x}_i | \rho_{(1:K)}, \boldsymbol{\nu}_{(1:K)}, \mathbf{C}_{(1:K)}\right) = \sum_{\boldsymbol{\pi}_i} p(\boldsymbol{\pi}_i) \prod_{i=1}^{K} \mathcal{N}\left(\mathbf{x}_i | \boldsymbol{\nu}_k, \mathbf{C}_k\right)^{\pi_{ik}}, \tag{1.1}$$

where $p(\pi_{ik} = 1) = \rho_k$. In this context, the latent variables π_{ik} can be interpreted as the i^{th} observations membership to the k^{th} cluster. The mixed membership model proposed by Heller et al. [2008] can directly be obtained from equation 1.1 by just making π a continuous variable, such that $\pi \in (0, 1)$. In the Gaussian case, this lead to the following likelihood:

$$\mathbf{x}_i | \boldsymbol{\pi}_{(1:N)}, \boldsymbol{\nu}_{(1:K)}, \mathbf{C}_{(1:K)} \sim \mathcal{N}\left(\mathbf{H}_i \mathbf{h}_i, \mathbf{H}_i\right), \tag{1.2}$$

where $\mathbf{h}_i = \sum_{k=1}^{K} \pi_{ik} \mathbf{C}_k^{-1} \boldsymbol{\nu}_k$ and $\mathbf{H}_i = \left(\sum_{k=1}^{K} \pi_{ik} \mathbf{C}_k^{-1}\right)^{-1}$. Therefore, from equation 1.2, we can see that in the Gaussian case, the likelihood proposed by Heller et al. [2008] is just the pdf of a normal distribution, where the natural parameters are a convex combination of the individual clusters' natural parameters. Section 2.3 contains a more comprehensive discussion on the differences between our proposed model and the mixed membership model proposed by Heller et al. [2008].

If we condition on $\boldsymbol{\pi}_1 \cdots \boldsymbol{\pi}_N$, then we can equivalently express the finite mixture model in equation 1.1 as

$$\mathbf{x}_i | \boldsymbol{\pi}_{(1:N)} = \sum_{k=1}^{K} \pi_{ik} \mathbf{f}_k, \tag{1.3}$$

where $\mathbf{f}_k \sim \mathcal{N}(\boldsymbol{\nu}_k, \mathbf{C}_k)$. In a Gaussian finite mixture model we can assume that the mixing components are independent, or that $\mathbf{f}_k \sim_{ind} \mathcal{N}(\boldsymbol{\nu}_k, \mathbf{C}_k)$. Thus we can rewrite the likelihood as

$$\mathbf{x}_i|\boldsymbol{\pi}_{(1:N)}, \boldsymbol{\nu}_{(1:K)}, \mathbf{C}_{(1:K)} \sim \mathcal{N}\left(\sum_{k=1}^{K} \pi_{ik}\boldsymbol{\nu}_k, \sum_{k=1}^{K} \pi_{ik}\mathbf{C}_k\right). \tag{1.4}$$

From equation 1.3, we can extend the Gaussian finite mixture model to arrive at our proposed mixed membership model by allowing each observation to come from a positive finite number of mixture components, which we will call *features* [Heller et al., 2008, Broderick et al., 2013, Marco et al., 2022a]. To generalize the finite mixture model framework, we will first introduce a new set of latent variables $\mathbf{z}_i = [Z_{i1} \cdots Z_{iK}]$ such that $Z_{ik} \in (0, 1)$ and $\sum_{k=1}^{K} Z_{ik} = 1$. Using this new set of latent variables, we arrive at the general from of our mixed membership model

$$\mathbf{x}_i|\mathbf{z}_{(1:N)} = \sum_{k=1}^{K} Z_{ik}\mathbf{f}_k. \tag{1.5}$$

In this context, the latent variables Z_{ik} can be interpreted as the i^{th} observations proportion of membership to the k^{th} feature. While the random variables $\mathbf{f}_k$ could be considered independent in the finite mixture model case, they can no longer be considered independent without making strong assumptions on the data generating process. Visualizations of the effects of the cross-covariance can be seen in figure 2.1. Thus in order to maintain an adequately expressive and flexible model, we will allow dependence between the K features by modelling the cross-covariance between the features. Let $\mathbf{C}^{(k,k')} = \text{Cov}(\mathbf{f_k}, \mathbf{f_{k'}})$ denote the cross-covariance between the feature k and feature k' and $\boldsymbol{\mathcal{C}}$ denote the collection of covariance and cross-covariance matrices. Thus, we arrive at the general form of the likelihood of our proposed mixed membership model:

$$\mathbf{x}_i|\mathbf{z}_{(1:N)}, \boldsymbol{\nu}_{(1:K)}, \boldsymbol{\mathcal{C}} \sim \mathcal{N}\left(\sum_{k=1}^{K} Z_{ik}\boldsymbol{\nu}_k, \sum_{k=1}^{K} Z_{ik}^2\mathbf{C}_k + \sum_{k=1}^{K}\sum_{k\neq k'} Z_{ik}Z_{ik'}\mathbf{C}^{(k,k')}\right). \tag{1.6}$$

Since we do not assume independence, a concise representation of the covariance structure is

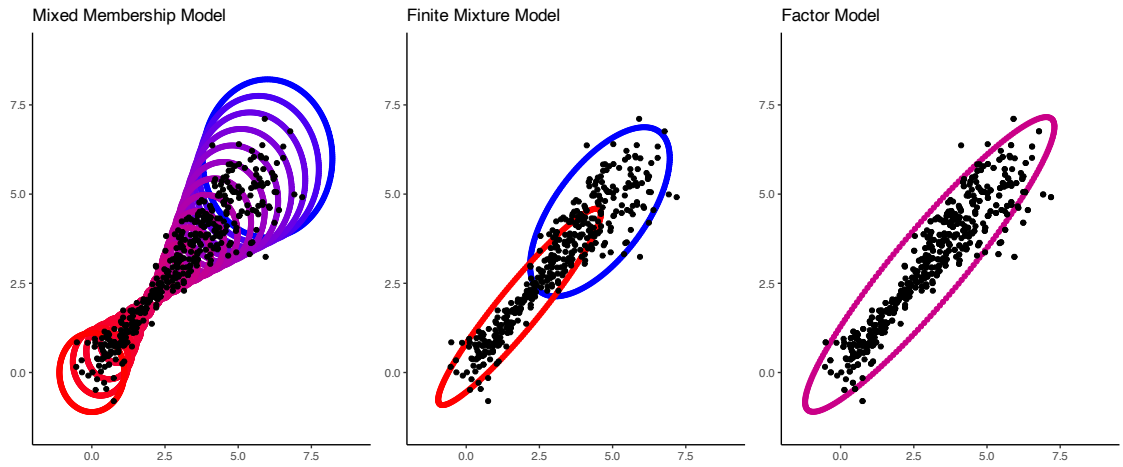


Figure 1.1: Visualization of the differences between mixed membership models, finite mixture models, and factor models.

needed in order to ensure scalability. If we were to implement a naïve characterization of the covariance structure, we would need $\mathcal{O}(K^2P^2)$ parameters just to represent the covariance structure, which does not scale well as we increase the dimension of the data (P) or the number of features (K). Instead, we will be using a joint decomposition of the K features, which will allow us to represent the covariance structure with $\mathcal{O}(KPM)$ parameters, where M is a user-determined variable that controls the accuracy of our representation of the covariance structure. While the mixed membership model was derived with $\mathbf{x}_i \in \mathbb{R}^P$, a similar derivation for square-integrable functions can be found in Section 1.2.

Figure 1.1 contains a visualization of the differences between the proposed multivariate Gaussian mixed membership model, finite mixture models, and factor analysis models. From Figure 1.1, we can see that the finite mixture model assumes that each observation comes from one of the two clusters. When fitting the finite mixture model, the goal is often to infer the mean and variance of the K Gaussian distributions used in the finite mixture model, as well as the probability that each observation comes from each cluster. While it is unknown which cluster each observation belongs to, the key assumption is that each observation was generated from one of the K clusters. Alternatively, Figure 1.1 illustrates that mixed membership models view membership as a spectrum instead of a binary concept, leading

to increased model expressivity. Similarly to finite mixture models, the goal is to infer the mean and covariance of the K Gaussian distributions, however we aim to also infer the cross-covariance between the K features, as well as each observation's proportion of membership to the K features. The ability to model the cross-covariance between features allows us to have a more flexible model, as illustrated in Figure 1.1, where distribution of observations that belong to both features equally (denoted by purple contours) have relatively small variances. While factor models are not used for clustering analysis, the proposed model has distinct similarities to a common factor model. A detailed discussion on the differences between factor models and our proposed mixed membership model can be found in Section A.4.

1.2 Overview of Covariate Adjusted Mixed Membership Models

Mixed membership models are unsupervised models that aim to explain the heterogeneity in a dataset through a set of latent underlying features. While we often have little previous knowledge on how the data are correlated, there are certain situations in which the distribution of the data is dependent on a covariate of interest, leading to the need for covariate-dependent clustering techniques. In the fields of statistics and machine learning, covariate-dependent clustering models can be found under numerous names, including *finite mixture of regressions* and *mixture of experts*. The term finite mixture of regressions [McLachlan et al., 2019, Faria and Soromenho, 2010, Grün et al., 2007, Khalili and Chen, 2007, Hyun et al., 2023, Devijver, 2015] refers to fitting a mixture model, where the mean structure is dependent on the covariates of interest through a regression framework. Mixture of experts models [Jordan and Jacobs, 1994, Bishop and Svenskn, 2002] are similar to finite mixture of regressions in that they assume that the likelihood is a weighted combination of probability distribution functions. However, in the mixture of experts model, the weights are dependent on the covariates of interest, adding an extra layer of flexibility compared to traditional finite of regressions models.

As discussed in Chapters 2 and 3, finite mixture models do a relatively poor job of explaining the variability of alpha oscillations do to the assumption that each observation is drawn from exactly one of the K clusters. Alternatively, mixed membership models assume a continuous mixing of the features, leading to more interpretable results. Therefore a finite mixture of regressions model may not be an appropriate model for inferring how age affects alpha oscillations as children grow, leading to the need for a covariate adjusted mixed membership model. In Chapter 4, we derive a covariate adjusted mixed membership model specifically for functional data. In this setting, we assume that the covariates of interest are scalar-valued or vector-valued, and the data which we would like to learn the allocation structure of are functional.

Functional data analysis (FDA) focuses on analyzing the sample paths of continuous stochastic processes $f : \mathcal{T} \to \mathbb{R}$, where $\mathcal{T}$ is a compact subset of $\mathbb{R}^d$. In FDA, we commonly assume that the random functions are elements of a Hilbert space, or more specifically that the random functions are square-integrable functions $\left(f \in L^2 \text{ or } \int_{\mathcal{T}} \mid f(t) \mid^2 \mathrm{d}t < \infty\right)$. In this dissertation, we will assume that the continuous stochastic processes are Gaussian processes (GP), meaning the distribution of function can be specified by a mean function, $\mu(t) = \mathbb{E}\left(f(t)\right)$, and a covariance function, $C(s,t) = \mathrm{Cov}\left(f(s), f(t)\right)$, for $t, s \in \mathcal{T}$.

Since mixed membership models can be considered a generalization of finite mixture models, we will show in this section how finite mixture of regressions and mixture of experts models relate to our proposed mixed membership models. For the theoretical developments discussed in this section, we will assume that the number of clusters or features, K, are known *a-priori*. Functional clustering generally assumes that each sample path is drawn from one of K underlying cluster-specific sub-processes [James and Sugar, 2003, Chiou and Li, 2007, Jacques and Preda, 2014]. Assuming that $f^{(1)}, \ldots, f^{(K)}$ are the K underlying cluster-specific sub-processes with corresponding mean functions $\mu^{(1)}, \ldots \mu^{(K)}$ and covariance

functions $C^{(1)}, \ldots, C^{(K)}$, we can arrive at the general form of a GP finite mixture model:

$$p\left(f_i \mid \rho^{(1:K)}, \mu^{(1:K)}, C^{(1:K)}\right) = \sum_{k=1}^{K} \rho^{(k)} \, \mathcal{GP}\left(f_i \mid \mu^{(k)}, C^{(k)}\right), \tag{1.7}$$

where $\rho^{(k)}$ ($\sum_{k=1}^{K} \rho^{(k)} = 1$) are the mixing proportions and f_i are the sample paths for $i = 1, \ldots, N$. Introducing the latent variables $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$, where $\boldsymbol{\pi}_i \sim_{iid} \text{Mult}(1; \rho^{(1)}, \ldots, \rho^{(K)})$, we can show that the likelihood can be written as

$$f_i \mid \boldsymbol{\pi}_i, \mu^{(1:K)}, C^{(1:K)} \sim \mathcal{GP}\left(\sum_{k=1}^{K} \pi_{ik}\mu^{(k)}, \sum_{k=1}^{K} \pi_{ik} C^{(k)}\right). \tag{1.8}$$

Using this formulation of the likelihood, we can interpret π_{ik} as a binary indicator of the i^{th} observation's membership to the k^{th} cluster. Let $\mathbf{x}_i = [X_{i1} \ldots X_{iR}]$ be the covariates of interest associated with the i^{th} observation. We will let $\mathbf{X}$ denote the design matrix (without an intercept column), where $\mathbf{x}_i$ is the i^{th} row of the design matrix. Extending the multivariate mixture of regressions model [McLachlan et al., 2019, Faria and Soromenho, 2010, Grün et al., 2007, Khalili and Chen, 2007, Hyun et al., 2023, Devijver, 2015] to a functional setting, we can represent the general form of mixture of regressions model for functional data as

$$f_i \mid \mathbf{X}, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N, \mu^{(1:K)}, C^{(1:K)} \sim \mathcal{GP}\left(\sum_{k=1}^{K} \pi_{ik}\mu^{(k)}(\mathbf{x}_i), \sum_{k=1}^{K} \pi_{ik} C^{(k)}\right). \tag{1.9}$$

In the multivariate setting, the mean is often modeled through a regression framework, leading to the functional form in the FDA setting of $\mu^{(k)}(\mathbf{x}_i, t) = \beta_0(t) + \sum_{r=1}^{R} X_{ir}\beta_r(t)$, where $\beta_0, \ldots, \beta_R \in L^2$ and $t \in \mathcal{T}$. An in-depth review on functional regression can be found in Section 4.2.4. Similarly to Equation 4.1, the mixture of experts model can be formulated as

$$p\left(f_i \mid \rho^{(1:K)}, \mu^{(1:K)}, C^{(1:K)}\right) = \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i, \alpha_k) \, \mathcal{GP}\left(f_i \mid \mu^{(k)}(\mathbf{x}_i), C^{(k)}\right). \tag{1.10}$$

From equation 1.10, we can see that the $\pi_{ik}(\mathbf{x}_i, \boldsymbol{\alpha}_k)$ act as mixing proportions, however they are dependent on the covariates of interest. In the mixture of experts model, we assume that $\pi_{ik}(\mathbf{x}_i, \boldsymbol{\alpha}_k) \propto \exp(\boldsymbol{\alpha}_k' \mathbf{x}_i)$, where $\boldsymbol{\alpha}_k$ is a learned set of parameters. Similarly to the mixture of regressions model, the mean component is model through a regression framework, such that $\mu^{(k)}(\mathbf{x}_i, t) = \beta_0(t) + \sum_{r=1}^{R} X_{ir}\beta_R(t)$, where $\beta_0, \ldots, \beta_R \in L^2$ and $t \in \mathcal{T}$. The mixture of experts model can be written in a similar form as Equation 1.9 with the introduction of the latent variables $\boldsymbol{\pi}_i$, however, the distribution of $\boldsymbol{\pi}_i$ now depends on $\mathbf{x}_i$. Similarly to Equation 1.5, we can rewrite the finite mixture model in Equation 1.8 as

$$f_i \mid \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N =_d \sum_{k=1}^{K} \pi_{ik} f^{(k)}. \tag{1.11}$$

By introducing a new set of latent variables $\mathbf{z}_i = (Z_{i1}, \ldots, Z_{iK})'$, where $Z_{ik} \in (0, 1)$ and $\sum_{k=1}^{K} Z_{ik} = 1$, we can arrive at the functional form of the of the functional mixed membership model:

$$f_i \mid \mathbf{Z}_1, \ldots, \mathbf{Z}_N =_d \sum_{k=1}^{K} Z_{ik} f^{(k)}. \tag{1.12}$$

Thus we can see that under the functional mixed membership model, each sample path is assumed to come from a convex combination of the underlying GPs, $f^{(k)}$. Unlike in the case of traditional clustering, the functional mixed membership model does not assume that the underlying GPs are mutually independent. Thus we will let $C^{(k,j)}$ represent the cross-covariance function between the k^{th} GP and the j^{th} GP, for $1 \leq k \neq j \leq K$. Letting $\boldsymbol{\mathcal{C}}$ be the collection of covariance and cross-covariance functions, we can specify the sampling model of the functional mixed membership model as

$$f_i \mid \mathbf{z}_1, \ldots, \mathbf{z}_N, \mu^{(1:K)}, \boldsymbol{\mathcal{C}} \sim \mathcal{GP}\left(\sum_{k=1}^{K} Z_{ik}\mu^{(k)}, \sum_{k=1}^{K} Z_{ik}^2 C^{(k)} + \sum_{k=1}^{K}\sum_{k' \neq k} Z_{ik} Z_{ik'} C^{(k,k')}\right). \tag{1.13}$$

Finite mixture models, as well as mixture of experts and finite mixture of regressions models, can be represented in the same functional form as the representation in Equation

1.11. However, for these covariate adjusted clustering models, the underlying stochastic processes, $f^{(k)}$, have an associated mean that depends on the covariates of interest, which we will denote as $\mu^{(k)}(\mathbf{x}_i)$. Similarly, by assuming the underlying stochastic processes in Equation 4.7 have a mean that depends on the covariates of interest, we can arrive at the sampling model of the covariate functional mixed membership model:

$$f_i \mid \mathbf{X}, \mathbf{z}_1, \ldots, \mathbf{z}_N, \mu^{(1:K)}(\mathbf{x}_i), \boldsymbol{\mathcal{C}} \sim \mathcal{GP}\left(\sum_{k=1}^{K} Z_{ik}\mu^{(k)}(\mathbf{x}_i), \sum_{k=1}^{K} Z_{ik}^2 C^{(k)} + \sum_{k=1}^{K}\sum_{k'\neq k} Z_{ik}Z_{ik'}C^{(k,k')}\right). \tag{1.14}$$

Similarly to finite mixture of regressions models, we will leverage work from the functional regression literature to model $\mu^{(1:K)}(\mathbf{x}_i)$. Therefore, while a covariate adjusted mixed membership model can be viewed as an extension of covariate-dependent clustering, it is also natural to consider it as a generalization of linear regression. In linear regression, we aim to model how the response relates to the covariates of interest. Linear regression can be considered population level inference, as it assumes that the covariates of interest have the same effect on each observation. However, in many complex modeling scenarios, there are often sub-groups of observations that have responses which have different relationships with the covariates of interest. This idea is one of the core ideas in *Precision Medicine*, where analyses often try to take into account individuals characteristics [Kosorok and Laber, 2019]. The need to account for heterogeneous covariate effects is exemplified in Kravitz et al. [2004] in the setting of evidence-based medicine, where they provide examples where population level analyses can lead to medical interventions that lead to adverse affects to sub-groups of the population. Finite mixture of regressions models aim to account for this heterogeneity in covariate effects, by allowing each observation to belong to a cluster with a unique covariate-dependent mean structure. Covariate adjusted mixed membership models can be thought of as the most granular model out of the three models, where each observation can be modeled on a spectrum, or unit-simplex. Therefore, you can estimate covariate effects at an individual level, as compared to a sub-group level in a mixture of regressions model

or population level in linear regression. This level of granularity is greatly desired in our neurodevelopmental case study, where we would like to see how children compare to their age-adjusted peers. Moreover, we are able to also specify the expected changes in alpha oscillations as children age at an individual level, which is of scientific interest.

CHAPTER 2

Flexible Regularized Estimation in High-Dimensional Mixed Membership Models

As stated in section 1, Heller et al. [2008] introduced a generalized framework for data that is assumed to have come from the exponential family of distributions. As seen in section 1.1, the mixed membership model proposed by Heller et al. [2008] can be naturally derived from finite mixture models, leading to the likelihood found in Equation 1.2. While this model is appealing due to the flexibility of modeling the features as any distribution from the exponentially family of distributions, using a Gaussian distribution to model the features can lead to hard to interpret models that are inflexible, leading to unnatural data generating assumptions. The same issue remains true in more recent generalizations of the same modeling framework [Hou-Liu and Browne, 2022]. In this chapter, we introduce a flexible and easily interpretable Gaussian mixed membership model for multivariate data.

This chapter starts by deriving a concise eigendecomposition of the features to ensure that our model is relatively scalable. To ensure a relatively simple sampling scheme, we leverage the multiplicative gamma process shrinkage prior proposed by Bhattacharya and Dunson [2011] and relax the orthogonality constraints on the eigenvectors. In section 2.1.4 we talk about potential identifiability issues and show that our model maintains a lot of the desired theoretical properties even though we relax the orthogonality constraint. Section 2.2 consists of two simulation studies and two case studies on real data. The first simulation study focuses on the recovery of our model parameters, while the second one focuses on the performance of various information criteria in choosing the number of features in our

proposed model. We conclude this chapter with a discussion on some of the key differences between our proposed model and the model proposed by Heller et al. [2008] in section 2.3.

2.1 Finite Mixture and Mixed Membership Models

Let $\mathbf{y}_1, \ldots, \mathbf{y}_N$ be the observed data, where $\mathbf{y}_i \in \mathbb{R}^P$ is the P-dimensional outcome for observational unit i, $(i = 1, 2, \ldots, N)$. We denote with K, the number of pure mixture components or *latent features* and defer to Section 2.2.2 for an in-depth discussion on the use of information criteria to select the number of features.

Under the framework of finite mixture models, a typical assumption is that each observation is drawn from one of K subpopulations. When $\mathbf{y}_i$ is continuous, a popular sampling model assumes a mixture of multivariate Gaussian distributions. In this case, the sampling model for each mixture component k is fully determined by a mean vector $\boldsymbol{\nu}_k \in \mathbb{R}^P$, and a covariance matrix $\mathbf{C}_k \in S_+^P$ (where S_+^P denotes the set of all symmetric positive semi-definite $P \times P$ matrices). Letting $\rho_k \in (0, 1)$, $(k = 1, 2, \ldots, K)$, be the marginal probability for any $\mathbf{y}_i$ to be drawn from component k, the final sampling model assumes

$$P\left(\mathbf{y}_i \mid \rho_{(1:K)}, \boldsymbol{\nu}_{(1:K)}, \mathbf{C}_{(1:K)}\right) = \sum_{k=1}^{K} \rho_k \mathcal{N}\left(\mathbf{y}_i \mid \boldsymbol{\nu}_k, \mathbf{C}_k\right), \tag{2.1}$$

s.t. $\sum_{k=1}^{K} \rho_k = 1$. The mixing proportion, ρ_k, quantify uncertain membership for any observation $\mathbf{y}_i$ to mixture component k. An equivalent representation of the finite mixture model in Equation 2.1, relies on the introduction of latent membership indicator variables, $\boldsymbol{\pi}_i = [\pi_{i1}, \ldots, \pi_{iK}] \sim \text{Cat}(K, \boldsymbol{\rho} = (\rho_1, \ldots, \rho_K))$, s.t.

$$P\left(\mathbf{y}_i \mid \boldsymbol{\pi}_i, \boldsymbol{\nu}_{(1:K)}, \mathbf{C}_{(1:K)}\right) = \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{y}_i \mid \boldsymbol{\nu}_k, \mathbf{C}_k\right)^{\pi_{ik}}, \tag{2.2}$$

where $\pi_{ik} \in \{0, 1\}$ is interpreted as the i^{th} observations membership indicator to the k^{th}

mixture component. While, in this setting, probabilistic finite mixture models provide a little room for ambiguity in interpretation, their generalizations to probability models describing *mixed* or *partial membership* are open to alternative conceptualizations [Galyardt, 2014, Gruhl and Erosheva, 2014]. A popular and direct generalization approach, simply replaces the binary membership indicator variables $\pi_{ik} \in \{0, 1\}$ with continuous membership scores $Z_{ik} \in (0, 1)$. Heller et al. [2008] and, similarly, Ghahramani et al. [2014], propose a direct application of membership scores $\mathbf{z}_i = [Z_{i1}, \cdots, Z_{ik}]$ to the latent membership representation in Equation 2.2, obtaining:

$$P\left(\mathbf{y}_i \mid \mathbf{z}_i, \boldsymbol{\nu}_{(1:K)}, \mathbf{C}_{(1:K)}\right) \propto \prod_{k=1}^{K} \mathcal{N}\left(\mathbf{y}_i \mid \boldsymbol{\nu}_k, \mathbf{C}_k\right)^{Z_{ik}}, \tag{2.3}$$

s.t $\sum_{k=1}^{K} Z_{ik} = 1$. From here, defining $\mathbf{h}_i = \sum_{k=1}^{K} Z_{ik}\mathbf{C}_k^{-1}\boldsymbol{\nu}_k$ and $\mathbf{H}_i = \left(\sum_{k=1}^{K} Z_{ik}\mathbf{C}_k^{-1}\right)^{-1}$ as convex combinations of the natural parameters, $\left(\mathbf{C}_k^{-1}\boldsymbol{\nu}_k,\ \mathbf{C}_k^{-1}\right)$, of a multivariate Gaussian distribution, we obtain:

$$\mathbf{y}_i \mid \mathbf{z}_i, \boldsymbol{\nu}_{(1:K)}, \mathbf{C}_{(1:K)} \sim \mathcal{N}\left(\mathbf{H}_i\mathbf{h}_i, \mathbf{H}_i\right). \tag{2.4}$$

In the following section we argue that while seemingly natural, this probabilistic conceptualization of mixed membership may prove too rigid for many applications, and propose an alternative representation based on convex combinations of dependent Gaussian random vectors.

2.1.1 Mixed Membership through Convex Combinations of Dependent Gaussian Features

The proposed probabilistic representation of mixed membership for continuous data starts with the introduction of K *dependent* latent Gaussian feature vectors $\mathbf{f}_k \sim \mathcal{N}(\boldsymbol{\nu}_k, \mathbf{C}_k)$, with cross-covariance $\mathrm{Cov}(\mathbf{f}_k, \mathbf{f}_{k'}) = \mathbf{C}^{(k,k')}$, for $k \neq k' = (1, 2, \ldots, K)$.

Like before, our representation maintains reliance on unit-specific mixed membership scores $\mathbf{z}_i = [Z_{i1}, \cdots, Z_{iK}]$, defined on the K-dimensional standard simplex Δ^K. However, in contrast to the representation found in Section 2.1, the application of our continuous membership scores does not rely on the latent membership representation of Equation 2.2, but exploits the direct convex combination of our latent Gaussian features. Specifically, we note that conditioning on the membership indicator variables $\boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N$, the model in Equation 2.2 may be restated through equivalence in distribution as follows:

$$\mathbf{y}_i \mid \boldsymbol{\pi}_i =_d \sum_{k=1}^{K} \pi_{ik} \mathbf{f}_k, \tag{2.5}$$

leading to the idea that $\mathbf{y}_i \mid \{\pi_{ik} = 1\} =_d \mathbf{f}_k \sim \mathcal{N}(\boldsymbol{\nu}_k, \mathbf{C}_k)$.

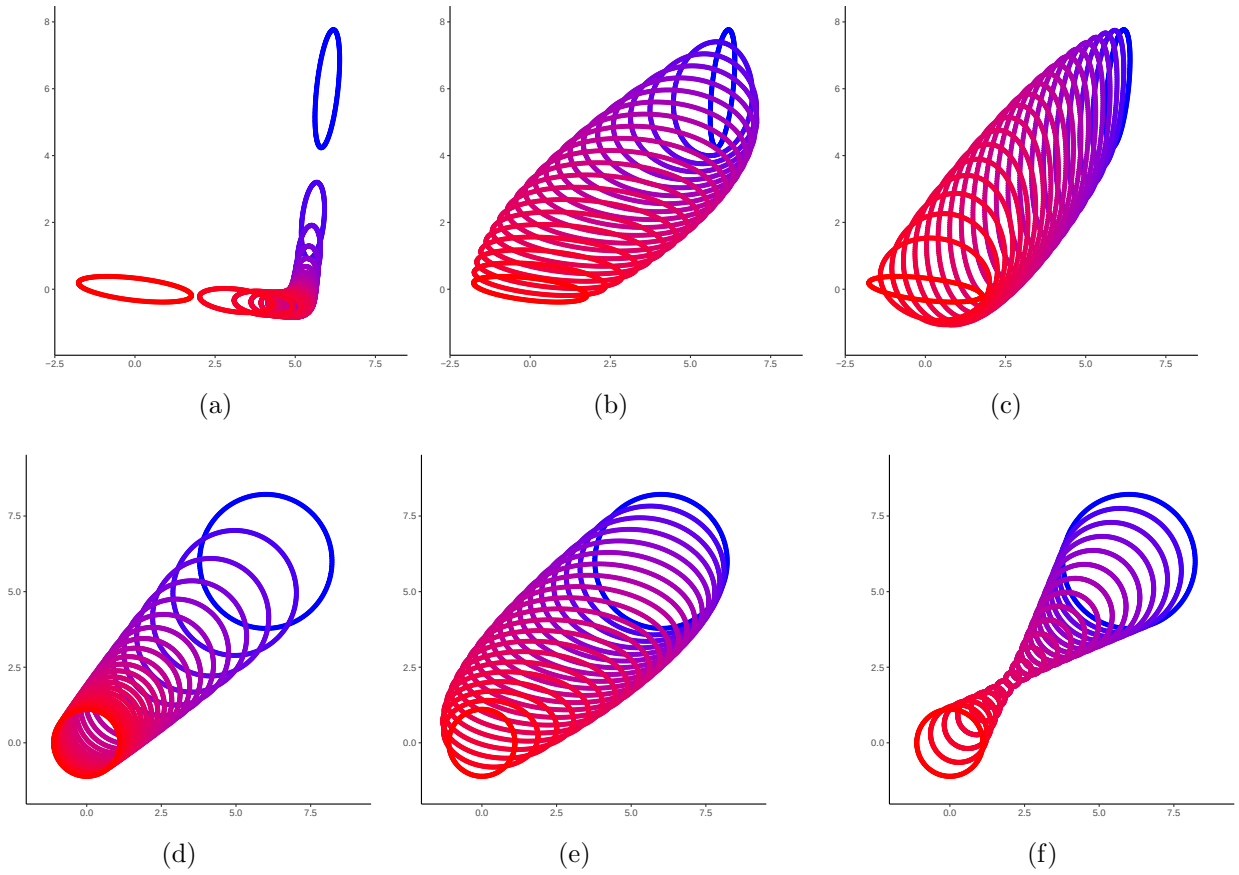
At this point, a direct application of the continuous membership scores to Equation 2.5, rather than Equation 2.2, leads to the natural definition of a sampling model through distributional equivalence with a convex combination of dependent Gaussian random vectors s.t.

$$\mathbf{y}_i \mid \mathbf{z}_i =_d \sum_{k=1}^{K} Z_{ik} \mathbf{f}_k. \tag{2.6}$$

In this context, the latent membership scores Z_{ik} can be interpreted, arguably more naturally, as the i^{th} observations proportion of membership to the k^{th} feature. Furthermore, denoting with $\boldsymbol{\mathcal{C}}$ the collection of covariance and cross-covariance matrices, we obtain a sampling model defined in terms of the original means and covariances, s.t.

$$\mathbf{y}_i \mid \mathbf{z}_i, \boldsymbol{\nu}_{(1:K)}, \boldsymbol{\mathcal{C}} \sim \mathcal{N}\left(\sum_{k=1}^{K} Z_{ik} \boldsymbol{\nu}_k, \sum_{k=1}^{K} Z_{ik}^2 \mathbf{C}_k + \sum_{k=1}^{K} \sum_{k \neq k'} Z_{ik} Z_{ik'} \mathbf{C}^{(k,k')} \right). \tag{2.7}$$

While the random variables $\mathbf{f}_k$ could be assumed independent, the inclusion of cross-covariance components allows for increased expressivity and flexibility of the ensuing sampling model. A comparative visualization of the representation offered in Equation 2.4 versus the model proposed in Equation 2.7, including the effects of cross-covariance components can be seen in



(a) (b) (c)

(d) (e) (f)

Figure 2.1: Subfigures 2.1(a) and 2.1(d) depict data generated from a two cluster Gaussian mixed membership model as specified in Heller et al. [2008]. Subfigures 2.1(b) and 2.1(c) show two examples of data generated with the same mean vectors and covariance functions as the clusters in subfigure 2.1(a), but with different cross-covariance functions. Similarly, subfigures 2.1(e) and 2.1(f) have the same mean vectors and covariance functions as the clusters in subfigure 2.1(d).

Figure 2.1. Subfigures 2.1(a) and 2.1(d) visualize the distribution of data generated from the Gaussian mixed membership model proposed by Heller et al. [2008]. Both scenarios have the same cluster-specific means, however they have different covariances associated with each cluster. We note that when the major axes of concentration for each cluster are close to orthogonal (Subfigure 2.1(a)), the generated data appear to lie on a manifold. Alternatively, Subfigures 2.1(b) and 2.1(c) show data generated from the proposed mixed membership model. In these two scenarios, each feature has the same mean and covariance as in Subfigure 2.1(a), but with varying cross-covariance matrices. Even in less extreme settings, e.g. when component mixtures are spherical (Subfigure 2.1(d)), our model generalizes sampling expressivity through the introduction of cross covariance components (Subfigures 2.1(e) and 2.1(f)).

From this simple visualization, it is evident that the mixed membership model specified by Equation 2.4 may lead to mean structures that are hard to interpret, especially in higher dimensions. This interpretability challenge is largely due to the individual feature covariances appearing in the mean term in Equation 2.4. Conversely, the mean structure in our model is a simple convex combination of individual feature means, where the weighting is directly determined by the membership proportions. More details on our analysis of alternative mixed membership representations are offered in Section 2.3.

Increased representational flexibility does, however, come at the cost of an increased dimensional parameter space. Therefore, a concise representation of the covariance structure is needed in order to ensure scalability and regularizability in estimation. In particular, if we were to implement a naïve characterization of the covariance structure, we would need $\mathcal{O}(K^2P^2)$ parameters just to represent the covariance structure, which does not scale well as we increase the dimension of the data (P) or the number of clusters (K). Instead, we will be using a joint decomposition of the K features described in Section 2.1.2, which will allow us to represent the covariance structure with $\mathcal{O}(KPM)$ parameters, where M is a user-determined variable that controls the accuracy of our representation of the covariance

structure. Once we have a concise representation of our K features, we will fully specify a Bayesian version of our proposed mixed membership model in Section 2.1.3.

2.1.2 Joint Feature Decomposition

In this section we construct a joint representation of the K features based on a multivariate eigendecomposition. This joint representation allows for a scalable representation of the underlying high-dimensional covariance structure.

Let $\mathbf{F} = [\mathbf{f}_1, \ldots, \mathbf{f}_K] \in \mathbb{R}^{P \times K}$, be a random matrix stacking the K latent Gaussian features. We define the corresponding mean matrix, $\boldsymbol{\mu} = [\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K]$, where $\boldsymbol{\nu}_k$ is the mean corresponding to the k^{th} feature. Let $\mathbf{C}^{(k,k')} = \text{Cov}\left(\mathbf{f}_k, \mathbf{f}_{k'}\right)$ for $1 \leq k, k' \leq K$. By vectorizing the matrix $\mathbf{F}$, we obtain

$$\text{Cov}(\text{vec}(\mathbf{F})) = \boldsymbol{\Sigma} = \begin{bmatrix} \mathbf{C}^{(1,1)} & \ldots & \mathbf{C}^{(1,K)} \\ \vdots & \ddots & \vdots \\ \mathbf{C}^{(K,1)} & \ldots & \mathbf{C}^{(K,K)} \end{bmatrix}. \tag{2.8}$$

Since $\boldsymbol{\Sigma}$ is a positive semi-definite matrix, we know that there exists a set of eigenvalues, $\lambda_1 \geq \lambda_2, \geq \cdots \geq \lambda_{KP} \geq 0$, and a set of eigenvectors, $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_{KP}$, such that

$$\boldsymbol{\Sigma}\boldsymbol{\Psi}_m = \lambda_m \boldsymbol{\Psi}_m.$$

Since the $\boldsymbol{\Psi}_m$ are eigenvectors, we know that they are orthonormal. We will define a set of parameters $\boldsymbol{\Phi}_m$ such that $\boldsymbol{\Phi}_m = \sqrt{\lambda_m}\boldsymbol{\Psi}_m$. Thus we can see that $\boldsymbol{\Phi}_m$ are still mutually orthogonal, but are scaled by the square root of the corresponding eigenvalue. Consider partitioning $\boldsymbol{\Phi}_m$ such that

$$\boldsymbol{\Phi}_m = \begin{bmatrix} \boldsymbol{\phi}_{1m} \\ \vdots \\ \boldsymbol{\phi}_{Km} \end{bmatrix}.$$

From Equation 2.8, by using a spectral decomposition on $\mathbf{\Sigma}$, we can see that

$$\mathbf{C}^{(k,k')} = \sum_{m=1}^{KP} \boldsymbol{\phi}_{km} \boldsymbol{\phi}'_{k'm}. \tag{2.9}$$

Since $\mathbf{\Phi}_m$ form a basis for $\mathbb{R}^{KP}$, we have that

$$\begin{aligned} \text{vec}(\mathbf{F}) - \text{vec}(\boldsymbol{\mu}) &= \mathbf{P} \circ (\text{vec}(\mathbf{F}) - \text{vec}(\boldsymbol{\mu})) \\ &= \sum_{m=1}^{KP} \lambda_m^{-1} \langle \text{vec}(\mathbf{F}) - \text{vec}(\boldsymbol{\mu}), \mathbf{\Phi}_m \rangle \mathbf{\Phi}_m, \end{aligned} \tag{2.10}$$

where $\mathbf{P}$ is the projection operator on $\mathbb{R}^{KP}$. Letting $\chi_m = \lambda_m^{-1} \langle \text{vec}(\mathbf{F}) - \text{vec}(\boldsymbol{\mu}), \mathbf{\Phi}_m \rangle$, we can see that

$$\begin{aligned} \mathbb{E}(\chi_m) &= \lambda_m^{-1} \mathbf{\Phi}'_m \mathbb{E} \left(\text{vec}(\mathbf{F}) - \text{vec}(\boldsymbol{\mu}) \right) = 0 \\ \text{Var}\left(\chi_m\right) &= \text{Cov}\left(\lambda_m^{-1} \langle \text{vec}(\mathbf{F}), \mathbf{\Phi}_m \rangle \right) \\ &= \lambda_m^{-2} \mathbf{\Phi}'_m \text{Cov}(\mathbf{F}) \mathbf{\Phi}_m \\ &= \lambda_m^{-2} \mathbf{\Phi}'_m \left(\sum_{j=1}^{KP} \mathbf{\Phi}_j \mathbf{\Phi}'_j \right) \mathbf{\Phi}_m = 1. \end{aligned}$$

Thus using the decomposition in Equation 2.10, we have that

$$\text{vec}(\mathbf{F}) = \text{vec}(\boldsymbol{\mu}) + \sum_{m=1}^{KP} \chi_m \mathbf{\Phi}_m,$$

where $\chi_m \sim \mathcal{N}(0, 1)$. In order to reduce the dimension of the model, we can often approximate $\mathbf{F}$ by only using the first M scaled eigenvectors, where $M \leq KP$. Thus, for sufficiently large M, we have

$$\text{vec}(\mathbf{F}) \approx \text{vec}(\boldsymbol{\mu}) + \sum_{m=1}^{M} \chi_m \mathbf{\Phi}_m.$$

Equivalently, we can express this in terms of the K partitioned vectors, such that

$$\mathbf{f}_k \approx \boldsymbol{\nu}_k + \sum_{m=1}^{M} \chi_m \boldsymbol{\phi}_{km}. \tag{2.11}$$

From Equation 2.11, we can see that each feature can be represented by a mean component, $\boldsymbol{\nu}_k$, and deviations from the mean controlled by the eigenstructure of the covariance structure of our model, $\boldsymbol{\phi}_{km}$. In this context, the hyperparameter M is user defined, and controls the approximation accuracy of the covariance structure. If $M = KP$, we are then able to recover the true covariance structure. However, in most applications, a relatively small M, allows for a good fit to large models, while still ensuring relatively good approximations to the true covariance structure. In the following section we discuss Bayesian estimation and adopt a regularization approach to selecting the effective dimension of M.

2.1.3 Sampling Model and Prior Distributions

In this section we combine the convolutional representation of mixed membership introduced in Section 2.1.1, with the multivariate eigen-approximation of Section 2.1.2 to define a probability model of mixed membership amenable to formal Bayesian analysis. Using our approximation in Equation 2.11, we let $\boldsymbol{\Theta}$ be the full collection of model parameters, and define the following sampling model:

$$\mathbf{y}_i \mid \boldsymbol{\Theta} \sim \mathcal{N}\left\{ \sum_{k=1}^{K} Z_{ik} \left(\boldsymbol{\nu}_k + \sum_{m=1}^{M} \chi_{im} \boldsymbol{\phi}_{km} \right),\ \sigma^2 \mathbf{I}_P \right\}, \tag{2.12}$$

where we assume $\chi_{im} \sim_{iid} \mathcal{N}(0,1)$, $(i = 1, 2, \ldots, N;\ m = 1, 2, \ldots, M)$. Marginally, integrating out χ_{im}, we recover an approximation to the model in (2.7):

$$\mathbf{y}_i \mid \boldsymbol{\Theta}_{-\chi} \sim \mathcal{N}\left\{ \sum_{k=1}^{K} Z_{ik} \boldsymbol{\nu}_k,\ \sum_{k=1}^{K} \sum_{k'=1}^{K} Z_{ik} Z_{ik'} \left(\sum_{m=1}^{M} \boldsymbol{\phi}_{km} \boldsymbol{\phi}'_{k'm} \right) + \sigma^2 \mathbf{I}_P \right\}, \tag{2.13}$$

where $\boldsymbol{\Theta}_{-\chi}$ denotes the full collection of model parameters excluding the χ_{im} parameters. Using Equation 2.9, we can see that the covariance in Equation 2.13 is a weighted sum of the individual feature covariances and the cross-covariances between the features, with an additional diagonal matrix to capture any residual noise. Similarly, the mean is a weighted sum of the individual feature means. Thus, the model in Equation 2.13 is what we would expect from a traditional additive model, however we only need $\mathcal{O}(KPM)$ parameters to estimate the covariance structure. Selecting the number of eigencomponents, M, allows one to balance computational cost and model flexibility.

From our derivations in Section 2.1.2, we know that $\boldsymbol{\Phi}_k$ are often assumed to be orthogonal. While aiding likelihood identifiability of the $\boldsymbol{\Phi}$ parameters, this constraint would require posterior simulation on a non-compact Stiefel manifold, which would be computationally challenging, and negatively affect mixing of algorithms using Markov chain Monte Carlo (MCMC) simulations. Section 2.1.4 shows that we can sample in an unconstrained space while still maintaining many of the desired theoretical properties of our model. While the $\boldsymbol{\Phi}$ parameters can no longer be interpreted as scaled eigenvectors, posterior samples of the eigenpairs can still be obtained via the posterior samples of the covariance matrix $\boldsymbol{\Sigma}$ in Equation 2.8.

Furthermore, the eigen-representation of our model in Equation 2.13, allows for adaptive regularization through shrinkage priors. Specifically, we know that the $\boldsymbol{\phi}_{km}$ should shrink in magnitude as they are scaled by the corresponding eigenvalues. Thus we leverage the multiplicative gamma process shrinkage prior proposed by Bhattacharya and Dunson [2011]. Letting ϕ_{kpm} be the p^{th} element of $\boldsymbol{\phi}_{km}$, we can specify our prior as:

$$\phi_{kpm}|\gamma_{kpm}, \tilde{\tau}_{mk} \sim \mathcal{N}\left(0, \gamma_{kpm}^{-1}\tilde{\tau}_{mk}^{-1}\right), \quad \gamma_{kpm} \sim \Gamma\left(\nu_\gamma/2, \nu_\gamma/2\right), \quad \tilde{\tau}_{mk} = \prod_{n=1}^{m} \delta_{nk}$$

$$\delta_{1k}|a_{1k} \sim \Gamma(a_{1k}, 1), \quad \delta_{jk}|a_{2k} \sim \Gamma(a_{2k}, 1), \quad a_{1k} \sim \Gamma(\alpha_1, \beta_1), \quad a_{2k} \sim \Gamma(\alpha_2, \beta_2),$$

where $1 \leq k \leq K$, $1 \leq p \leq P$, $1 \leq m \leq M$, and $2 \leq j \leq M$. Since the $\boldsymbol{\Phi}$ parameters can

be thought of as eigenvectors scaled by the square root of their corresponding eigenvalues, we would like a prior that promotes shrinkage as m increases. By setting $\alpha_2 > \beta_2$, we have that $\mathbb{E}(\delta_{jk}) > 1$, which will promote shrinkage in $\mathbf{\Phi}_i$ as m increases.

The model is completed through a conjugate prior for the mean parameters, s.t.

$$\boldsymbol{\nu}_k|\tau_k \sim \mathcal{N}\left(\mathbf{0}_P, \tau_k \mathbf{I}_P\right) \text{ and } \tau_k \sim IG(\alpha, \beta),$$

and, following Heller et al. [2008] and Ghahramani et al. [2014], we assume that

$$\mathbf{z}_i|\boldsymbol{\pi}, \alpha_3 \sim Dir(\alpha_3 \boldsymbol{\pi}), \quad \boldsymbol{\pi} \sim Dir(\mathbf{c}), \quad \alpha_3 \sim exp(b)$$

for $i = 1, \ldots, N$. Lastly, we will assume that $\sigma^2 \sim IG(\alpha_0, \beta_0)$.

2.1.4 Identifiability and Posterior Consistency

Mixed membership models are subject to non-identifiability problems similar to the ones affecting finite mixture models. As in finite mixture models, a common source of non-identifiability in mixed membership models is what is commonly known as the *label switching* problem. To solve this issue, we leverage the work of Stephens [2000] and implement relabelling algorithms to post-process the posterior samples after running MCMC simulations.

A second form of non-identifiability stems from the additional flexibility of the allocation parameters, namely that $Z_{ik} \in (0, 1)$ rather than just being a binary variable like in a finite mixture model. Consider a two feature model, where $\mathbf{\Theta}_0$ denotes the set of "true" parameters (i.e. $(Z_{ik})_0$ denotes the true value of Z_{ik}). Let $Z^*_{i1} = \frac{1}{3}(Z_{i1})_0$ and $Z^*_{i2} = (Z_{i2})_0 + \frac{2}{3}(Z_{i1})_0$ (transformation preserves the constraint that $Z^*_{i1} + Z^*_{i2} = 1$). If we let $\boldsymbol{\nu}^*_1 = 3(\boldsymbol{\nu}_1)_0 - 2(\boldsymbol{\nu}_2)_0$, $\boldsymbol{\nu}^*_2 = (\boldsymbol{\nu}_2)_0$, $\boldsymbol{\phi}^*_{1m} = 3(\boldsymbol{\phi}_{1m})_0 - 2(\boldsymbol{\phi}_{2m})_0$, $\boldsymbol{\phi}^*_{2m} = (\boldsymbol{\phi}_{2m})_0$, $\chi^*_{im} = (\chi_{im})_0$, and $(\sigma^2)^* = \sigma^2_0$, then from Equation 2.12, we have that $P(\mathbf{y}_i|\mathbf{\Theta}_0) = P(\mathbf{y}_i|\mathbf{\Theta}^*)$. We will refer to this type of non-identifiability as the *rescaling problem*. To mitigate the effects of the rescaling problem, we

derived the membership rescale algorithm (Algorithm 2). In a two feature mixed membership model, the membership rescale algorithm ensures that at least one observation completely lies in each of the features. In the case when we have more than two features, we leverage the work of Chen et al. [2022], which essentially rescales the allocation parameters such that they cover as much of the unit simplex as possible.

A final source of non-identifiability arises due to the additional flexibility of the allocation parameters. Specifically, when $(\boldsymbol{\nu}_{k'})_0 \propto (\boldsymbol{\phi}_{km})_0$ in Equation 2.12, we can see that the mean vector and covariance matrices are unidentifiable. An underestimate of $(\boldsymbol{\phi}_{km})_0$ will typically lead to additional variability in the allocation parameters. This type of non-identifiability needs to be taken into account when looking at the recovery of parameters, as in the first simulation study (Section 2.2.1), however it is often of little practical importance since the uncertainty is still being captured by the model.

In Section 2.1.3, we formulated a model where the $\boldsymbol{\Phi}_m$ parameters are no longer mutually orthogonal. Therefore, the $\boldsymbol{\Phi}_m$ parameters can no longer be interpreted as scaled eigenvectors. However, the relaxation of the orthogonality constraint facilitates easier sampling schemes and the use of "black-box" samplers to obtain samples from the posterior distribution. Relaxation of the orthogonality constraint tends to also lead to better mixing of the Markov chain. Assuming we can still recover the mean and covariance structure, we can still obtain posterior samples of the eigenvectors by reconstructing posterior draws of the covariance matrix and then calculating the eigenvectors of the sampled covariance matrices. The remaining part of this section will focus on proving that we can recover the mean and covariance structure. However, due to the identifiability issues described earlier in this section, we will be proving weak posterior consistency conditional on knowing the mixing allocation parameters.

Since our main goal is to prove that we can recover the mean and covariance structure, we will be proving weak posterior consistency using the likelihood in Equation 2.13 (integrating out the χ parameters). Since the $\boldsymbol{\phi}_{km}$ are not identifiable, we will prove posterior

consistency with respect to $\boldsymbol{\Sigma}_{kk'} := \sum_{p=1}^{KP} \left(\boldsymbol{\phi}_{kp}\boldsymbol{\phi}'_{k'p}\right)$. Let $\boldsymbol{\Pi}$ be the prior distribution on $\boldsymbol{\omega} := \{\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K, \boldsymbol{\Sigma}_{11}, \ldots, \boldsymbol{\Sigma}_{1K}, \ldots, \boldsymbol{\Sigma}_{KK}, \sigma^2\}$. We will denote the set of true parameters as

$$\boldsymbol{\omega}_0 = \left\{(\boldsymbol{\nu}_1)_0, \ldots, (\boldsymbol{\nu}_K)_0, (\boldsymbol{\Sigma}_{11})_0, \ldots, (\boldsymbol{\Sigma}_{1K})_0, \ldots, (\boldsymbol{\Sigma}_{KK})_0, \sigma_0^2\right\}.$$

In order to prove weak posterior consistency, we will have to make the following two assumptions:

Assumption 1. *The variables Z_{ik} are known a-priori for $i = 1, \ldots, N$ and $k = 1, \ldots, K$.*

Assumption 2. *The true parameter modeling the random noise is positive ($\sigma_0^2 > 0$).*

Under assumptions 1 and 2, we would like to prove that the posterior distribution $\boldsymbol{\Pi}_N(.|\mathbf{y}_1, \ldots, \mathbf{y}_N)$, is weakly consistent at $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$. In order to do that, we will have to show that there is positive prior probability around the set of true parameters. To do this, we will first define some quantities related to the Kullback–Leibler (KL) divergence. Following the notation of Choi and Schervish [2007], we will define the following quantities:

$$\Lambda_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) = \log\left(\frac{f_i(\mathbf{y}_i; \boldsymbol{\omega}_0)}{f_i(\mathbf{y}_i; \boldsymbol{\omega})}\right), \quad K_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) = \mathbb{E}_{\boldsymbol{\omega}_0}(\Lambda_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})), \quad V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) = \mathrm{Var}_{\boldsymbol{\omega}_0}(\Lambda_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})),$$

where $f_i(\mathbf{y}_i; \boldsymbol{\omega}_0)$ is the likelihood when we have the parameters $\boldsymbol{\omega}_0$. To simplify the notation, we will let

$$\begin{aligned}
\boldsymbol{\mu}_i &= \sum_{k=1}^{K} Z_{ik}\boldsymbol{\nu}_k, \\
\boldsymbol{\Sigma}_i &= \sum_{k=1}^{K}\sum_{k'=1}^{K} Z_{ik}Z_{ik'}\left(\sum_{p=1}^{KP}\left(\boldsymbol{\phi}_{kp}\boldsymbol{\phi}'_{k'p}\right)\right) + \sigma^2\mathbf{I}_P = \mathbf{U}_i'\mathbf{D}_i\mathbf{U}_i + \sigma^2\mathbf{I}_P,
\end{aligned}$$

where $\mathbf{U}_i'\mathbf{D}_i\mathbf{U}_i$ is the spectral decomposition of $\sum_{k=1}^{K}\sum_{k'=1}^{K} Z_{ik}Z_{ik'}\left(\sum_{p=1}^{KP}\left(\boldsymbol{\phi}_{kp}\boldsymbol{\phi}'_{k'p}\right)\right)$. Let $\boldsymbol{\Omega}_\epsilon(\boldsymbol{\omega}_0)$ be the set of parameters such that $\boldsymbol{\Omega}_\epsilon(\boldsymbol{\omega}_0) := \{\boldsymbol{\omega} : K_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) < \epsilon \text{ for all i}\}$. Thus we have that $\boldsymbol{\Omega}_\epsilon(\boldsymbol{\omega}_0)$ is the set of parameters such that the KL divergence is less than ϵ. We

will let $\mathcal{B}(\boldsymbol{\omega}_0)$ be the set of parameters such that

$$\mathcal{B}(\boldsymbol{\omega}_0) := \left\{ \boldsymbol{\omega} : \frac{1}{a}\left((d_{il})_0 + \sigma_0^2\right) \leq d_{il} + \sigma^2 \leq a\left((d_{il})_0 + \sigma_0^2\right), \|\left(\boldsymbol{\mu}_i\right)_0 - \boldsymbol{\mu}_i\| \leq b \right\}$$

for some $a, b \in \mathbb{R}$, where d_{il} is the l^{th} diagonal element of $\mathbf{D}_i$.

Lemma 1. *Let $\mathcal{C}(\boldsymbol{\omega}_0, \epsilon) := \mathcal{B}(\boldsymbol{\omega}_0) \cap \boldsymbol{\Omega}_\epsilon(\boldsymbol{\omega}_0)$. Thus for $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$ and $\epsilon > 0$, there exists $a > 1$ and $b > 0$ such that*

1. $\sum_{i=1}^{\infty} \frac{V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})}{i^2} < \infty$, *for any* $\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)$,
2. $\boldsymbol{\Pi}\left(\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)\right) > 0$.

Lemma 1 shows us that there are neighborhoods around $\boldsymbol{\omega}_0$ that have positive prior probability. Since $\mathbf{y}_1, \ldots, \mathbf{y}_N$ are not identically distributed, the first condition of lemma 1 is needed in order to prove weak posterior consistency.

Lemma 2. *Under assumptions 1 and 2, the posterior distribution, $\boldsymbol{\Pi}_N(.|\mathbf{y}_1, \ldots, \mathbf{y}_N)$, is weakly consistent at $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$.*

Lemma 2 states that given the known allocation structure, we are able to recover the mean and covariance structure. Thus while we cannot directly make inference on the eigenvectors of the covariance structure, we can still recover the covariance structure without enforcing the orthogonality constraint on the $\boldsymbol{\Phi}$ parameters. While the allocation parameters are usually not known, empirical evidence in Section 2.2.1 shows that the mean and covariance structure converge as we get more information.

2.2 Simulations and Case Studies

We conducted two simulation studies to explore the empirical performance of our model and two case studies to illustrate the application of mixed membership models to substantive

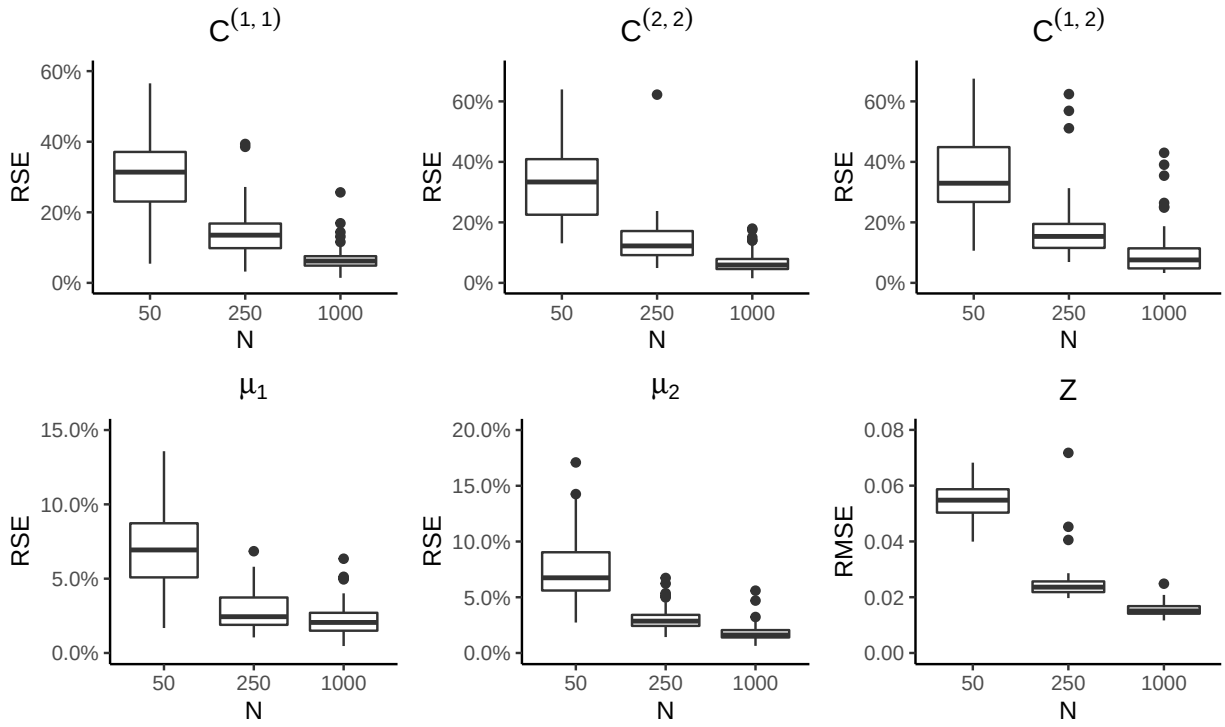


Figure 2.2: The relative squared error (RSE) for the mean and covariance structure evaluated under various sample sizes. To evaluate the recovery of the allocation parameters, we used the root mean squared error (RMSE).

scientific applications in functional brain imaging and cancer genomics. The first simulation study in Section 2.2.1 explores the empirical convergence of our model on simulated data. We then look at how information criteria can aid in choosing the number of features in Section 2.2.2. In Section 2.2.3, we look at using a mixed membership model to model electroencephalography (EEG) signals in children with Autism spectrum disorder (ASD) and typically developing (TD) children. Lastly, in Section 2.2.4, we look at how a mixed membership model can be used to classify different breast cancer subtypes using targeted gene-expression data.

2.2.1 Simulation Study 1: Operating Characteristics under Increasing Sample Size

This simulation study focuses on how well we can recover the mean, covariance, and allocation structures under different sample sizes. In this simulation study, we will specifically look at the case when we have 50, 250, and 1000 observations. For each of the three different sample sizes, we generated 50 different datasets with observations $\mathbf{y}_i \in \mathbb{R}^{10}$. The data was generated from a two feature model with $M = 4$. More information on how the simulation study was conducted can be found in Section A.3.1.

To measure how well we can recover the mean vector, covariance matrices, and cross-covariance matrices, we will use the relative squared error (RSE). The RSE is defined as

$$\text{RSE} = \frac{\|f - \hat{f}\|_2}{\|f\|_2} \times 100\%,$$

where $\|\cdot\|_2$ is the Euclidean norm if f is a vector and the Frobenius norm if f is a matrix. In this simulation study, we will use the posterior median as our estimate, $\hat{f}$. To measure how well we can recover the allocation parameters, we will use the root mean squared error (RMSE).

From Figure 2.2, we can see that our recovery of the mean and covariance structure

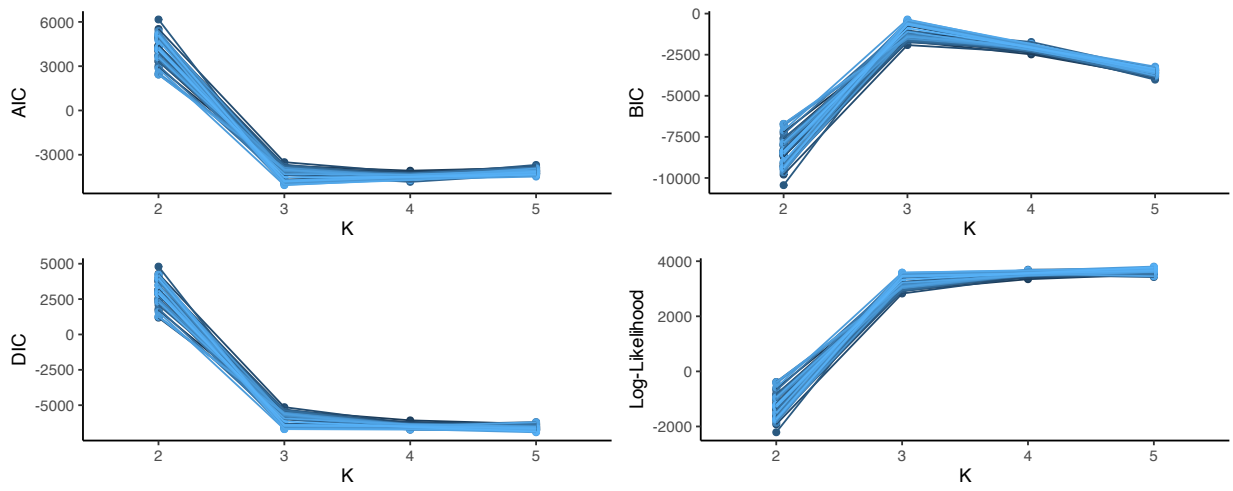


Figure 2.3: Information Criteria evaluated on our proposed Bayesian mixed membership model for $K = 2, 3, 4$, and 5. The information criteria were evaluated on 50 different simulated data sets, where the true number of features was 3.

improves as we obtain more data. While the RSE was relatively large in the simulations when $N = 50$, the credible intervals were also significantly wider to account for the uncertainty in the estimate. Specifically, the average credible interval width for the mean vectors when 50 observations were observed were roughly 4.8 times wider than the average credible interval width for the mean vectors when 1000 observations were observed. Overall, this simulation study provides empirical support that the mean and covariance structure converges to the true parameters, even when the allocation parameters are unknown.

2.2.2 Simulation Study 2: Information Criteria for the Number of Latent Features

In the finite mixed membership setting, inference and interpretation depends on the chosen number of features. To aid in this choice, practitioners often use information criteria (IC) to employ a data-driven approach for the selection of the number of clusters or features for their model. In this section, we study how IC such as AIC [Akaike, 1974], BIC [Schwarz, 1978], and DIC [Celeux et al., 2006, Spiegelhalter et al., 2002] perform in choosing the optimal number

of mixtures. In this section, we will also observe how heuristics such as the "elbow-method" perform in choosing the optimal number of clusters.

To evaluate the performance of the IC, we simulated 50 different data sets, where the true number of features was 3. For each data set, four mixed membership models were fit using a varying number of features ($K = 2, 3, 4$, and 5). Once the models were fit, we calculated the BIC, AIC, and DIC for each one of the models and evaluated the performance of these IC. Definitions of the IC used in this section, as well as detailed information on how the simulation study was conducted, can be found in Section A.3.2.

From Figure 2.3, we note that the BIC exhibited the most reliable performance, selecting the true number of features all 50 times. Since AIC does not penalize excess parameters as much as BIC, we can see that AIC sometimes selected a model with 4 features over a model with 3 features. Overall, the AIC selected the true number of features only 23 times out of the 50 different datasets, and selected a model with 4 features in 27 of the 50 datasets. In this setting, the DIC was shown to be the least reliable IC, selecting a model with 4 or more features all 50 times. As discussed in Marco et al. [2022a], the average log-likelihood can aid in the selection of the optimal number of features in mixed membership models. From Figure 2.3, we can see that there is a distinct "elbow" at $K = 3$. Using the elbow method, we would correctly identify the correct number of features all 50 times. Thus, based on the simulation results, we recommend using the "elbow-method" in conjunction with BIC to select the optimal number of features for our proposed model.

2.2.3 A Case Study on Functional Brain Imaging through EEG

Autism spectrum disorder (ASD) is a disorder that is characterized by social communication deficits and/or unusual sensory-motor behaviors [Lord et al., 2018, Edition, 2013]. While once a more narrowly defined disorder, autism is now viewed as a wide spectrum of symptoms, ranging from very mild symptoms to severe symptoms that may require life-long care. In this case study, we analyze electroencephalogram (EEG) data collected on 39 typically developing

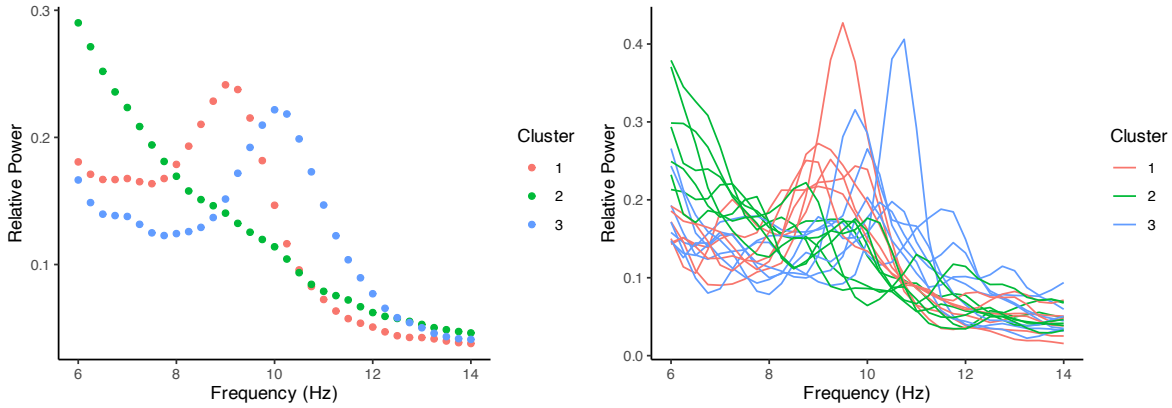


Figure 2.4: (Left Panel) Recovered means from fitting a k-means model with 3 clusters. (Right Panel) Data from the T8 electrode of 20 individuals with varying clinical diagnosis (TD vs ASD), colored by the estimated cluster membership.

(TD) children and 58 children diagnosed with ASD , between the ages of 2 and 12 years old [Dickinson et al., 2018]. EEG data in the present analyses are considered spontaneous, in that they reflect intrinsic functional brain dynamics under task-free conditions. EEG data were recorded for two minutes using a 128-channel HydroCel Geodesic Sensory net. During the recording, children were seated in front of a computer monitor displaying floating bubbles, a commonly used approach for collecting resting EEG data in developmental populations [Dawson et al., 1995, Stroganova et al., 1999, Tierney et al., 2012, McEvoy et al., 2015]. The signal from the sensors were filtered to remove signals outside of the 0.1 to 100 Hz band range, and were then interpolated to match the international 10-20 system 25 channel montage. A fast Fourier transform was then applied to the data to transform the data into the frequency domain. In this analysis, we consider the relative power in the frequency domain, meaning each function is scaled to integrate to 1. Visualizations of the data, as well as the results from a k-means analysis [Lloyd, 1982], can be seen in Figure 2.4.

When neuroscientists evaluate resting-state EEG data, one area of interest is the location of a single prominent peak in the spectral density located in the alpha band of frequencies (6-14 Hz), called the *peak alpha frequency* (PAF). The emergence of this peak has been shown to be a biomarker of neural development in typically developing children [Rodríguez-Martínez

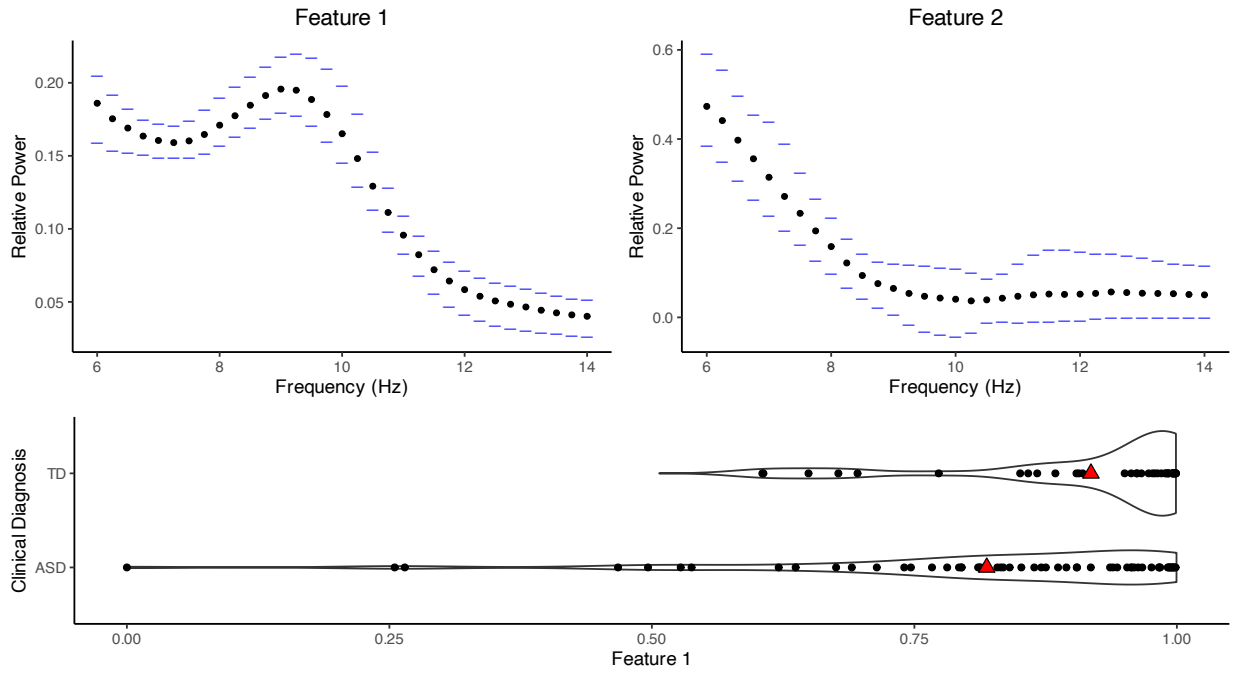


Figure 2.5: (Top Panels) Posterior median estimates of the recovered features, with corresponding 95% credible intervals. (Bottom Panel) Posterior median estimates of the membership to the first feature, stratified by clinical cohort. The red triangles represent the mean membership to the first feature.

et al., 2017]. A clear alpha peak gradually emerges over the first year of life in typically developing children, and increases in frequency over childhood before reaching stability in adolescence/early adulthood [Rodríguez-Martínez et al., 2017, Scheffler et al., 2019]. On the other hand, both the emergence of peak alpha frequency and developmental frequency shifts are shown to be atypical in children diagnosed with ASD. More specifically, children are less likely to display a clear alpha peak than age-matched peers, and do not show the same age-related increase that is well-documented in typical children [Scheffler et al., 2019].

In this case study, we conducted a multivariate analysis of the EEG analysis by using the average power of the 25 electrodes at 33 frequencies of interest in the alpha frequency band (from 6 Hz to 14 Hz with 0.25 Hz step sizes). By using the information criteria discussed in Section 2.2.2, we found that a mixed membership model with 2 features seemed to be optimal. Thus we fit a 2 feature mixed membership model ($K = 2$) with 5 eigenvectors

($M = 5$), and ran our chain for 500,000 iterations. To save on computational resources, the Markov chain was thinned, saving every 10^{th} iteration.

From Figure 2.5, we can see that the first functional feature can be interpreted as a distinct alpha peak. On the other hand, the second feature can be interpreted as a $1/f$ trend, or *pink noise*. These two features help to differentiate between periodic (alpha waves) and aperiodic ($1/f$ trend) neural activity patterns, which coexist in the EEG spectra. Loading highly on feature 2 suggests that the 1/f trend is the most prominent in an individual's EEG recording, suggesting a clear alpha peak has not yet emerged. From Figure 2.5, we can see that typically developing children seem to heavily load on the first feature, representing individuals with a distinct alpha peak. On the other hand, children with ASD seem to have relatively high heterogeneity in their loadings. We can see that on average children with ASD tend to have a more attenuated alpha peak (more loading on feature 2), with some individuals having no discernible alpha peak at all. These results closely match the results obtained by Marco et al. [2022a], who used a functional mixed membership model to model the spectral density as a random function.

Given that the presence and/or emergence of an alpha peak is a developmental biomarker, objectively quantifying the presence of a peak is important. By allowing individuals to partially belong to both features, we are able to objectively quantify the presence of a clear alpha peak over and above $1/f$ aperiodic patterns. This approach offers a novel way to quantify the progression of alpha peak emergence in individuals where the peak has not yet reached maturity. Visualizations of the recovered covariance structure can be found in Section A.3.3.

Figures 2.4 and 2.5 clearly depict the differences between traditional clustering models, like a k-means model [Lloyd, 1982], and our proposed mixed membership model. Information criteria determined that 3 clusters were the optimal number of clusters for a k-means analysis, illustrating how the increased flexibility of mixed membership models can potentially represent data using fewer clusters than traditional clustering. In the k-means analysis, the

first and third clusters can be interpreted as distinct alpha peaks, while the second cluster can be interpreted as primarily $1/f$ aperiodic patterns. While the cluster means found in the k-means analysis are relatively interpretable, we loose the ability to quantify the ratio of periodic signals to aperiodic signals, which is of significant scientific interest. From a modeling standpoint, traditional clustering models assume that each observation comes from one of the three clusters, meaning that children who have a developing alpha peak are not represented by any of the three clusters. Alternatively, the mixed membership model assumes that each observation can be represented by a continuous mixture of periodic and aperiodic neural activity patterns, leading to a more natural sampling model that is supported by previous scientific literature.

2.2.4 A Case Study on Molecular Subtypes in Breast Cancer

As of 2015, breast cancer is the 29^{th} leading cause of death in the world, with an estimated 534,000 deaths per year [Wang et al., 2016]. While there has been a significant increase in the number of deaths between 2005 and 2015 (21.3%), the age-adjusted death rate has decreased by 6.8% mainly due to our increased understanding of how to treat breast cancer. In the last two decades, 5 molecular subtypes of breast cancer have been discovered; each with a different prognosis, risk factors, and treatment sensitivity [Prat et al., 2015]. Parker et al. [2009] discovered that the cancer subtype can be accurately classified by centroid-based prediction methods using gene expression data from 50 genes (known as PAM50). Following Xu et al. [2016], we will use the PAM50 dataset to fit a mixed membership model on patients with LumA, Basal, and Her2 cancer subtypes. When restricting the PAM50 dataset to these 3 cancer subtypes, we have 115 patients with breast cancer ($N = 115$), and have 50 genes of interest ($P = 50$). For this case study, we fit a 3 feature mixed membership model ($K = 3$) with 4 eigenvectors to approximate the covariance structure ($M = 4$).

From Figure 2.6, we can see that each feature corresponds to a specific breast cancer subtype. While the data is still separable by cancer subtype, we can see that there is

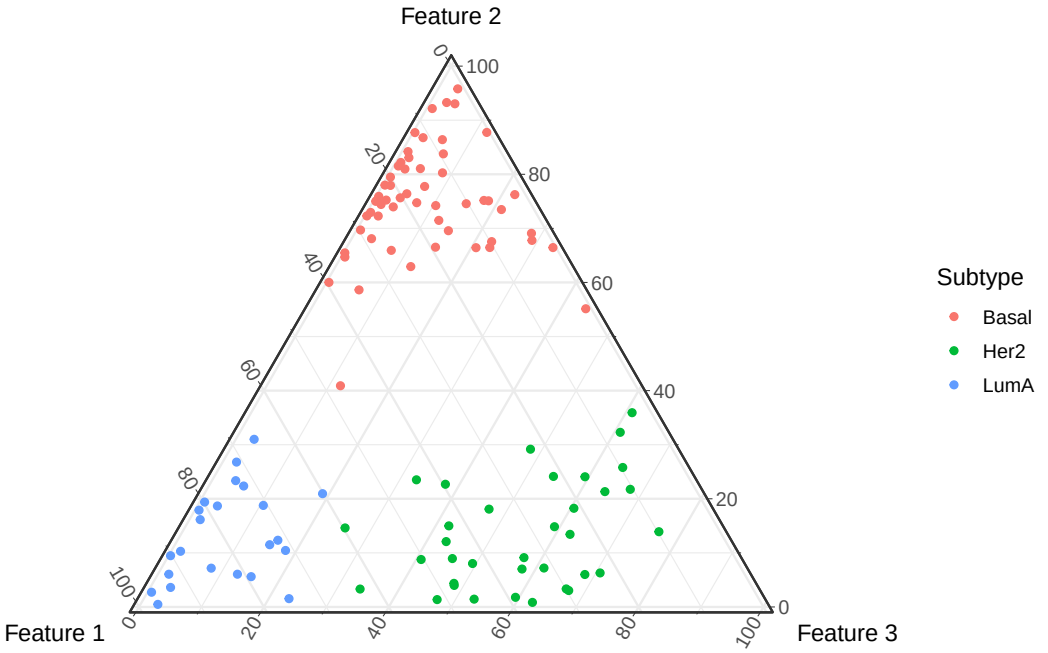


Figure 2.6: Estimated feature membership stratified by cancer subtype.

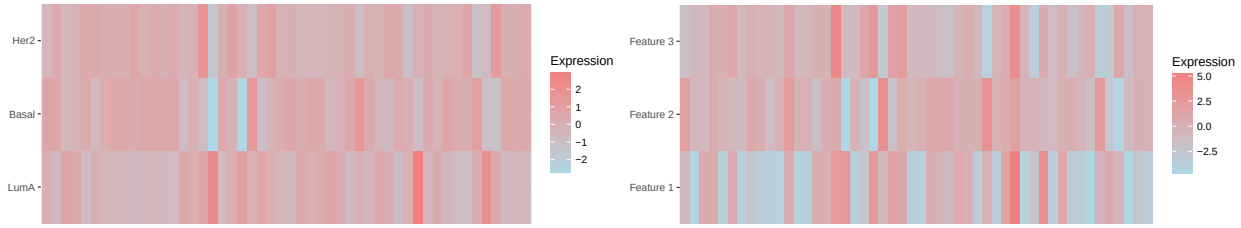


Figure 2.7: Cluster centroids for the model constructed by Parker et al. [2009](left) and the feature means for our mixed membership model (right).

considerable overlap between features 1 and 3 for some of the classified Her2 cancer subtype patients. The added flexibility of a mixed membership model is crucial as physicians can offer more personalized treatment plans for each patient. Since each cancer subtype has particular risk factors and treatment sensitivity, physicians that have patients that load heavily on two or more features will be aware that they should be monitoring all of the risk factors in the corresponding features. Treatment plans can also be customized to account for the treatment sensitivity of two or more cancer subtypes.

From Figure 2.7, we can see that there is more heterogeneity in the mean structure of

our model than in the cluster centroids found by Parker et al. [2009]. Conceptually, these mean structures represent two different ideas. In the centroid-based clustering, the data is assumed to only belong to only one of the cancer subtypes. Therefore, the mean structure in the centroid-based clustering model represents the average gene expression values for an individual belonging to that cluster. On the other hand, our model assumes that each individual can simultaneously belong to multiple cancer subtype groups. Therefore, the mean for a particular feature represents the average gene expression values for an individual that solely belongs to that feature. Thus the mean structure for the mixed membership model is more heterogeneous because they represent the "extreme" cases, or cases that are the most dissimilar from the other cancer subtypes. Overall, the added flexibility offered by our proposed mixed membership model allows clinicians to make more personalized treatment plans for patients, potentially leading to better clinical outcomes. Visualizations of the correlation structure of the genes can be found in Section A.3.4.

2.3 Discussion

This chapter introduces a flexible, yet scalable mixed membership model for continuous multivariate data. Mixed membership models, sometimes referred to as partial membership models or admixture models, can be thought of as a generalization of clustering where each observation can partially belong to multiple clusters. In Section 2.1, we derive our mixed membership model by extending the framework of finite mixture models. To have a scalable framework, we use a spectral decomposition of the K features, leveraging the multiplicative gamma shrinkage prior to ensure that the scaled eigenvectors are stochastically shrunk. To facilitate the use of simple sampling methods, we removed the constraint the the scaled eigenvectors, $\mathbf{\Phi}_m$, had to be mutually orthogonal. Within this context, we proved that the model had conditional weak posterior consistency of our mean and covariance structures, allowing us to facilitate relatively simple sampling schemes. Compared to previous works on mixed

membership models, our proposed model has an easily interpretable mean and covariance structure, allowing practitioners to easily and effectively communicate their findings from the model.

For the case of continuous data, the mixed membership model representation proposed by Heller et al. [2008] and discussed by [Ghahramani et al., 2014, Gruhl and Erosheva, 2014] is seemingly natural and generally applicable to data assumed to be sampled from a multivariate exponential family of distributions. We note that for the case of Normally distributed mixture components, this framework may prove to be too rigid to model more complex datasets. Subfigures 2.1(a) and 2.1(d) contain visualizations of the distribution of data generated from the Gaussian mixed membership model proposed by Heller et al. [2008]. Both scenarios have the same cluster-specific means, however they have different covariances associated with each cluster. In this setting, the generated data, follows the direction associated with the eigenvectors of the pure mixture covariance matrices, which can lead to unwieldy implied trajectories for plausible data realizations, e.g. data lying on a manifold (Subfigure 2.1(a)). Alternatively, Subfigures 2.1(b) and 2.1(c) show data generated from our proposed mixed membership model. In these two scenarios, each feature has the same mean and covariance as in Subfigure 2.1(a), but they have different cross-covariance matrices. In these cases, we can see a more natural trajectory of the data, where an assumption of an Euclidean metric seems appropriate. In less extreme settings, e.g. assuming spherical contours for the pure mixture components (Subfigure 2.1(d)), the representation in Equation 2.4 still leads to natural trajectories for the generated data, where the variance monotonically grows as an observation moves from the red to the blue cluster. Crucially, the proposed mixed membership model in Equation 2.7 recovers the model in Equation 2.4 as a special case, by setting the cross-covariance matrix equal to the zero matrix. The inclusion of cross-covariance elements, however, can express more flexible sampling scenarios as shown in Subfigures 2.1(e) and 2.1(f). Specifically, in Subfigure 2.1(e) we show the distribution of data where the symmetric part of the cross-covariance matrix has positive eigenvalues,

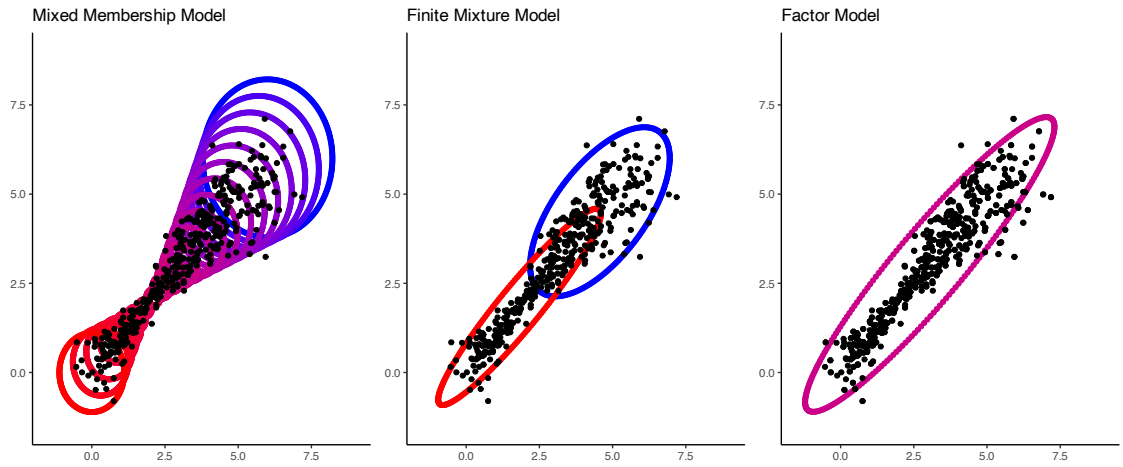


Figure 2.8: Comparative visualization of the differences between mixed membership models, finite mixture models, and factor models. Each of the models were fit on the same set of data, illustrated by the black dots.

while 2.1(f) shows the distribution of data where the symmetric part of the cross-covariance matrix has negative eigenvalues. Overall, the inclusion of explicit cross-covariance elements and the additive nature of our model leads to more natural trajectories of the implied model and an increased expressivity of the implied data generating process.

Mixed membership models for continuous data, as encoded in our representation in Equation 2.5, are related to latent factor models as they rely on similar additive structures. Nevertheless, mixed membership models obtain an alternative decomposition of the data variation, leading to a different interpretation of the model parameters. An illustration of the differences between Gaussian finite mixture models, factor models, and our proposed mixed membership model can be seen in Figure 2.8. The principal difference between factor models and mixed membership models being that mixed membership models restrict mixing between latent features to be defined on the standard simplex. As a consequence, the ensuing model defines margins which are not constrained to be elliptically symmetric, like in the case of factor models. An in-depth discussion of how factor analysis compares to mixed membership modeling is provided in Section A.4.

While flexibility is crucial to ensure that the model adequately fits the data, interpretabil-

ity of a model is paramount when communicating scientific findings. Under the Gaussian mixed membership model framework described by Heller et al. [2008], each observation is distributed normally, where the natural parameters of the distribution are a convex combination of the natural parameters associated with each cluster. The natural parameter representation can be particularly hard to work with because the mean structure is influenced by the covariance matrices of each individual cluster, leading to unnatural trajectories of the memberships, as seen in Subfigure 2.1(a). While the mean structure can be easily visualized in low dimensions, visualization of the mean structure can be particularly challenging in high dimensional model, leading to challenges in communicating the overall scientific findings. In our proposed model, each observation is also normally distributed, however, the mean parameter is simply a convex combination of the individual features' mean parameters. Similarly, the covariance of each observation can be written as a weighted sum of the covariance and cross-covariance functions of the features. Overall this simple structure allows practitioners to easily describe the distribution of observations that belong to multiple features, while maintaining a relatively flexible model.

From a practitioner's perspective, the main challenge to fitting a mixed membership model is choosing the number of features (K). As shown in Section 2.2.2, information criteria and simple heuristics such as the "elbow" method can be very informative in choosing the number of features in a mixed membership model. In traditional finite mixture models, Rousseau and Mengersen [2011] and Nguyen [2013] have shown that under certain conditions, an overfit mixture model would have a posterior distribution where only the "true" parameters would have positive weight, and the rest of the parameters would converge to zero. In both manuscripts, they assumed that the parameters were identifiable if we disregard the non-identifiability caused by the label-switching problem. However, as discussed in Section 2.1.4, the continuous nature of the allocation parameters create multiple non-identifiability problems. These types of non-identifiability problems make applying the results from Rousseau and Mengersen [2011] and Nguyen [2013] non-trivial. An alternative

non-parametric approach can be used by leveraging the Indian Buffet Process [Griffiths and Ghahramani, 2011], allowing for an infinite number of potential clusters. However, implementing and conducting inference across changing-dimensions is a non-trivial task and still an active area of research. Other than the number of features, users also are tasked with choosing the number of eigenvectors (M) to use in the mixed membership model. The number of eigenvectors controls how accurate we can approximate the covariance structure of the model. If $M = KP$, then we have a fully saturated model and we can have an exact representation of the covariance structure. In large models, setting $M = KP$ may be computationally intractable, which means that we will have to approximated the covariance structure by using a low-rank approximation of the covariance structure. Thus the choice of M is a user-defined choice that primarily depends on the computational budget afforded to fitting the model.

Due to the added flexibility of mixed membership model in general, we are often faced with multiple modal posterior distributions. Due to the potential multiple modal nature of the posterior distribution, we implemented tempered transitions (Section A.2.3) to ensure adequate exploration of the posterior distribution by allowing the chain to traverse areas of low posterior probability. While tempered transitions theoretically allow you to move between modes of the posterior distribution, they can be computationally expensive and sometimes tricky to tune. To limit the computational burden we suggest a mixture of tempered transitions and untempered transitions when performing MCMC. To speed up convergence of the Markov chain, we pick an informative starting position for our parameters using the multiple start algorithm (Algorithm 1) . The added flexibility of the allocation parameters also causes the *rescaling problem* described in Section 2.1.4. In order to make interpretation easier for practitioners, we recommend using the membership rescale algorithm (Algorithm 2). In the two feature case, this algorithm ensures that at least one observation belongs entirely to one feature. In the case where we have more than two features, the algorithm can be reformulated as an optimization problem. However, in practice, we found

that the membership rescale algorithm is seldomly needed when we have 3 or more features. An R package for our proposed Gaussian mixed membership model is available for download at `https://github.com/ndmarco/BayesFMMM`.

CHAPTER 3

Functional Mixed Membership Models

Functional data analysis (FDA) is concerned with the statistical analysis of realizations of a random function. In the functional clustering framework, the random function is often conceptualized as a mixture of stochastic processes, so that each realization is assumed to be sampled from one of $K < \infty$ cluster-specific sub-processes. The literature on functional clustering is well established and several analytical strategies have been proposed to handle different sampling designs and to ensure increasing levels of model flexibility. In the setting of sparsely observed functional data, James and Sugar [2003] proposed a mixed effects modeling framework for efficient dimension reduction after projection of the observed data onto a basis system characterized by natural cubic splines. Alternatively, Chiou and Li [2007] made use of functional principal component analysis to help reduce dimensionality by inferring cluster specific means and eigenfuctions. Jacques and Preda [2014] proposed a similar strategy for clustering multivariate functional data. Petrone et al. [2009] extended the previous work on functional mixture models by allowing hybrid species, where subintervals of the domain within each functional observation can belong to a different cluster. This type of clustering can be thought of as local clustering, where the cluster an observation belongs to changes depending on where you are in the domain of the function. Alternatively, mixed membership models can be thought of as a generalization of global clustering, where the membership allocation parameters do not change with respect to the domain of the random function.

In this chapter, we introduce the concept of mixed membership functions to the broader field of functional data analysis. Assuming a known number of latent functional features,

we show how a functional mixed membership process is naturally defined through a simple extension of finite Gaussian process (GP) mixture models, by allowing mixing proportions to be indexed at the level of individual sample paths. Some sophistication is introduced through the multivariate treatment of the underlying functional features to ensure sampling models of adequate expressivity. Specifically, we represent a family of multivariate GPs through the multivariate Karhunen-Loève construction of Happ and Greven [2018]. Using this idea, we jointly model the mean and covariance structures of the underlying functional features, and derive a sampling model through the ensuing finite dimensional marginal distributions. Since naïve GP mixture models assume that observations can only belong to one cluster, they only require the univariate treatment of the underlying GPs. However, when sample paths are allowed partial membership in multiple clusters, we need to either assume independence between clusters or estimate the cross-covariance functions. Since the assumption of independence does not often hold in practice, we propose a model that allows us to jointly model the covariance and cross-covariance functions through the eigenfunctions of the multivariate covariance operator. Typically, this representation would require sampling on a constrained space to ensure that the eigenfunctions are mutually orthogonal. However, in this context, we are able to relax these constraints and keep many of the desirable theoretical properties of our model by extending the multiplicative gamma process shrinkage prior proposed by Bhattacharya and Dunson [2011].

The remainder of the chapter is organized as follows. Section 3.1 formalizes the concept of functional mixed membership as a natural extension of functional clustering. Section 3.2 discusses Bayesian inference through posterior simulation, and establishes conditions for weak posterior consistency. Section 3.3 carries out two simulation studies to assess operating characteristics on engineered data, and illustrates the application of the proposed model to a case study involving functional brain imaging through electroencephalography (EEG). Finally, we conclude our exposition with a critical discussion of methodological challenges and related literature in Section 3.4.

3.1 Functional Mixed Membership

Functional data analysis (FDA) focuses on methods to analyze the sample paths of an underlying continuous stochastic process $f : \mathcal{T} \to \mathbb{R}$, where $\mathcal{T}$ is a compact subset of $\mathbb{R}^d$. In the iid setting, a typical characterization of these stochastic processes relies on the definition of the mean function, $\mu(t) = \mathbb{E}\left(f(t)\right)$, and the covariance function, $C(s,t) = \mathrm{Cov}\left(f(s), f(t)\right)$, where $t, s \in \mathcal{T}$. In this paper, we focus on jointly modeling K underlying stochastic processes, where each stochastic process represents one cluster. In this section, we will proceed under the assumption that the number of clusters, K, is known *a priori*. While K is fixed for the technical developments of the paper, Section 3.3.2 discusses the use of various information criteria for selection of the optimal number of features.

Let $f^{(k)} : \mathcal{T} \to \mathbb{R}$ denote the underlying stochastic process associated with the k^{th} cluster. Let $\mu^{(k)}$ and $C^{(k)}$ be, respectively, the mean and covariance function corresponding to the k^{th} stochastic process. To ensure desirable properties such as decompositions, FDA often makes the assumption that the random functions are elements in a Hilbert space. Specifically, we will assume $f^{(k)} \in L^2(\mathcal{T})$ $\left(\int_{\mathcal{T}} |f^{(k)}(t)|^2 \mathrm{d}t < \infty\right)$. Within this context, a typical model-based representation of functional clustering assumes that each underlying process is a latent Gaussian process, $f^{(k)} \sim \mathcal{GP}\left(\mu^{(k)}, C^{(k)}\right)$, with sample paths f_i $(i = 1, 2, \ldots, N)$, assumed to be independently drawn from a finite GP mixture

$$P\left(f_i \mid \rho^{(1:K)}, \mu^{(1:K)}, C^{(1:K)}\right) = \sum_{k=1}^{K} \rho^{(k)}\, \mathcal{GP}\left(f_i \mid \mu^{(k)}, C^{(k)}\right);$$

where $\rho^{(k)}$ is the mixing proportion (fraction of sample paths) for component k, such that $\sum_{k=1}^{K} \rho^{(k)} = 1$. Equivalently, by introducing latent variables $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$, such that $\boldsymbol{\pi}_i \sim_{iid} \mathrm{Multinomial}(1;\, \rho^{(1)}, \ldots, \rho^{(K)})$, it is easy to show that,

$$f_i \mid \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N, \mu^{(1:K)}, C^{(1:K)} \sim \mathcal{GP}\left(\sum_{k=1}^{K} \pi_{ik}\mu^{(k)}, \sum_{k=1}^{K} \pi_{ik} C^{(k)}\right). \tag{3.1}$$

Mixed membership is naturally defined by generalizing the finite mixture model to allow each sample path to belong to a finite positive number of mixture components, which we call *functional features* [Heller et al., 2008, Broderick et al., 2013]. By introducing continuous latent variables $\mathbf{z}_i = (Z_{i1}, \ldots, Z_{iK})'$, where $Z_{ik} \in (0, 1)$ and $\sum_{k=1}^{K} Z_{ik} = 1$, the general form of the proposed mixed membership model follows the following convex combination:

$$f_i \mid \mathbf{z}_1, \ldots, \mathbf{z}_N = \sum_{k=1}^{K} Z_{ik} f^{(k)}. \tag{3.2}$$

In equation 3.1, since each observation only belongs to one cluster, the Gaussian processes, $f^{(k)}$, are most commonly assumed to be mutually independent. However, assuming that the Gaussian processes, $f^{(k)}$, are independent in equation 3.2 leads to strong assumptions on the data generating process. The same assumption is, however, too restrictive for the model in equation 3.2. Thus, we will let $C^{(k,k')}$ denote the cross-covariance function between $f^{(k)}$ and $f^{(k')}$, and denote with $\boldsymbol{\mathcal{C}}$ the collection of covariance and cross-covariance functions. Therefore, the sampling model for the proposed mixed membership scheme can be written as

$$f_i \mid \mathbf{z}_1, \ldots, \mathbf{z}_N, \mu^{(1:K)}, \boldsymbol{\mathcal{C}} \sim \mathcal{GP}\left(\sum_{k=1}^{K} Z_{ik}\mu^{(k)}, \sum_{k=1}^{K} Z_{ik}^2 C^{(k)} + \sum_{k=1}^{K}\sum_{k'\neq k} Z_{ik}Z_{ik'}C^{(k,k')}\right). \tag{3.3}$$

Given a finite evaluation grid $\mathbf{t}_i = (t_{i1}, \ldots, t_{in_i})'$, the process' finite dimensional marginal distribution can be characterized such that

$$f_i(\mathbf{t}_i) \mid \mathbf{z}_1, \ldots, \mathbf{z}_N, \mu^{(1:K)}, \boldsymbol{\mathcal{C}} \sim \mathcal{N}\left(\sum_{k=1}^{K} Z_{ik}\mu^{(k)}(\mathbf{t}_i), C_i(\mathbf{t}_i, \mathbf{t}_i)\right), \tag{3.4}$$

where $C_i(\mathbf{t}_i, \mathbf{t}_i) = \sum_{k=1}^{K} Z_{ik}^2 C^{(k)}(\mathbf{t}_i, \mathbf{t}_i) + \sum_{k=1}^{K}\sum_{k'\neq k} Z_{ik}Z_{ik'}C^{(k,k')}(\mathbf{t}_i, \mathbf{t}_i)$. Since we do not assume that the functional features are independent, a concise characterization of the K stochastic processes is needed in order to ensure that our model is scalable and computationally tractable. In Section 3.1.1, we review the multivariate Karhunen-Loève theorem

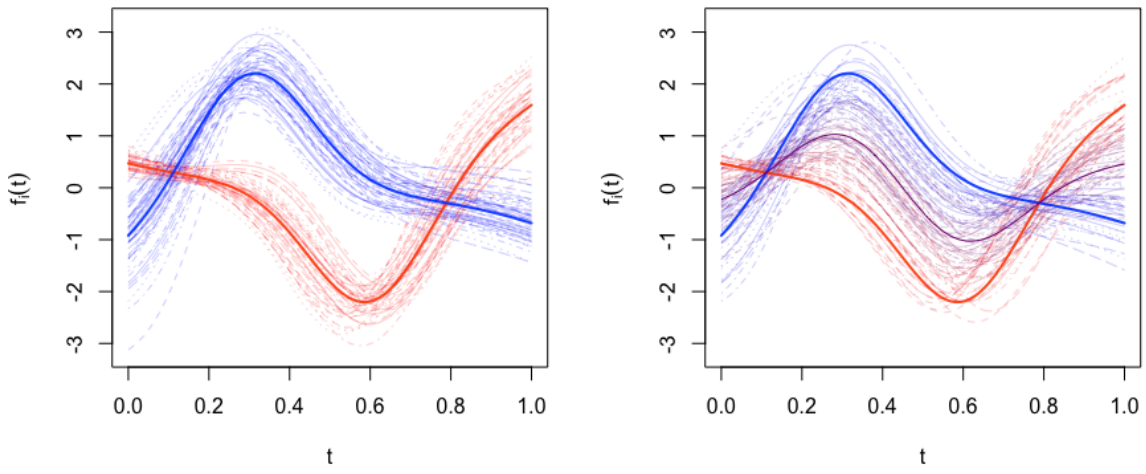


Figure 3.1: Generative model illustration. (Left panel) Data generated under the functional clustering framework. (Right panel) Data generated under a mixed membership framework.

[Happ and Greven, 2018], which will provide a joint characterization of the latent functional features, $\left(f^{(1)}, f^{(2)}, \ldots, f^{(K)}\right)$. Using this joint characterization, we are able to specify a scalable sampling model for the finite-dimensional marginal distribution found in equation 3.4, which is described in full detail in Section 3.1.2.

3.1.1 Multivariate Karhunen-Loève Characterization

To aid computation, we will make the assumption that $f^{(k)}$ is a smooth function and is in the P-dimensional subspace, $\boldsymbol{\mathcal{S}} \subset L^2(\mathcal{T})$, spanned by a set of linearly independent square-integrable basis functions, $\{b_1, \ldots, b_P\}$ $(b_p : \mathcal{T} \to \mathbb{R})$. While the choice of basis functions are user-defined, in practice B-splines (or the tensor product of B-splines for $\mathcal{T} \subset \mathbb{R}^d$ for $d \geq 2$) are a common choice due to their flexibility. Alternative basis systems can be selected in relation to application-specific considerations.

The multivariate Karhunen-Loève (KL) theorem, proposed by Ramsay and Silverman [2005b], can be used to jointly decompose our K stochastic processes. In order for our

model to be able to handle higher dimensional functional data, such as images, we will use the extension of the multivariate KL theorem for different dimensional domains proposed by Happ and Greven [2018]. While they state the theorem in full generality, we will only be considering the case when $f^{(k)} \in \boldsymbol{\mathcal{S}}$. In this section, we will show that the multivariate KL theorem for different dimensional domains still holds under our assumption that $f^{(k)} \in \boldsymbol{\mathcal{S}}$, given that the conditions in lemma 4 are satisfied.

We will start by defining the multivariate function $\mathbf{f}(\mathbf{t})$ in the following way:

$$\mathbf{f}(\mathbf{t}) := \left(f^{(1)}\left(t^{(1)}\right), f^{(2)}\left(t^{(2)}\right), \ldots, f^{(K)}\left(t^{(K)}\right)\right), \tag{3.5}$$

where $\mathbf{t}$ is a K-tuple such that $\mathbf{t} = \left(t^{(1)}, t^{(2)}, \ldots, t^{(K)}\right)$, where $t^{(1)}, t^{(2)}, \ldots, t^{(K)} \in \mathcal{T}$. This construction allows for the joint representation of K stochastic processes, each at different points in their domain. Since $f^{(k)} \in \boldsymbol{\mathcal{S}}$, we have that $\mathbf{f} \in \boldsymbol{\mathcal{H}} := \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{S}} \times \cdots \times \boldsymbol{\mathcal{S}} := \boldsymbol{\mathcal{S}}^K$. We define the corresponding mean of $\mathbf{f}(\mathbf{t})$, such that

$$\boldsymbol{\mu}(\mathbf{t}) := \left(\mu^{(1)}\left(t^{(1)}\right), \mu^{(2)}\left(t^{(2)}\right), \ldots, \mu^{(K)}\left(t^{(K)}\right)\right),$$

where $\boldsymbol{\mu}(\mathbf{t}) \in \boldsymbol{\mathcal{H}}$, and the mean-centered cross-covariance functions as

$$C^{(k,k')}(s,t) := \mathrm{Cov}\left(f^{(k)}(s) - \mu^{(k)}(s), f^{(k')}(t) - \mu^{(k')}(t)\right),$$

for $s, t \in \mathcal{T}$.

Lemma 3. *$\boldsymbol{\mathcal{S}}$ is a closed linear subspace of $L^2(\mathcal{T})$.*

Proofs for all results are given in Appendix Chapter B. From lemma 3, since $\boldsymbol{\mathcal{S}}$ is a closed linear subspace of $L^2(\mathcal{T})$, we have that $\boldsymbol{\mathcal{S}}$ is a Hilbert space with respect to the inner

product $\langle f, g\rangle_{\boldsymbol{\mathcal{S}}} = \int_{\mathcal{T}} f(t)g(t)\mathrm{d}t$ where $f, g \in \boldsymbol{\mathcal{S}}$. By defining the inner product on $\boldsymbol{\mathcal{H}}$ as

$$\langle \mathbf{f}, \mathbf{g}\rangle_{\boldsymbol{\mathcal{H}}} := \sum_{k=1}^{K} \int_{\mathcal{T}} f^{(k)}\left(t^{(k)}\right) g^{(k)}\left(t^{(k)}\right) \mathrm{d}t^{(k)}, \tag{3.6}$$

where $\mathbf{f}, \mathbf{g} \in \boldsymbol{\mathcal{H}}$, we have that $\boldsymbol{\mathcal{H}}$ is the direct sum of Hilbert spaces. Since $\boldsymbol{\mathcal{H}}$ is the direct sum of Hilbert spaces, $\boldsymbol{\mathcal{H}}$ is a Hilbert space [Reed and Simon, 1972]. Thus we have constructed a subspace $\boldsymbol{\mathcal{H}}$ using our assumption that $f^{(k)} \in \boldsymbol{\mathcal{S}}$, which satisfies proposition 1 in Happ and Greven [2018]. Letting $\mathbf{f} \in \boldsymbol{\mathcal{H}}$, we can define the covariance operator $\boldsymbol{\mathcal{K}} : \boldsymbol{\mathcal{H}} \to \boldsymbol{\mathcal{H}}$ element-wise in the following way:

$$\left(\boldsymbol{\mathcal{K}}\mathbf{f}\right)^{(k)}(\mathbf{t}) := \left\langle C^{(\cdot,k)}\left(\cdot, t^{(k)}\right), \mathbf{f}\right\rangle_{\boldsymbol{\mathcal{H}}} = \sum_{k'=1}^{K} \int_{\mathcal{T}} C^{(k',k)}\left(s, t^{(k)}\right) f^{(k')}(s)\mathrm{d}s. \tag{3.7}$$

Since we made the assumption $f^{(k)} \in \boldsymbol{\mathcal{S}}$, we can simplify equation 3.7. Since $\boldsymbol{\mathcal{S}}$ is the span of the basis functions, we have that $f^{(k)}(t) - \mu^{(k)}(t) = \sum_{p=1}^{P} \theta_{(k,p)} b_p(t) = \mathbf{B}'(t)\boldsymbol{\theta}_k$, for some $\boldsymbol{\theta}_k \in \mathbb{R}^P$ and $\mathbf{B}'(t) = [b_1(t) \cdots b_P(t)]$. Thus we can see that

$$C^{(k,k')}(s,t) = \mathrm{Cov}\left(\mathbf{B}'(s)\boldsymbol{\theta}_k, \mathbf{B}'(t)\boldsymbol{\theta}_{k'}\right) = \mathbf{B}'(s)\mathrm{Cov}\left(\boldsymbol{\theta}_k, \boldsymbol{\theta}_{k'}\right)\mathbf{B}(t), \tag{3.8}$$

and we can rewrite equation 3.7 as $\left(\boldsymbol{\mathcal{K}}\mathbf{f}\right)^{(k)}(\mathbf{t}) = \sum_{k'=1}^{K} \int_{\mathcal{T}} \mathbf{B}'(s)\mathrm{Cov}\left(\boldsymbol{\theta}_{k'}, \boldsymbol{\theta}_k\right)\mathbf{B}\left(t^{(k)}\right) f^{(k')}(s)\mathrm{d}s$, for some $\mathbf{f} \in \mathcal{H}$. The following lemma establishes conditions under which $\boldsymbol{\mathcal{K}}$ is a bounded and compact operator, which is a necessary condition for the multivariate KL decomposition to exist.

Lemma 4. *$\boldsymbol{\mathcal{K}}$ is a bounded and compact operator if we have that*

1. *the basis functions, $b_1, \ldots, b_P$, are uniformly continuous*
2. *there exists $M \in \mathbb{R}$ such that $\left|Cov\left(\theta_{(k,p)}, \theta_{(k',p')}\right)\right| \leq M$.*

Assuming the conditions specified in lemma 4 hold, by the Hilbert-Schmidt theorem,

since $\boldsymbol{\mathcal{K}}$ is a bounded, compact, and self-adjoint operator, we know that there exists real eigenvalues $\lambda_1, \ldots, \lambda_{KP}$ and a complete set of eigenfunctions $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_{KP} \in \boldsymbol{\mathcal{H}}$ such that $\boldsymbol{\mathcal{K}}\boldsymbol{\Psi}_p = \lambda_p \boldsymbol{\Psi}_p$, for $p = 1, \ldots, KP$. Since $\boldsymbol{\mathcal{K}}$ is a non-negative operator, we know that $\lambda_p \geq 0$ for $p = 1, \ldots, KP$ (Happ and Greven [2018], proposition 2). From theorem VI.17 of Reed and Simon [1972], we have that the positive, bounded, self-adjoint, and compact operator $\boldsymbol{\mathcal{K}}$ can be written as $\boldsymbol{\mathcal{K}}\mathbf{f} = \sum_{p=1}^{KP} \lambda_p \langle \boldsymbol{\Psi}_p, \mathbf{f} \rangle_{\boldsymbol{\mathcal{H}}} \boldsymbol{\Psi}_p$. Thus from equation 3.7, we have that

$$\sum_{k'=1}^{K} \int_{\mathcal{T}} C^{(k',k)}\left(s, t^{(k)}\right) f^{(k')}(s) \mathrm{d}s = \sum_{k'=1}^{K} \int_{\mathcal{T}} \left(\sum_{p=1}^{KP} \lambda_p \Psi_p^{(k')}(s) \Psi_p^{(k)}\left(t^{(k)}\right) \right) f^{(k')}(s) \mathrm{d}s,$$

where $\Psi_p^{(k)}(t^{(k)})$ is the k^{th} element of $\boldsymbol{\Psi}_p(\mathbf{t})$. Thus we can see that the covariance kernel can be written as the finite sum of eigenfunctions and eigenvalues,

$$C^{(k,k')}(s,t) = \sum_{p=1}^{KP} \lambda_p \Psi_p^{(k)}(s) \Psi_p^{(k')}(t). \tag{3.9}$$

Since we are working under the assumption that the random function, $\mathbf{f} \in \boldsymbol{\mathcal{H}}$, we can use a modified version of the Multivariate KL theorem. Considering that $\{\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_{KP}\}$ form a complete basis for $\boldsymbol{\mathcal{H}}$ and $\mathbf{f}(\mathbf{t}) - \boldsymbol{\mu}(\mathbf{t}) \in \boldsymbol{\mathcal{H}}$, we have

$$\mathbf{f}(\mathbf{t}) - \boldsymbol{\mu}(\mathbf{t}) \;=\; \mathbf{P}_{\boldsymbol{\mathcal{H}}} \circ (\mathbf{f} - \boldsymbol{\mu})(\mathbf{t}) \;=\; \sum_{p=1}^{KP} \langle \boldsymbol{\Psi}_p, \mathbf{f} - \boldsymbol{\mu} \rangle_{\boldsymbol{\mathcal{H}}} \boldsymbol{\Psi}_p(\mathbf{t}),$$

where $\mathbf{P}_{\boldsymbol{\mathcal{H}}}$ is the projection operator onto $\boldsymbol{\mathcal{H}}$. Letting $\rho_p = \langle \boldsymbol{\Psi}_p, \mathbf{f} - \boldsymbol{\mu} \rangle_{\boldsymbol{\mathcal{H}}}$, we have that $\mathbf{f}(\mathbf{t}) - \boldsymbol{\mu}(\mathbf{t}) = \sum_{p=1}^{KP} \rho_p \boldsymbol{\Psi}_p(\mathbf{t})$, where $\mathbb{E}(\rho_p) = 0$ and $\mathrm{Cov}(\rho_p, \rho_{p'}) = \lambda_p \delta_{pp'}$ (Happ and Greven [2018], proposition 4).

Since $\boldsymbol{\Psi}_p \in \boldsymbol{\mathcal{H}}$ and $\boldsymbol{\mu} \in \boldsymbol{\mathcal{H}}$, there exists $\boldsymbol{\nu}_k \in \mathbb{R}^P$ and $\boldsymbol{\phi}_{kp} \in \mathbb{R}^P$ such that $\boldsymbol{\mu}^{(k)}(\mathbf{t}) = \boldsymbol{\nu}_k' \mathbf{B}\left(t^{(k)}\right)$ and $\sqrt{\lambda_p} \boldsymbol{\Psi}_p^{(k)}(\mathbf{t}) = \boldsymbol{\phi}_{kp}' \mathbf{B}\left(t^{(k)}\right) := \boldsymbol{\Phi}_p^{(k)}(\mathbf{t})$. These scaled eigenfunctions, $\boldsymbol{\Phi}_p^{(k)}(\mathbf{t})$, are used over $\boldsymbol{\Psi}_p^{(k)}(\mathbf{t})$ because they fully specify the covariance structure of the latent features, as described in 3.1.2. From a modeling prospective, the scaled eigenfunction param-

eterization is advantageous as it admits a prior model based on the multiplicative gamma process shrinkage prior proposed by Bhattacharya and Dunson [2011], allowing for adaptive regularized estimation [Shamshoian et al., 2022]. Thus we have that

$$\mathbf{f}^{(k)}(\mathbf{t}) = \boldsymbol{\mu}^{(k)}(\mathbf{t}) + \sum_{p=1}^{KP} \chi_p \boldsymbol{\Phi}_p^{(k)}(\mathbf{t}) = \boldsymbol{\nu}_k' \mathbf{B}\left(t^{(k)}\right) + \sum_{p=1}^{KP} \chi_p \boldsymbol{\phi}_{kp}' \mathbf{B}\left(t^{(k)}\right), \tag{3.10}$$

where $\chi_p = \left\langle (\lambda_p)^{-1/2} \boldsymbol{\Psi}_p, \mathbf{f} - \boldsymbol{\mu} \right\rangle_{\boldsymbol{\mathcal{H}}}$, $\mathbb{E}(\chi_p) = 0$, and $\text{Cov}(\chi_p, \chi_{p'}) = \delta_{pp'}$. Equation 3.10 gives us a way to jointly decompose any realization of the K stochastic processes, $\mathbf{f} \in \boldsymbol{\mathcal{H}}$, as a finite weighted sum of basis functions.

Corollary 1. *If $\chi_p \sim_{iid} \mathcal{N}(0,1)$, then the random function $\mathbf{f}(\mathbf{t})$ follows a multivariate $\mathcal{GP}$ with means $\boldsymbol{\mu}^{(k)}(\mathbf{t}) = \boldsymbol{\nu}_k' \mathbf{B}\left(t^{(k)}\right)$, and cross-covariance functions $C^{(k,k')}(s,t) = \mathbf{B}'(s) \sum_{p=1}^{KP} (\boldsymbol{\phi}_{kp} \boldsymbol{\phi}_{k'p}') \mathbf{B}(t)$.*

3.1.2 Functional Mixed Membership Process

In this section, we will describe a Bayesian additive model that allows for the constructive representation of mixed memberships for functional data. Our model allows for direct inference on the mean and covariance structures of the K stochastic processes, which are often of scientific interest. To aid computational tractability, we will use the joint decomposition described in Section 3.1.1. In this section, we will make the assumption that $f^{(k)} \in \boldsymbol{\mathcal{S}}$, and that the conditions from lemma 4 hold.

We aim to model the mixed membership of N observed sample paths, $\{\mathbf{Y}_i(.)\}_{i=1}^N$, to K latent functional features $\mathbf{f} = \left(f^{(1)}, f^{(2)}, \ldots, f^{(K)}\right)$. Assuming each path is observed over n_i evaluation points $\mathbf{t}_i = [t_{i1} \cdots t_{in_i}]'$, without loss of generality we define a sampling model for the finite dimensional marginals of $\mathbf{Y}_i(\mathbf{t}_i)$. To allow for path-specific partial membership, we extend the finite mixture model in equation 3.1 and introduce path-specific mixing proportions $Z_{ik} \in (0,1)$, such that $\sum_{k=1}^K Z_{ik} = 1$, for $i = 1, 2, \ldots, N$.

Let $\mathbf{S}(\mathbf{t}_i) = [\mathbf{B}(t_1) \cdots \mathbf{B}(t_{n_i})] \in \mathbb{R}^{P \times n_i}$, $\chi_{im} \sim \mathcal{N}(0,1)$, and $\mathbf{\Theta}$ denote a collection of model parameters. Let $M \leq KP$ be the number of eigenfunctions used to approximate the covariance structure of the K stochastic processes. Using the decomposition of $f^{(k)}(t)$ in equation 3.10 and assuming a normal distribution on $\mathbf{Y}_i(\mathbf{t}_i)$, we obtain:

$$\mathbf{Y}_i(\mathbf{t}_i)|\mathbf{\Theta} \sim \mathcal{N}\left\{\sum_{k=1}^{K} Z_{ik}\left(\mathbf{S}'(\mathbf{t}_i)\boldsymbol{\nu}_k + \sum_{m=1}^{M} \chi_{im}\mathbf{S}'(\mathbf{t}_i)\boldsymbol{\phi}_{km}\right),\ \sigma^2\mathbf{I}_{n_i}\right\}. \tag{3.11}$$

If we integrate out the latent χ_{im} variables, we obtain a more transparent form for the proposed functional mixed membership process. Specifically, we have

$$\mathbf{Y}_i(\mathbf{t}_i)|\mathbf{\Theta}_{-\chi} \sim \mathcal{N}\left\{\sum_{k=1}^{K} Z_{ik}\mathbf{S}'(\mathbf{t}_i)\boldsymbol{\nu}_k,\ \mathbf{V}_i + \sigma^2\mathbf{I}_{n_i}\right\}, \tag{3.12}$$

where $\mathbf{\Theta}_{-\chi}$ is the collection of our model parameters excluding the χ_{im} variables, and $\mathbf{V}_i = \sum_{k=1}^{K}\sum_{k'=1}^{K} Z_{ik}Z_{ik'}\left\{\mathbf{S}'(\mathbf{t}_i)\sum_{m=1}^{M}\left(\boldsymbol{\phi}_{km}\boldsymbol{\phi}'_{k'm}\right)\mathbf{S}(\mathbf{t}_i)\right\}$. Thus, for a sample path, we have that the mixed membership mean is a convex combination of the functional feature means, and the mixed membership covariance, is a weighted sum of the covariance and cross-covariance functions between different functional features, following from the multivariate KL characterization in Section 3.1.1. Furthermore, from equation 3.9 it is easy to show that, for large enough M, we have $\mathbf{S}'(\mathbf{t}_i)\sum_{m=1}^{M}\left(\boldsymbol{\phi}_{kp}\boldsymbol{\phi}'_{k'p}\right)\mathbf{S}(\mathbf{t}_i) \approx C^{(k,k')}(\mathbf{t}_i, \mathbf{t}_i)$,with equality when $M = KP$.

Mixed membership models can be thought of as a generalization of clustering. As such, these stochastic schemes are characterized by an inherent lack of likelihood identifiability. A typical source of non-indentifiability is the common *label switching* problem. To deal with the *label switching* problem, a relabelling algorithm can be derived for this model directly from the work of Stephens [2000]. A second source of non-identifiabilty stems from allowing Z_{ik} to be continuous random variables. Specifically, consider a model with 2 features, and let $\mathbf{\Theta}_0$ be the set of "true" parameters. Let $Z^*_{i1} = 0.5(Z_{i1})_0$ and $Z^*_{i2} = (Z_{i2})_0 + 0.5(Z_{i1})_0$. If we

let $\boldsymbol{\nu}_1^* = 2(\boldsymbol{\nu}_1)_0 - (\boldsymbol{\nu}_2)_0$, $\boldsymbol{\nu}_2^* = (\boldsymbol{\nu}_2)_0$, $\boldsymbol{\phi}_{1m}^* = 2(\boldsymbol{\phi}_{1m})_0 - (\boldsymbol{\phi}_{2m})_0$, $\boldsymbol{\phi}_{2m}^* = (\boldsymbol{\phi}_{2m})_0$, $\chi_{im}^* = (\chi_{im})_0$, and $(\sigma^2)^* = \sigma_0^2$, we have that $P(Y_i(t)|\boldsymbol{\Theta}_0) = P(Y_i(t)|\boldsymbol{\Theta}^*)$ (equation 3.11). Thus we can see that our model is not identifiable, and we will refer to this problem as the *rescaling problem*. To address the *rescaling problem*, we developed the Membership Rescale Algorithm (Algorithm 4). This algorithm will rescale the allocation parameters so that at least one observation will completely belong to each of the two functional features. Section B.3.4 also briefly reviews the work of Chen et al. [2022] which focuses on identifiability when we have more than two functional features. A third non-identifiability problem may arise numerically as a form of concurvity, i.e. when $\boldsymbol{\nu}_{k'} \propto \boldsymbol{\phi}_{km}$ in equation 3.11. Typically, overestimation of the magnitude of $\boldsymbol{\phi}_{km}$, may result in small variance estimates for the allocation parameters (smaller credible intervals). This phenomenon does need to be considered when studying how well we can recover the model parameters, as in Section 3.3.1, but it is typically of little practical relevance in applications.

3.1.3 Prior Distributions and Model Specification

The sampling model in Section 3.1.2, allows a practitioner to select how many eigenfunctions are to be used in the approximation of the covariance function. In the case where of $M = KP$, we have a fully saturated model and can represent any realization $\mathbf{f} \in \boldsymbol{\mathcal{H}}$. In equation 3.10, $\boldsymbol{\Phi}$ parameters are mutually orthogonal (where orthogonality is defined by the inner product defined in equation 3.6), and have a magnitude proportional to the square root of the corresponding eigenvalue, λ_p. Thus, a modified version of the multiplicative gamma process shrinkage prior proposed by Bhattacharya and Dunson [2011] will be used as our prior for $\boldsymbol{\phi}_{km}$. By using this prior, we promote shrinkage across the $\boldsymbol{\phi}_{km}$ coefficient vectors, with increasing prior shrinkage towards zero as m increases.

To facilitate MCMC sampling from the posterior target we will remove the assumption that $\boldsymbol{\Phi}$ parameters are mutually orthogonal. Even though $\boldsymbol{\Phi}$ parameters can no longer be thought of as scaled eigenfunctions, posterior inference can still be conducted on the

eigenfunctions by post-processing posterior samples of $\boldsymbol{\phi}_{km}$ (given that we can recover the true covariance operator). Specifically, given posterior samples from $\boldsymbol{\phi}_{km}$, we obtain posterior realizations for the covariance function, and then calculate the eigenpairs of the posterior samples of the covariance operator using the method described in Happ and Greven [2018]. Thus, letting ϕ_{kpm} be the p^{th} element of $\boldsymbol{\phi}_{km}$, we have

$$\phi_{kpm}|\gamma_{kpm}, \tilde{\tau}_{mk} \sim \mathcal{N}\left(0, \gamma_{kpm}^{-1}\tilde{\tau}_{mk}^{-1}\right), \quad \gamma_{kpm} \sim \Gamma\left(\nu_\gamma/2, \nu_\gamma/2\right), \quad \tilde{\tau}_{mk} = \prod_{n=1}^{m} \delta_{nk},$$

$$\delta_{1k}|a_{1k} \sim \Gamma(a_{1k}, 1), \quad \delta_{jk}|a_{2k} \sim \Gamma(a_{2k}, 1), \quad a_{1k} \sim \Gamma(\alpha_1, \beta_1), \quad a_{2k} \sim \Gamma(\alpha_2, \beta_2),$$

where $1 \leq k \leq K$, $1 \leq p \leq P$, $1 \leq m \leq M$, and $2 \leq j \leq M$. In order for us to promote shrinkage across the M matrices, we need that $\delta_{jk} > 1$. Thus letting $\alpha_2 > \beta_2$, we have that $\mathbb{E}(\delta_{jk}) > 1$, which will promote shrinkage. In this construction, we allow for different rates of shrinkage across different functional features, which is particularly important in cases where the covariance functions of different features have different magnitudes. In cases where the magnitudes of the covariance functions are different, we would expect the δ_{mk} to be relatively smaller in the k associated with the functional feature with a large covariance function.

To promote adaptive smoothing in the mean function, we will use a first order random walk penalty proposed by Lang and Brezger [2004]. The first order random walk penalty penalizes differences in adjacent B-spline coefficients. In the case where $\mathcal{T} \subset \mathbb{R}$, we have that $P(\boldsymbol{\nu}_k|\tau_k) \propto exp\left(-\frac{\tau_k}{2}\sum_{p=1}^{P-1}\left(\nu'_{pk} - \nu_{(p+1)k}\right)^2\right)$, for $k = 1, \ldots, K$, where $\tau_k \sim \Gamma(\alpha, \beta)$ and ν_{pk} is the p^{th} element of $\boldsymbol{\nu}_k$. Since we have that $Z_{ik} \in (0, 1)$ and $\sum_{k=1}^{K} Z_{ik} = 1$, it is natural to consider prior Dirichlet sampling for $\mathbf{z}_i = [Z_{i1} \cdots Z_{iK}]$. Therefore, following Heller et al. [2008], we have

$$\mathbf{z}_i \mid \boldsymbol{\pi}, \alpha_3 \sim_{iid} Dir(\alpha_3\boldsymbol{\pi}), \quad \boldsymbol{\pi} \sim Dir(\mathbf{c}), \quad \alpha_3 \sim Exp(b)$$

for $i = 1, \ldots, N$. Lastly, we will use a conjugate prior for our random error component of

the model, such that $\sigma^2 \sim IG(\alpha_0, \beta_0)$. While we relax the assumption of orthogonality, in Section B.3.1, we outline an alternative sampling scheme where we impose the condition that the $\mathbf{\Phi}$ parameters form orthogonal eigenfunctions using the work of Kowal et al. [2017].

3.2 Posterior Inference

Statistical inference is based on Markov chain Monte Carlo samples from the posterior distribution, by using the Metropolis-within-Gibbs algorithm. While the naïve sampling scheme is relatively simple, ensuring good exploration of the posterior target can be challenging due to the potentially multimodal nature of the posterior distribution. Specifically, some sensitivity of results to the starting values of the chain can be observed for some data. Section B.3.2 outlines an algorithm for the selection of informed starting values. Furthermore, to mitigate sensitivity to chain initialization, we also implemented a tempered transition scheme, which improves the mixing of the Markov chain by allowing for transitions between modal configuration of the target. Implementation details for the proposed tempered transition scheme are reported in Section B.3.3. Additional information on sampling schemes, as well as how to construct simultaneous confidence intervals can be found in Section B.4.

3.2.1 Weak Posterior Consistency

In the previous section we saw that the $\mathbf{\Phi}$ parameters are not assumed to be mutually orthogonal. By removing this constraint, we facilitate MCMC sampling from the posterior target and can perform inference on the eigenpairs of the covariance operator, as long as we can recover the covariance structure. Due to the complex identifiability issues mentioned in Section 3.1.2, establishing weak posterior consistency with unknown allocation parameters unattainable, even if we include the orthogonality constraint on the $\mathbf{\Phi}$ parameters. However, the model becomes identifiable when we condition on the allocation parameters. In this section, we show that we can achieve conditional weak posterior consis-

tency without enforcing the orthogonality constraint of the $\mathbf{\Phi}$ parameter. Therefore, we will show that conditional on the allocation parameters, relaxing the orthogonality constraint does not affect our ability to recover the mean and covariance structure of the K underlying stochastic processes from equation 3.12. Let $\mathbf{\Pi}$ be the prior distribution on $\boldsymbol{\omega} := \{\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_K, \mathbf{\Sigma}_{11}, \ldots, \mathbf{\Sigma}_{1K}, \ldots, \mathbf{\Sigma}_{KK}, \sigma^2\}$, where $\mathbf{\Sigma}_{kk'} := \sum_{p=1}^{KP} \left(\boldsymbol{\phi}_{kp}\boldsymbol{\phi}'_{k'p}\right)$. We will be proving weak posterior consistency with respect to the parameters $\mathbf{\Sigma}_{kk'}$ because the parameters $\boldsymbol{\phi}_{kp}$ are non-identifiable. Since the covariance and cross-covariance structure are completely specified by the $\mathbf{\Sigma}_{kk'}$ parameters, the lack of identifiability of the $\boldsymbol{\phi}_{kp}$ parameters bears no importance on inferential considerations. We will denote the true set of parameter values as $\boldsymbol{\omega}_0 = \{(\boldsymbol{\nu}_1)_0, \ldots, (\boldsymbol{\nu}_K)_0, (\mathbf{\Sigma}_{11})_0, \ldots, (\mathbf{\Sigma}_{1K})_0, \ldots, (\mathbf{\Sigma}_{KK})_0, \sigma_0^2\}$. In order to prove weak posterior consistency, will make the following assumptions:

Assumption 3. *The observed realizations $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ are observed on the same grid of $R > KP$ points in the domain, say $\{t_1, \ldots, t_R\}$.*

Assumption 4. *The variables Z_{ik} are known a-priori for $i = 1, \ldots, N$ and $k = 1, \ldots, K$.*

Assumption 5. *The true parameter modeling the random noise is positive ($\sigma_0^2 > 0$).*

In order to prove weak posterior consistency, we will first specify the following quantities related to the Kullback–Leibler (KL) divergence. Following the notation of Choi and Schervish [2007], we will define the following quantities

$$\Lambda_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) = \log\left(\frac{f_i(\mathbf{Y}_i; \boldsymbol{\omega}_0)}{f_i(\mathbf{Y}_i; \boldsymbol{\omega})}\right), \quad K_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) = \mathbb{E}_{\boldsymbol{\omega}_0}(\Lambda_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})), \quad V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) = \mathrm{Var}_{\boldsymbol{\omega}_0}(\Lambda_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})),$$

where $f_i(\mathbf{Y}_i; \boldsymbol{\omega}_0)$ is the likelihood under $\boldsymbol{\omega}_0$. To simplify the notation, we will define the

following two quantities

$$\boldsymbol{\mu}_i = \sum_{k=1}^{K} Z_{ik}\mathbf{S}'(\mathbf{t})\boldsymbol{\nu}_k,$$
$$\boldsymbol{\Sigma}_i = \sum_{k=1}^{K}\sum_{k'=1}^{K} Z_{ik}Z_{ik'}\left(\mathbf{S}'(\mathbf{t})\sum_{p=1}^{KP}\left(\boldsymbol{\phi}_{kp}\boldsymbol{\phi}'_{k'p}\right)\mathbf{S}(\mathbf{t})\right) + \sigma^2\mathbf{I}_R = \mathbf{U}'_i\mathbf{D}_i\mathbf{U}_i + \sigma^2\mathbf{I}_R,$$

where $\mathbf{U}'_i\mathbf{D}_i\mathbf{U}_i$ is the corresponding spectral decomposition. Let d_{il} be the l^{th} diagonal element of $\mathbf{D}_i$. Let $\boldsymbol{\Omega}_\epsilon(\boldsymbol{\omega}_0)$ be the set of parameters such that the KL divergence is less than some $\epsilon > 0$ ($\boldsymbol{\Omega}_\epsilon(\boldsymbol{\omega}_0) := \{\boldsymbol{\omega} : K_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) < \epsilon \text{ for all i}\}$). Let $a, b \in \mathbb{R}$ be such that $a > 1$ and $b > 0$, and define $\mathcal{B}(\boldsymbol{\omega}_0) := \left\{\boldsymbol{\omega} : \frac{1}{a}\left((d_{il})_0 + \sigma_0^2\right) \leq d_{il} + \sigma^2 \leq a\left((d_{il})_0 + \sigma_0^2\right), \|\left(\boldsymbol{\mu}_i\right)_0 - \boldsymbol{\mu}_i\| \leq b\right\}$.

Lemma 5. *Let $\mathcal{C}(\boldsymbol{\omega}_0, \epsilon) := \mathcal{B}(\boldsymbol{\omega}_0) \cap \boldsymbol{\Omega}_\epsilon(\boldsymbol{\omega}_0)$. Thus for $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$ and $\epsilon > 0$, there exists $a > 1$ and $b > 0$ such that (1) $\sum_{i=1}^{\infty} \frac{V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})}{i^2} < \infty$, for any $\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)$ and (2) $\boldsymbol{\Pi}\left(\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)\right) > 0$.*

Lemma 5 shows that the prior probability of our parameters being arbitrarily close (where the measure of closeness is defined by the KL divergence) to the true parameters is positive. Since $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ are not identically distributed, condition (1) in lemma 5 is need in order to prove lemma 6.

Lemma 6. *Under assumptions 3-5, the posterior distribution, $\boldsymbol{\Pi}_N(.|\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$, is weakly consistent at $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$.*

All proofs are provided in the supplemental appendix. Lemma 6 shows us that conditional on the allocation parameters, we are able to recover the covariance structure of the K stochastic processes. Thus, Relaxing the orthogonality constraint on the $\boldsymbol{\phi}_{km}$ parameters does not affect our ability to perform posterior inference on the main functions of scientific interest. Inference on the eigenstructure can still be performed by calculating the eigenvalues and eigenfunctions of the covariance operator using the MCMC samples of the $\boldsymbol{\phi}_{km}$ parameter. Finally, we point out that, in most cases, the parameters Z_{ik} are unknown. While

theoretical guarantees for consistent estimation of the latent mixed allocation parameters is still elusive, we provide some empirical evidence of convergence in Section 3.3.1.

3.3 Case Studies and Experiments on Simulated Data

3.3.1 Simulation Study 1

In this simulation study, we examine how well our model can recover the mean and covariance functions when we vary the number of observed functions. The model used in this section will be a mixed membership model with 2 functional features ($K = 2$), that can be represented by a basis constructed of 8 b-splines ($P = 8$), and uses 3 eigenfunctions ($M = 3$). We consider the case when we have 40, 80, and 160 observed functional observations, where each observation is uniformly observed at 100 time points. We then simulated 50 datasets for each of the three cases ($N = 40, 80, 160$). Section B.2.1 goes into further detail of how the model parameters were drawn and how estimates of all quantities of interest were calculated in this simulation. To measure how well we recovered the functions of interest, we estimated the relative mean integrated square error (R-MISE) of the mean, covariance, and cross-covariance functions, where R-MISE $= \frac{\int\{f(t)-\hat{f}(t)\}^2 \mathrm{d}t}{\int f(t)^2 \mathrm{d}t} \times 100\%$. In this case, the $\hat{f}$ used to estimate the R-MISE is the estimated posterior median of the function f. To measure how well we recovered the allocation parameters, Z_{ik}, we calculated the root-mean-square error (RMSE).

From Figure 3.2, we can see that we have good recovery of the mean structure with as little as 40 functional observations. While the R-MISE of the mean functions improve as we increase the number of functional observations (N), this improvement will likely have little practical impact. However, when looking at the recovery of the covariance and cross-covariance functions, we can see that the R-MISE noticeably decreases as more functional observations are added. As the recovery of the mean and covariance structures improve, the recovery of the allocation structure ($\mathbf{Z}$) improves. Visualizations of the recovered covariance

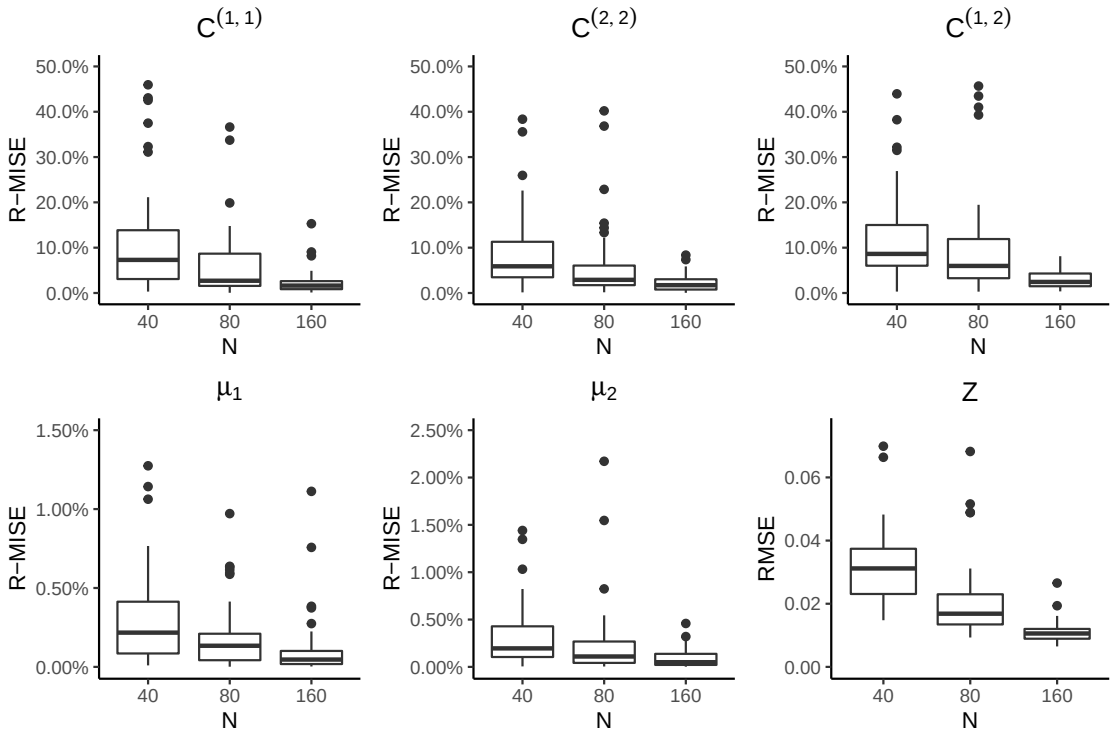


Figure 3.2: R-MISE values for the latent feature means and cross-covariances, as well as RMSE values for the allocation parameters, evaluated as we increase sample size (number of functional observations).

structures for one of the datasets can be found in Section B.2.1.

3.3.2 Simulation Study 2

Choosing the number of latent features can be challenging, especially when no prior knowledge is available for this quantity. Information criteria, such as the AIC, BIC, or DIC, are often used to aid practitioners in the selection of a data-supported value for K. In this simulation, we evaluate how various types of information criteria perform in recovering the true number of latent features. To do this, we simulate 10 different data-sets, each with 200 functional observations, from a mixed membership model with three functional features. We then calculate these information criteria on the 10 data-sets for mixed membership models where $K = 2, 3, 4, 5$. In addition to examining how AIC, BIC, and DIC perform, we will

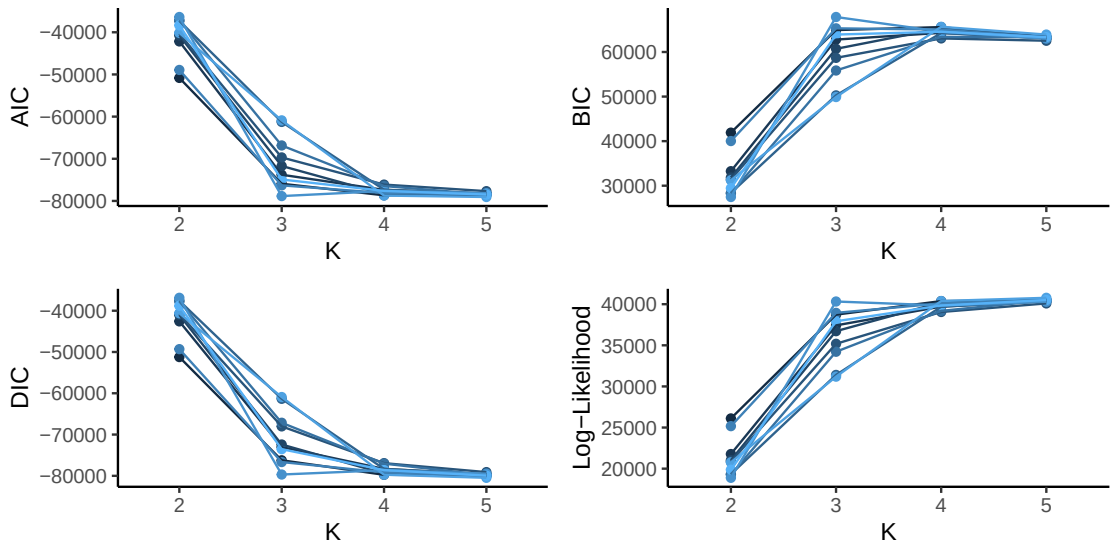


Figure 3.3: AIC, BIC, DIC, and the average log-likelihood evaluated for each of the 10 simulated data-sets.

also look at the performance of simple heuristics such as the "elbow-method". Additional information on how the simulation study was conducted, as well as definitions of the information criterion, can be found in Section B.2.2. As specified in Section B.2.2, the optimal model should have the largest BIC, smallest AIC, and smallest DIC.

From the simulation results presented in Figure 3.3, we can see that on average, each information criterion overestimated the number of functional features in the model. While the three information criteria seem to greatly penalize models that do not have enough features, they do not seem punish overfit models to a great enough extent. Figure 3.3 also shows the average log-likelihood of the models. As expected, the log-likelihood increases as we add more features, however, we can see that there is an elbow at $K = 3$ for most of the models. Using the "elbow-method" led to selecting the correct number of latent functional features 8 times out of 10, while BIC picked the correct number of latent functional features twice. DIC and AIC were found to be the least reliable information criteria, only choosing the correct number of functional features once. Thus, through empirical consideration, we suggest using the "elbow-method" along with the information criteria discussed in this

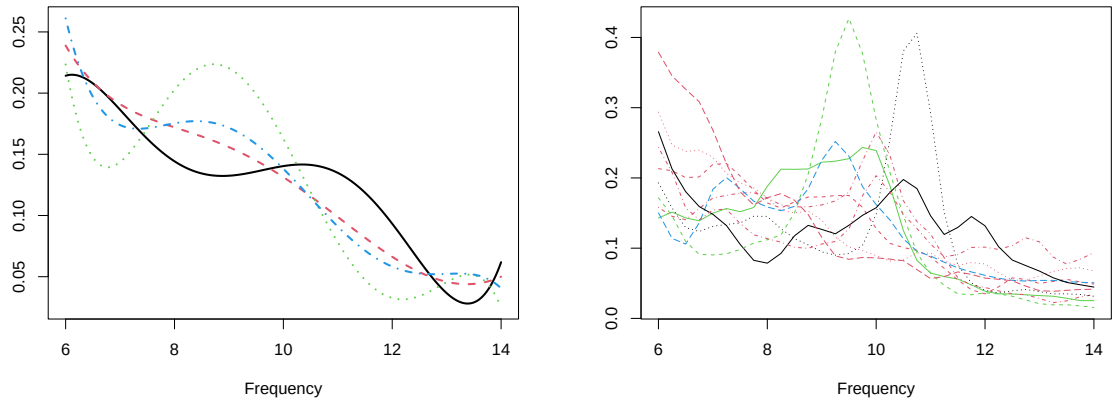


Figure 3.4: Preliminary Data Clustering. (Left Panel) Recovered means of Model-based functional clustering with 4 clusters. (Right Panel) Alpha frequency patterns for a sample of EEG recordings from the T8 electrode of children (TD and ASD). Individual observations are color-coded to match the estimated cluster membership.

section to aid a final selection for the number of latent features to be interpreted in analyses. While formal considerations of model-selection consistency are out of scope for the current contribution, we maintain that some of these techniques are best interpreted in the context of data exploration, with a potential for great improvement in interpretation if semi-supervised considerations allow for an *a-priori* informed choice of K.

3.3.3 A Case Study of EEG in ASD

Autism spectrum disorder (ASD) is a term used to describe individuals with a collection of social communication deficits and restricted or repetitive sensory-motor behaviors [Lord et al., 2018]. While once considered a very rare disorder with very specific symptoms, today the definition is more broad and is now thought of as a spectrum. Some individuals with ASD may have minor symptoms, while other may have very severe symptoms and require lifelong support. To diagnose a child with ASD, pediatricians and psychiatrists often administer a variety of test and come to a diagnosis based off of the test results and reports from the parents or caregivers. In this case study, we will be using electroencephalogram (EEG) data

that was previously analyzed by Scheffler et al. [2019] in the context of regression. The data-set consists of EEG recordings of 39 typically developing (TD) children and 58 children with ASD between the ages of 2 and 12 years old, which were analyzed in the frequency domain. Additional information on how the study was conducted can be found in Section B.2.3.

Scheffler et al. [2019] found that the T8 electrode, corresponding to the right temporal region, had the highest average contribution to the log-odds of ASD diagnosis, so we will specifically be using data from the T8 electrode in our mixed membership model. We focus our analysis to the alpha band of frequencies (6 to 14 Hz), whose patterns at rest are thought to play a role in neural coordination and communication between distributed brain regions. As clinicians examine sample paths for the two cohorts, shown in Figure 3.4 (right panel), they are often interested in the location of a single prominent peak in the spectral density located within the alpha frequency band called the peak alpha frequency (PAF). This quantity has been previously liked to neural development in TD children [Rodríguez-Martínez et al., 2017]. Scheffler et al. [2019] found that as TD children grow, the peak becomes more prominent and the PAF shifts to a higher frequency. Conversely, a discernible PAF pattern is attenuated for most children with ASD when compared to their TD counterpart.

A visual examination of the data in Figure 3.4 (right panel) anticipates the potential inadequacy of cluster analysis in this application, as path-specific heterogeneity does not seem to define well separated sub-populations. In fact, if we cluster our data using the model in Pya Arnqvist et al. [2021] with $K = 4$ (BIC-selection), we find cluster means of dubious interpretability, and poor separation of sample paths between clusters.

In contrast to classical clustering, we use a mixed membership model with only 2 functional features ($K = 2$), (AIC-BIC-selection). We note that the enhanced flexibility of mixed membership models, induces parsimony in the number of pure mixture components supported by the data. In particular, the mean function of the first feature, depicted in Figure 3.5, can be interpreted as $1/f$ noise, or *pink noise*. This component noise is expected

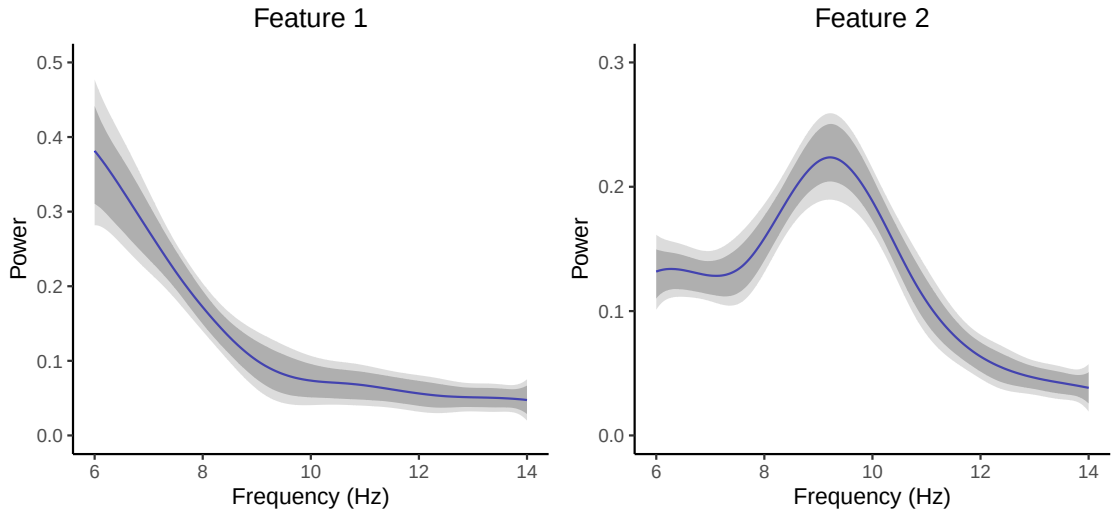


Figure 3.5: Posterior median and 95% credible (pointwise credible interval in dark gray and simultaneous credible interval in light gray) of the mean function for each functional feature.

to be found in every individual to some extent, but we can see that the first feature has no discernible alpha peak. The mean function of the second feature captures a distinct alpha peak, typically observed in EEGs of older neurotypical individuals. These two features help differentiation between periodic (alpha waves) and aperiodic ($1/f$ trend) neural activity patterns, which coexist in the EEG spectrum.

In this context, a model of *uncertain membership* would be necessarily inadequate to describe the observed sample path heterogeneity, as we would not naturally think of subjects in our sample to belong to one or the other cluster. Instead, assuming *mixed membership* between the two feature processes is likely to represent a more realistic and interpretable data generating mechanism, as we conceptualize periodic and aperiodic neural activity patterns to mix continuously.

From Figure 3.6, we find that TD children are highly likely to load heavily on feature 2 (well defined PAF), whereas ASD children exhibit a higher level of heterogeneity. These loadings suggest that clear alpha peaks take longer to emerge in children with ASD, when compared to their typically developing counterparts.

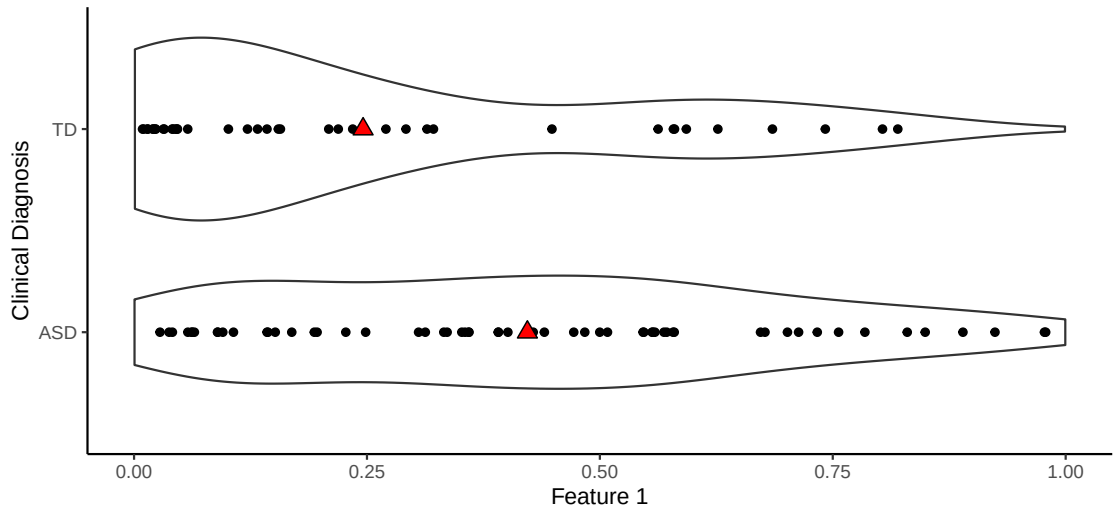


Figure 3.6: Posterior median for the membership to feature 1, stratified by clinical cohort. The red triangles represent the mean (feature-1)-membership for each clinical group.

Overall, our findings confirm related evidence in the scientific literature on developmental neuroimaging, but offer a completely novel point of view in quantifying group membership as part of a spectrum. An in-depth comparison between our method and alternative methods such as FPCA and functional clustering are reported in Section B.2.3. An extended analysis of multi-channel EEG for the same case study is reported in Section B.2.4, investigating how spatial patterns vary across the scalp.

3.4 Discussion

This manuscript introduces the concept of functional mixed membership to the broader field of functional data analysis. Mixed membership is defined as a natural generalization of the concept of *uncertain membership*, which underlies the many approaches to functional clustering discussed in the literature. Our discussion is carried out within the context of Bayesian analysis. In this paper, a coherent and flexible sampling model is introduced by defining a mixed membership Gaussian process through projections on the linear subspace spanned by a suitable set of basis functions. Within this context, we leverage the multivariate KL

formulation [Happ and Greven, 2018], to define a model ensuring weak conditional posterior consistency. Inference is carried out through standard MCMC with the inclusion of tempered transitions to improve Markov chain exploration over differential modal configurations.

Our work is closely related to the approach introduced by Heller et al. [2008], who extended the theory of mixed membership models in the Bayesian framework for multivariate exponential family data. For Gaussian marginal distributions, their representation implies multivariate normal observations, with the natural parameters modeled as convex combinations of the individual cluster's natural parameters. While intuitively appealing, this idea has important drawbacks. Crucially, the differential entropy of observations in multiple clusters is constrained to be smaller than the minimum differential entropy across clusters, which may not be realistic in many sampling scenarios.

The main computational challenge associated MCMC simulations from the posterior target has to do with the presence of multiple modal configurations for the model parameters, which is typical of mixture models. To ensure good mixing and irreducible exploration of the target, our implementation included tempered transitions (Section B.3.3) allowing the Markov chain to cross areas of low posterior density. We also recommend non-naïve chain initialization by using the Multiple Start Algorithm (Section B.3.2). As it is often the case for non-trivial posterior simulations, careful consideration is needed in tuning temperature ladders and the associated tuning parameters.

Information criteria are often used to aid the choice of K in the context of mixture models. However, in Section 3.3.2, we saw that they often overestimated the number of features in our model. In simulations, we observed that using the "elbow-method" could lead to the selection of the correct number of features with good frequency. The literature has discussed non-parametric approaches to feature allocation models by using, for example, the Indian Buffet Processes [Griffiths and Ghahramani, 2011], but little is known about operating characteristics of these proposed procedures and little has been discussed about the ensuing need to carry out statistical inference across changing-dimensions. Rousseau and

Mengersen [2011], as well as Nguyen [2013], proved that under certain conditions, overfitted mixture model have a posterior distribution that converges to a set of parameters where the only components with positive weight will be the "true" parameters (the rest will be zero). However, by allowing the membership parameters to be a continuous random variable we introduce a stronger type of non-identifiability, which led to the *rescaling problem* discussed in Section 3.1.2. Therefore, more work on the posterior convergence of overfitted models is needed for mixed membership models.

To remove the interpretability problems caused by the *rescaling problem*, we recommend using the Membership Rescale Algorithm (Algorithm 4) when using a model with only two latent features. The Membership Rescale Algorithm ensures that the entire range of the simplex is used, guaranteeing that at least one observation completely lies within each latent feature. Maximizing the volume of the convex polytope constructed by the allocation parameters can be more challenging when we have more than 2 functional features, but it can be reformulated as solving an optimization problem (Section B.3.4). In practice, we found that rescaling was seldomly needed when working with 3 or more functional features. An R package for fitting functional mixed membership models is available for download at `https://github.com/ndmarco/BayesFMMM`.

CHAPTER 4

Covariate Adjusted Functional Mixed Membership Models

4.1 Introduction

Clustering analysis is a unsupervised, exploratory task that aims to group similar observations into "clusters" to explain heterogeneity found in a dataset. While we often have little previous knowledge on how the data are correlated, there are certain situations in which the distribution of the data is dependent on a covariate of interest, leading to the need for covariate-dependent clustering. Covariate-dependent clustering has been popular in the field of clinical trials [Müller et al., 2011], where conditioning on factors like dose response, tumor grade, and other clinical factors are often of interest when performing clustering. In addition to clinical trial settings, covariate-dependent clustering has also become popular in fields like genetics [Qin and Self, 2006], flow cytometry [Hyun et al., 2023], neuroimaging [Guha and Guhaniyogi, 2022, Binkiewicz et al., 2017], and spatial statistics [Page and Quintana, 2016].

In the fields of statistics and machine learning, covariate-dependent clustering models can be found under numerous names, including *finite mixture of regressions*, *mixture of experts*, and *covariate adjusted product partition models*. The term finite mixture of regressions [McLachlan et al., 2019, Faria and Soromenho, 2010, Grün et al., 2007, Khalili and Chen, 2007, Hyun et al., 2023, Devijver, 2015] refers to fitting a mixture model, where the mean structure is dependent on the covariates of interest through a regression framework. Mixture of experts models [Jordan and Jacobs, 1994, Bishop and Svenskn, 2002] are similar to finite

mixture of regressions in that they assume that the likelihood is a weighted combination of probability distribution functions. However, in the mixture of experts model, the weights are dependent on the covariates of interest, adding an extra layer of flexibility compared to traditional finite of regressions models. Lastly, covariate adjusted product partition models [Müller et al., 2011, Park and Dunson, 2010] are a Bayesian non-parametric version of covariate adjusted clustering. Similarly to mixture of experts models, covariate adjusted product partition models can make the cluster partitions dependent on the covariates of interest.

In this manuscript, we extend the class of functional mixed membership models proposed by Marco et al. [2022b] to allow for features to depend on covariate information. Mixed membership models [Erosheva et al., 2004, Heller et al., 2008, Gruhl and Erosheva, 2014, Marco et al., 2022b,c] can be thought of as a generalization of traditional clustering, where each observations is allowed to partially belong to multiple clusters or features. While there have been many advancements in covariate adjusted clustering models for multivariate data, to our knowledge there has been little work on incorporating covariate information in functional mixed membership models or functional clustering models. Two important exceptions are Yao et al. [2011], whom specified a finite mixture of regressions model where the covariates are functional data and the data that are clustered is scalar, and Gaffney and Smyth [1999] whom proposed a functional mixture of regressions model where the function is modeled by a deterministic linear function of the covariates. In this manuscript we consider the case where we have multivariate covariates and the data we cluster are functional.

This manuscript is primarily motivated by functional brain imaging studies on children with autism spectrum disorder (ASD) through electroencephalography (EEG). Specifically, Marco et al. [2022b] analyzed how alpha oscillations differ between children with ASD and typically developing (TD) children. Since alpha oscillations are known to change as children develop, the need for an age-dependent mixed membership model is crucial to ensure that shifts in the alpha oscillations do not confound measures of alpha power [Haegens et al., 2014]. Unlike mixture of experts models and covariate adjusted product partition models, we aim

to specify a mixed membership models in which the allocation structure does not depend on the covariates of interest. While previous studies have shown that the alpha peak shifts to a higher frequency and becomes more attenuated [Scheffler et al., 2019, Rodríguez-Martínez et al., 2017], the results can be confounded if this effect isn't observed in all individuals. In our model, since we assume that each individual's allocation parameters do not change with age, we can infer how alpha oscillations change as children age at an individual level, making a covariate adjusted mixed membership model ideal for this case study.

This manuscript starts with a brief review of functional clustering and covariate-dependent clustering frameworks such as mixture of experts and mixture of regressions models. Using these previous frameworks as a reference, we derive the general form of our covariate adjusted mixed membership model. In Section 4.2.1 we review the Multivariate Karhunen-Loève (KL) theorem, which allows us to have a concise representation of the K latent functional features. In Section 4.2.2, we leverage the KL decomposition to completely specify the covariate adjusted functional mixed membership model. A review of the identifiability issues that occur in mixed membership models, as well as sufficient conditions to ensure identifiability in a two feature covariate adjusted mixed membership model can be found in Section 4.2.3. Section 4.3 covers a simulation study which explores the empirical convergence properties of the mean, covariance, and allocation structure of the proposed model. Section 4.4 illustrates the usefulness of the covariate adjusted functional mixed membership model by analyzing EEG data from children with ASD and TD children. Lastly, we conclude this manuscript with discussion on some of the challenges of fitting these models, as well as possible theoretical challenges when working with covariate adjusted mixed membership models with 3 or more features.

4.2 Covariate Adjusted Functional Mixed Membership Model

Functional data analysis (FDA) focuses on analyzing the sample paths of continuous stochastic processes $f : \mathcal{T} \to \mathbb{R}$, where $\mathcal{T}$ is a compact subset of $\mathbb{R}^d$. In FDA, we commonly assume that the random functions are elements of a Hilbert space, or more specifically that the random functions are square-integrable functions $\left(f \in L^2 \text{ or } \int_{\mathcal{T}} \mid f(t) \mid^2 \mathrm{d}t < \infty\right)$. In this manuscript, we will assume that the continuous stochastic processes are Gaussian processes (GP), meaning the distribution of function can be specified by a mean function, $\mu(t) = \mathbb{E}\left(f(t)\right)$, and a covariance function, $C(s,t) = \mathrm{Cov}\left(f(s), f(t)\right)$, for $t, s \in \mathcal{T}$. Since mixed membership models can be considered a generalization of finite mixture models, we will show in this section how finite mixture of regressions and mixture of experts models relate to our proposed mixed membership models. While we don't review covariate adjusted product partition models due to the significant differences in both theory and implementation, detailed explanations can be found in Müller et al. [2011] and Park and Dunson [2010]. For the theoretical developments discussed in this section, we will assume that the number of clusters or features, K, are known *a-priori*. While the number of clusters or features are often unknown, the use of information criterion or simple heuristic methods such as the "elbow" method have shown to be informative in choosing the number of features in a mixed membership model Marco et al. [2022b].

Functional clustering generally assumes that each sample path is drawn from one of K underlying cluster-specific sub-processes [James and Sugar, 2003, Chiou and Li, 2007, Jacques and Preda, 2014]. Assuming that $f^{(1)}, \ldots, f^{(K)}$ are the K underlying cluster-specific sub-processes with corresponding mean functions $\mu^{(1)}, \ldots \mu^{(K)}$ and covariance functions $C^{(1)}, \ldots, C^{(K)}$, we can arrive at the general form of a GP finite mixture model:

$$p\left(f_i \mid \rho^{(1:K)}, \mu^{(1:K)}, C^{(1:K)}\right) = \sum_{k=1}^{K} \rho^{(k)} \, \mathcal{GP}\left(f_i \mid \mu^{(k)}, C^{(k)}\right), \tag{4.1}$$

where $\rho^{(k)}$ ($\sum_{k=1}^{K} \rho^{(k)} = 1$) are the mixing proportions and f_i are the sample paths for $i = 1, \ldots, N$. Introducing the latent variables $\boldsymbol{\pi}_i = (\pi_{i1}, \ldots, \pi_{iK})$, where $\boldsymbol{\pi}_i \sim_{iid} \text{Mult}(1;\, \rho^{(1)}, \ldots, \rho^{(K)})$, we can show that the likelihood can be written as

$$f_i \mid \boldsymbol{\pi}_i, \mu^{(1:K)}, C^{(1:K)} \sim \mathcal{GP}\left(\sum_{k=1}^{K} \pi_{ik}\mu^{(k)}, \sum_{k=1}^{K} \pi_{ik}C^{(k)}\right). \tag{4.2}$$

Using this formulation of the likelihood, we can interpret π_{ik} as a binary indicator of the i^{th} observation's membership to the k^{th} cluster. Let $\mathbf{x}_i = [X_{i1} \ldots X_{iR}]$ be the covariates of interest associated with the i^{th} observation. We will let $\mathbf{X}$ denote the design matrix (without an intercept column), where $\mathbf{x}_i$ is the i^{th} row of the design matrix. Extending the multivariate mixture of regressions model [McLachlan et al., 2019, Faria and Soromenho, 2010, Grün et al., 2007, Khalili and Chen, 2007, Hyun et al., 2023, Devijver, 2015] to a functional setting, we can represent the general form of mixture of regressions model for functional data as

$$f_i \mid \mathbf{X}, \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N, \mu^{(1:K)}, C^{(1:K)} \sim \mathcal{GP}\left(\sum_{k=1}^{K} \pi_{ik}\mu^{(k)}(\mathbf{x}_i), \sum_{k=1}^{K} \pi_{ik}C^{(k)}\right). \tag{4.3}$$

In the multivariate setting, the mean is often modeled through a regression framework, leading to the functional form in the FDA setting of $\mu^{(k)}(\mathbf{x}_i, t) = \beta_0(t) + \sum_{r=1}^{R} X_{ir}\beta_r(t)$, where $\beta_0, \ldots, \beta_R \in L^2$ and $t \in \mathcal{T}$. Similarly to Equation 4.1, the mixture of experts model can be formulated as

$$P\left(f_i \mid \mathbf{X}, \boldsymbol{\alpha}_{(1:K)}, \mu^{(1:K)}, C^{(1:K)}\right) = \sum_{k=1}^{K} \pi_{ik}(\mathbf{x}_i, \boldsymbol{\alpha}_k)\, \mathcal{GP}\left(f_i \mid \mu^{(k)}(\mathbf{x}_i), C^{(k)}\right). \tag{4.4}$$

From equation 4.4, we can see that the $\pi_{ik}(\mathbf{x}_i, \boldsymbol{\alpha}_k)$ act as mixing proportions, however they are dependent on the covariates of interest. In the mixture of experts model, we assume that $\pi_{ik}(\mathbf{x}_i, \boldsymbol{\alpha}_k) \propto \exp(\boldsymbol{\alpha}_k'\mathbf{x}_i)$, where $\boldsymbol{\alpha}_k$ is a learned set of parameters. Similarly to the mixture of regressions model, the mean component is model through a regression framework, such

that $\mu^{(k)}(\mathbf{x}_i, t) = \beta_0(t) + \sum_{r=1}^{R} X_{ir}\beta_R(t)$, where $\beta_0, \ldots, \beta_R \in L^2$ and $t \in \mathcal{T}$. The mixture of experts model can be written in a similar form as Equation 4.3 with the introduction of the latent variables $\boldsymbol{\pi}_i$, however, the distribution of $\boldsymbol{\pi}_i$ now depends on $\mathbf{x}_i$. To arrive at the functional mixed membership model specified in Marco et al. [2022b], we can rewrite the finite mixture model in Equation 4.2 as

$$f_i \mid \boldsymbol{\pi}_1, \ldots, \boldsymbol{\pi}_N =_d \sum_{k=1}^{K} \pi_{ik} f^{(k)}. \tag{4.5}$$

By introducing a new set of latent variables $\mathbf{z}_i = (Z_{i1}, \ldots, Z_{iK})'$, where $Z_{ik} \in (0, 1)$ and $\sum_{k=1}^{K} Z_{ik} = 1$, we can arrive at the functional form of the of the functional mixed membership model:

$$f_i \mid \mathbf{Z}_1, \ldots, \mathbf{Z}_N =_d \sum_{k=1}^{K} Z_{ik} f^{(k)}. \tag{4.6}$$

Thus we can see that under the functional mixed membership model, each sample path is assumed to come from a convex combination of the underlying GPs, $f^{(k)}$. Unlike in the case of traditional clustering, the functional mixed membership model does not assume that the underlying GPs are mutually independent. Thus we will let $C^{(k,j)}$ represent the cross-covariance function between the k^{th} GP and the j^{th} GP, for $1 \leq k \neq j \leq K$. Letting $\boldsymbol{\mathcal{C}}$ be the collection of covariance and cross-covariance functions, we can specify the sampling model of the functional mixed membership model as

$$f_i \mid \mathbf{z}_1, \ldots, \mathbf{z}_N, \mu^{(1:K)}, \boldsymbol{\mathcal{C}} \sim \mathcal{GP}\left(\sum_{k=1}^{K} Z_{ik}\mu^{(k)}, \sum_{k=1}^{K} Z_{ik}^2 C^{(k)} + \sum_{k=1}^{K}\sum_{k' \neq k} Z_{ik} Z_{ik'} C^{(k,k')}\right). \tag{4.7}$$

Finite mixture models, as well as mixture of experts and finite mixture of regressions models, can be represented in the same functional form as the representation in Equation 4.5. However, for these covariate adjusted clustering models, the underlying stochastic processes, $f^{(k)}$, have an associated mean that depends on the covariates of interest, which we will denote as $\mu^{(k)}(\mathbf{x}_i)$. Similarly, by assuming the underlying stochastic processes in Equation 4.7 have

a mean that depends on the covariates of interest, we can arrive at the sampling model of the covariate functional mixed membership model:

$$f_i \mid \mathbf{X}, \mathbf{z}_1, \ldots, \mathbf{z}_N, \mu^{(1:K)}(\mathbf{x}_i), \boldsymbol{\mathcal{C}} \sim \mathcal{GP}\left(\sum_{k=1}^{K} Z_{ik}\mu^{(k)}(\mathbf{x}_i), \sum_{k=1}^{K} Z_{ik}^2 C^{(k)} + \sum_{k=1}^{K}\sum_{k'\neq k} Z_{ik}Z_{ik'}C^{(k,k')}\right). \tag{4.8}$$

In section 4.2.1, we will review the Multivariate Karhunen-Loève (KL) theorem [Ramsay and Silverman, 2005b, Happ and Greven, 2018], which will allow us to have a joint approximation of the covariance structure of the K underlying GPs. Using the joint approximation, we are able to concisely represent the K GPs, facilitating inference on higher dimensional functions such as surfaces. Using the KL decomposition, we are able to fully specify our proposed covariate adjusted functional mixed membership model (Equation 4.8) in section 4.2.2.

4.2.1 Multivariate Karhunen-Loève Characterization

In the previous section, we showed how our proposed covariate adjusted functional mixed membership model relates to other covariate adjusted models such as the mixture of regressions and mixture of experts models. While Equation 4.8 shows the proposed form of our model, the mean functions and covariance functions that we are left to estimate are infinite dimensional parameters. Thus in order to be able to estimate the mean functions and covariance functions, we will assume that the underlying GPs are smooth and lie in the P-dimensional subspace, $\boldsymbol{\mathcal{S}} \subset L^2(\mathcal{T})$, spanned by a set of linearly independent square-integrable basis functions, $\{b_1, \ldots, b_P\}$ ($b_p : \mathcal{T} \to \mathbb{R}$). While the choice of basis functions are user-defined, the basis functions must be uniformly continuous in order to satisfy Lemma 2.2 of Marco et al. [2022b]. In this manuscript, we will primarily use B-splines for all case studies and simulation studies.

The assumption that $f^{(k)} \in \boldsymbol{\mathcal{S}}$ allows us to turn an infinite-dimensional problem into a finite-dimensional problem, making traditional inference tractable. While tractable, modeling the covariance functions and cross-covariance functions separately leads to a model that

needs $\mathcal{O}(K^2P^2)$ parameters to model the covariance structure. While the number of clusters, K, and the number of basis functions, P, may be relatively small for simple problems, the number of basis functions needed to model higher dimensional functional data such as surfaces becomes large and computationally intractable. Thus we will use the multivariate KL decomposition [Ramsay and Silverman, 2005b, Happ and Greven, 2018] to reduce the number of parameters needed to estimate the covariance surface to $\mathcal{O}(KPM)$, where M is the number of eigenfunctions used to approximate the covariance structure. A more detailed derivation of a KL decomposition for functions in $\boldsymbol{\mathcal{S}}$ can be found in Marco et al. [2022b].

To achieve a concise joint representation of the K latent GPs, we will define a multivariate GP, which we will denote $\mathbf{f}(\mathbf{t})$ such that

$$\mathbf{f}(\mathbf{t}) := \left(f^{(1)}\left(t^{(1)}\right), f^{(2)}\left(t^{(2)}\right), \ldots, f^{(K)}\left(t^{(K)}\right)\right),$$

such that $\mathbf{t} = \left(t^{(1)}, t^{(2)}, \ldots, t^{(K)}\right)$ and $t^{(1)}, t^{(2)}, \ldots, t^{(K)} \in \mathcal{T}$. Since $\boldsymbol{\mathcal{S}} \subset L^2$ is a Hilbert space, with the inner-product defined as the L^2 inner-product, we can see that $\mathbf{f} \in \boldsymbol{\mathcal{H}} := \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{S}} \times \cdots \times \boldsymbol{\mathcal{S}} := \boldsymbol{\mathcal{S}}^K$, where $\boldsymbol{\mathcal{H}}$ is defined as the direct sum of the Hilbert spaces $\boldsymbol{\mathcal{S}}$, making $\boldsymbol{\mathcal{H}}$ a Hilbert space as well. Since $\boldsymbol{\mathcal{H}}$ is a Hilbert space and the covariance operator of $\mathbf{f}$, denoted $\boldsymbol{\mathcal{K}}$, is a positive, bounded, self-adjoint, and compact operator, we know there exists a complete set of eigenfunctions $\boldsymbol{\Psi}_1, \ldots, \boldsymbol{\Psi}_{KP} \in \boldsymbol{\mathcal{H}}$ and corresponding eigenvalues $\lambda_1 \geq \cdots \geq \lambda_{KP} \geq 0$ such that $\boldsymbol{\mathcal{K}}\boldsymbol{\Psi}_p = \lambda_p \boldsymbol{\Psi}_p$, for $p = 1, \ldots, KP$. Using the eigenpairs of the covariance operator, we can rewrite $f^{(k)}$ as

$$f^{(k)}(t) = \mu^{(k)}(\mathbf{x}, t) + \sum_{m=1}^{KP} \chi_m \left(\sqrt{\lambda_m}\Psi_m^{(k)}(t)\right),$$

where $\chi_m \sim \mathcal{N}(0,1)$ and $\Psi_m^{(k)}(t)$ is the k^{th} element of $\boldsymbol{\Psi}_m(\mathbf{t})$. Since $\Psi_m^{(k)}(t) \in \boldsymbol{\mathcal{S}}$, we know there exists $\phi_m \in \mathbb{R}^P$ such that $\sqrt{\lambda_m}\Psi_m^{(k)}(t) = \boldsymbol{\phi}'_{km}\mathbf{B}(t)$, where $\mathbf{B}'(t) := [b_1(t), b_2(t), \ldots, b_P(t)]$. Similarly, since $\mu^{(k)}(\mathbf{x}, t) \in \boldsymbol{\mathcal{S}}$, we can introduce a mapping $\mathbf{g} : \mathbb{R}^R \to \mathbb{R}^P$, such that

$\mu^{(k)}(\mathbf{x}, t) = (\mathbf{g}_k(\mathbf{x}))' \mathbf{B}(t)$. Therefore, we arrive at the general form of our decomposition:

$$f^{(k)}(t) = (\mathbf{g}_k(\mathbf{x}))' \mathbf{B}(t) + \sum_{m=1}^{KP} \chi_m \boldsymbol{\phi}'_{km} \mathbf{B}(t). \tag{4.9}$$

Using this decomposition, our covariance and mean structures can be recovered such that $C^{(k,k')}(s,t) = \mathbf{B}'(s) \left(\sum_{m=1}^{KP} \boldsymbol{\phi}_{km} \boldsymbol{\phi}'_{k'm} \right) \mathbf{B}(t)$ and $\mu^{(k)}(\mathbf{x}, t) = (\mathbf{g}_k(\mathbf{x}))' \mathbf{B}(t)$, for $1 \leq k, k' \leq K$ and $s, t \in \mathcal{T}$. To reduce the dimensionality of the problem, we will only use the first M eigenpairs to approximate the K stochastic processes. While traditional functional principal component analysis (FPCA) will choose the number of eigenfunctions based off of the proportion of variance explained, the same strategy cannot be employed in functional mixed membership models, because the allocation parameters of the model are typically not known. Therefore, we suggest picking as large of M as your computational budget allows, in order to get the best approximation to the covariance structure.

4.2.2 Model and Prior Specification

In this section, we will be fully specifying the covariate adjusted functional mixed membership model utilizing a truncated version of the KL decomposition specified in Equation 4.9. We start by first specifying how the covariates of interest will influence the mean function of the functional mixed membership model, denoted as $\mathbf{g}_k(\mathbf{x})$ in Equation 4.9. Following the previous works of mixture of regressions and mixture of expert models, we will model the dependence of the covariates on the mean structure using a regression framework. Under the standard functional regression framework, we have that $\mu^{(k)}(\mathbf{x}_i, t) = \beta_{k0} + \sum_{r=1}^{R} X_{ir} \beta_{kr}(t)$ for $k = 1, \ldots, K$. Since we assumed that $\mu^{(k)}(\mathbf{x}_i, t) \in \boldsymbol{\mathcal{S}}$ for $k = 1, \ldots, K$, we know that $\beta_{k0}, \ldots, \beta_{kR} \in \boldsymbol{\mathcal{S}}$. Therefore, there exists $\boldsymbol{\nu}_k \in \mathbb{R}^P$ and $\boldsymbol{\eta}_k \in \mathbb{R}^{P \times R}$ such that

$$\mu^{(k)}(\mathbf{x}_i, t) = \boldsymbol{\nu}'_k \mathbf{B}(t) + (\boldsymbol{\eta}_k \mathbf{x}'_i)' \mathbf{B}(t). \tag{4.10}$$

Under a standardized set of covariates, $\boldsymbol{\nu}_k$ specifies the population level mean of the k^{th} feature and $\boldsymbol{\eta}_k$ encodes the covariate dependence of the k^{th} feature. Thus, assuming the mean structure specified in Equation 4.10 and using a truncated version of the KL decomposition specified in Equation 4.9, we can specify the sampling model of our covariate adjusted functional mixed membership model.

Let $\{\mathbf{Y}_i(.)\}_{i=1}^N$ be the observed sample paths that we want to model to the K features, $f^{(k)}$, conditionally on the covariates of interest, $\mathbf{x}_i$. Since the observed sample paths are observed at only a finite number of points, we will let $\mathbf{t}_i = [t_{i1}, \ldots, t_{in_i}]'$ denote the time points at which the i^{th} function was observed over. Without loss of generality, we will define the sampling distribution over the finite dimensional marginals of $\mathbf{Y}_i(\mathbf{t}_i)$. Using the general form of our proposed model defined in Equation 4.8, we have

$$\mathbf{Y}_i(\mathbf{t}_i) \mid \boldsymbol{\Theta}, \mathbf{X} \sim \mathcal{N}\left\{\sum_{k=1}^{K} Z_{ik}\left(\mathbf{S}'(\mathbf{t}_i)\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k\mathbf{x}_i'\right) + \sum_{m=1}^{M}\chi_{im}\mathbf{S}'(\mathbf{t}_i)\boldsymbol{\phi}_{km}\right),\ \sigma^2\mathbf{I}_{n_i}\right\}, \quad (4.11)$$

where $\mathbf{S}(\mathbf{t}_i) = [\mathbf{B}(t_1)\cdots\mathbf{B}(t_{n_i})] \in \mathbb{R}^{P\times n_i}$ and $\boldsymbol{\Theta}$ is the collection of the model parameters. As defined in section 4.2, Z_{ik} are variables that lie on the unit simplex, such that $Z_{ik} \in (0,1)$ and $\sum_{k=1}^K Z_{ik} = 1$. From this characterization, we can see that each observation is modeled as a convex combination of realizations from the K features with additional Gaussian noise, represented by σ^2. If we integrate out the χ_{im} variables, for $i = 1, \ldots, N$ and $m = 1, \ldots, M$, we arrive the following likelihood:

$$\mathbf{Y}_i(\mathbf{t}_i) \mid \boldsymbol{\Theta}_{-\chi}, \mathbf{X} \sim \mathcal{N}\left\{\sum_{k=1}^{K} Z_{ik}\mathbf{S}'(\mathbf{t}_i)\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k\mathbf{x}_i'\right),\ \mathbf{V}(\mathbf{t}_i, \mathbf{z}_i) + \sigma^2\mathbf{I}_{n_i}\right\}, \quad (4.12)$$

where $\boldsymbol{\Theta}_{-\chi}$ is the collection of the model parameters excluding the χ_{im} parameters ($i = 1, \ldots, N$ and $m = 1, \ldots, M$) and $\mathbf{V}(\mathbf{t}_i, \mathbf{z}_i) = \sum_{k=1}^K\sum_{k'=1}^K Z_{ik}Z_{ik'}\left\{\mathbf{S}'(\mathbf{t}_i)\sum_{m=1}^M\left(\boldsymbol{\phi}_{km}\boldsymbol{\phi}_{k'm}'\right)\mathbf{S}(\mathbf{t}_i)\right\}$. Equation 4.12 illustrates that the proposed covariate adjusted functional mixed membership model can be expressed as an additive model. The mean structure is a convex combination

of the feature specific means, while the covariance can be written as a weighted sum of covariance functions and cross-covariance functions.

To have an adequately expressive and scalable model, we approximate the covariance surface of the K features using M scaled *pseudo-eigenfunctions*. In this framework, orthogonality will not be imposed on the $\boldsymbol{\phi}'_{km}\mathbf{B}(t)$ parameters, making them *pseudo-eigenfunctions* instead of the eigenfunctions described in section 4.2.1. From a modeling prospective, this allows us to sample on an unconstrained space, facilitating better Markov chain mixing and easier sampling schemes. While direct analysis on the eigenfunctions are no longer available, a formal analysis can still be conducted by reconstructing the posterior samples of the covariance surface and calculating eigenfunctions from the posterior samples. To avoid overfitting of the covariance functions, we follow Marco et al. [2022b] by using the multiplicative gamma process shrinkage prior proposed by Bhattacharya and Dunson [2011] to achieve adaptive regularized estimation of the covariance structure. Therefore, letting ϕ_{kpm} be the p^{th} element of $\boldsymbol{\phi}_{km}$, we have that

$$\phi_{kpm} \mid \gamma_{kpm}, \tilde{\tau}_{mk} \sim \mathcal{N}\left(0, \gamma_{kpm}^{-1}\tilde{\tau}_{mk}^{-1}\right), \quad \gamma_{kpm} \sim \Gamma\left(\nu_\gamma/2, \nu_\gamma/2\right), \quad \tilde{\tau}_{mk} = \prod_{n=1}^{m} \delta_{nk},$$

$$\delta_{1k} \mid a_{1k} \sim \Gamma(a_{1k}, 1), \quad \delta_{jk} \mid a_{2k} \sim \Gamma(a_{2k}, 1), \quad a_{1k} \sim \Gamma(\alpha_1, \beta_1), \quad a_{2k} \sim \Gamma(\alpha_2, \beta_2),$$

where $1 \leq k \leq K$, $1 \leq p \leq P$, $1 \leq m \leq M$, and $2 \leq j \leq M$. By letting $\alpha_2 > \beta_2$, we can show that $\mathbb{E}(\tilde{\tau}_{mk}) > \mathbb{E}(\tilde{\tau}_{m'k})$ for $1 \leq m < m' \leq M$, leading to the prior on ϕ_{kpm} having stochastically decreasing variance as m increases. This will have a regularizing effect on the posterior draws of ϕ_{kpm}, making ϕ_{kpm} more likely to be close to zero as m increases.

In functional data analysis, we often desire smooth mean functions to safeguard against overfit models. Thus, we utilize a first order random walk penalty proposed by Lang and Brezger [2004] on our $\boldsymbol{\nu}_k$ and $\boldsymbol{\nu}_k$ parameters to promoted adaptive smoothing of the mean function of the features in our model. Therefore, we have that $P(\boldsymbol{\nu}_k \mid \tau_{\boldsymbol{\nu}_k}) \propto exp\left(-\frac{\tau_{\boldsymbol{\nu}_k}}{2}\sum_{p=1}^{P-1}\left(\nu'_{pk} - \nu_{(p+1)k}\right)^2\right)$, for $k = 1, \ldots, K$, where $\tau_{\boldsymbol{\nu}_k} \sim \Gamma(\alpha_{\boldsymbol{\nu}}, \beta_{\boldsymbol{\nu}})$ and ν_{pk} is the p^{th}

element of $\boldsymbol{\nu}_k$. Similarly, we have that $P(\{\eta_{prk}\}_{p=1}^{P} \mid \tau_{\boldsymbol{\eta}_{rk}}) \propto exp\left(-\frac{\tau_{\boldsymbol{\eta}_{rk}}}{2}\sum_{p=1}^{P-1}\left(\eta'_{prk} - \eta_{(p+1)rk}\right)^2\right)$, for $k = 1, \ldots, K$ and $r = 1, \ldots, R$, where $\tau_{\boldsymbol{\eta}_{rk}} \sim \Gamma(\alpha_{\boldsymbol{\eta}}, \beta_{\boldsymbol{\eta}})$ and η_{prk} is the p^{th} row and r^{th} column of $\boldsymbol{\eta}_k$. Following previous mixed membership models [Heller et al., 2008, Marco et al., 2022b,c], we will assume that $\mathbf{z}_i \mid \boldsymbol{\pi}, \alpha_3 \sim_{iid} Dir(\alpha_3\boldsymbol{\pi})$, $\boldsymbol{\pi} \sim Dir(\mathbf{c})$, and $\alpha_3 \sim Exp(b)$. Lastly, we will assume that $\sigma^2 \sim IG(\alpha_0, \beta_0)$. The posterior distributions of the parameters in our model, as well as a sampling scheme with tempered transitions, can be found in Section C.2.2 of the Supplementary Materials. A covariate adjusted model where the covariance is also dependent on the covariates of interest can found in Section C.4.1 of the Supplementary Materials.

4.2.3 Model Identifiability

Mixed membership models face multiple identifiability issues due to their increased flexibility over traditional clustering models [Chen et al., 2022, Marco et al., 2022b,c]. Similarly to traditional clustering models, mixed membership models also face the common *label switching* problem, where the an equivalent model can be formulated by permuting the labels or allocation parameters. Even though this is one source of non-identifiability, relabelling algorithms can be formulated from the work of Stephens [2000]. More complex identifiability problems arise since the allocation parameters are now continuous variables on the unit simplex, rather than binary variables like in clustering. These identifiability issues are discussed in further detail in Chen et al. [2022] and Marco et al. [2022b]. These identifiability issues can be easily solved in a 2 feature mixed membership models by assuming the *separability condition* holds. The separability condition assumes that at least one observation belongs completely to each of the K features [Pettit, 1990, Donoho and Stodden, 2003, Arora et al., 2012, Azar et al., 2001, Chen et al., 2022]. While the separability condition can also be assumed in mixed membership models that have over 3 features, they make strong geometric assumptions on the data generating process. Weaker geometric assumptions that ensure an identifiable model in mixed membership models with 3 or more features are discussed in

Chen et al. [2022], however implementing these constrains are non-trivial.

When considering a two feature mixed membership model, the separability condition is an assumption that ensures identifiability up to a permutation of the labels without making strong assumptions. When extending multivariate mixed membership models to functional data and introducing a covariate-dependent mean structure, ensuring an identifiable mean and covariance structure requires further assumptions. Lemma 7 states the conditions needed in order to have an identifiable mean and covariance structure up to a permutation of the labels. Proof of Lemma 7 can be found in Section 1 of the Supporting Materials. The first assumption in Lemma 7 is similar to the assumptions needed in a regression setting in order to ensure identifiability. We note that while the individual $\boldsymbol{\phi}_{km}$ parameters are unidentifiable, an eigen analysis can still be conducted by constructing posterior draws of the covariance structure and calculating the eigenvalues and eigenfunctions of the posterior draws. While the assumptions are relatively minor for ensuring identifiability in a two feature model, ensuring identifiability in models with 3 or more features requires stronger assumptions. Section 4.3 provides empirical evidence that the mean and covariance structure converge to the truth as we have more observations.

Lemma 7. *Consider a two feature ($K = 2$) covariate adjusted model as specified in Equation 4.12. The parameters $\boldsymbol{\nu}_k$, $\boldsymbol{\eta}_k$, Z_{ik}, $\sum_{m=1}^{M}(\boldsymbol{\phi}_{km}\boldsymbol{\phi}'_{k'm})$, and σ^2 are identifiable up to a permutation of the labels (i.e. label switching), for $k, k' = 1, 2$, $n = 1, \ldots, N$, and $m = 1, \ldots, M$, given the following assumptions:*

1. $\mathbf{X}$ *is full column rank with* $\mathbf{1}$ *not in the column space of* $\mathbf{X}$.

2. *The separability condition holds on the allocation matrix (there exists $\tilde{i}_1, \tilde{i}_2$ such that $Z_{\tilde{i}_1 1} = 1$ and $Z_{\tilde{i}_2 2} = 1$). Moreover, there exists at least 2 observations with allocation parameters that lie in the interior of the unit simplex*
   (i.e. $\mathbf{z}_i \in \left\{\mathbf{z} \in \mathbb{R}^2 \mid \sum_{k=1}^{2} Z_k = 1, 0 < Z_k < 1\right\}$).

3. *The sample paths $\mathbf{Y}_i(\mathbf{t}_i)$ are sampled such that $n_i \geq P$, and furthermore, there exists*

a sample path $\mathbf{Y}_i(\mathbf{t}_i)$ *such that* $n_i > 4M$.

4.2.4 Relationship to Function-on-Scalar Regression

Function-on-scalar regression is a common method in FDA which allows the mean structure of the continuous stochastic process to be covariate-dependent. In function-on-scalar regression, we often assume that the response is a GP, and that the covariates of interest are scalar-valued or vector-valued. A comprehensive review of the broader area of functional regression can be found in Ramsay and Silverman [2005a] and Morris [2015]. While there have been many advancements and generalizations [Krafty et al., 2008, Reiss et al., 2010, Goldsmith et al., 2015, Kowal and Bourgeois, 2020] of function-on-scalar regression since the initial papers of Faraway [1997] and Brumback and Rice [1998], the general form of function-on-scalar regression can be expressed as follows:

$$Y(t) = \mu(t) + \sum_{r=1}^{R} X_r \beta_r(t) + \epsilon(t); \quad t \in \mathcal{T}, \tag{4.13}$$

where $Y(t)$ is the response function evaluated at t, $\beta_r(\cdot)$ is the functional coefficient representing the effect that the r^{th} covariate (X_r) has on the mean structure, and ϵ is a mean-zero Gaussian process with covariance function $\mathcal{C}$. The function $\mu : \mathcal{T} \to \mathbb{R}$ in Equation 4.13 represents the mean of the GP when all of the covariates, X_{ir} are set to zero. Unlike the traditional setting for multiple linear regression in finite-dimensional vector spaces, function-on-scalar regression requires the estimation of the infinite dimensional functions μ and $\beta_1, \ldots, \beta_R$ from a finite number of observed sample paths at a finite number of points ($\mathbf{Y}_i(\mathbf{t}_i)$ for $i = 1, \ldots, N$ and $\mathbf{t}_i = [t_{i1}, \ldots, t_{in_i}]'$).

In order to make inference tractable, we assume that the data lie in the span of a finite set of basis functions, which will allow us to expand μ and $\beta_1, \ldots, \beta_R$ as a finite sum of the basis functions. The set of basis functions can be specified by using data-driven basis functions, or by specifying the basis functions *a-priori*. If the basis functions are specified

a-priori, common choices of basis functions are B-splines and wavelets due to their flexibility, as well as Fourier series for periodic functions. Alternatively, if the use of data-driven basis functions is desired, a common choice is to use the eigenfunctions of the covariance operator as basis functions. In order to estimate the eigenfunctions, functional principal component analysis [Shang, 2014] is often performed and the obtained estimates of the eigenfunctions are used. Functional principal component analysis faces a similar problem in that the objective is to estimate eigenfunctions using only a finite number of sample paths observed at a finite number of points. To solve this problem, Rice and Silverman [1991] proposes using a basis of splines to estimate smooth eigenfunctions, while Yao et al. [2005] proposes using local linear smoothers to estimate the smooth eigenfunctions. Therefore even using data-driven basis functions require smoothing assumptions, suggesting similar results between data-driven basis functions and a reasonable set of *a-priori* specified basis functions paired with a penalty to prevent overfitting.

Specifying the basis functions *a-priori* $(b_1(t), \ldots, b_P(t))$, and letting $\mathbf{Y}_i(\mathbf{t}_i)$ be the observed sample paths at points $\mathbf{t}_i = [t_{i1}, \ldots, t_{in_i}]'$ $(i = 1, \ldots, N)$, we can simplify Equation 4.13 to get

$$\mathbf{Y}_i(\mathbf{t}_i) = \mathbf{S}'(\mathbf{t}_i)\tilde{\boldsymbol{\nu}} + \mathbf{S}'(\mathbf{t}_i)\tilde{\boldsymbol{\eta}}\mathbf{x}_i' + \boldsymbol{\epsilon}_i(\mathbf{t}_i), \tag{4.14}$$

where $\tilde{\boldsymbol{\nu}} \in \mathbb{R}^P$, $\tilde{\boldsymbol{\eta}} \in \mathbb{R}^{P \times R}$, and $\mathbf{S}(\mathbf{t}_i) = [\mathbf{B}(t_1) \cdots \mathbf{B}(t_{n_i})] \in \mathbb{R}^{P \times n_i}$ are the set of basis function evaluated at the time points of interest. As specified in the previous sections, $\mathbf{B}'(t) := [b_1(t), b_2(t), \ldots, b_P(t)]$. Equation 4.14 shows that the function $\mu(.)$ evaluated at the points $\mathbf{t}_i$ can be represented by $\mathbf{S}'(\mathbf{t}_i)\tilde{\boldsymbol{\nu}}$, and similarly the functional coefficients $\beta_1(\cdot), \ldots, \beta_R(\cdot)$ can be represented by $\mathbf{S}'(\mathbf{t}_i)\tilde{\boldsymbol{\eta}}$. Therefore, we are left to estimate $\tilde{\boldsymbol{\nu}}$, $\tilde{\boldsymbol{\eta}}$, and the parameters associated with the covariance function of $\epsilon(\cdot)$, denoted $\mathcal{C}$.

The covariance function $\mathcal{C}$ represents the within-function covariance structure of the data. In the simplest case, we often assume that $\boldsymbol{\epsilon}_i(\mathbf{t}_i) \sim \mathcal{N}(\mathbf{0}, \tilde{\sigma}^2\mathbf{I}_{n_i})$, meaning we only need $\tilde{\sigma}^2$ to specify $\mathcal{C}$. In more complex models [Faraway, 1997, Krafty et al., 2008], we may make less restrictive assumptions and assume that the covariance $\boldsymbol{\epsilon}_i(\mathbf{t}_i) \sim \mathcal{N}(\mathbf{0}_{n_i}, \tilde{V}(\mathbf{t}_i) + \tilde{\sigma}^2\mathbf{I}_{n_i})$, where

$\tilde{V}(\cdot)$ is a low dimensional approximation of a smooth covariance surface using an truncated eigen-decomposition. While functional regression usually assumes that the functions are independent, functional mixed membership models have been proposed to model between-function variation, or cases where observations can be correlated [Morris and Carroll, 2006, Staicu et al., 2010].

Assuming a relatively general covariance structure as in Krafty et al. [2008], the function-on-scalar model assumes the following distributional assumptions on our sample paths:

$$\mathbf{Y}_i(\mathbf{t}_i) \mid \tilde{\boldsymbol{\nu}}, \tilde{\boldsymbol{\eta}}, \tilde{\mathbf{V}}(\mathbf{t}_i), \tilde{\sigma}^2, \mathbf{X} \sim \mathcal{N}\left\{\mathbf{S}'(\mathbf{t}_i)\left(\tilde{\boldsymbol{\nu}} + \tilde{\boldsymbol{\eta}}\mathbf{x}_i'\right),\ \tilde{\mathbf{V}}(\mathbf{t}_i) + \tilde{\sigma}^2\mathbf{I}_{n_i}\right\}. \tag{4.15}$$

From Equation 4.15 it is apparent that the proposed covariate adjusted mixed membership model specified in Equation 4.12 is closely related to function-on-scalar regression. The key difference between the two models is that the covariate adjusted mixed membership model does not assume a common mean and covariate structure across all observations conditionally on the covariates of interest. Instead, the covariate adjusted mixed membership model allows each observation to be modeled as a convex combination of K underlying features. Each feature is assumed allowed to have different different mean and covariance structures, meaning that the covariates are not assumed to have the same affect on all observations. By allowing this type of heterogeneity in our model, we are able to conduct a more granular analysis and identify subgroups that interact differently with the covariates of interest. Alternatively, if there are subgroups of the population that interact differently with the covariates of interest, then the results from a function-on-scalar regression model will be confounded, as the effects will likely be averaged out in the analysis.

Figures 4.1 and 4.2 illustrate the differences in the results from fitting a function-on-scalar regression model and a covariate adjusted mixed membership model. From Figure 4.2, we can see that age does not have a common effect on the alpha oscillations of developing children. For children that load heavily on feature 1 we see that age has relatively little

effect on their alpha oscillations. Conversely, we see that the alpha oscillations of children who load heavily on feature 2 are expected to have relatively large changes as they age. As seen in Figure 4.1, the results from a function-on-scalar regression analysis average out the effects of age on alpha oscillations. Perhaps the greatest advantage in performing a covariate adjusted mixed membership analysis over a function-on-scalar regression analysis is that we can make individualized inference. Covariate adjusted mixed membership models allow us to compare how the alpha oscillations of an individual child compares to their age-adjusted peers. Moreover, a covariate adjusted mixed membership model allows us to infer an individual's expected changes in alpha oscillations as they age, while a function-on-scalar regression model only allows us to infer the changes at a population level. The mixture of experts model and the mixture of regressions model, described in Section 4.2, can be thought of as a tool that provides a moderate level of granularity, where we are able to infer the changes at a sub-population level, but not at an individual level.

4.3 Simulation Study

In this section, we explore the empirical convergence properties of the proposed covariate adjusted functional mixed membership models. In this simulation study, we generate data from a covariate adjusted functional mixed membership model and see how well the proposed framework can recover the true mean, covariance, and allocation structures. To evaluate how how well we can recover the mean, covariance, and cross covariance functions, we use calculate the relative mean integrated square error (R-MISE), which is defined as $\text{R-MISE} = \frac{\int\{f(t)-\hat{f}(t)\}^2\mathrm{d}t}{\int f(t)^2\mathrm{d}t} \times 100\%$ or $\text{R-MISE} = \frac{\int_t\int_x\{f(t,x)-\hat{f}(t,x)\}^2\mathrm{d}x\mathrm{d}t}{\int_t\int_x f(t,x)^2\mathrm{d}x\mathrm{d}t} \times 100\%$ in the case of a covariate adjusted model. In this simulation study, $\hat{f}(t)$ will be the posterior median obtained from our posterior samples. To measure how well we recover the allocation structure, Z_{ik}, we calculated the root-mean-square error (RMSE). In addition to studying the empirical convergence properties of correctly specified models, we also included a scenario

where we fit a covariate adjusted functional mixed membership model, when the generating truth had no covariate dependence. Conversely, we also studied the scenario where we fit a functional mixed membership model with no covariates, when the generating truth was generated from a covariate adjusted functional mixed membership model with one covariate. Additional details on how the simulation was conducted can be found in Section C.3.1 of the Supplementary Materials.

Table 4.1 contains summary statistics of the performance metrics from each of the 5 scenarios considered in this simulation study. We can see that our model does a good job in recovering the mean structure with relatively few observations under a correctly specified model. On the other hand, a relatively large number of observations are needed to recover the covariance structure when we have a correctly specified model. This simulation study also shows that we pay a penalty in terms of statistical efficiency when we over-specify a model, however the over-specified model still shows signs of convergence to the true parameters. Conversely, an under-specified model seems to never be able to recover the mean or covariance structure, and shows no signs of converging to the true parameters as we get more observations.

4.4 Autism Spectrum Disorder Case Study

Autism spectrum disorder (ASD) is a developmental disorder characterized by social communication deficits and restrictive and/or repetitive behaviors [American Psychiatric Association et al., 2013]. While once more narrowly defined, autism is now seen as a spectrum, with some individuals having very mild symptoms, to others that require lifelong support [Lord et al., 2018]. In this case study, we will be using electroencephalogram (EEG) data that was obtained in a resting-state EEG study conducted by Dickinson et al. [2018]. The study consisted of 58 children who have been diagnosed with ASD between the ages of 2 and 12 years old, and 39 age-matched typically developing (TD) children, or children who have

Table 4.1: The median RISE/RSE, as well as the 10^{th} and 90^{th} percentiles, from 50 simulated data sets under a variety of conditions. The left column contains the true number of parameters used to simulate the column, as well as the number of covariates used when fitting the covariate adjusted functional mixed membership models.

Truth / Model (# Covariates)	Parameter	$N = 60$	$N = 120$	$N = 240$
2/2	μ_1	1.9% (0.3%, 24.7%)	1.1% (0.2%, 10.4%)	0.3%(0.1%, 8.8%)
	μ_2	1.5% (0.4%,14.5%)	1.0% (0.2%, 10.5%)	0.2% (0.1%, 10.9%)
	$C^{(1,1)}$	156.1% (2.1%, 112219.4%)	110.3% (0.1%, 1806067.0%)	6.1% (0.1%, 362938.9%)
	$C^{(2,2)}$	88.1% (1.8%, 60673.8%)	416.2% (1.9%, 1008651.0%)	4.9% (0.5%, 22725.8%)
	$C^{(1,2)}$	431.2% (3.5%, 35924.4%)	433.7% (2.2%, 246646.3%)	22.3% (0.6%, 29231.3%)
	Z	0.047 (0.020, 0.099)	0.030 (0.013, 0.074)	0.013 (0.008, 0.054)
		$N = 50$	$N = 100$	$N = 200$
1/1	μ_1	1.5% (0.2%, 7.6%)	0.8% (0.1%, 4.9%)	1.1%(0.2%,5.4%)
	μ_2	1.6% (0.3%,5.7%)	1.2% (0.2%, 7.6%)	1.2% (0.2%, 5.4%)
	$C^{(1,1)}$	218.5% (26.0%, 11299.6%)	30.8% (14.4%, 308.4%)	37.1% (9.5%, 421.2%)
	$C^{(2,2)}$	204.4% (22.5%, 2603.4%)	40.2% (8.3%, 597.6%)	25.5% (5.7%, 157.7%)
	$C^{(1,2)}$	219.8% (42.9%, 1912.9%)	89.1% (21.2%, 403.0%)	60.6% (13.0%, 350.2%)
	Z	0.067 (0.047, 0.085)	0.056 (0.042, 0.081)	0.051 (0.040, 0.065)
1/0	μ_1	382.2% (153.4%, 961.9%)	650.7% (91.1%, 1511.0%)	1076.7%(94.8%,2339.0%)
	μ_2	394.6% (117.5%,1292.3%)	751.4% (69.0%, 1721.0%)	885.1% (145.0%, 2313.0%)
	$C^{(1,1)}$	1581365.0% (81644.7%, 23059352.5%)	1328559.4% (64656.5%, 40230314.1%)	1348112.9% (98035.6%, 65828353.0%)
	$C^{(2,2)}$	730829.2% (133764.2%, 9829513.4%)	1015747.1% (86551.9%, 17361755.8%)	802590.5% (44704.4%, 21037857.8%)
	$C^{(1,2)}$	1271237.9% (90303.1%, 9356418.4%)	1917180.3% (91394.3%, 20373022.9%)	1392890.2% (81254.1%, 19419032.6%)
	Z	0.202 (0.180, 0.217)	0.172 (0.157, 0.184)	0.144 (0.121, 0.156)
		$N = 40$	$N = 80$	$N = 160$
0/1	μ_1	2.3% (0.3%, 36.7%)	2.5% (0.2%, 33.6%)	1.9%(0.2%,20.4%)
	μ_2	4.1% (0.3%,36.1%)	1.9% (0.3%, 21.6%)	3.8% (0.2%, 26.1%)
	$C^{(1,1)}$	27.1% (7.7%, 703.6%)	19.1% (3.3%, 95.5%)	20.3% (3.1%, 64.9%)
	$C^{(2,2)}$	28.9% (9.4%, 319.1%)	19.0% (3.7%, 206.9%)	13.5% (3.0%, 74.8%)
	$C^{(1,2)}$	31.4% (8.8%, 353.3%)	24.2% (7.7%, 61.2%)	26.9% (4.9%, 67.1%)
	Z	0.0957 (0.070, 0.148)	0.083 (0.061, 0.107)	0.068 (0.048, 0.088)
0/0	μ_1	0.23% (0.04%, 1.23%)	0.12% (0.01%, 0.35%)	0.04%(0.01%,0.31%)
	μ_2	0.27% (0.09%,0.88%)	0.12% (0.02%, 0.42%)	0.04% (0.01%, 0.31%)
	$C^{(1,1)}$	3.5% (0.9%, 16.0%)	1.9% (0.3%, 7.4%)	1.3% (0.3%, 4.4%)
	$C^{(2,2)}$	4.5% (0.6%, 18.0%)	1.6% (0.3%, 8.0%)	1.1% (0.2%, 4.5%)
	$C^{(1,2)}$	5.3% (1.1%, 19.9%)	2.0% (0.6%, 9.5%)	1.3% (0.6%, 5.4%)
	Z	0.032 (0.023, 0.049)	0.018 (0.013, 0.024)	0.011 (0.009, 0.015)

never been diagnosed with ASD. The children were instructed to view bubbles on a monitor in a dark, sound-attenuated room for 2 minutes, while EEG recordings ere taken. The EEG recordings were obtained using a 128-channel HydroCel Geodesic Sensory net, and were then interpolated to match the international 10-20 system 25 channel montage. The data were filtered using a band pass of 0.1 to 100 Hz, and then were transformed into the frequency domain using a fast Fourier transform. Lastly, to obtain the relative power, we scaled the functions so that they integrate to 1. Visualizations of the functional data obtained from the T8 electrode can be seen in Figure 4.1.

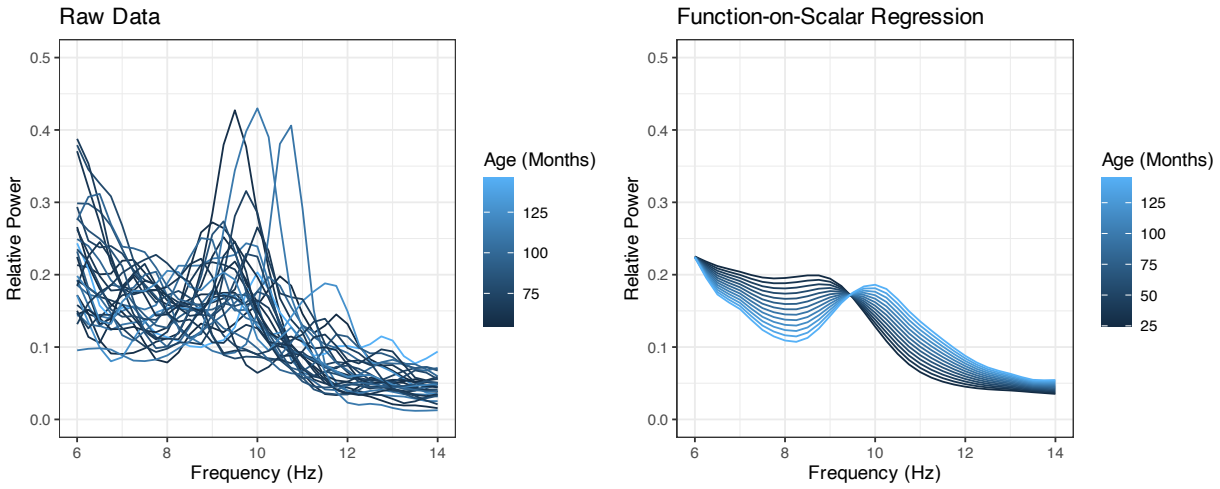


Figure 4.1: (Left Panel) Alpha frequency patterns for a sample of EEG recordings from the T8 electrode of 30 individuals (ASD and TD) with varying ages. (Right Panel) Estimated affects of age on alpha oscillations obtained by fitting a function-on-scalar model.

In this case study, we will specifically be analyzing the alpha alpha band of frequencies (neural activity between 6Hz and 14Hz), and comparing how alpha oscillations differ between children with ASD and TD children. The alpha band of frequencies have been shown to play a role in neural coordination and communication between distributed brain regions [Fries, 2005, Klimesch et al., 2007]. Alpha oscillations are composed of periodic and aperiodic neural activity patterns that coexist in the EEG spectra. Neuroscientists are primarily interested in the periodic signals, specifically in the location of a single prominent peak in the spectral

density located in the alpha band of frequencies, called the *peak alpha frequency* (PAF). The peak alpha frequency has shown to be a biomarker of neural development in typically developing children [Rodríguez-Martínez et al., 2017]. Studies have shown that the alpha peak becomes more prominent and shifts to a higher frequency within the first years of life for TD children [Rodríguez-Martínez et al., 2017, Scheffler et al., 2019]. Compared to TD children, the emergence of a distinct alpha peak and developmental shifts have been shown to be atypical in children with ASD [Dickinson et al., 2018, Scheffler et al., 2019, Marco et al., 2022b,c].

In Section 4.4.1, we conduct a formal analysis on how age affects alpha oscillations by fitting a covariate adjusted functional mixed membership model with age as the only covariate. This analysis is extended in Section 4.4.2 where an analysis on how developmental shifts differ based on diagnostic group. To conduct this analysis, a covariate adjusted functional mixed membership mode was fit with age, diagnostic group, and an interaction between age and diagnostic group as the covariates.

4.4.1 Alpha Oscillations Stratified by Age

In this study, we will primarily be looking at the T8 electrode, which is located in the right temporal region of the scalp. By using the T8 electrode, we will directly be able to compare the results from our covariate adjusted model to the results found in Marco et al. [2022b], which used a non-adjusted functional mixed membership model on this data. From Figure 4.2, we can see that the first feature mainly consists of aperiodic neural activity patterns, which are commonly referred to as a $1/f$ trend or pink noise. The second feature on the other hand can be interpreted as a distinct alpha peak, which is considered a periodic neural activity. We can see that as children that load heavily on the second feature age, the alpha peak becomes larger in magnitude and the PAF shifts to a higher frequency, which has been observed in many other studies [Haegens et al., 2014, Rodríguez-Martínez et al., 2017, Scheffler et al., 2019]. As stated in Haegens et al. [2014], this shift in PAF can confound

the measures of alpha power, thus demonstrating the need for a covariate adjusted model. From a clinical perspective, it is also valuable to compare children's alpha power conditional on age since we know that are developmental changes in alpha oscillations as children age. From Figure 4.2, we can also see that on average children with ASD have a less attenuated alpha peaks when compared to their age adjusted TD counterpart.

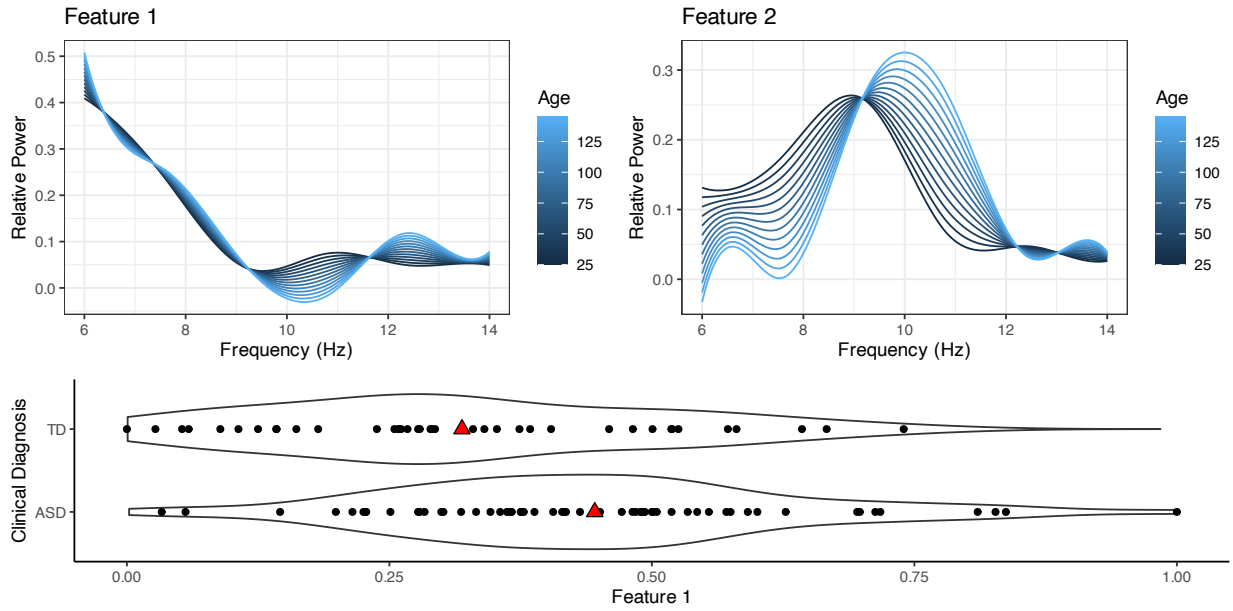


Figure 4.2: (Top Panels) Estimates of the mean functions of the two functional features conditional on Age. (Bottom Panel) Estimates of the allocation parameters found by fitting a covariate adjusted functional mixed membership model. Diagnostic group level means of allocation to the first feature is depicted as a red triangle.

In this analysis, we assume that the alpha oscillations of children with ASD and TD children can be represented as a continuous mixture of the same two features shown in Figure 4.2. Just as shifts in PAF can confound the measures of alpha power [Haegens et al., 2014], this assumption can also confound the results found in this section if the assumption is shown to be incorrect. In Section 4.4.2, we relax this assumption and allow for the features to differ based on diagnostic group.

4.4.2 Alpha Oscillations Stratified by Age and Diagnostic Group

Previous studies have shown that both the emergence of alpha peaks and the developmental shifts in frequency are atypical in children with ASD [Dickinson et al., 2018, Scheffler et al., 2019, Marco et al., 2022b,c]. Therefore, assuming that the alpha oscillations of children with ASD and TD children can be represented by the same two features may not be realistic. In this section, we will be fitting a covariate adjusted functional mixed membership model with age, clinical diagnosis, and an interaction of age and clinical diagnosis as the covariates of interest. By including the interaction between age and diagnostic group, we allow for differences in the developmental changes of alpha oscillations between diagnostic groups.

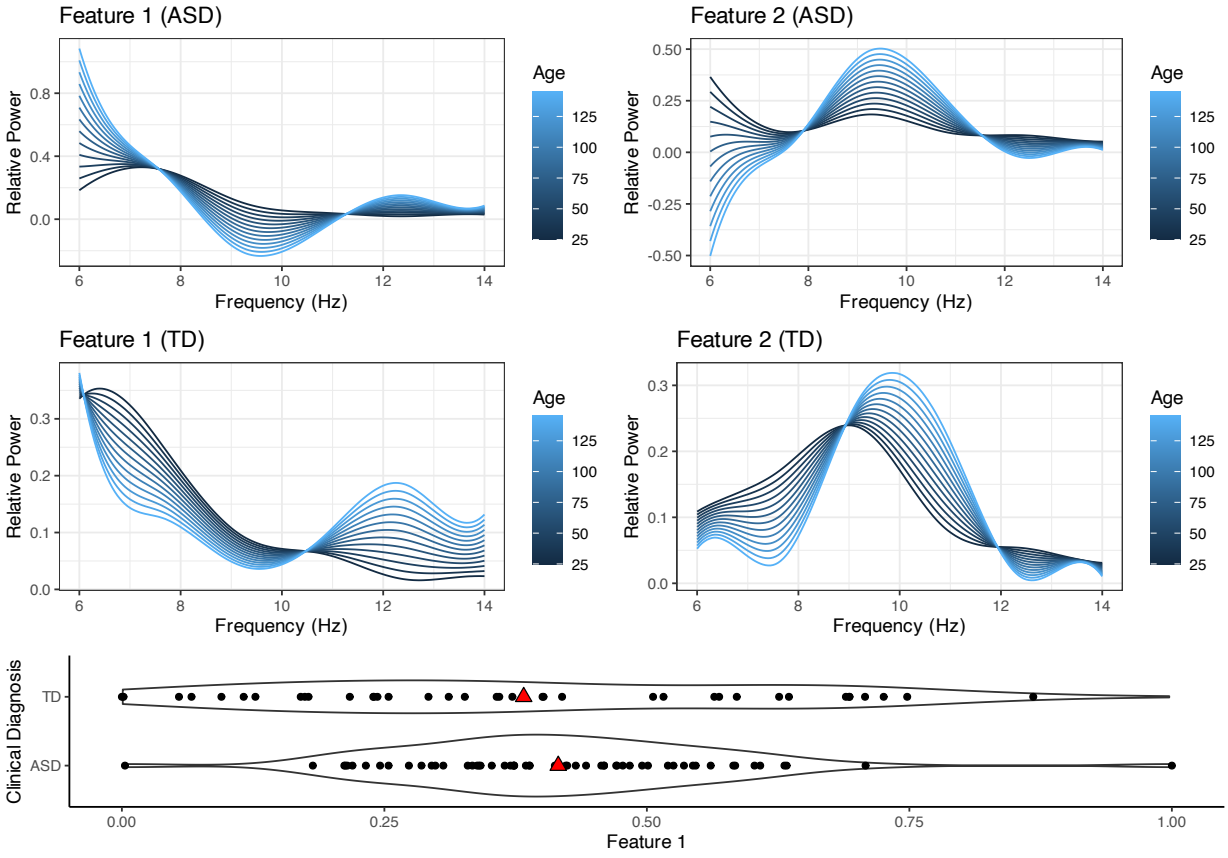


Figure 4.3: (Top Panels) Estimates of the mean functions of the first two features for ASD children conditional on Age. (Middle Panels) Estimates of the mean functions of the first two features for TD children conditional on Age. (Bottom panel) Estimates of the allocation parameters by clinical Diagnosis, where the red triangles depict the group level means.

By fitting a covariate adjusted functional mixed membership model with age and diagnostic group, we can see that the two features seems to greatly differ between children with ASD and TD children (Figure 4.3). TD children that load heavily on feature 2 have a relatively distinct alpha peak and tend to have a shift in PAF to a higher magnitude as they age, reaffirming results found in the neuroscience literature [Haegens et al., 2014, Rodríguez-Martínez et al., 2017, Scheffler et al., 2019]. Alternatively, TD children that load heavily on feature 1 tend to have more aperiodic alpha frequencies and do not develop a prominent alpha peak even as they age. Compared to the two features found for TD children, the two features found for ASD children are more abstruse. Similarly to the second feature for TD children, the second feature for children with ASD can be interpreted as individuals with a distinct alpha peak. Compared to the second feature of TD children, we can see that the second feature's alpha peak is less prominent and there is no sign of a developmental shift in PAF as children age. The second feature for children with ASD can also be interpreted as a signal with mainly aperiodic neural activity, with no discernible alpha peak. We can see that the changes to feature 1 as children with ASD age seem to directly oppose the changes to feature 2 as children with ASD age. This phenomenon leads to sample paths that are less featured in children with ASD that have roughly equal weighting of both features. This is apparent in Figure 4.4, where the average alpha oscillations stratified by developmental group are shown for various ages. This figure was created by using the estimated sample path for individuals with the group average allocations (depicted by the red triangles in Figure 4.3). From Figure 4.3, we can see that on average, children with ASD have a stronger $1/f$ trend and a less attenuated alpha peak than their age-adjusted TD counterpart.

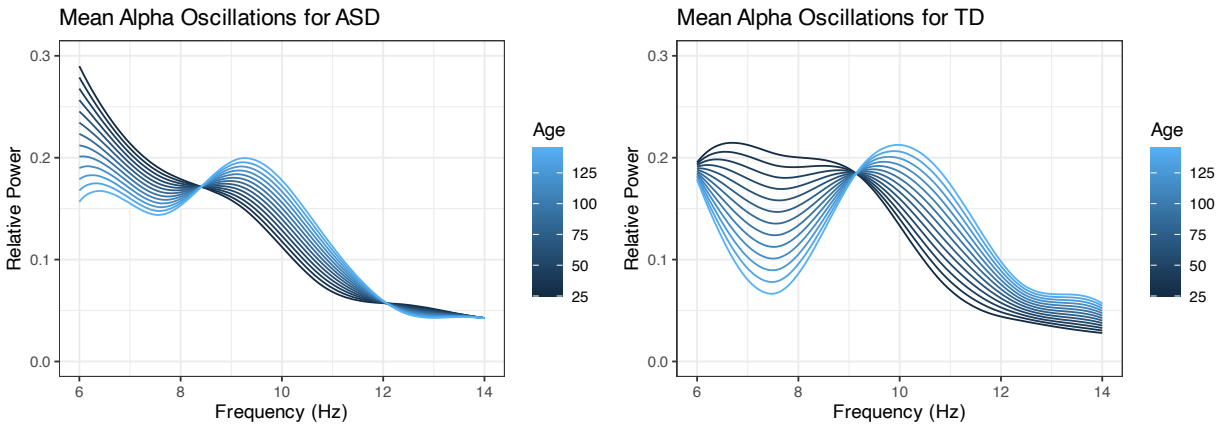


Figure 4.4: Estimated average developmental trajectory of alpha oscillations stratified by diagnostic group.

As discussed in Section 4.2.4, the proposed covariate adjusted mixed membership model can be thought of as a generalization of function-on-scalar regression. Figure 4.5 contains the estimated developmental trajectories obtained by fitting a function-on-scalar regression model using the package "refund" package in R [Goldsmith et al., 2016]. The results from function-on-scalar regression coincide with the estimated average developmental trajectory of alpha oscillations obtained from our covariate adjusted mixed membership model visualized in Figure 4.4. However, the function-on-scalar analysis does not allow practitioners to conduct analysis at an individual level. Compared to function-on-scalar regression, covariate adjusted mixed membership models allow practitioners to quantitatively compare the alpha oscillations of an individual to peers of the same age. Moreover, covariate adjusted mixed membership models are able to specify individualized predictions of the estimated changes of the alpha oscillations based off of an individual's estimate allocation parameters. Overall, the added flexibility of covariate adjusted mixed membership models allows scientists to have greater insight into the developmental changes of alpha oscillations when compared to function-on-scalar regression.

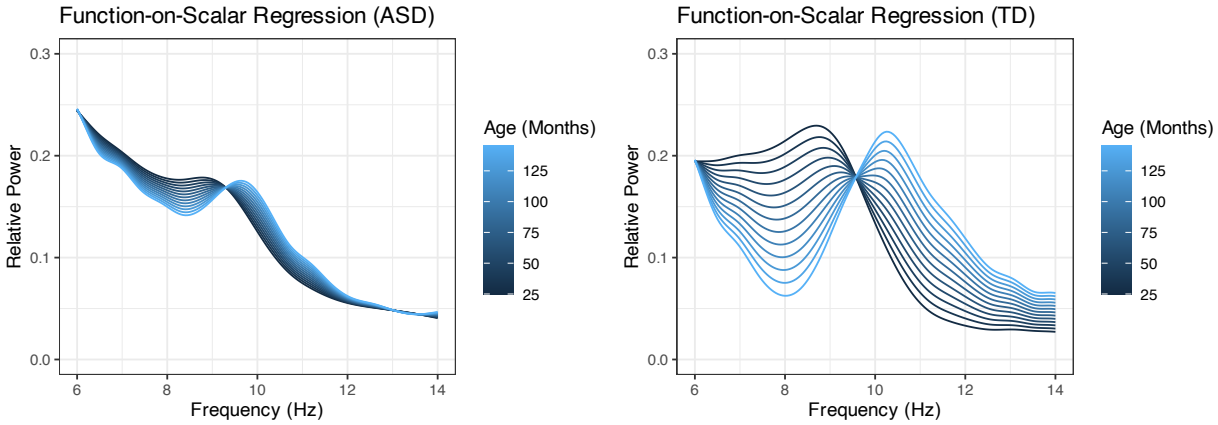


Figure 4.5: Estimated population level developmental trajectories stratified by diagnostic group, obtained by fitting a function-on-scalar regression model. The model included age, diagnostic group, and an interaction between age and diagnostic group as the covariates of interest.

4.4.3 Comparison Between Mixed Membership Models

In this section, we extended the analysis on alpha oscillations conducted by Marco et al. [2022b] to allow for a covariate depended mixed membership model. While previous studies [Haegens et al., 2014, Rodríguez-Martínez et al., 2017, Scheffler et al., 2019] have shown that alpha oscillations are dependent on age, little is known about how alpha oscillations differ between children with ASD and TD children conditional on age. While it is enticing to add more covariates to our model, the small sample sizes often found in neurodevelopmental studies limit our ability to fit models with a large amount of covariates. Thus in order to prevent having overfit models, we can perform cross-validated methods such as conditional predictive ordinates (CPO) [Pettit, 1990, Chen et al., 2012, Lewis et al., 2014]. CPO for our model can be defined as $P\left(\mathbf{Y}_i(\mathbf{t}_i) \mid \{\mathbf{Y}_j(\mathbf{t}_j)\}_{j\neq i}\right)$. Unlike traditional cross-validation methods, CPO requires no extra sampling to be conducted. Following Chen et al. [2012] and Lewis et al. [2014], an estimate of CPO for our model can be obtained through the following

MCMC approximation:

$$\widehat{CPO}_i = \left(\frac{1}{N_{MC}} \sum_{r=1}^{N_{MC}} \frac{1}{P\left(\mathbf{Y}_i(\mathbf{t}_i) \mid \hat{\boldsymbol{\Theta}}^r_{-\chi}, \mathbf{x}_i\right)} \right)^{-1}, \tag{4.16}$$

where $\hat{\boldsymbol{\Theta}}^r_{-\chi}$ are the samples from the r^{th} MCMC iteration, N_{MC} are the number of MCMC iterations (not including burn-in), and $P\left(\mathbf{Y}_i(\mathbf{t}_i) \mid \hat{\boldsymbol{\Theta}}^r_{-\chi}, \mathbf{x}_i\right)$ is specified in Equation 4.12. While CPO is a measure of how well the model fits each individual observation, the pseudomarginal likelihood (PML), defined as $\widehat{PML} = \prod_{i=1}^{N} \widehat{CPO}_i$, is an overall measure of how well the model fits the entire dataset. Using CPO and PML, we will compare the unadjusted functional mixed membership model [Marco et al., 2022b] as well as the two covariate adjusted functional mixed membership models fit in this section.

In this section, we will let M_0 denote the unadjusted functional mixed membership model from Marco et al. [2022b], M_1 denote the age adjusted functional mixed membership model, and M_2 denote the age and diagnostic group adjusted functional mixed membership model. From Figure 4.6, we can see that the age adjusted functional mixed membership model tends to fit the data slightly better than the unadjusted model (M_0 log (PML) = 6543.9, M_1 log (PML) = 6657.9). The covariate adjusted model with age and diagnostic group as covariates had a slightly worse fit than the covariate adjusted model with just age alone, suggesting that more data may be needed in order to conduct such an analysis (M_2 log (PML) = 6616.1). While the fit may be slightly worse for the covariate adjusted model with age and diagnostic group as covariates, this model does give us useful insight into how the two features differ between children with ASD and TD children. Moreover, besides the one TD individual, we can see that the model tends to fit children with ASD worse than TD children, supporting that the alpha oscillations are more irregular compared to TD individuals.

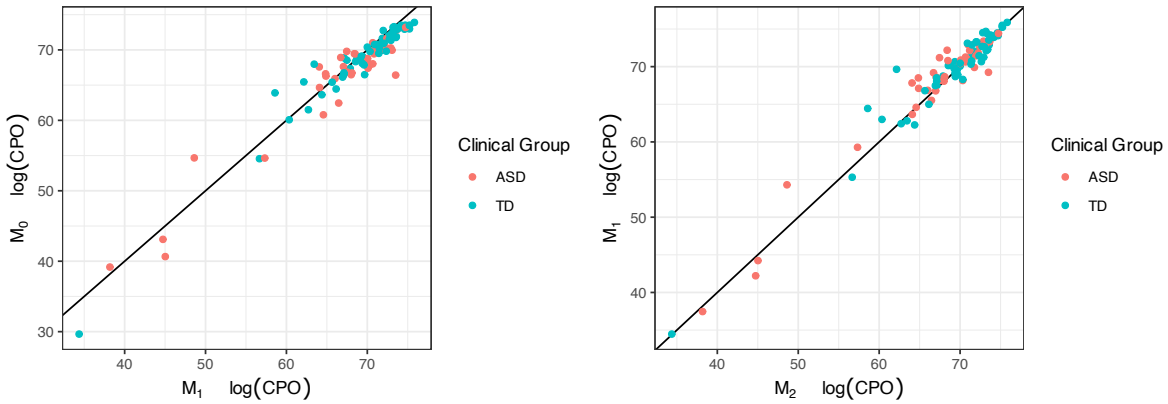


Figure 4.6: CPO comparisons between different models on the log scale. M_0 denotes the unadjusted functional mixed membership model, while M_1 denotes model stratified by age and M_2 denotes the model stratified by age and diagnostic group.

4.5 Conclusion

In this manuscript, we extended the functional mixed membership model framework proposed by Marco et al. [2022b] to allow for covariate dependence. This work was primarily motivated by a neurodevelopmental study on alpha oscillations, where alpha oscillations are known to change as children age. While mixed membership models provided a novel way to quantify the emergence and presence of developmental biomarker known as an alpha peak [Marco et al., 2022b,c], it has been shown that not accounting for developmental shifts in the alpha peak can confound measures of the peak alpha frequency [Haegens et al., 2014], leading to a need for a covariate adjusted functional mixed membership model. In Section 4.2, we derive the covariate adjusted functional mixed membership models and compare the framework to common covariate-dependent clustering frameworks such as mixture of experts and mixture of regressions models. Once we completely specify the model and priors, we provide a set of sufficient conditions in Lemma 7 to ensure that a two feature covariate adjusted mixed membership model has identifiable mean, covariance, and allocation structures. Section 4.3 contains a simulation study exploring the empirical convergence properties of our model. We conclude by revisiting the neurodevelopmental case study on alpha oscillations

in Section 4.4, where we use covariate adjusted functional mixed membership models to gain novel insight into developmental differences of alpha oscillations between children with ASD and TD children.

While we primarily focus on a mixed membership model where the covariate dependence encoded in the mean structure, we can encode covariate dependence into the covariance structure by introducing covariate-dependent pseudo-eigenfunctions, as discussed in Section C.4.1 of the Supplementary Materials. While formulating such a model is relatively straightforward, the amount of data needed to fit such a model would be relatively large compared to just model where the mean depends on the covariates of interest. This is supported by the simulation study conducted in Section 4.3, where even with 200 functional observations, a covariate adjusted mixed membership model with one covariate (with no covariate dependence on the covariance structure) has relatively high R-MISE for the covariance and cross-covariance functions. Therefore, in order to justify using a functional mixed membership that encodes covariate dependence into the mean and covariance structure, one would not only need a lot of data, but also a justification for sacrificing statistical efficiency in return for a more expressive covariance structure.

While the framework described in this manuscript works for any number of functional features, the guarantees for model identifiability described in Lemma 7 only hold for models with 2 functional features. Chen et al. [2022] describes in great detail the identifiability challenges that exists in mixed membership models, especially when considering mixed membership models with 3 or more features. One way to ensure identifiability is to assume the separability condition [Pettit, 1990, Donoho and Stodden, 2003, Arora et al., 2012, Azar et al., 2001, Chen et al., 2022], which assumes that for each feature, there exists at least one observation that completely belongs to that feature. In the setting of a two feature model, this is very weak assumption that does not impact the flexibility of the model. However, in models with 3 or more features, this assumption makes strong geometric assumptions on the model. However, if we are able to assume the separability condition, a generalization of

Lemma 7 can be derived, but we will need at least $\frac{k^2-k}{2}+1$ observations with allocation parameters that lie in the interior of the unit simplex in order to ensure identifiability. Weaker assumptions utilizing the work of Chen et al. [2022] can be made to ensure identifiability, however an extension to the proposed work is non-trivial. While choosing the number of features is one of the main challenges when fitting a mixed membership model, Marco et al. [2022b] and Marco et al. [2022c] discuss how information criteria can be informative in choosing the correct number of clusters. Similarly, definitions of information criteria, such as BIC, or simple heuristics such as the "elbow-method" can be informative in choosing the number of features in the covariate adjusted model. Conducting an unadjusted analysis prior to a covariate adjusted analysis can also be informative in choosing the number of latent features. Most importantly, we maintain that the most interpretable model is the optimal model, as covariate adjusted functional mixed membership models are an unsupervised technique that aims to explain data heterogeneity conditional on covariates of interest.

As observed in unadjusted mixed membership models [Marco et al., 2022b,c], the posterior distribution often has multiple modes, leading to poor posterior exploration using traditional sampling methods. Thus we use a similar algorithm to Algorithm 1 described in the supplement of Marco et al. [2022b] to pick a good starting point for our Markov chain. In addition to finding good initial starting points, we also outline a tempered-transition sampling scheme [Pritchard et al., 2000, Behrens et al., 2012] in Section 2.2 of the Supplementary Materials, which allows us to traverse areas of low posterior probability. An R package for the proposed covariate adjusted functional mixed membership model is available for download at `https://github.com/ndmarco/BayesFMMM`.

CHAPTER 5

Future Extensions

Mixed membership models can be thought of as generalizations of traditional clustering models such as finite mixture models, where membership is thought of as a spectrum rather than a binary variable. Due to the added flexibility of mixed membership models, ensuring identifiability becomes more challenging. One way to ensure identifiability is to assume the *separability condition*, which assumes that for each feature at least one observation belongs completely to that feature. In the case when we only have two features, assuming the separability condition does not make any major assumptions on the sampling distribution. However, in cases when we have three or more features, the separability condition makes strong geometric assumptions on the sampling distribution. Chen et al. [2022] discusses weaker geometric assumptions that also ensure an identifiable model. While the assumptions are weaker, implementing them and performing Markov chain Monte Carlo under those constraints is not trivial. While challenging, the implementation of these constraints into a mixed membership model would be a significant step forward in the mixed membership literature from both an application perspective, as well as a theoretical perspective. From a theoretical perspective, we would have an identifiable model, which is a necessary step in ensuring posterior consistency without conditioning on the allocation variables. Once identifiability is established, research can also be conducted on theoretical results such as posterior contraction rates.

In this dissertation, we proposed mixed membership models for multivariate data and functional data (for both one-dimensional functions and higher-dimensional functions). One

type of data structure that often arises in neurodevelopmental studies is data that is both longitudinal and functional. One way this type data could arise would be from EEG experiments where a subject will visit multiple times over a certain time period, and at each visit an EEG recording was obtained. While our model can handle functional data, our model does not account for the structured dependence created from repeat measurements of the same subject, leading to a need for a mixed membership models that can handle longitudinal functional data. The proposed models can be also be extended to handle a combination of functional data and multivariate Gaussian data. Similarly to Happ and Greven [2018], we can obtain a eigen decomposition of the combination of functional data and multivariate data by first jointly representing the data as a direct sum of vector spaces. Once the joint representation is formalized, an eigen decomposition should be obtainable, allowing for the specification of a mixed membership model.

In Chapter 4, we derived a covariate adjusted mixed membership model where the covariates of interest were scalars and the data which we wanted to learn the allocation structure of were functional. Similar models can be formulated where the covariates of interest are functional. Instead of leveraging the literature of function-on-scalar regression, we would leverage the vast amount of work on function-on-function regression. When dealing with function-on-function regression, we assume that

$$Y(t) = \mu(t) + \sum_{r=1}^{R} \int_{s \in \mathcal{T}_{x_r}} X_r(s)\beta_r(t,s)\mathrm{d}s + \epsilon(t); \quad t \in \mathcal{T},$$

where $\beta_r(t,s)$ is the r^{th} coefficient surface and $\mathcal{T}_{x_r}$ is the domain of the function $X_r(\cdot)$. Similarly to Chapter 4, we can assume the basis functions can be represented by basis functions. By leveraging the function-on-function regression framework, a covariate adjusted mixed membership model can be formulated for covariates which are functional. Similarly, covariate adjusted mixed membership models for multivariate data can be formulated, and are implemented in the "BayesFMMM" package (`https://github.com/ndmarco/BayesFMMM`).

APPENDIX A

Appendix: Flexible Regularized Estimation in High-Dimensional Mixed Membership Models

A.1 Proofs

A.1.1 Proof of Lemma 2.1

We will start be explicitly defining the functions $\Lambda_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})$, $K_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})$, and $V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})$. Thus we have

$$\begin{aligned}
\Lambda_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) =& \log\left(\frac{|(\boldsymbol{\Sigma}_i)_0|^{-1/2}\exp\left\{-\frac{1}{2}\left(\mathbf{y}_i-(\boldsymbol{\mu}_i)_0\right)'(\boldsymbol{\Sigma}_i)_0^{-1}\left(\mathbf{y}_i-(\boldsymbol{\mu}_i)_0\right)\right\}}{|\boldsymbol{\Sigma}_i|^{-1/2}\exp\left\{-\frac{1}{2}\left(\mathbf{y}_i-\boldsymbol{\mu}_i\right)'(\boldsymbol{\Sigma}_i)^{-1}\left(\mathbf{y}_i-\boldsymbol{\mu}_i\right)\right\}}\right) \\
=& -\frac{1}{2}\left[\log\left(|(\boldsymbol{\Sigma}_i)_0|\right)-\log\left(|\boldsymbol{\Sigma}_i|\right)\right] \\
& -\frac{1}{2}\left[\left(\mathbf{y}_i-(\boldsymbol{\mu}_i)_0\right)'(\boldsymbol{\Sigma}_i)_0^{-1}\left(\mathbf{y}_i-(\boldsymbol{\mu}_i)_0\right)-\left(\mathbf{y}_i-\boldsymbol{\mu}_i\right)'(\boldsymbol{\Sigma}_i)^{-1}\left(\mathbf{y}_i-\boldsymbol{\mu}_i\right)\right] \\
=& -\frac{1}{2}\left[\sum_{l=1}^{P}\log\left((d_{il})_0+\sigma_0^2\right)-\log\left(d_{il}+\sigma^2\right)\right] \\
& -\frac{1}{2}\left[\left(\mathbf{y}_i-(\boldsymbol{\mu}_i)_0\right)'(\boldsymbol{\Sigma}_i)_0^{-1}\left(\mathbf{y}_i-(\boldsymbol{\mu}_i)_0\right)-\left(\mathbf{y}_i-\boldsymbol{\mu}_i\right)'(\boldsymbol{\Sigma}_i)^{-1}\left(\mathbf{y}_i-\boldsymbol{\mu}_i\right)\right] \qquad \text{(A.1)}
\end{aligned}$$

$$
\begin{aligned}
K_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) =& -\frac{1}{2}\left[\sum_{l=1}^{P} \log\left((d_{il})_0 + \sigma_0^2\right) - \log\left(d_{il} + \sigma^2\right)\right] \\
& -\frac{1}{2}\mathbb{E}_{\boldsymbol{\omega}_0}\left[(\mathbf{y}_i - (\boldsymbol{\mu}_i)_0)' (\boldsymbol{\Sigma}_i)_0^{-1} (\mathbf{y}_i - (\boldsymbol{\mu}_i)_0) - (\mathbf{y}_i - \boldsymbol{\mu}_i)' (\boldsymbol{\Sigma}_i)^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)\right] \\
=& -\frac{1}{2}\left[\sum_{l=1}^{P} \log\left((d_{il})_0 + \sigma_0^2\right) - \log\left(d_{il} + \sigma^2\right)\right] \\
& -\frac{1}{2}\left[P - \left(\text{tr}\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Sigma}_i)_0\right) + ((\boldsymbol{\mu}_i)_0 - \boldsymbol{\mu}_i)' (\boldsymbol{\Sigma}_i)^{-1} ((\boldsymbol{\mu}_i)_0 - \boldsymbol{\mu}_i)\right)\right] \qquad \text{(A.2)}
\end{aligned}
$$

$$
\begin{aligned}
V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) =& \frac{1}{4}\text{Var}_{\boldsymbol{\omega}_0}\left[(\mathbf{y}_i - (\boldsymbol{\mu}_i)_0)' (\boldsymbol{\Sigma}_i)_0^{-1} (\mathbf{y}_i - (\boldsymbol{\mu}_i)_0) - (\mathbf{y}_i - \boldsymbol{\mu}_i)' (\boldsymbol{\Sigma}_i)^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)\right] \\
=& \frac{1}{4}\text{Var}_{\boldsymbol{\omega}_0}\left[\mathbf{y}_i'\left((\boldsymbol{\Sigma}_i)_0^{-1} + \boldsymbol{\Sigma}_i^{-1}\right)\mathbf{y}_i - 2\mathbf{y}_i'\left((\boldsymbol{\Sigma}_i)_0^{-1}(\boldsymbol{\mu}_i)_0 + \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i\right)\right]
\end{aligned}
$$

Letting $\mathbf{M}_v = (\boldsymbol{\Sigma}_i)_0^{-1} + \boldsymbol{\Sigma}_i^{-1}$, and $\mathbf{m}_v = (\boldsymbol{\Sigma}_i)_0^{-1}(\boldsymbol{\mu}_i)_0 + \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$, we have

$$
\begin{aligned}
V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) =& \frac{1}{4}\text{Var}_{\boldsymbol{\omega}_0}\left[(\mathbf{y}_i - \mathbf{M}_v^{-1}\mathbf{m}_v)'\mathbf{M}_v(\mathbf{y}_i - \mathbf{M}_v^{-1}\mathbf{m}_v)\right] \\
=& \frac{1}{4}\left[2\text{tr}\left(\mathbf{M}_v(\boldsymbol{\Sigma}_i)_0\mathbf{M}_v(\boldsymbol{\Sigma}_i)_0\right) + 4\left((\boldsymbol{\mu}_i)_0 - \mathbf{M}_v^{-1}\mathbf{m}_v\right)'(\boldsymbol{\Sigma}_i)_0\left((\boldsymbol{\mu}_i)_0 - \mathbf{M}_v^{-1}\mathbf{m}_v\right)\right] \\
=& \frac{1}{2}\left[P + 2\text{tr}\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Sigma}_i)_0\right) + \text{tr}\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Sigma}_i)_0\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Sigma}_i)_0\right)\right] \\
& + ((\boldsymbol{\mu}_i)_0 - \boldsymbol{\mu}_i)'\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Sigma}_i)_0\boldsymbol{\Sigma}_i^{-1}\right)((\boldsymbol{\mu}_i)_0 - \boldsymbol{\mu}_i). \qquad \text{(A.3)}
\end{aligned}
$$

Let $\boldsymbol{\Omega}_\epsilon(\boldsymbol{\omega}_0) = \{\boldsymbol{\omega} : K_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) < \epsilon \text{ for all i}\}$ for some $\epsilon > 0$. We will assume that $\sigma_0^2 > 0$. Consider the set $\mathcal{B}(\boldsymbol{\omega}_0) = \{\boldsymbol{\omega} : \frac{1}{a}((d_{il})_0 + \sigma_0^2) \leq d_{il} + \sigma^2 \leq a((d_{il})_0 + \sigma_0^2), \|(\boldsymbol{\mu}_i)_0 - \boldsymbol{\mu}_i\| \leq b\}$ for some $a, b \in \mathbb{R}$ such that $a > 1$ and $b > 0$. Thus for a fixed $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$ and any $\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon) := \mathcal{B}(\boldsymbol{\omega}_0) \cap \boldsymbol{\Omega}_\epsilon(\boldsymbol{\omega}_0)$, we can bound $V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})$. We will let $\lambda_r(\mathbf{A})$ denote the r^{th} eigenvalue of the matrix $\mathbf{A}$, and $\lambda_{max}(\mathbf{A})$ denote the largest eigenvalue of $\mathbf{A}$. Thus we have

$$
\text{tr}\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Sigma}_i)_0\right) \leq P\lambda_{max}\left(\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{\Sigma}_i)_0\right) \leq \frac{Pa}{\sigma_0^2}\left(\max_l(d_{il} + \sigma_0^2)\right)
$$

$$\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\right) \leq \operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\right)^2 \leq \left(\frac{Pa}{\sigma_0^2}\left(\max_l(d_{il}+\sigma_0^2)\right)\right)^2$$

$$\begin{aligned}\left(\left(\boldsymbol{\mu}_i\right)_0-\boldsymbol{\mu}_i\right)'\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\boldsymbol{\Sigma}_i^{-1}\right)\left(\left(\boldsymbol{\mu}_i\right)_0-\boldsymbol{\mu}_i\right) &\leq b^2\lambda_{max}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\boldsymbol{\Sigma}_i^{-1}\right)\\ &\leq \frac{a^2b^2}{\sigma_0^4}\max_l(d_{il}+\sigma_0^2)\end{aligned}$$

Thus we can see that for any $\boldsymbol{\omega}\in\mathcal{C}(\boldsymbol{\omega}_0,\epsilon)$,

$$\begin{aligned}V_i(\boldsymbol{\omega}_0,\boldsymbol{\omega}) \leq& \frac{1}{2}\left[P+2\left(\frac{Pa}{\sigma_0^2}\left(\max_l(d_{il}+\sigma_0^2)\right)\right)+\left(\frac{Pa}{\sigma_0^2}\left(\max_l(d_{il}+\sigma_0^2)\right)\right)^2\right]\\ &+\frac{a^2b^2}{\sigma_0^4}\max_l(d_{il}+\sigma_0^2)\\ =& M_V.\end{aligned}$$

If we can bound $\lambda_{max}\left((d_{il})_0+\sigma_0\right)$, then we have that $V_i(\boldsymbol{\omega}_0,\boldsymbol{\omega})$ is bounded. Let $\|\cdot\|_F$ be the Frobenius norm. Using the triangle inequality, we have

$$\begin{aligned}\|(\boldsymbol{\Sigma}_i)_0\|_F &\leq \sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{p=1}^{KP} Z_{ij}Z_{ik}\|(\boldsymbol{\phi}_{kp})_0(\boldsymbol{\phi}_{jp})_0'\|_F+\sigma_0^2\|\mathbf{I}_P\|_F\\ &\leq \sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{p=1}^{KP}\|(\boldsymbol{\phi}_{kp})_0(\boldsymbol{\phi}_{jp})_0'\|_F+\sigma_0^2\|\mathbf{I}_P\|_F\\ &= \sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{p=1}^{KP}\sqrt{\operatorname{tr}\left((\boldsymbol{\phi}_{jp})_0(\boldsymbol{\phi}_{kp})_0'(\boldsymbol{\phi}_{kp})_0(\boldsymbol{\phi}_{jp})_0'\right)}+\sqrt{P}\sigma_0^2\\ &= \sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{p=1}^{KP}\sqrt{\operatorname{tr}\left(\boldsymbol{\phi}_{kp}'(\boldsymbol{\phi}_{kp})_0(\boldsymbol{\phi}_{jp})_0'(\boldsymbol{\phi}_{jp})_0\right)}+\sqrt{P}\sigma_0^2\\ &= \sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{p=1}^{KP}\|(\boldsymbol{\phi}_{jp})_0\|_2\|(\boldsymbol{\phi}_{kp})_0\|_2+\sqrt{P}\sigma_0^2\\ &= M_{\boldsymbol{\Sigma}_0}<\infty,\end{aligned}$$

for all $i \in \mathbb{N}$. Therefore, we know that $\lambda_{max}\left((d_{il})_0 + \sigma_0\right) \leq M_{\boldsymbol{\Sigma}_0}$, as the Frobenius is the squareroot of the sum of the squared eigenvalues for a square matrix. Therefore, we have for all $i \in \mathbb{N}$ and $\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)$, we have that

$$\frac{V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})}{i^2} \leq \frac{M_V}{i^2}.$$

Since $\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$, we have that $\sum_{i=1}^{\infty} \frac{M_V}{i^2} = \frac{M_V \pi^2}{6} < \infty$. Thus we have

$$\sum_{i=1}^{\infty} \frac{V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})}{i^2} < \infty. \tag{A.4}$$

We will next show that for $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$ and $\epsilon > 0$, $\boldsymbol{\Pi}(\mathcal{C}(\boldsymbol{\omega}_0), \epsilon) > 0$. Fix $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$. While the $(\boldsymbol{\phi}_{jp})_0$ may not be identifiable (for any orthogonal matrix $\mathbf{H}$, $(\boldsymbol{\phi}_{jp})_0 \mathbf{H}\mathbf{H}'(\boldsymbol{\phi}_{kp})_0 = (\boldsymbol{\phi}_{jp})_0'(\boldsymbol{\phi}_{kp})_0$), let $(\boldsymbol{\phi}_{jp})_0$ be such that $\sum_{p=1}^{KP}(\boldsymbol{\phi}_{jp})_0'(\boldsymbol{\phi}_{kp})_0 = (\boldsymbol{\Sigma}_{jk})_0$. Thus we can define the following sets:

$$\begin{aligned}
\boldsymbol{\Omega}_{\boldsymbol{\phi}_{jp}} &= \left\{\boldsymbol{\phi}_{jp} : (\boldsymbol{\phi}_{jp})_0 \leq \boldsymbol{\phi}_{jp} \leq (\boldsymbol{\phi}_{jp})_0 + \epsilon_1 \mathbf{1}\right\} \\
\boldsymbol{\Omega}_{\boldsymbol{\nu}_k} &= \left\{\boldsymbol{\nu}_k : (\boldsymbol{\nu}_k)_0 \leq \boldsymbol{\nu}_k \leq (\boldsymbol{\nu}_k)_0 + \epsilon_2 \mathbf{1}\right\} \\
\boldsymbol{\Omega}_{\sigma^2} &= \left\{\sigma^2 : \sigma_0^2 \leq \sigma^2 \leq (1+\epsilon_1)\sigma_0^2\right\}.
\end{aligned}$$

We define $\boldsymbol{\epsilon}_{1jp}$ and $\boldsymbol{\epsilon}_{2k}$ such that each element of $\boldsymbol{\epsilon}_{1jp}$ is between 0 and ϵ_1, and each element of $\boldsymbol{\epsilon}_{2k}$ is between 0 and ϵ_2. Therefore $(\boldsymbol{\phi}_{jp})_0 + \boldsymbol{\epsilon}_{1jp} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}_{jp}}$ and $(\boldsymbol{\nu}_k)_0 + \boldsymbol{\epsilon}_{2k} \in \boldsymbol{\Omega}_{\boldsymbol{\nu}_k}$. We will define

$$\boldsymbol{\Omega}_{\boldsymbol{\Sigma}_{jk}} := \left\{ \sum_{p=1}^{KP} \boldsymbol{\phi}_{jp}' \boldsymbol{\phi}_{kp} \middle| \boldsymbol{\phi}_{jp} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}_{jp}}, \boldsymbol{\phi}_{kp} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}_{kp}} \right\}.$$

Thus for $\boldsymbol{\Sigma}_i$ such that $\boldsymbol{\phi}_{jp} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}_{jp}}$ and $\sigma^2 \in \boldsymbol{\Omega}_{\sigma^2}$, we have that

$$\begin{aligned}
\boldsymbol{\Sigma}_i =& \sum_{k=1}^{K}\sum_{j=1}^{K} Z_{ik}Z_{ij}\left(\sum_{p=1}^{KP}\left(\left((\boldsymbol{\phi}_{kp})_0 + \boldsymbol{\epsilon}_{1kp}\right)\left((\boldsymbol{\phi}_{jp})_0 + \boldsymbol{\epsilon}_{1jp}\right)'\right)\right) + (1+\epsilon_\sigma)\sigma_0^2\mathbf{I}_P \\
=&(\boldsymbol{\Sigma}_i)_0 + \sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{p=1}^{KP} Z_{ik}Z_{ij}\left(\left(\boldsymbol{\epsilon}_{1kp}\right)\left(\boldsymbol{\phi}_{jp}\right)_0'\right) \\
&+ \sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{p=1}^{KP} Z_{ik}Z_{ij}\left(\left(\boldsymbol{\phi}_{kp}\right)_0\left(\boldsymbol{\epsilon}_{1jp}\right)'\right) \\
&+ \sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{p=1}^{KP} Z_{ik}Z_{ij}\left(\left(\boldsymbol{\epsilon}_{1kp}\right)\left(\boldsymbol{\epsilon}_{1jp}\right)'\right) + \epsilon_\sigma\sigma_0^2\mathbf{I}_P \\
=&(\boldsymbol{\Sigma}_i)_0 + \tilde{\boldsymbol{\Sigma}}_i,
\end{aligned}$$

for some ϵ_{kp} and ϵ_σ such that $0 < \epsilon_\sigma \leq \epsilon_1$. Thus, letting $\boldsymbol{\zeta}_{jkp} = \left(\left(\boldsymbol{\epsilon}_{1kp}\right)\left(\boldsymbol{\phi}_{jp}\right)_0'\right) + \left(\left(\boldsymbol{\phi}_{kp}\right)_0\left(\boldsymbol{\epsilon}_{1jp}\right)'\right)$, we have

$$\begin{aligned}
\left|\left|Z_{ik}Z_{ij}\boldsymbol{\zeta}_{jkp}\right|\right|_F^2 \leq& \left|\left|\boldsymbol{\zeta}_{jkp}\right|\right|_F^2 \\
=&\mathrm{tr}\left(\left(\left(\boldsymbol{\epsilon}_{1kp}\right)\left(\boldsymbol{\phi}_{jp}\right)_0'\right)\left(\left(\boldsymbol{\phi}_{jp}\right)_0\left(\boldsymbol{\epsilon}_{1kp}\right)'\right)\right) \\
&+ \mathrm{tr}\left(\left(\left(\boldsymbol{\epsilon}_{1kp}\right)\left(\boldsymbol{\phi}_{jp}\right)_0'\right)\left(\left(\boldsymbol{\epsilon}_{1jp}\right)\left(\boldsymbol{\phi}_{kp}\right)_0'\right)\right) \\
&+ \mathrm{tr}\left(\left(\left(\boldsymbol{\phi}_{kp}\right)_0\left(\boldsymbol{\epsilon}_{1jp}\right)'\right)\left(\left(\boldsymbol{\phi}_{jp}\right)_0\left(\boldsymbol{\epsilon}_{1kp}\right)'\right)\right) \\
&+ \mathrm{tr}\left(\left(\left(\boldsymbol{\phi}_{kp}\right)_0\left(\boldsymbol{\epsilon}_{1jp}\right)'\right)\left(\left(\boldsymbol{\epsilon}_{1jp}\right)\left(\boldsymbol{\phi}_{kp}\right)_0'\right)\right) \\
\leq&\epsilon_1^2\mathrm{tr}\left(\left(\boldsymbol{\phi}_{jp}\right)_0'\left(\boldsymbol{\phi}_{jp}\right)_0\mathbf{1}'\mathbf{1}\right) \\
&+ 2\mathrm{tr}\left(\left(\boldsymbol{\phi}_{jp}\right)_0'\left(\boldsymbol{\epsilon}_{1jp}\right)\left(\boldsymbol{\phi}_{kp}\right)_0'\left(\boldsymbol{\epsilon}_{1kp}\right)\right) \qquad \text{(A.5)} \\
&+ \epsilon_1^2\mathrm{tr}\left(\mathbf{1}'\mathbf{1}\left(\boldsymbol{\phi}_{kp}\right)_0'\left(\boldsymbol{\phi}_{kp}\right)_0\right).
\end{aligned}$$

Using the Cauchy-Schwarz inequality, we can simplify Equation A.5, such that

$$\begin{aligned}(A.5) &= 2\langle(\boldsymbol{\phi}_{jp})_0, \boldsymbol{\epsilon}_{1jp}\rangle\langle(\boldsymbol{\phi}_{kp})_0, \boldsymbol{\epsilon}_{1kp}\rangle \\ &\leq 2\|(\boldsymbol{\phi}_{jp})_0\|_2\|\boldsymbol{\epsilon}_{1jp}\|_2\|(\boldsymbol{\phi}_{kp})_0\|_2\|\boldsymbol{\epsilon}_{1kp}\|_2 \\ &\leq 2\epsilon_1^2 P\|(\boldsymbol{\phi}_{jp})_0\|_2\|(\boldsymbol{\phi}_{kp})_0\|_2.\end{aligned}$$

Letting

$$\tilde{M}_{jkp} = P\left[\|(\boldsymbol{\phi}_{jp})_0\|_2^2 + \|(\boldsymbol{\phi}_{kp})_0\|_2^2 + 2\|(\boldsymbol{\phi}_{jp})_0\|_2\|(\boldsymbol{\phi}_{kp})_0\|_2\right],$$

we have

$$\left|\left|Z_{ik}Z_{ij}\boldsymbol{\zeta}_{jkp}\right|\right|_F^2 \leq \epsilon_1^2\tilde{M}_{jkp}.$$

In a similar fashion, we can show that

$$\|Z_{ik}Z_{ij}\left(\left(\left(\boldsymbol{\epsilon}_{1kp}\right)\left(\boldsymbol{\epsilon}_{1jp}\right)'\right)\right)\|_F^2 \leq \epsilon_1^2 P$$

and

$$\|\epsilon_\sigma\sigma_0^2\mathbf{I}_P\|_F^2 \leq \epsilon_1^2\sigma_0^4 P.$$

By using the triangle inequality we have

$$\|\tilde{\boldsymbol{\Sigma}}_i\|_F \leq \epsilon_1\left(\sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{p=1}^{KP}\left(\sqrt{\tilde{M}_{jkp}}\right) + JK^2P^{3/2} + \sigma_0^2\sqrt{P}\right) := \epsilon_1 M_{\boldsymbol{\Sigma}} \tag{A.6}$$

for all $i \in \mathbb{N}$. By the Wielandt-Hoffman Theorem (Golub and Van Loan [2013] Theorem 8.1.4), we have that

$$\sum_{p=1}^{P}\left(\lambda_p\left((\boldsymbol{\Sigma}_i)_0 + \tilde{\boldsymbol{\Sigma}}_i\right) - \lambda_p\left((\boldsymbol{\Sigma}_i)_0\right)\right)^2 \leq \|\tilde{\boldsymbol{\Sigma}}_i\|_F^2,$$

which implies that

$$\max_p \left| \lambda_p \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i \right) - \lambda_p \left((\mathbf{\Sigma}_i)_0 \right) \right| \leq \| \tilde{\mathbf{\Sigma}}_i \|_F \tag{A.7}$$

where $\lambda_p(\mathbf{A})$ are the eigenvalues of the matrix $\mathbf{A}$. By using Equation A.6, we can bound the log-determinant of the ratio of the two covariance matrices as follows

$$\begin{aligned} \log \left(\frac{|\mathbf{\Sigma}_i|}{|(\mathbf{\Sigma}_i)_0|} \right) &= \log \left(\frac{\prod_{p=1}^{P} \lambda_p \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i \right)}{\prod_{p=1}^{P} \lambda_p \left((\mathbf{\Sigma}_i)_0 \right)} \right) \\ &\leq \log \left(\prod_{p=1}^{P} \frac{((d_{ip})_0 + \sigma_0^2) + \epsilon_1 M_{\mathbf{\Sigma}}}{(d_{ip})_0 + \sigma_0^2} \right) \\ &\leq P \log \left(1 + \frac{\epsilon_1 M_{\mathbf{\Sigma}}}{\sigma_0^2} \right). \end{aligned} \tag{A.8}$$

We can also bound $\operatorname{tr}\left(\mathbf{\Sigma}_i^{-1} (\mathbf{\Sigma}_i)_0 \right)$. To do this, we will first consider the spectral norm, defined as $\|\mathbf{A}\|_2 = \sqrt{\mathbf{A}^* \mathbf{A}}$ for some matrix $\mathbf{A}$. In the case where $\mathbf{A}$ is symmetric, we have that $\|\mathbf{A}\|_2 = \max_r |\lambda_r(\mathbf{A})|$. By the submultiplicative property of induced norms, we have that

$$\max_p |\lambda_p(\mathbf{AB})| = \|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_2 = \max_p |\lambda_p(\mathbf{A})| \max_p |\lambda_p(\mathbf{B})|, \tag{A.9}$$

for two symmetric matrices $\mathbf{A}$ and $\mathbf{B}$. By using the Sherman–Morrison–Woodbury formula, we can see that

$$\begin{aligned} \mathbf{\Sigma}_i^{-1} &= \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i \right)^{-1} \\ &= (\mathbf{\Sigma}_i)_0^{-1} - (\mathbf{\Sigma}_i)_0^{-1} \tilde{\mathbf{\Sigma}}_i \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i \right)^{-1}. \end{aligned}$$

Thus, we have that

$$\mathbf{\Sigma}_i^{-1} (\mathbf{\Sigma}_i)_0 = \mathbf{I}_R - (\mathbf{\Sigma}_i)_0^{-1} \tilde{\mathbf{\Sigma}}_i \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i \right)^{-1} (\mathbf{\Sigma}_i)_0 . \tag{A.10}$$

Using Equation A.9, we would like to bound the magnitude of the eigenvalues of

$(\mathbf{\Sigma}_i)_0^{-1}\tilde{\mathbf{\Sigma}}_i \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i\right)^{-1} (\mathbf{\Sigma}_i)_0$. We know that

$$\max_p \left|\lambda_p \left(\left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i\right)^{-1}\right)\right| \leq \frac{1}{\sigma_0^2}$$

and

$$\max_p \left|\lambda_p(\tilde{\mathbf{\Sigma}}_i)\right| \leq \epsilon_1 M_{\mathbf{\Sigma}},$$

with the second inequality coming from Equation A.6. From Equation A.10 and basic properties of the trace, we have that

$$\begin{aligned}
\operatorname{tr}\left(\mathbf{\Sigma}_i^{-1}(\mathbf{\Sigma}_i)_0\right) &= \operatorname{tr}\left(\mathbf{I}_P - (\mathbf{\Sigma}_i)_0^{-1}\tilde{\mathbf{\Sigma}}_i \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i\right)^{-1} (\mathbf{\Sigma}_i)_0\right) \\
&= \operatorname{tr}(\mathbf{I}_P) - \operatorname{tr}\left(\tilde{\mathbf{\Sigma}}_i \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i\right)^{-1} (\mathbf{\Sigma}_i)_0 (\mathbf{\Sigma}_i)_0^{-1}\right) \\
&= \operatorname{tr}(\mathbf{I}_P) - \operatorname{tr}\left(\tilde{\mathbf{\Sigma}}_i \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i\right)^{-1}\right)
\end{aligned}$$

Thus, using the fact that the trace of a matrix is the sum of its eigenvalues, we have that

$$\operatorname{tr}\left(\mathbf{\Sigma}_i^{-1}(\mathbf{\Sigma}_i)_0\right) \leq P + P \max_p \left|\lambda_p \left(\tilde{\mathbf{\Sigma}}_i \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i\right)^{-1}\right)\right|.$$

Using the submultiplicative property stated in Equation A.9, we have

$$\operatorname{tr}\left(\mathbf{\Sigma}_i^{-1}(\mathbf{\Sigma}_i)_0\right) \leq P + \frac{P\epsilon_1 M_{\mathbf{\Sigma}}}{\sigma_0^2}. \tag{A.11}$$

Lastly, we can bound the quadratic term in $K_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})$ in the following way:

$$\begin{aligned}
\left((\boldsymbol{\mu}_i)_0 - \boldsymbol{\mu}_i\right)' \left(\boldsymbol{\Sigma}_i\right)^{-1} \left((\boldsymbol{\mu}_i)_0 - \boldsymbol{\mu}_i\right) &\leq ||(\boldsymbol{\mu}_i)_0 - \boldsymbol{\mu}_i||_2^2 \max_p \lambda_p(\left(\boldsymbol{\Sigma}_i\right)^{-1}) \\
&\leq \frac{1}{\sigma^2} \sum_{k=1}^{K} \|(\boldsymbol{\nu}_k)_0 - \boldsymbol{\nu}_k\|_2^2 \\
&= \frac{1}{\sigma^2} \sum_{k=1}^{K} \boldsymbol{\epsilon}'_{2k} \boldsymbol{\epsilon}_{2k} \\
&\leq \frac{KP\epsilon_2^2}{\sigma_0^2}. \qquad \text{(A.12)}
\end{aligned}$$

Thus letting

$$\epsilon_1 < \min \left\{ \frac{\sigma_0^2}{M_{\boldsymbol{\Sigma}}} \left(\exp\left(\frac{2\epsilon}{3P}\right) - 1 \right), \frac{2\epsilon\sigma_0^2}{3PM_{\boldsymbol{\Sigma}}} \right\} \qquad \text{(A.13)}$$

and

$$\epsilon_2 < \sqrt{\frac{2\sigma_0^2\epsilon}{3KP}}, \qquad \text{(A.14)}$$

we have from Equations A.8, A.11, and A.12 that

$$K_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) < \epsilon \text{ for all } \boldsymbol{\omega} \in \boldsymbol{\Omega}_1$$

where $\boldsymbol{\Omega}_1 := \left(\bigtimes_{j=1}^{K} \bigtimes_{k=1}^{K} \boldsymbol{\Omega}_{\boldsymbol{\Sigma}_{jk}}\right) \times \left(\bigtimes_{k=1}^{K} \boldsymbol{\Omega}_{\boldsymbol{\nu}_k}\right) \times \boldsymbol{\Omega}_{\sigma^2}$. Letting $a > \max\left\{1 + \frac{\epsilon_1 M_{\boldsymbol{\Sigma}}}{\sigma_0^2}, \left(1 - \frac{\epsilon_1 M_{\boldsymbol{\Sigma}}}{\sigma_0^2}\right)^{-1}\right\}$ and $b > \sqrt{KP\epsilon_2^2}$ in the definition of $\mathcal{C}(\boldsymbol{\omega}_0, \epsilon)$, we have that $\boldsymbol{\Omega}_1 \subset \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)$. Let $H_{\boldsymbol{\phi}}$ be the set of hyper-parameters corresponding to the $\boldsymbol{\phi}$ parameters, and let $\boldsymbol{\Pi}(\boldsymbol{\eta}_{\boldsymbol{\phi}})$ be the prior distribution on $\boldsymbol{\eta}_{\boldsymbol{\phi}} \in H_{\boldsymbol{\phi}}$. Thus we have that

$$\begin{aligned}
\boldsymbol{\Pi}\left(\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)\right) \geq & \int_{H_{\boldsymbol{\phi}}} \prod_{j=1}^{K} \prod_{p=1}^{KP} \prod_{r=1}^{P} \int_{(\phi_{jrp})_0}^{(\phi_{jrp})_0 + \epsilon_1} \sqrt{\frac{\gamma_{jrp}\tilde{\tau}_{pj}}{2\pi}} \exp\left\{ -\frac{\gamma_{jrp}\tilde{\tau}_{pj}}{2} \phi_{jrp}^2 \right\} \mathrm{d}\phi_{jrp} \mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\eta}_{\boldsymbol{\phi}}) \\
& \times \prod_{k=1}^{K} \int_0^{\infty} \int_{(\boldsymbol{\nu}_k)_0}^{(\boldsymbol{\nu}_k)_0 + \epsilon_2 \mathbf{1}} \left(\frac{\tau_k}{2\pi}\right)^{P/2} \exp\left\{ \frac{\tau_k}{2} \boldsymbol{\nu}'_k \boldsymbol{\nu}_k \right\} \mathrm{d}\boldsymbol{\nu}_k \mathrm{d}\boldsymbol{\Pi}(\tau_k) \\
& \times \int_{\sigma_0^2}^{(1+\epsilon_1)\sigma_0^2} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} (\sigma^2)^{-\alpha_0 - 1} \exp\left\{ -\frac{\beta_0}{\sigma^2} \right\} \mathrm{d}\sigma^2.
\end{aligned}$$

Restricting the hyper-parameters of $\boldsymbol{\phi}$ to only a subset of the support, say $\tilde{H}_{\boldsymbol{\phi}}$, where

$$\tilde{H}_{\boldsymbol{\phi}} = \left\{ \boldsymbol{\eta}_{\boldsymbol{\phi}} : \frac{1}{10} \leq \gamma_{jrp} \leq 10, 1 \leq \delta_{pj} \leq 2, 1 \leq a_{1j} \leq 10, 1 \leq a_{2j} \leq 10 \right\},$$

we can see that there exists a $M_{\phi_{jrp}} > 0$ such that

$$\sqrt{\frac{\gamma_{jrp}\tilde{\tau}_{pj}}{2\pi}} \exp\left\{ -\frac{\gamma_{jrp}\tilde{\tau}_{pj}}{2}\phi_{jrp}^2 \right\} \geq M_{\phi_{jrp}},$$

for all $\phi_{jrp} \in [(\phi_{jrp})_0, (\phi_{jrp})_0 + \epsilon_1]$. Similarly, we can find a lower bound $M_{\tilde{H}_{\boldsymbol{\phi}}} > 0$, such that

$$\int_{\tilde{H}_{\boldsymbol{\phi}}} \mathrm{d}(\boldsymbol{\eta}_{\boldsymbol{\phi}}) \geq M_{\tilde{H}_{\boldsymbol{\phi}}}.$$

Similarly, if we bound τ_k such that $\frac{1}{10} \leq \tau_k \leq 10$, it is easy to see that there exists constants $M_{\boldsymbol{\nu}_k}, M_{\tau_k}, M_{\sigma^2} > 0$ such that

$$\left(\frac{\tau_k}{2\pi}\right)^{P/2} \exp\left\{ \frac{\tau_k}{2}\boldsymbol{\nu}_k'\boldsymbol{\nu}_k \right\} \geq M_{\boldsymbol{\nu}_k},$$

for all $\boldsymbol{\nu}_k \in [(\boldsymbol{\nu}_k)_0, (\boldsymbol{\nu}_k)_0 + \epsilon_2 \mathbf{1}]$,

$$\int_{\frac{1}{10}}^{10} \boldsymbol{\Pi}(\tau_k) \geq M_{\tau_k},$$

and

$$\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\sigma^2)^{-\alpha_0-1} \exp\left\{ -\frac{\beta_0}{\sigma^2} \right\} \geq M_{\sigma^2}$$

for all $\sigma^2 \in [\sigma_0^2, (1+\epsilon_1)\sigma_0^2]$. Therefore we have that

$$\begin{aligned}\boldsymbol{\Pi}\left(\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)\right) &\geq M_{\tilde{H}_{\boldsymbol{\phi}}} \prod_{j=1}^{K} \prod_{p=1}^{KP} \prod_{r=1}^{P} \epsilon_1 M_{\phi_{jrp}} \\ &\times \prod_{k=1}^{K} M_{\tau_k} \epsilon_2^P M_{\boldsymbol{\nu}_k} \\ &\times \epsilon_1 \sigma_0^2 M_{\sigma_0^2} \\ &> 0.\end{aligned}$$

Therefore, for $\epsilon > 0$, there exists a and b such that $\sum_{i=1}^{\infty} \frac{V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})}{i^2} < \infty$ for any $\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)$ and $\boldsymbol{\Pi}\left(\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)\right) > 0$.

A.1.2 Proof of Lemma 2.2

Following the notation of Ghosal and Van der Vaart [2017], we will let $P_{\boldsymbol{\omega}_0}^{(N)}$ denote the joint distribution of $\mathbf{y}_1, \ldots, \mathbf{y}_N$ at $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$. In order to show that the posterior distribution, $\boldsymbol{\Pi}_N(.|\mathbf{y}_1, \ldots, \mathbf{y}_N)$, is weakly consistent at $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$, we need to show that $\boldsymbol{\Pi}_N(\mathcal{U}^c|\mathbf{y}_1, \ldots, \mathbf{y}_N) \to 0$ a.s. $[P_{\boldsymbol{\omega}_0}]$ for every weak neighborhood, $\mathcal{U}$ of $\boldsymbol{\omega}_0$. Following a similar notation to Ghosal and Van der Vaart [2017], let ψ_N be measurable mappings, $\psi_N : \boldsymbol{\mathcal{S}}^N \times \boldsymbol{\mathcal{Z}}^N \to [0, 1]$, where $\boldsymbol{\mathcal{Z}}$ is the sample space of $\{Z_{i1}, \ldots, Z_{iK}\}$. Let $\psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)$ be the corresponding test function, and $P_{\boldsymbol{\omega}}^N \psi_N = \mathbb{E}_{P_{\boldsymbol{\omega}}^N} \psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) = \int \psi_N \mathrm{d}P_{\boldsymbol{\omega}}^N$, where $P_{\boldsymbol{\omega}}^N$ denotes the joint distribution on $\mathbf{y}_1, \ldots, \mathbf{y}_N$ with parameters $\boldsymbol{\omega}$. Suppose there exists tests ψ_N such that $P_{\boldsymbol{\omega}_0}^N \psi_N \to 0$,

and $\sup_{\boldsymbol{\omega}\in\mathcal{U}^c} P_{\boldsymbol{\omega}}^N(1-\psi_N) \to 0$. Since $\psi_N(\mathbf{y}_1,\ldots,\mathbf{y}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N) \in [0,1]$, we have that

$$\begin{aligned}\boldsymbol{\Pi}_n(U^c|\mathbf{y}_1,\ldots,\mathbf{y}_N) &\leq \boldsymbol{\Pi}_n(U^c|\mathbf{y}_1,\ldots,\mathbf{y}_N) + \psi_N(\mathbf{y}_1,\ldots,\mathbf{y}_N)\left(1-\boldsymbol{\Pi}_n(U^c|\mathbf{y}_1,\ldots,\mathbf{y}_N)\right) \\ &= \psi_N(\mathbf{y}_1,\ldots,\mathbf{y}_N) + \frac{\left(1-\psi_N(\mathbf{y}_1,\ldots,\mathbf{y}_N)\right)\int_{U^c}\prod_{i=1}^N \frac{f_i(\mathbf{y}_i;\boldsymbol{\omega})}{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})}{\int_{\boldsymbol{\Omega}}\prod_{i=1}^N \frac{f_i(\mathbf{y}_i;\boldsymbol{\omega})}{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})}.\end{aligned} \tag{A.15}$$

To show that $\boldsymbol{\Pi}_n(U^c|\mathbf{y}_1,\ldots,\mathbf{y}_N) \to 0$, it is sufficient to show the following three conditions:

1. $\psi_N(\mathbf{y}_1,\ldots,\mathbf{y}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N) \to 0$ a.s. $[P_{\boldsymbol{\omega}_0}]$,
2. $e^{\beta_1 N}\left(1-\psi_N(\mathbf{y}_1,\ldots,\mathbf{y}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N)\right)\int_{\mathcal{U}^c}\prod_{i=1}^N \frac{f_i(\mathbf{y}_i;\boldsymbol{\omega})}{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega}) \to 0$ a.s. $[P_{\boldsymbol{\omega}_0}]$ for some $\beta_1 > 0$,
3. $e^{\beta N}\left(\int_{\boldsymbol{\Omega}}\prod_{i=1}^N \frac{f_i(\mathbf{y}_i;\boldsymbol{\omega})}{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\right) \to \infty$ a.s. $[P_{\boldsymbol{\omega}_0}]$ for all $\beta > 0$.

We will start by proving (c). Fix $\beta > 0$. Thus we have

$$e^{\beta N}\left(\int_{\boldsymbol{\Omega}}\prod_{i=1}^N \frac{f_i(\mathbf{y}_i;\boldsymbol{\omega})}{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\right) = e^{\beta N}\left(\int_{\boldsymbol{\Omega}}\exp\left[-\sum_{i=1}^N \log\left(\frac{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}{f_i(\mathbf{y}_i;\boldsymbol{\omega})}\right)\right]\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\right).$$

By Fatou's Lemma, we have

$$\begin{aligned}&\liminf_{N\to\infty}\int_{\boldsymbol{\Omega}}\exp\left[\beta N - \sum_{i=1}^N \log\left(\frac{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}{f_i(\mathbf{y}_i;\boldsymbol{\omega})}\right)\right]\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega}) \\ \geq &\int_{\boldsymbol{\Omega}}\liminf_{N\to\infty}\exp\left[\beta N - \sum_{i=1}^N \log\left(\frac{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}{f_i(\mathbf{y}_i;\boldsymbol{\omega})}\right)\right]\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\end{aligned}$$

Let $\beta > \epsilon > 0$ and $a, b > 0$ be defined such that lemma 3.1 holds. Since $\mathcal{C}(\boldsymbol{\omega}_0,\epsilon) \subset \boldsymbol{\Omega}$, we

have that

$$\int_{\boldsymbol{\Omega}} \liminf_{N\to\infty} \exp\left[\beta N - \sum_{i=1}^{N} \log\left(\frac{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}{f_i(\mathbf{y}_i;\boldsymbol{\omega})}\right)\right] \mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})$$
$$\geq \int_{\mathcal{C}(\boldsymbol{\omega}_0,\epsilon)} \liminf_{N\to\infty} \exp\left[\beta N - \sum_{i=1}^{N} \log\left(\frac{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}{f_i(\mathbf{y}_i;\boldsymbol{\omega})}\right)\right] \mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})$$

By Kolmogorov's strong law of large numbers for non-identically distributed random variables, we have that

$$\frac{1}{N}\sum_{i=1}^{N} \left(\Lambda_i(\boldsymbol{\omega}_0,\boldsymbol{\omega}) - K_i(\boldsymbol{\omega}_0,\boldsymbol{\omega})\right) \to 0$$

a.s. $[P_{\boldsymbol{\omega}_0}]$. Thus for each $\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0,\epsilon)$, with $P_{\boldsymbol{\omega}_0}$-probability 1,

$$\frac{1}{N}\sum_{i=1}^{N} \Lambda_i(\boldsymbol{\omega}_0,\boldsymbol{\omega}) \to \mathbb{E}(\overline{K_i(\boldsymbol{\omega}_0,\boldsymbol{\omega})}) < \epsilon < B,$$

since $\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0,\epsilon)$. Therefore, we have that

$$\int_{\mathcal{C}(\boldsymbol{\omega}_0,\epsilon)} \liminf_{N\to\infty} \exp\left[\beta N - \sum_{i=1}^{N} \log\left(\frac{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}{f_i(\mathbf{y}_i;\boldsymbol{\omega})}\right)\right] \mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega}) \geq \int_{\mathcal{C}(\boldsymbol{\omega}_0,\epsilon)} \inf_{N\to\infty} \exp\left\{N(\beta-\epsilon)\right\} \mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega}).$$

Since $\beta - \epsilon > 0$, and $\boldsymbol{\Pi}\left(\theta \in \mathcal{C}(\boldsymbol{\omega}_0,\epsilon)\right) > 0$ (lemma 3.1), we have that

$$e^{\beta N}\left(\int_{\boldsymbol{\Omega}} \prod_{i=1}^{N} \frac{f_i(\mathbf{y}_i;\boldsymbol{\omega})}{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)} \mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\right) \to \infty \tag{A.16}$$

a.s. $[P_{\boldsymbol{\omega}_0}]$ for all $\beta > 0$. We will now show that exists measurable mappings such that $P^N_{\boldsymbol{\omega}_0}\psi_N \to 0$ and $\sup_{\boldsymbol{\omega}\in\mathcal{U}^c} P^N_{\boldsymbol{\omega}}(1-\psi_N) \to 0$. Consider weak neighborhoods $\mathcal{U}$ of $\boldsymbol{\omega}_0$ of the form

$$\mathcal{U} = \left\{\boldsymbol{\omega} : \left|\int f_i \mathrm{d}P_{\boldsymbol{\omega}} - \int f_i \mathrm{d}P_{\boldsymbol{\omega}_0}\right| < \epsilon_i, \;\; i = 1,2,\ldots,r\right\}, \tag{A.17}$$

where $r \in \mathbb{N}$, $\epsilon_i > 0$, and f_i are continuous functions such that $f_i : \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{Z}} \to [0,1]$. As shown in Ghosh and Ramamoorthi [2003], for any particular f_i and $\epsilon_i > 0$, $\left|\int f_i \mathrm{d}P_{\boldsymbol{\omega}} - \int f_i \mathrm{d}P_{\boldsymbol{\omega}_0}\right| <$

ϵ_i iff $\int f_i \mathrm{d}P_{\boldsymbol{\omega}} - \int f_i \mathrm{d}P_{\boldsymbol{\omega}_0} < \epsilon_i$ and $\int (1-f_i)\mathrm{d}P_{\boldsymbol{\omega}} - \int (1-f_i)\mathrm{d}P_{\boldsymbol{\omega}_0} < \epsilon$. Since $\tilde{f}_i := (1-f_i)$ is still a continuous function such that $\tilde{f}_i : \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{Z}} \to [0,1]$, we can rewrite Equation A.17 as

$$\mathcal{U} = \cap_{i=1}^{2r} \left\{ \boldsymbol{\omega} : \int g_i \mathrm{d}P_{\boldsymbol{\omega}} - \int g_i \mathrm{d}P_{\boldsymbol{\omega}_0} < \epsilon_i \right\}, \tag{A.18}$$

where g_i are continuous functions such that $g_i : \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{Z}} \to [0,1]$ and $\epsilon_i > 0$. Following Ghosal and Van der Vaart [2017], it can be shown by Hoeffding's inequality that using the test function $\tilde{\psi}$, defined as

$$\tilde{\psi}_{iN}(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) := \mathbb{1}\left\{ \frac{1}{N} \sum_{j=1}^{N} g_i(\mathbf{y}_j, \mathbf{z}_j) > \int g_i \mathrm{d}P_{\boldsymbol{\omega}_0} + \frac{\epsilon_i}{2} \right\}, \tag{A.19}$$

leads to

$$\int \tilde{\psi}_{iN}(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) \mathrm{d}P_{\boldsymbol{\omega}_0} \leq e^{-N\epsilon_i^2/2}$$

and

$$\int \left(1 - \tilde{\psi}_{iN}(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right) \mathrm{d}P_{\boldsymbol{\omega}} \leq e^{-N\epsilon_i^2/2}$$

for any $\boldsymbol{\omega} \in \mathcal{U}^c$. Let $\boldsymbol{\psi}_n = \max_i \tilde{\psi}_{iN}$ be our test function and $\epsilon = \min_i \epsilon_i$. Using the fact that $\mathbb{E}(\max_i \tilde{\psi}_{iN}) \leq \sum_i \mathbb{E}(\tilde{\psi}_{iN})$ and $\mathbb{E}(1 - \max_i \tilde{\psi}_{iN}) \leq \mathbb{E}(1 - \tilde{\psi}_{iN})$, we have

$$\int \psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) \mathrm{d}P_{\boldsymbol{\omega}_0} \leq (2r) e^{-N\epsilon^2/2} \tag{A.20}$$

and

$$\int \left(1 - \psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right) \mathrm{d}P_{\boldsymbol{\omega}} \leq e^{-N\epsilon^2/2}, \tag{A.21}$$

for any $\boldsymbol{\omega} \in \mathcal{U}^c$. Using Markov's inequality on Equation A.20, we have that

$$\begin{aligned} P\left(\psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) \geq e^{-nC}\right) &\leq \frac{\mathbb{E}\left(\psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)}{e^{-NC}} \\ &\leq (2r) e^{-N(\epsilon^2/2 - C)} \end{aligned}$$

Thus letting $C < \epsilon^2/2$, we have that $\sum_{N=1}^{\infty} P\left(\psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) \geq e^{-NC}\right) < \infty$. Thus by the Borel-Cantelli lemma, we know that

$$P\left(\limsup_{N\to\infty} P\left(\psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) \geq e^{-NC}\right)\right) = 0$$

Thus we have that $\psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) \to 0$ a.s. $[P_{\boldsymbol{\omega}_0}]$ (Condition (a)). To prove condition (b), we will first start by taking the expectation with respect to $P_{\boldsymbol{\omega}_0}$:

$$\begin{aligned}
&\mathbb{E}_{P^N_{\boldsymbol{\omega}_0}}\left(e^{\beta N}\left(1-\psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)\int_{\mathcal{U}^c}\prod_{i=1}^{N}\frac{f_i(\mathbf{y}_i;\boldsymbol{\omega})}{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\right)\\
=&\int_{\boldsymbol{\mathcal{S}}^N}\left(e^{\beta N}\left(1-\psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)\int_{\mathcal{U}^c}\prod_{i=1}^{N}\frac{f_i(\mathbf{y}_i;\boldsymbol{\omega})}{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\right)\mathrm{d}P^N_{\boldsymbol{\omega}_0}\\
=&\int_{\mathcal{U}^c}\left(\prod_{i=1}^{N}\int_{\mathcal{S}}e^{\beta N}\left(1-\psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)f_i(\mathbf{y}_i;\boldsymbol{\omega})\mathrm{d}\mathbf{y}_i\right)\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\\
=&e^{\beta N}\int_{\mathcal{U}^c}\mathbb{E}_{P^N_{\boldsymbol{\omega}}}\left(1-\psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\\
\leq&e^{\beta_1 N}e^{-N\epsilon^2/2},
\end{aligned}$$

where the last inequality is from Equation A.21. Thus by Markov's inequality and letting $\beta_1 < \epsilon^2/2$, we have that

$$\begin{aligned}
&P\left(e^{\beta N}\left(1-\psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)\int_{\mathcal{U}^c}\prod_{i=1}^{N}\frac{f_i(\mathbf{y}_i;\boldsymbol{\omega})}{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega}) \geq e^{-N((\epsilon^2/2-\beta_1)/2)}\right)\\
\leq&\frac{\mathbb{E}_{P^N_{\boldsymbol{\omega}_0}}\left(e^{\beta N}\left(1-\psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)\int_{\mathcal{U}^c}\prod_{i=1}^{N}\frac{f_i(\mathbf{y}_i;\boldsymbol{\omega})}{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\right)}{e^{-N((\epsilon^2/2-\beta_1)/2)}}\\
\leq&e^{-N((\epsilon^2/2-\beta_1)/2)}
\end{aligned}$$

Letting E_N be the event that $e^{\beta N}\left(1-\psi_N(\mathbf{y}_1, \ldots, \mathbf{y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)\int_{\mathcal{U}^c}\prod_{i=1}^{N}\frac{f_i(\mathbf{y}_i;\boldsymbol{\omega})}{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})$ $\geq e^{-N((\epsilon^2/2-\beta_1)/2)}$, we have that $\sum_{i=1}^{\infty} P(E_N) < \infty$. Thus by the Borel-Cantelli lemma, we

have that

$$e^{\beta N}\left(1-\psi_N(\mathbf{y}_1,\ldots,\mathbf{y}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N)\right)\int_{\mathcal{U}^c}\prod_{i=1}^{N}\frac{f_i(\mathbf{y}_i;\boldsymbol{\omega})}{f_i(\mathbf{y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\to 0$$

a.s. $[P_{\boldsymbol{\omega}_0}]$ for $0<\beta_1<\epsilon^2/2$. Therefore, we have proved conditions (a), (b), and (c). Thus by letting β in condition (c) be such that $\beta=\beta_1$, where $0<\beta_1<\epsilon^2/2$, we can see that $\boldsymbol{\Pi}_N(\mathcal{U}^c|\mathbf{y}_1,\ldots,\mathbf{y}_N)\to 0$ a.s. $[P_{\boldsymbol{\omega}_0}]$ for every weak neighborhood, $\mathcal{U}$ of $\boldsymbol{\omega}_0$.

A.2 Computation

A.2.1 Posterior Distributions and Computation

In this section, we will discuss the computational strategy used to perform Bayesian inference. In cases where the posterior distribution is a known distribution, a Gibbs update will be performed. We will let $\boldsymbol{\Theta}$ be the collection of all parameters, and $\boldsymbol{\Theta}_{-\zeta}$ be the collection of all parameters, excluding the ζ parameter. We will first start with the $\boldsymbol{\phi}_{km}$ parameters, for $j=1,\ldots,K$ and $m=1,\ldots,M$. Let $\mathbf{D}_{km}=\tilde{\tau}_{mk}^{-1}diag\left(\gamma_{k1m}^{-1},\ldots,\gamma_{kPm}^{-1}\right)$. By letting

$$\mathbf{m}_{jm}=\frac{1}{\sigma^2}\sum_{i=1}^{N}\left(\chi_{im}\left(\mathbf{y}_iZ_{ij}-Z_{ij}^2\boldsymbol{\nu}_j-Z_{ij}^2\sum_{n\neq m}\chi_{in}\boldsymbol{\phi}_{jn}-\sum_{k\neq j}Z_{ij}Z_{ik}\left[\boldsymbol{\nu}_k+\sum_{n=1}^{M}\chi_{in}\boldsymbol{\phi}_{kn}\right]\right)\right)$$

and

$$\mathbf{M}_{jm}^{-1}=\frac{1}{\sigma^2}\sum_{i=1}^{N}\left(Z_{ij}^2\chi_{im}^2\right)\mathbf{I}_P+\mathbf{D}_{km}^{-1},$$

we have that

$$\boldsymbol{\phi}_{jm}|\boldsymbol{\Theta}_{-\boldsymbol{\phi}_{jm}},\mathbf{y}_1,\ldots,\mathbf{y}_N\sim\mathcal{N}(\mathbf{M}_{jm}\mathbf{m}_{jm},\mathbf{M}_{jm}).$$

The posterior distribution of δ_1 is

$$\delta_{1k}|\boldsymbol{\Theta}_{-\delta_{1k}}, \mathbf{y}_1, \ldots, \mathbf{y}_N \sim \Gamma\left(a_{1k} + (PM/2), 1 + \frac{1}{2}\sum_{r=1}^{P}\gamma_{k,r,1}\phi_{k,r,1}^2 + \frac{1}{2}\sum_{m=2}^{M}\sum_{r=1}^{P}\gamma_{k,r,m}\phi_{k,r,m}^2\left(\prod_{j=2}^{m}\delta_j\right)\right).$$

The posterior distribution for δ_{ik}, for $i = 2, \ldots, M$, is

$$\delta_{ik}|\boldsymbol{\Theta}_{-\delta_{ik}}, \mathbf{y}_1, \ldots, \mathbf{y}_N \sim \Gamma\left(a_2 + (P(M-i+1)/2), 1 + \frac{1}{2}\sum_{m=i}^{M}\sum_{r=1}^{P}\gamma_{k,r,m}\phi_{k,r,m}^2\left(\prod_{j=1;j\neq i}^{m}\delta_j\right)\right).$$

The posterior distribution for a_{1k} is not a commonly known distribution, however we have that

$$P(a_{1k}|\boldsymbol{\Theta}_{-a_{1k}}, \mathbf{y}_1, \ldots, \mathbf{y}_N) \propto \frac{1}{\Gamma(a_{1k})}\delta_{1k}^{a_{1k}-1}a_{1k}^{\alpha_1-1}exp\left\{-a_{1k}\beta_1\right\}.$$

Since this is not a known kernel of a distribution, we will have to use Metropolis-Hastings algorithm. Consider the proposal distribution $Q(a'_{1k}|a_{1k}) = \mathcal{N}\left(a_{1k}, \epsilon_1\beta_1^{-1}, 0, +\infty\right)$ (Truncated Normal) for some small $\epsilon_1 > 0$. Thus the probability of accepting any step is

$$A(a'_{1k}, a_{1k}) = \min\left\{1, \frac{P\left(a'_{1k}|\boldsymbol{\Theta}_{-a'_{1k}}, \mathbf{y}_1, \ldots, \mathbf{y}_N\right)}{P\left(a_{1k}|\boldsymbol{\Theta}_{-a_{1k}}, \mathbf{y}_1, \ldots, \mathbf{y}_N\right)}\frac{Q\left(a_{1k}|a'_{1k}\right)}{Q\left(a'_{1k}|a_{1k}\right)}\right\}.$$

Similarly for a_{2k}, we have

$$P(a_{2k}|\boldsymbol{\Theta}_{-a_{2k}}, \mathbf{y}_1, \ldots, \mathbf{y}_N) \propto \frac{1}{\Gamma(a_{2k})^{M-1}}\left(\prod_{i=2}^{M}\delta_{ik}^{a_{2k}-1}\right)a_{2k}^{\alpha_2-1}exp\left\{-a_{2k}\beta_2\right\}.$$

We will use a similar proposal distribution, such that $Q(a'_{2k}|a_{2k}) = \mathcal{N}\left(a_{2k}, \epsilon_2\beta_2^{-1}, 0, +\infty\right)$ for

some small $\epsilon_2 > 0$. Thus the probability of accepting any step is

$$A(a'_{2k}, a_{2k}) = \min\left\{1, \frac{P\left(a'_{2k}|\boldsymbol{\Theta}_{-a'_{2k}}, \mathbf{y}_1, \ldots, \mathbf{y}_N\right)}{P\left(a_{2k}|\boldsymbol{\Theta}_{-a_{2k}}, \mathbf{y}_1, \ldots, \mathbf{y}_N\right)} \frac{Q\left(a_{2k}|a'_{2k}\right)}{Q\left(a'_{2k}|a_{2k}\right)}\right\}.$$

For the $\gamma_{j,r,m}$ parameters, for $j = 1, \ldots K$, $r = 1, \ldots, P$, and $m = 1, \ldots, M$, we have

$$\gamma_{j,r,m}|\boldsymbol{\Theta}_{-\gamma_{j,r,m}}, \mathbf{y}_1, \ldots, \mathbf{y}_N \sim \Gamma\left(\frac{\nu_\gamma + 1}{2}, \frac{\phi^2_{j,r,m}\tilde{\tau}_{mj} + \nu_\gamma}{2}\right).$$

The posterior distribution for the $\mathbf{z}_i$ parameters are not a commonly known distribution, so we will have to use the Metropolis-Hastings algorithm. We know that

$$\begin{aligned} p(\mathbf{z}_i|\boldsymbol{\Theta}_{-\mathbf{z}_i}, \mathbf{y}_1, \ldots, \mathbf{y}_N) \propto & \prod_{k=1}^{K} Z_{ik}^{\alpha_3\pi_k - 1} \\ & \times exp\left\{-\frac{1}{2\sigma^2}\left(\mathbf{y}_i - \sum_{k=1}^{K} Z_{ik}\left(\boldsymbol{\nu}_k + \sum_{n=1}^{M}\chi_{in}\boldsymbol{\phi}_{kn}\right)\right)'\right. \\ & \left.\left(\mathbf{y}_i - \sum_{k=1}^{K} Z_{ik}\left(\boldsymbol{\nu}_k + \sum_{n=1}^{M}\chi_{in}\boldsymbol{\phi}_{kn}\right)\right)\right\}. \end{aligned}$$

We will use $Q(\mathbf{z}'_i|\mathbf{z}_i) = Dir(a_{\mathbf{z}}\mathbf{z}_i)$ for some large $a_{\mathbf{z}} \in \mathbb{R}^+$ as the proposal distribution. Thus the probability of accepting a proposed step is

$$A(\mathbf{z}'_i, \mathbf{z}_i) = \min\left\{1, \frac{P\left(\mathbf{z}'_i|\boldsymbol{\Theta}_{-\mathbf{z}'_i}, \mathbf{y}_1, \ldots, \mathbf{y}_N\right)}{P\left(\mathbf{z}_i|\boldsymbol{\Theta}_{-\mathbf{z}_i}, \mathbf{y}_1, \ldots, \mathbf{y}_N\right)} \frac{Q\left(\mathbf{z}_i|\mathbf{z}'_i\right)}{Q\left(\mathbf{z}'_i|\mathbf{z}_i\right)}\right\}.$$

Similarly, a Gibbs update is not available for an update of the $\boldsymbol{\pi}$ parameters. We have that

$$\begin{aligned} p(\boldsymbol{\pi}|\boldsymbol{\Theta}_{-\boldsymbol{\pi}}, \mathbf{y}_1, \ldots, \mathbf{y}_N) \propto & \prod_{k=1}^{K} \pi_k^{c_k - 1} \\ & \times \prod_{i=1}^{N} \frac{1}{B(\alpha_3\boldsymbol{\pi})} \prod_{k=1}^{K} Z_{ik}^{\alpha_3\pi_k - 1}. \end{aligned}$$

Letting out proposal distribution be such that $Q(\boldsymbol{\pi}'|\boldsymbol{\pi}) = Dir(a_{\boldsymbol{\pi}}\boldsymbol{\pi})$, for some large $a_{\boldsymbol{\pi}} \in \mathbb{R}^+$,

we have that our probability of accepting any proposal is

$$A(\boldsymbol{\pi}', \boldsymbol{\pi}) = \min\left\{1, \frac{P\left(\boldsymbol{\pi}'|\boldsymbol{\Theta}_{-\boldsymbol{\pi}'}, \mathbf{y}_1, \ldots, \mathbf{y}_N\right)}{P\left(\boldsymbol{\pi}|\boldsymbol{\Theta}_{-\boldsymbol{\pi}}, \mathbf{y}_1, \ldots, \mathbf{y}_N\right)} \frac{Q\left(\boldsymbol{\pi}|\boldsymbol{\pi}'\right)}{Q\left(\boldsymbol{\pi}'|\boldsymbol{\pi}\right)}\right\}.$$

The posterior distribution of α_3 is also not a commonly known distribution, so we will use the Metropolis-Hastings algorithm to sample from the posterior distribution. We have that

$$\begin{aligned} p(\alpha_3|\boldsymbol{\Theta}_{-\alpha_3}, \mathbf{y}_1, \ldots, \mathbf{y}_N) \propto e^{-b\alpha_3} \\ \times \prod_{i=1}^{N} \frac{1}{B(\alpha_3\boldsymbol{\pi})} \prod_{k=1}^{K} Z_{ik}^{\alpha_3\pi_k - 1}. \end{aligned}$$

Using a proposal distribution such that $Q(\alpha_3'|\alpha_3) = \mathcal{N}(\alpha_3, \sigma^2_{\alpha_3}, 0, +\infty)$ (Truncated Normal), we are left with the probability of accepting a proposed state as

$$A(\alpha_3', \alpha_3) = \min\left\{1, \frac{P\left(\alpha_3'|\boldsymbol{\Theta}_{-\alpha_3'}, \mathbf{y}_1, \ldots, \mathbf{y}_N\right)}{P\left(\alpha_3|\boldsymbol{\Theta}_{-\alpha_3}, \mathbf{y}_1, \ldots, \mathbf{y}_N\right)} \frac{Q\left(\alpha_3|\alpha_3'\right)}{Q\left(\alpha_3'|\alpha_3\right)}\right\}.$$

Letting

$$\mathbf{B}_j = \left(\frac{1}{\tau_j}\mathbf{I}_P + \frac{1}{\sigma^2}\mathbf{I}_P \sum_{i=1}^{N} Z_{ij}^2\right)^{-1}$$

and

$$\mathbf{b}_j = \frac{1}{\sigma^2}\sum_{i=1}^{N} Z_{ij}\left(\mathbf{y}_i - \left(\sum_{k\neq j} Z_{ik}\boldsymbol{\nu}_k\right) - \left(\sum_{k=1}^{K}\sum_{m=1}^{M} Z_{ik}\chi_{im}\boldsymbol{\phi}_{km}\right)\right),$$

we have that

$$\boldsymbol{\nu}_j|\boldsymbol{\Theta}_{-\boldsymbol{\nu}_j}, \mathbf{y}_1, \ldots, \mathbf{y}_N \sim \mathcal{N}(\mathbf{B}_j\mathbf{b}_j, \mathbf{B}_j),$$

for $j = 1, \ldots, K$. Thus we can perform a Gibbs update to update our $\boldsymbol{\nu}_\cdot$ parameters. The τ_l parameters, for $l = 1, \ldots K$, can also be updated by using a Gibbs update since the posterior distribution is:

$$\tau_l|\boldsymbol{\Theta}_{-\tau_l}, \mathbf{y}_1, \ldots, \mathbf{y}_N \sim IG\left(\alpha + P/2, \beta + \frac{1}{2}\boldsymbol{\nu}_l'\boldsymbol{\nu}_l\right).$$

The parameter σ^2 can be updated by using a Gibbs update. If we let

$$\beta_\sigma = \frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{y}_i - \sum_{k=1}^{K} Z_{ik}\left(\boldsymbol{\nu}_k + \sum_{m=1}^{M}\chi_{im}\boldsymbol{\phi}_{km}\right)\right)'\left(\mathbf{y}_i - \sum_{k=1}^{K} Z_{ik}\left(\boldsymbol{\nu}_k + \sum_{m=1}^{M}\chi_{im}\boldsymbol{\phi}_{km}\right)\right)$$

then we have

$$\sigma^2|\boldsymbol{\Theta}_{-\sigma^2}, \mathbf{y}_1, \ldots, \mathbf{y}_N \sim IG\left(\alpha_0 + \frac{PN}{2}, \beta_0 + \beta_\sigma\right),$$

where n_i are the number of time points observed for the i^{th} observed function. Lastly, we can update the χ_{im} parameters, for $i = 1, \ldots, N$ and $m = 1, \ldots, M$, using a Gibbs update. If we let

$$\mathbf{w}_{im} = \frac{1}{\sigma^2}\left(\left(\sum_{k=1}^{K} Z_{ik}\boldsymbol{\phi}_{km}\right)'\left(\mathbf{y}_i - \sum_{k=1}^{K} Z_{ik}\left(\boldsymbol{\nu}_k + \sum_{n\neq m}\chi_{in}\boldsymbol{\phi}_{kn}\right)\right)\right)$$

and

$$\mathbf{W}_{im}^{-1} = 1 + \frac{1}{\sigma^2}\left(\left(\sum_{k=1}^{K} Z_{ik}\boldsymbol{\phi}_{km}\right)'\left(\sum_{k=1}^{K} Z_{ik}\boldsymbol{\phi}_{km}\right)\right),$$

then we have that

$$\chi_{im}|\boldsymbol{\zeta}_{-\chi_{im}}, \mathbf{y}_1, \ldots, \mathbf{y}_N \sim \mathcal{N}(\mathbf{W}_{im}\mathbf{w}_{im}, \mathbf{W}_{im}).$$

In our paper, we have relaxed the assumption that the $\boldsymbol{\Phi}$ are mutually orthogonal parameters. We have shown that we can still maintain many of the desirable properties, while not having to sample in a constrained space. This relaxation makes implementation easier, and may actually help with mixing of the Markov chain. However, we realize that users may want to enforce that the $\boldsymbol{\Phi}$ parameters are orthogonal and therefore can be interpreted as scaled eigenvectors. Using the approach described by Kowal et al. [2017], we will describe how to sample in this constrained space.

In order to impose the orthogonality constraint, we have that

$$\boldsymbol{\Phi}_i'\boldsymbol{\Phi}_j = \sum_{k=1}^{K}\boldsymbol{\phi}_{ik}'\boldsymbol{\phi}_{jk} = 0,$$

for some i such that $1 \leq i \leq KP$ and for all $j \neq i$. Letting

$$\mathbf{L}_{-ip} = \begin{bmatrix} \boldsymbol{\phi}_{i1} \\ \vdots \\ \boldsymbol{\phi}_{i(p-1)} \\ \boldsymbol{\phi}_{i(p+1)} \\ \vdots \\ \boldsymbol{\phi}_{i(KP)} \end{bmatrix} \quad \text{and} \quad \mathbf{c}_{-ip} = \begin{bmatrix} \sum_{k \neq i} \boldsymbol{\phi}'_{kp} \boldsymbol{\phi}_{k1} \\ \vdots \\ \sum_{k \neq i} \boldsymbol{\phi}'_{kp} \boldsymbol{\phi}_{k(p-1)} \\ \sum_{k \neq i} \boldsymbol{\phi}'_{kp} \boldsymbol{\phi}_{k(p+1)} \\ \vdots \\ \sum_{k \neq i} \boldsymbol{\phi}'_{kp} \boldsymbol{\phi}_{k(KP)} \end{bmatrix},$$

we can write the constraint as

$$\boldsymbol{\phi}_{ip} \mathbf{L}_{-ip} = -\mathbf{c}_{-ip},$$

for $1 \leq i \leq KP$ and $1 \leq p \leq K$. Using the results in Kowal et al. [2017], we have that $\boldsymbol{\phi}_{ip} \sim \mathcal{N}(\tilde{\mathbf{M}}_{ip}\mathbf{m}_{ip}, \tilde{\mathbf{M}}_{ip})$, where

$$\tilde{\mathbf{M}}_{ip} = \mathbf{M}_{ip} - \mathbf{M}_{ip}\mathbf{L}_{-ip}\left(\mathbf{L}'_{-ip}\mathbf{M}_{ip}\mathbf{L}_{-ip}\right)^{-1}\left(\mathbf{L}'_{-ip}\mathbf{M}_{ip} + \mathbf{c}_{-ip}\right).$$

Like in Kowal et al. [2017], $\mathbf{M}_{ip}$ and $\mathbf{m}_{ip}$ are such that when we relax the orthogonal constraints, we have $\boldsymbol{\phi}_{ip} \sim \mathcal{N}(\mathbf{M}_{ip}\mathbf{m}_{ip}, \mathbf{M}_{ip})$. By using this alternate sampling scheme, one can ensure the orthogonality of the $\boldsymbol{\Phi}$ parameters.

A.2.2 Multiple Start Algorithm

Due to the flexible nature of our model, we often end up with multimodal posterior distributions, which makes posterior inference challenging. In addition to tempered transitions (described in Section A.2.3), we implement an algorithm called the multiple start algorithm (MSA) in order to obtain a good starting position for our Markov chain. The MSA, Algorithm 1, starts by first trying to recover the mean and allocation structure. Once a suitable starting point for the mean and allocation parameters are found, we then estimate to covariance structure conditioned on the starting point for the mean and allocation parameters.

Algorithm 1 Multiple Start Algorithm

Require: n_try1, n_try2, $Y, K, n\_MCMC1, n\_MCMC2, \ldots$

$P \leftarrow$ `BPMM_Nu_Z(Y, K, n_MCMC1, ...)` ▷ Returns the likelihood and estimates for $\boldsymbol{\nu}$ and $\mathbf{Z}$

max_likelihood $\leftarrow P$["likelihood"]

$i \leftarrow 1$

while $i \leq$ n_try1 **do**

  $P_i \leftarrow$ `BPMM_Nu_Z(Y, K, n_MCMC1, ...)`

  if max_likelihood $< P_i$["likelihood"] **then**

    max_likelihood $\leftarrow P$["likelihood"]

    $P \leftarrow P_i$

  end if

  $i \leftarrow i + 1$

end while

$\theta \leftarrow$ `BPMM_Theta(P, Y, K, n_MCMC2, ...)` ▷ Returns estimates for the rest of the parameters

max_likelihood $\leftarrow \theta$["likelihood"]

$i \leftarrow 1$

while $i \leq$ n_try2 **do**

  $\theta_i \leftarrow$ `BPMM_Theta(P, Y, K, n_MCMC2, ...)`

  if max_likelihood $< \theta_i$["likelihood"] **then**

    max_likelihood $\leftarrow \theta_i$["likelihood"]

    $\theta \leftarrow \theta_i$

  end if

  $i \leftarrow i + 1$

end while

return (θ, P) ▷ Returns estimates for all model parameters

We can see that the MSA primarily calls two functions, `BPMM_Nu_Z(...)` and `BPMM_Theta(...)`. The first function, `BPMM_Nu_Z(...)`, finds initial starting points for the $\boldsymbol{z}_i$ parameters, $\boldsymbol{\nu}_k$ parameters, and related hyperparameters, while setting χim and $\boldsymbol{\phi}_{km}$ equal to 0 (or $\mathbf{0}$). The second function, `BPMM_Theta(...)`, finds initial starting points for the χ_{im} parameters, $\boldsymbol{\phi}$ parameters, σ^2, and related hyperparameters, conditioning on the initial starting point of the $\mathbf{z}_i$ and $\boldsymbol{\nu}_k$ parameters. In order to get the best results, we recommend standardizing the raw data before performing inference. The multiple start algorithm can be easily implemented in R using the accompanying software package to this paper.

A.2.3 Tempered Transitions

As stated in the previous section, the posterior distribution may often be multimodal, which often causes traditional MCMC methods to get stuck in local modes. In order to be able to move across modes, we implement tempered transitions, which will allow us to traverse areas of low posterior probability.

Following the works of Behrens et al. [2012] and Pritchard et al. [2000], we will only temper the likelihood of the model, which can often be written as

$$p(x) \propto \pi(x) exp\left(-\beta_h h(x)\right), \tag{A.22}$$

where β_h controls how much the distribution is tempered. We will assume $1 = \beta_0 < \cdots < \beta_h < \cdots < \beta_{N_t}$ and that the hyperparameters N_t and β_{N_t} are user specified. For larger and more complex models, we will often need a larger N_t, however our tempered transitions will be more computationally expensive with larger N_t. We will assume that the parameters β_h

follow a geometric scheme. We can rewrite our likelihood to fit Equation A.22:

$$
\begin{aligned}
p_h(\mathbf{y}_i|\boldsymbol{\Theta}) \propto & exp\left\{-\beta_h\left(\frac{1}{2}log(\sigma^2)+\frac{1}{2\sigma^2}\left(\mathbf{y}_i-\sum_{k=1}^{K}Z_{ik}\left(\boldsymbol{\nu}_k+\sum_{m=1}^{M}\chi_{im}\boldsymbol{\phi}_{km}\right)\right)'\right.\right. \\
& \left.\left.\left(\mathbf{y}_i-\sum_{k=1}^{K}Z_{ik}\left(\boldsymbol{\nu}_k+\sum_{m=1}^{M}\chi_{im}\boldsymbol{\phi}_{km}\right)\right)\right)\right\} \\
= & \left(\sigma^2\right)^{-\beta_h/2} exp\left\{-\frac{\beta_h}{2\sigma^2}\left(\mathbf{y}_i-\sum_{k=1}^{K}Z_{ik}\left(\boldsymbol{\nu}_k+\sum_{m=1}^{M}\chi_{im}\boldsymbol{\phi}_{km}\right)\right)'\right. \\
& \left.\left(\mathbf{y}_i-\sum_{k=1}^{K}Z_{ik}\left(\boldsymbol{\nu}_k+\sum_{m=1}^{M}\chi_{im}\boldsymbol{\phi}_{km}\right)\right)\right\}.
\end{aligned}
$$

Let $\boldsymbol{\Theta}_h$ be the set of parameters generated from the model using the tempered likelihood associated with β_h. The tempered transition algorithm can be summarized by the following steps:

1. Start with initial state $\boldsymbol{\Theta}_0$.
2. Transition from $\boldsymbol{\Theta}_0$ to $\boldsymbol{\Theta}_1$ using the tempered likelihood associated with β_1.
3. Continue in this manner until we transition from $\boldsymbol{\Theta}_{N_t-1}$ to $\boldsymbol{\Theta}_{N_t}$ using the tempered likelihood associated with β_{N_t}.
4. Transition from $\boldsymbol{\Theta}_{N_t}$ to $\boldsymbol{\Theta}_{N_t+1}$ using the tempered likelihood associated with β_{N_t}.
5. Continue in this manner until we transition from $\boldsymbol{\Theta}_{2N_t-1}$ to $\boldsymbol{\Theta}_{2N_t}$ using β_1.
6. Accept transition from $\boldsymbol{\Theta}_0$ to $\boldsymbol{\Theta}_{2N_t}$ with probability

$$
\min\left\{1, \prod_{h=0}^{N_t-1}\frac{\prod_{i=1}^{N}p_{h+1}(\mathbf{y}_i)|\boldsymbol{\Theta}_h)}{\prod_{i=1}^{N}p_h(\mathbf{y}_i|\boldsymbol{\Theta}_h)}\prod_{h=N_t+1}^{2N_t}\frac{\prod_{i=1}^{N}p_h(\mathbf{y}_i)|\boldsymbol{\Theta}_h)}{\prod_{i=1}^{N}p_{h+1}(\mathbf{y}_i)|\boldsymbol{\Theta}_h)}\right\}.
$$

Since we only temper the likelihood, many of the posterior distributions from Section A.2.1 can be used. Thus we will only have to modify the posterior distributions for the $\boldsymbol{\nu}$, σ^2, χ,

$\boldsymbol{\phi}$, and $\mathbf{Z}$ parameters. We will start with the $(\boldsymbol{\phi})_h$ parameters. Letting

$$(\mathbf{m}_{km})_h = \frac{\beta_h}{(\sigma^2)_h} \sum_{i=1}^{N} \left((\chi_{im})_h \left(\mathbf{y}_i (Z_{ij})_h - (Z_{ij}) h^2 (\boldsymbol{\nu}_j)_h - (Z_{ij})_h^2 \sum_{n \neq m} (\chi_{in})_h (\boldsymbol{\phi}_{jn})_h \right. \right.$$
$$\left. \left. - \sum_{k \neq j} (Z_{ij})_h (Z_{ik})_h \left[(\boldsymbol{\nu}_k)_h + \sum_{n=1}^{M} (\chi_{in})_h (\boldsymbol{\phi}_{kn})_h \right] \right) \right),$$

and

$$(\mathbf{M}_{km})_h^{-1} = \frac{\beta_h}{(\sigma^2)_h} \sum_{i=1}^{N} \left((Z_{ij})_h^2 (\chi_{im})_h^2 \right) \mathbf{I}_P + (\mathbf{D}_{km})_h^{-1},$$

we have that

$$(\boldsymbol{\phi}_{km})_h | \boldsymbol{\Theta}_{-(\phi_{km})_h}, \mathbf{y}_1, \ldots, \mathbf{y}_N \sim \mathcal{N}((\mathbf{M}_{km})_h (\mathbf{m}_{km})_h, (\mathbf{M}_{km})_h).$$

The posterior distribution of the $(\mathbf{Z})_h$ parameters are still not commonly known distributions, so we have to use the Metropolis-Hastings algorithm. Thus, we have that

$$p((\mathbf{z}_i)_h | (\boldsymbol{\Theta}_{-(\mathbf{z}_i)_h})_h, \mathbf{y}_1, \ldots, \mathbf{y}_N) \propto \prod_{k=1}^{K} (Z_{ik})_h^{(\alpha_3)_h (\pi_k)_h - 1}$$
$$\times exp \left\{ -\frac{\beta_h}{2(\sigma^2)_h} \left(\mathbf{y}_i - \sum_{k=1}^{K} (Z_{ik})_h \left((\boldsymbol{\nu}_k)_h + \sum_{n=1}^{M} (\chi_{in})_h (\boldsymbol{\phi}_{kn})_h \right) \right)' \right.$$
$$\left. \left(\mathbf{y}_i - \sum_{k=1}^{K} (Z_{ik})_h \left((\boldsymbol{\nu}_k)_h + \sum_{n=1}^{M} (\chi_{in})_h (\boldsymbol{\phi}_{kn})_h \right) \right) \right\}.$$

We will use $Q((\mathbf{z}_i)_h' | (\mathbf{z}_i)_{h-1}) = Dir(a_{\mathbf{z}} (\mathbf{z}_i)_{h-1})$ for some large $a_{\mathbf{z}} \in \mathbb{R}^+$ as the proposal distribution. Thus the probability of accepting a proposed step is

$$A((\mathbf{z}_i)_h', (\mathbf{z}_i)_{h-1}) = \min \left\{ 1, \frac{P\left((\mathbf{z}_i)_h' | (\boldsymbol{\Theta}_{-(\mathbf{z}_i)_h'})_h, \mathbf{y}_1, \ldots, \mathbf{y}_N \right)}{P\left((\mathbf{z}_i)_{h-1} | \boldsymbol{\Theta}_{-(\mathbf{z}_i)_{h-1}}, \mathbf{y}_1, \ldots, \mathbf{y}_N \right)} \frac{Q\left((\mathbf{z}_i)_{h-1} | (\mathbf{z}_i)_h' \right)}{Q\left((\mathbf{z}_i)_h' | (\mathbf{z}_i)_{h-1} \right)} \right\}.$$

Next, letting

$$(\mathbf{B}_j)_h = \left(\frac{\beta_h}{(\tau_j)_h} \mathbf{I}_P + \frac{1}{(\sigma^2)_h} \mathbf{I}_P \sum_{i=1}^{N} (Z_{ij})_h^2 \right)^{-1}$$

and

$$\mathbf{b}_j = \frac{\beta_h}{(\sigma^2)_h} \sum_{i=1}^{N} (Z_{ij})_h \left(\mathbf{y}_i - \left(\sum_{k \neq j} (Z_{ik})_h (\boldsymbol{\nu}_k)_h \right) - \left(\sum_{k=1}^{K} \sum_{m=1}^{M} (Z_{ik})_h (\chi_{im})_h (\boldsymbol{\phi}_{km})_h \right) \right)$$

we have that

$$(\boldsymbol{\nu}_j)_h | \boldsymbol{\Theta}_{-(\boldsymbol{\nu}_j)_h}, \mathbf{y}_1, \ldots, \mathbf{y}_N \sim \mathcal{N}((\mathbf{B}_j)_h (\mathbf{b}_j)_h, (\mathbf{B}_j)_h).$$

The posterior distribution for $(\sigma^2)_h$ is a distribution that can be easily sampled from, so we can use Gibbs sampling to get posterior draws. Letting

$$(\beta_\sigma)_h = \frac{\beta_h}{2} \sum_{i=1}^{N} \left[\left(\mathbf{y}_i - \sum_{k=1}^{K} (Z_{ik})_h \left((\boldsymbol{\nu}_k)_h + \sum_{m=1}^{M} (\chi_{im})_h (\boldsymbol{\phi}_{km})_h \right) \right)' \left(\mathbf{y}_i - \sum_{k=1}^{K} (Z_{ik})_h \left((\boldsymbol{\nu}_k)_h + \sum_{m=1}^{M} (\chi_{im})_h (\boldsymbol{\phi}_{km})_h \right) \right) \right]$$

we have

$$(\sigma^2)_h | \boldsymbol{\Theta}_{-(\sigma^2)_h}, \mathbf{y}_1, \ldots, \mathbf{y}_N \sim IG \left(\alpha_0 + \frac{\beta_h PN}{2}, \beta_0 + (\beta_\sigma)_h \right).$$

Lastly, we can sample from $(\chi)_h$ using a Gibbs sampler to get posterior draws. Letting

$$(\mathbf{w}_{im})_h = \frac{\beta_h}{(\sigma^2)_h} \left(\left(\sum_{k=1}^{K} (Z_{ik})_h (\boldsymbol{\phi}_{km})_h \right)' \left(\mathbf{y}_i - \sum_{k=1}^{K} (Z_{ik})_h \left((\boldsymbol{\nu}_k)_h + \sum_{n \neq m} (\chi_{in})_h (\boldsymbol{\phi}_{kn})_h \right) \right) \right)$$

and

$$(\mathbf{W}_{im}^{-1})_h = 1 + \frac{\beta_h}{(\sigma^2)_h} \left(\left(\sum_{k=1}^{K} (Z_{ik})_h (\boldsymbol{\phi}_{km})_h \right)' \left(\sum_{k=1}^{K} (Z_{ik})_h (\boldsymbol{\phi}_{km})_h \right) \right),$$

then we have that

$$(\chi_{im})_h | \boldsymbol{\zeta}_{-(\chi_{im})_h}, \mathbf{y}_1, \ldots, \mathbf{y}_N \sim \mathcal{N}((\mathbf{W}_{im})_h (\mathbf{w}_{im})_h, (\mathbf{W}_{im})_h).$$

As stated before, complex models will often require large N_t to accept the tempered transition proposed states. Unfortunately, this can be very computationally expensive, which is why we recommend using a mixture of tempered transitions and standard sampling techniques as described in Section A.2.1. From proposition 1 of Roberts and Rosenthal [2007], we know that an independent mixture of tempered transitions and untempered transitions will preserve the stationary distribution of the Markov chain.

A.2.4 Membership Rescale Algorithm

As discussed in Section 2.1.4, our model can be unidentifiable. To make the clusters more interpretable, we will rescale the allocation parameters Z_{ik} such that in the two feature model, at least one observation belongs completely to each feature. This specific assumption that one observation belongs entirely in each feature is known as the *seperability* condition [Papadimitriou et al., 1998, McSherry, 2001, Azar et al., 2001, Chen et al., 2022]. Thus in order to ensure identifiability, algorithm 2 can be used when we only have two features. In the case where there are more than two features, the assumption of seperability can be relatively strong, and weaker geometric assumptions, such as the *sufficiently scattered* condition [Huang et al., 2016, Jang and Hero, 2019, Chen et al., 2022] can be used to ensure identifiability. From Chen et al. [2022], we have that an allocation matrix $\mathbf{Z}$ is sufficiently scattered if:

1. $\text{cone}(\mathbf{Z}')^* \subseteq \mathcal{K}$
2. $\text{cone}(\mathbf{Z}')^* \cap bd\mathcal{K} \subseteq \{\lambda \mathbf{e}_f, f = 1, \ldots, k, \lambda \geq 0\}$

where $\mathcal{K} := \left\{\mathbf{x} \in \mathbb{R}^K | \|\mathbf{x}\|_2 \leq \mathbf{x}'\mathbf{1}_K\right\}$, $bd\mathcal{K} := \left\{\mathbf{x} \in \mathbb{R}^K | \|\mathbf{x}\|_2 = \mathbf{x}'\mathbf{1}_K\right\}$,
$\text{cone}(\mathbf{Z}')^* := \left\{\mathbf{x} \in \mathbb{R}^K | \mathbf{x}\mathbf{Z}' \geq 0\right\}$, and $\mathbf{e}_f$ is a vector with the i^{th} element equal to 1 and zero elsewhere. The first condition can be interpreted as the allocation parameters should from

a convex polytope that contains the dual cone $\mathcal{K}^*$. Thus we have that

$$\text{Conv}(\mathbf{Z}') \subseteq \mathcal{K}^*,$$

where $\mathcal{K}^* := \left\{\mathbf{x} \in \mathbb{R}^K | \mathbf{x}'\mathbf{1}_K \geq \sqrt{k-1}\|\mathbf{x}\|_2\right\}$ and $\text{Conv}(\mathbf{Z}') := \{\mathbf{x} \in \mathbb{R}^K | \mathbf{x} = \mathbf{Z}'\lambda, \lambda \in \Delta^N\}$, where Δ^k denotes the N-dimensional simplex. Ensuring that these conditions are met in our proposed model is non-trivial. The major non-identifiabilty problem we wish to solve is the *rescaling problem* discussed in Section 2.1.4. Therefore, we will focus on trying to promote allocation structures such that the first condition is satisfied. Similarly to the case of two functional features, we aim to find a linear transformation such that the convex polytope of our transformed allocation parameters covers the most area. Thus letting $\mathbf{T} \in \mathbb{R}^K \times \mathbb{R}^K$ be our transformation matrix, we aim to solve the following optimization problem:

$$\begin{aligned} &\max_{\mathbf{T}} \ |\text{Conv}(\mathbf{T}\mathbf{Z}')| \\ &s.t. \ \ \mathbf{z}_i\mathbf{T} \in \mathcal{C} \ \ \forall \, i, \end{aligned}$$

where $|\text{Conv}(\mathbf{T}\mathbf{Z}')|$ denotes the volume of the convex polytope constructed by the allocation parameters. Since the second condition is likely not met, we cannot ensure that our model is identifiable. However, the model is more interpretable, making inference easier for the end user. Once the transformation matrix ($\mathbf{T}$) is found, then we can rescale the allocation parameters ($\mathbf{z}_i$) and the corresponding mean and covariance parameters ($\boldsymbol{\nu}_k$ and $\boldsymbol{\phi}_{km}$). From empirical evidence, we found that the membership rescale algorithm is almost always needed in the case when we have only two features, however when we have more than 2 features, the rescaling may often not be needed.

Algorithm 2 Membership Rescale Algorithm

Require: $\mathbf{Z}, \boldsymbol{\nu}, \boldsymbol{\Phi}, M$

```
T ← matrix(0, 2, 2)                          ▷ Initialize inverse transformation matrix (2 x 2)
i ← 1
while i ≤ 2 do
    max_ind ← max_ind(Z[, i])                ▷ Find index of max entry in i^th column
    T[i, ] ← (Z[max_ind, ])
    i ← i + 1
end while
Z_t ← Z × inv(T)                             ▷ Transform the Z parameters
ν_t ← T × ν                                  ▷ Transform the ν parameters
i ← 1
while i ≤ M do
    Φ_t[, , i] ← T × Φ[, , i]                ▷ Transform the Φ parameters
    i ← i + 1
end while
return (Z_t, ν_t, Φ_t)
```

A.3 Simulations and Case Studies

A.3.1 Simulation Study 1

In this simulation study, we empirically study the convergence properties of our model on simulated data. In this simulation study, we considered a two feature mixed membership model, where the observed data are 10-dimensional vectors ($y \in \mathbb{R}^{10}$). Since we have an identifiability problem between the mean and covariance parameters, we will have to ensure that the $\boldsymbol{\nu}$ parameters are orthogonal to the $\boldsymbol{\phi}$ parameters. Thus in order to generate the dataset, we first will generate the model parameters in the following way:

$$\boldsymbol{\nu}_k \sim \mathcal{N}(\mathbf{0}_{10}, 9\mathbf{I}_{10}),$$

$$\chi_{im} \sim \mathcal{N}(0, 1),$$

for $k = 1, 2$ and $m = 1, \ldots, 4$. In order to ensure that the $\boldsymbol{\phi}$ parameters are orthogonal to the mean parameters, we will let $\mathbf{B}^{\perp} := \{\mathbf{b}_1, \ldots, \mathbf{b}_8\}$ be a set of basis vectors such that

the $\mathbf{B}^{\perp}$ spans the subspace orthogonal to the ν parameters. Thus we can generate the $\boldsymbol{\phi}$ parameters in the following way:

$$\boldsymbol{\phi}_{km} = \mathbf{q}_{km}\mathbf{B}^{\perp},$$

where $\mathbf{q}_{k1} \sim \mathcal{N}(\mathbf{0}_8, \mathbf{I}_8)$, $\mathbf{q}_{k2} \sim \mathcal{N}(\mathbf{0}_8, 0.49\mathbf{I}_8)$, $\mathbf{q}_{k3} \sim \mathcal{N}(\mathbf{0}_8, 0.25\mathbf{I}_8)$, and $\mathbf{q}_{k4} \sim \mathcal{N}(\mathbf{0}_8, 0.09\mathbf{I}_8)$. The allocation parameters $\mathbf{z}_i$ were drawn from a mixture of Dirichlet distributions. Roughly 30% of the allocation parameters were dawn from a Dirichlet distribution with $\alpha_1 = 10$ and $\alpha_2 = 1$. Another roughly 30% were drawn from a Dirichlet distribution with $\alpha_1 = 1$ and $\alpha_2 = 10$. The final roughly 40% of the allocation parameters were drawn from a Dirichlet distribution with $\alpha_1 = \alpha_2 = 1$. Lastly, σ^2 was set to 0.01 for this simulation study. Once all of the parameters were drawn, we generated a dataset and repeated this process 50 times for each of the three different sample sizes ($i = 50, 250, 1000$). A mixed membership model was then fit for each of the datasets using 200,000 MCMC iterations saving only every 100 iterations (n_try1 = 150, n_try2 = 10, n_MCMC1 = 4000, n_MCMC2 = 4000, M = 4). Lastly, convergence metrics were calculated and displayed in Figure 2.2.

A.3.2 Simulation Study 2

In this simulation study, we evaluate the performance of various information criteria in choosing the number of features in our proposed mixed membership model. To evaluate the information criteria (BIC, AIC, and DIC), we fit multiple mixed membership models with as little as 2 features to as many as 5 features ($K = 2, \ldots, 5$) on 50 different datasets. The datasets were generated from our proposed mixed membership model with 3 features. In order to generate the datasets, we first randomly generated the parameters of our model. For each of the 50 datasets, the parameters were drawn in the following way:

$$\boldsymbol{\nu}_k \sim \mathcal{N}(\mathbf{0}_{20}, 10\mathbf{I}_{20}),$$

$$\boldsymbol{\phi}_{i1} \sim \mathcal{N}(\mathbf{0}_{20}, \mathbf{I}_{20}),$$

$$\boldsymbol{\phi}_{i2} \sim \mathcal{N}(\mathbf{0}_{20}, 0.5\mathbf{I}_{20}),$$

$$\boldsymbol{\phi}_{i3} \sim \mathcal{N}(\mathbf{0}_{20}, 0.2\mathbf{I}_{20}),$$

$$\chi_{im} \sim \mathcal{N}(0, 1),$$

where $i = 1, \ldots, 200$ and $m = 1, \ldots, 3$. Similarly to simulation study 1, the allocation parameters were drawn from a mixture of Dirichlet distributions. Roughly 20% of the allocation parameters were dawn from a Dirichlet distribution with $\alpha_1 = 10$ and $\alpha_2 = 1$. Another roughly 20% were drawn from a Dirichlet distribution with $\alpha_1 = 1$ and $\alpha_2 = 10$. The final roughly 60% of the allocation parameters were drawn from a Dirichlet distribution with $\alpha_1 = \alpha_2 = 1$. Similarly, we set σ^2 equal to 0.01 for all 50 datasets. Once all of the parameters were drawn, the 200 observation ($\mathbf{y}_i \in \mathbb{R}^{20}$) datasets were drawn. Once the data sets were created, we fit 4 models ($K = 2, \ldots, 5$) using a MCMC with 100,000 iterations, saving only every 10 iterations, with the following hyperparameters: n_try1 = 50, n_try2 = 5, n_MCMC1 = 4000, n_MCMC2 = 10000, M = 4.

The first IC we considered for this simulation study is the Bayesian information criterion (BIC). The BIC, proposed by Schwarz [1978], is defined as:

$$\text{BIC} = 2 \log P\left(\mathbf{Y}|\hat{\boldsymbol{\Theta}}\right) - d \log(N)$$

where d is the number of parameters, $\hat{\boldsymbol{\Theta}}$ is the collection of maximum likelihood estimators (MLE) of our parameters, and $\mathbf{Y} = \{\mathbf{y}_i\}_{i=1}^N$ is the collection of our observed vectors. In the case of our mixed membership model, we have that

$$\text{BIC} = 2 \log P\left(\mathbf{Y}|\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\Phi}}, \hat{\sigma}^2, \hat{\mathbf{Z}}, \hat{\boldsymbol{\chi}}\right) - d \log(N) \tag{A.23}$$

where $d = (N + P)K + 2MKP + 4K + (N + K)M + 2$.

Similarly, the Akaike IC (AIC), proposed by Akaike [1974], can be written as

$$\text{AIC} = -2 \log P\left(\mathbf{Y}|\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\Phi}}, \hat{\sigma}^2, \hat{\mathbf{Z}}, \hat{\boldsymbol{\chi}}\right) + 2d. \tag{A.24}$$

Following the work of Roeder and Wasserman [1997], we will use the posterior mean instead of the MLE for our estimates of BIC and AIC. Due to identifiability problems, the posterior mean of the mean component in Equation 2.12 for each observation, as well as the posterior mean of σ^2, will be used to estimate the BIC and AIC instead of estimates of the posterior mean for each individual parameter.

The modified Deviance IC (DIC), proposed by Celeux et al. [2006], is advantageous to the original DIC proposed by Spiegelhalter et al. [2002] when we have a posterior distribution with multiple modes, and when identifiability may be a problem. The modified DIC (referred to as DIC_3 in Celeux et al. [2006]) is specified as:

$$\text{DIC} = -4\mathbb{E}_{\boldsymbol{\Theta}}[\log f(\mathbf{Y}|\boldsymbol{\Theta})|\mathbf{Y}] + 2log\hat{f}(\mathbf{Y}) \tag{A.25}$$

where $\hat{f}(\mathbf{y}_i) = \frac{1}{N_{MC}} \sum_{l=1}^{N_{MC}} P\left(\mathbf{y}_i|\boldsymbol{\nu}^{(l)}, \boldsymbol{\Phi}^{(l)}, (\sigma^2)^{(l)}, \mathbf{Z}^{(l)}\right)$, $\hat{f}(\mathbf{Y}) = \prod_{i=1}^{N} \hat{f}(\mathbf{y}_i)$, and N_{MC} is the number of MCMC samples used for estimating $\hat{f}(\mathbf{y}_i)$. We can approximate $\mathbb{E}_{\boldsymbol{\Theta}}[\log f(\mathbf{Y}|\boldsymbol{\Theta})|\mathbf{Y}]$ by using the MCMC samples, such that

$$\mathbb{E}_{\boldsymbol{\Theta}}[\log f(\mathbf{Y}|\boldsymbol{\Theta})|\mathbf{Y}] \approx \frac{1}{N_{MC}} \sum_{l=1}^{N_{MC}} \sum_{i=1}^{N} \log\left[P\left(\mathbf{y}_i|\boldsymbol{\nu}^{(l)}, \boldsymbol{\Phi}^{(l)}, \left(\sigma^2\right)^{(l)}, \mathbf{Z}^{(l)}\right)\right].$$

A.3.3 EEG Case Study

In this case study, we analyze resting-state EEG data from typically developing (TD) children and children with Autism spectrum disorder (ASD) [Dickinson et al., 2018]. For this case study, we fit a 2 feature mixed membership model and a 3 feature mixed membership model with 5 eigenvectors ($M = 5$) for each model. Using AIC and BIC to help inform our

choice on the number of features, we find that the 2 feature model seems to be a better model for the data ($AIC_2 = -12905.6$, $AIC_3 = -12204.5$, $BIC_2 = 9236.7$, $BIC_3 = 7328.0$, $DIC_2 = -14010.5$, $DIC_3 = -14197.9$). To get a good starting position, we used the multiple start algorithm (Algorithm 1) with n_try1 = 50, n_try2= 50, n_MCMC1 = 8000, and n_MCMC2 = 8000. Once we had our initial starting position, we ran a Markov chain for 500,000 iterations, saving only every 10 iterations. Figure A.1 shows the recovered covariance structure from our mixed membership model. We can see that the covariance structure for feature 1 accounts for the shift in the alpha peak that was found in Scheffler et al. [2019]. On the other hand, we can see that most of the variation in feature 2 is in the low frequency range, which is where we expect the most pink noise.

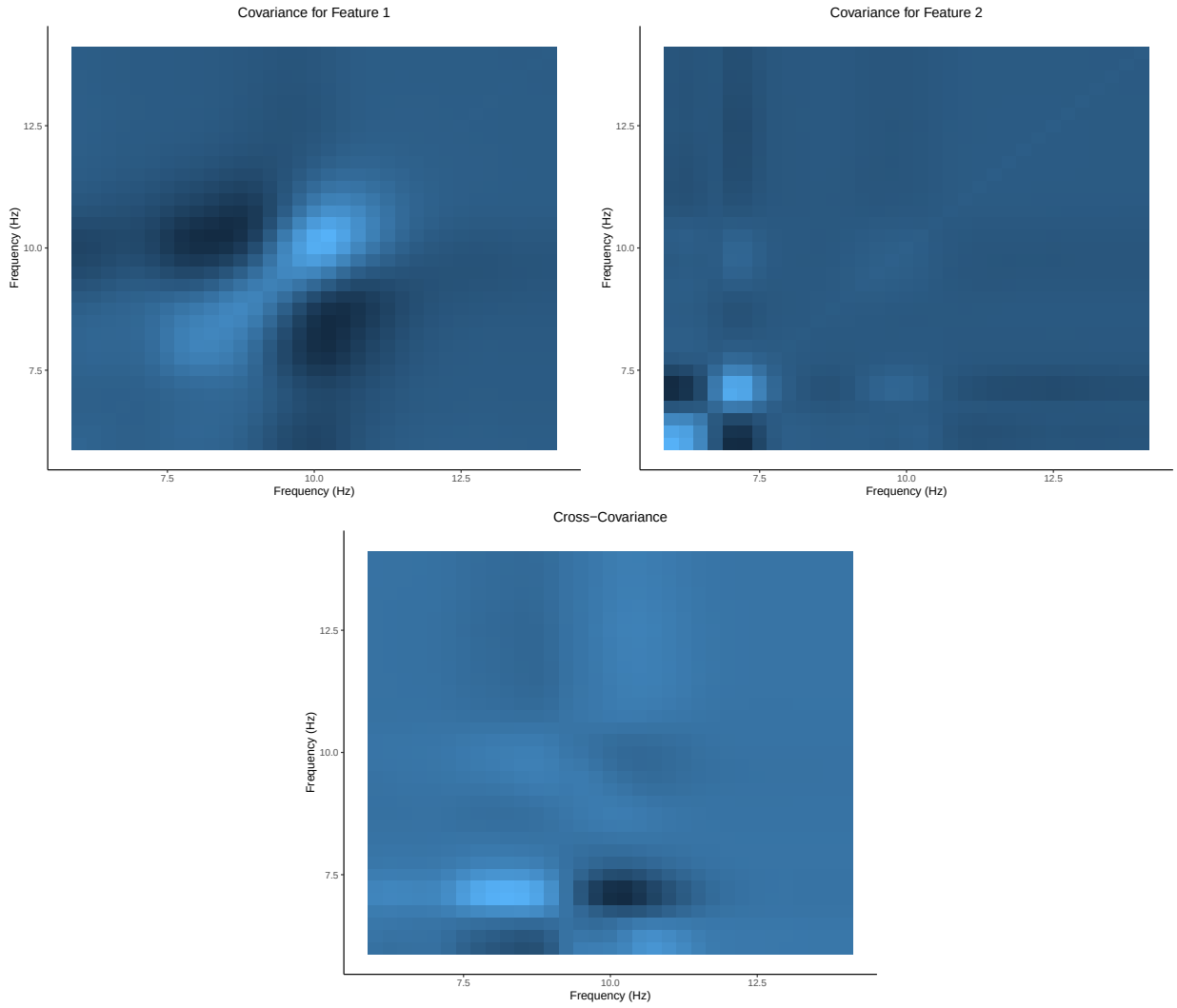


Figure A.1: Visualization of the covariance structure for the two feature mixed membership model. Light blue represents positive covariance, while dark blue represents negative covariance.

A.3.4 Molecular Subtypes of Breast Cancer

For this case study, we used the data provided in Parker et al. [2009], and only used the observations labeled as LumA, Her2, or Basal ($N = 115$). The data set contained some missing values, so we used MICE [Van Buuren and Groothuis-Oudshoorn, 2011] to impute the missing data. To get a good starting position, we used the multiple start algorithm

(Algorithm 1) with n_try1 = 50, n_try2 = 6, n_MCMC1 = n_MCMC2 = 10000, and K = 3. Using the informed starting position, we then ran our Markov chain for 500,000 iterations, saving every 10^{th} iteration. The parameters were then rescaled for ease of interpretation by using the membership rescale algorithm (Algorithm 2). From Figure A.2, we can see the correlation structure in each of the 3 features. We can see that there is relatively high correlation between many of the genes in feature 1 (corresponding to the LumA cancer subtype).

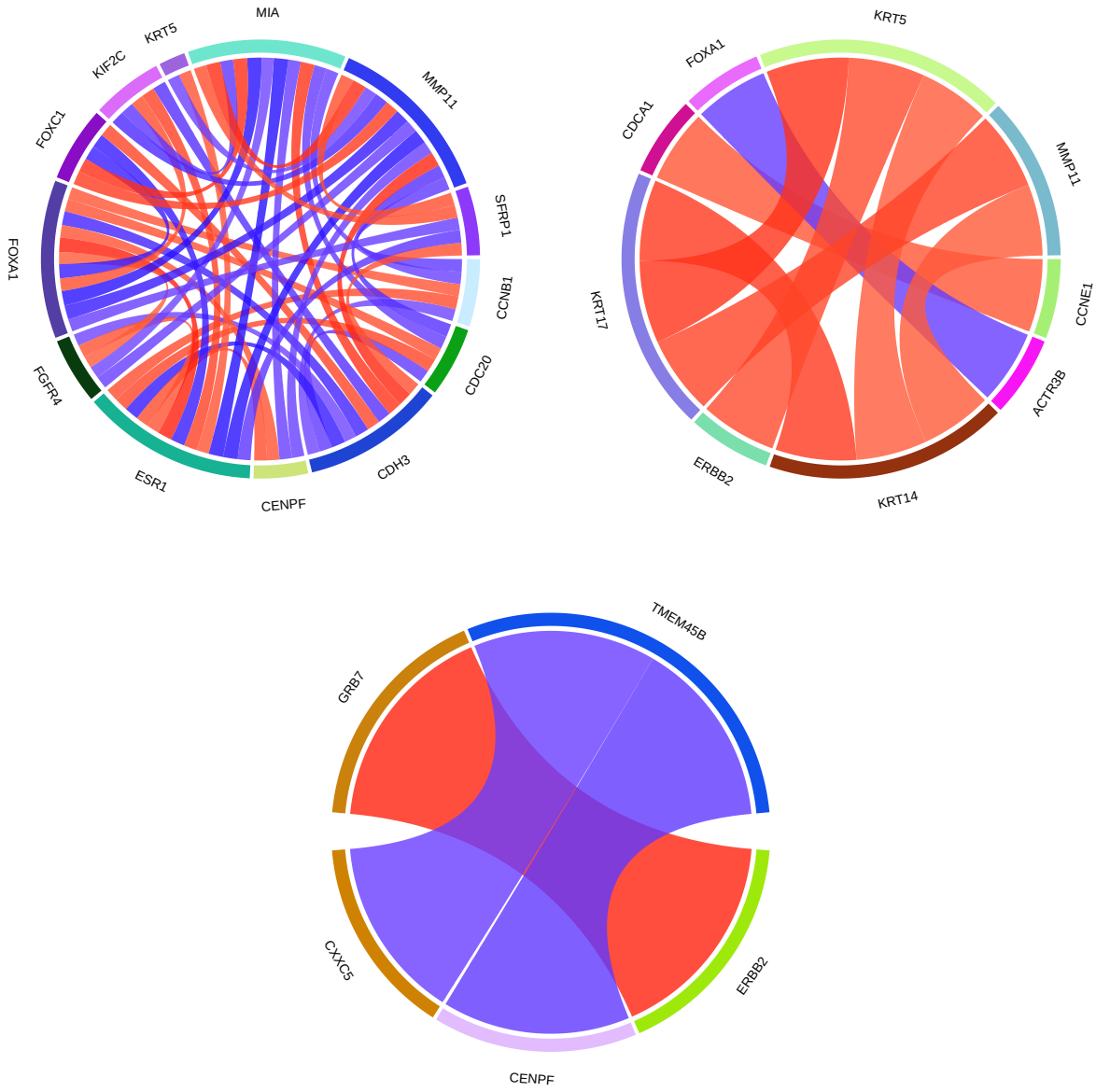


Figure A.2: Visualization of the correlation structure of the each feature (Feature 1: Top Left, Feature 2: Top Right, Feature 3: Bottom Middle). Positive correlation is depicted by a red chord, while negative correlation is depicted by a blue chord. Pairwise correlations of less than 0.8 were omitted from the diagrams above.

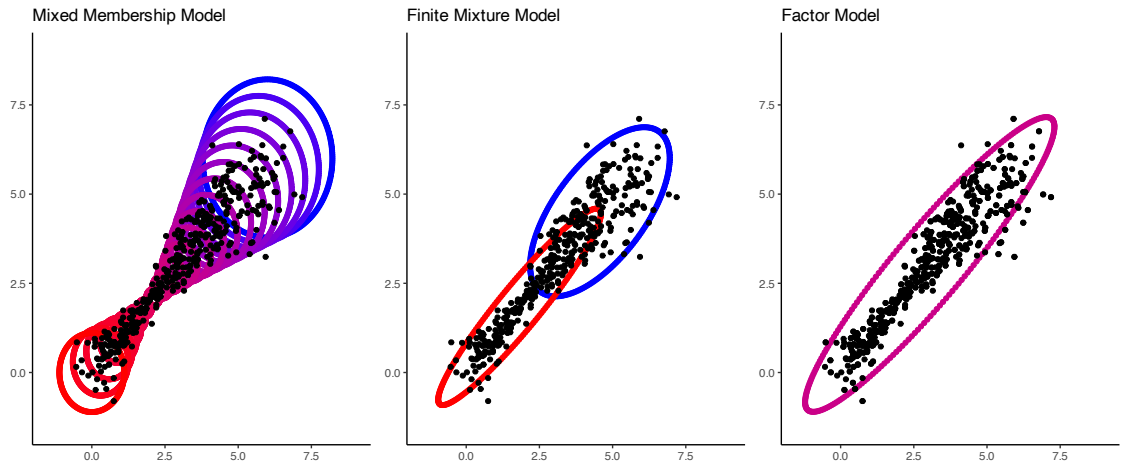


Figure A.3: Comparative visualization of the differences between mixed membership models, finite mixture models, and factor models. Each of the models were fit on the same set of data, illustrated by the black dots.

A.4 Factor Models and Mixed Membership Models

Mixed membership models for continuous data, as encoded in our representation in (2.5), are related to latent factor models, as they rely on similar additive structures. Nevertheless, mixed membership models obtain an alternative decomposition of the data variance, leading to a different interpretation of the model parameters. An illustration of the differences between Gaussian finite mixture models, factor models, and our proposed mixed membership model can be seen in Figure A.3.

Factor models are a common tool used in multivariate analysis to model dependence in high-dimensions through a lower-dimensional linear combination of latent *factors* [Bernardo et al., 2003, Carvalho et al., 2008, Bhattacharya and Dunson, 2011]. The general form of a factor model can be written as

$$\mathbf{y}_i - \boldsymbol{\mu} = \mathbf{B}\boldsymbol{\lambda}_i + \boldsymbol{\nu}_i,$$

where $\mathbf{B} \in \mathbb{R}^{P\times K}$ is known as a matrix of *factor loadings* and $\boldsymbol{\lambda}_i \sim \mathcal{N}_P(\mathbf{0}, \mathbf{I}_P)$ are known as *latent factors*. The parameters $\boldsymbol{\nu}_i \sim \mathcal{N}_P(\mathbf{0}, \boldsymbol{\Sigma})$ are parameters accounting for random error, where $\boldsymbol{\Sigma}$ is a $P \times P$ diagonal matrix. Integrating out the latent factors, factor models

generally assume the following distribution on our data:

$$\mathbf{y}_i \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{B}\mathbf{B}' + \boldsymbol{\Sigma}).$$

Using this parameterization, we can see how factor models are useful in estimating high dimensional covariance matrices using a low-dimensional representation using factors.

Factor models can be written in an alternative representation that look similar to the mixed membership model in Equation 2.13. Treating the latent factors in a similar fashion as the allocation parameters in Equation 2.13, we arrive at

$$\mathbf{y}_i | \boldsymbol{\lambda}_i \sim \mathcal{N}\left(\boldsymbol{\mu} + \sum_{k=1}^{K} \lambda_{ik}\mathbf{b}_k, \boldsymbol{\Sigma}\right),$$

where λ_{ik} is the k^{th} element of $\boldsymbol{\lambda}_i$ and $\mathbf{b}_k$ is the k^{th} column of $\mathbf{B}$. If we try to interpret λ_{ik} in a similar way as the allocation parameters in our proposed mixed membership model, we have that the mean of the k^{th} feature becomes $\boldsymbol{\mu} + \mathbf{b}_k$. While the form of factor models may seem similar to our proposed mixed membership model, there are two key differences between the models. The first, and most important difference, is that λ_i do not lie on the unit simplex. This constraint greatly affects the estimation of the feature specific means, more than just a simple rescaling of the means. Constraining the allocation parameters also helps extremely with interpretability. Since $\mathbf{z}_i$ lie on the unit simplex, we can interpret the elements Z_{ik} as the i^{th} observation's proportion of membership to the k^{th} feature. On the other hand, $\sum_{k=1}^{K} \lambda_{ik}$ is not necessarily equal to 1, meaning we cannot interpret the λ_{ik} parameters in a similar fashion. Moreover, the λ_{ik} parameters can be negative, making interpretability of the λ_{ik} parameters challenging. The second key difference is that the factor model conditional on the latent factors has the same covariance, $\boldsymbol{\Sigma}$. Thus using a factor model, we cannot estimate the correlation structure stratified by feature (i.e. Figures 1 and 2 in the Supplementary Materials).

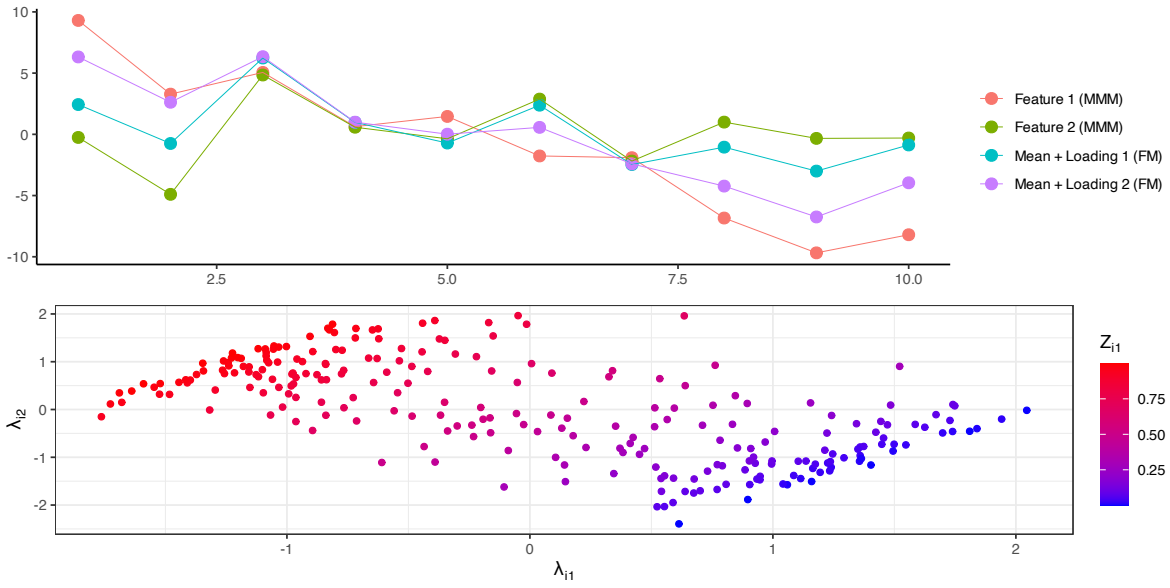


Figure A.4: Comparison between a factor model and our mixed membership model, fit on simulated data. The top subfigure illustrates the difference in the mean components between the two models, while the bottom subfigure illustrates the difference between the latent factors of a factor model and allocation parameters of a mixed membership model.

An illustration of the differences between factor models and mixed membership models can be seen in Figure A.4. To compare the differences between factor models and mixed membership models, we simulated 250 data points ($\mathbf{y}_i \in \mathbb{R}^{10}$) and fit a factor analysis model with 2 factors, as well as a mixed membership model with 2 features. Even though factor models and mixed membership models have a similar additive mean structure, we can see that the estimated means significantly differ due to the added constraint on the allocation parameters in a mixed membership model. Figure A.4 also illustrates that the allocation parameters ($\mathbf{z}_i$) are closely related to both factors in a factor model. However, trying to interpret the factors as membership to a cluster or feature is challenging because the factors lie $\mathbb{R}^2$, which is an unconstrained space. On the other hand, the allocation parameters can simply be represented on the unit interval, allowing for easy interpretation. Therefore, while there are similarities between factor models and mixed membership models, we can see that there are substantial differences between the two models. We maintain that while factor

models are a useful tool to estimated the covariance structure of high dimensional data, they are not well suited for the clustering-type problems discussed in this manuscript.

APPENDIX B

Appendix: Functional Mixed Membership Models

B.1 Proofs

B.1.1 Proof of Lemma 3

Proof. We will first show that $\boldsymbol{\mathcal{S}}$ is a linear subspace of $L^2(\mathcal{T})$. Let $w_1, w_2 \in \boldsymbol{\mathcal{S}}$, and let $\alpha_1, \alpha_2 \in \mathbb{R}$. Since $\boldsymbol{\mathcal{S}}$ is the space spanned by the square-integrable basis functions $b_1, \dots b_P$ $\left(\boldsymbol{\mathcal{S}} = \left\{\sum_{p=1}^{P} a_p b_p : a_i \in \mathbb{R}\right\}\right)$, we can write $w_1 = \sum_{p=1}^{P} \delta_p b_p$ and $w_2 = \sum_{p=1}^{P} \beta_p b_p$ for some $\delta_p, \beta_p \in \mathbb{R}$. Therefore we have that

$$\alpha_1 w_1 + \alpha_2 w_2 = \alpha_1 \left(\sum_{p=1}^{P} \delta_p b_p\right) + \alpha_2 \left(\sum_{p=1}^{P} \beta_p b_p\right).$$

Letting $\gamma_p = \alpha_1 \delta_p + \alpha_2 \beta_p$, we have that

$$\alpha_1 w_1 + \alpha_2 w_2 = \sum_{p=1}^{P} \gamma_p b_p \in \boldsymbol{\mathcal{S}}.$$

Therefore, by definition, we know that $\boldsymbol{\mathcal{S}}$ is a linear subspace of $L^2(\mathcal{T})$. Next, we will show that $\boldsymbol{\mathcal{S}}$ is a closed linear subspace. Let f_n be a Cauchy sequence in $\boldsymbol{\mathcal{S}}$. Thus by definition, for some $\epsilon > 0$, there exists a $m \in \mathbb{N}$ such that for $i, j > m$ we have

$$\|f_i - f_j\|_{\boldsymbol{\mathcal{S}}} < \epsilon. \tag{B.1}$$

Since $f_i, f_j \in \boldsymbol{\mathcal{S}}$, we know that $f_i, f_j \in \text{span}\{b_1, \ldots, b_P\}$. Thus using the Gram–Schmidt process, we know that there exists an orthonormal set of functions such that span $\{b_1, \ldots, b_P\}$ = span $\left\{\tilde{b}_1, \ldots, \tilde{b}_P\right\}$. Thus can expand f_i and f_j such that $f_i = \sum_{p=1}^{P} \alpha_{ip}\tilde{b}_p$ and $f_j = \sum_{p=1}^{P} \alpha_{jp}\tilde{b}_p$. Thus we can rewrite equation B.1 as

$$\begin{aligned}
\|f_i - f_j\|_{\boldsymbol{\mathcal{S}}} &= \left(\left\langle \sum_{p=1}^{P}(\alpha_{ip} - \alpha_{jp})\tilde{b}_p, \sum_{p=1}^{P}(\alpha_{ip} - \alpha_{jp})\tilde{b}_p \right\rangle\right)^{1/2} \\
&= \left(\sum_{p=1}^{P}\left\langle (\alpha_{ip} - \alpha_{jp})\tilde{b}_p, (\alpha_{ip} - \alpha_{jp})\tilde{b}_p \right\rangle\right)^{1/2} \\
&= \left(\sum_{p=1}^{P}\left|\left|(\alpha_{ip} - \alpha_{jp})\tilde{b}_p\right|\right|_{\boldsymbol{\mathcal{S}}}\right)^{1/2} \\
&= \left(\sum_{p=1}^{P}\int_{\mathcal{T}}\left((\alpha_{ip} - \alpha_{jp})\tilde{b}_p(t)\right)^2 \mathrm{d}t\right)^{1/2} \\
&= \left(\sum_{p=1}^{P}(\alpha_{ip} - \alpha_{jp})^2\int_{\mathcal{T}}\tilde{b}_p(t)^2\mathrm{d}t\right)^{1/2}. \qquad \text{(B.2)}
\end{aligned}$$

Since $\tilde{b}_p(t)$ are orthonormal, we know that $\int_{\mathcal{T}} \tilde{b}_p(t)^2\mathrm{d}t = 1$. Thus from equations B.1 and B.2, for $i, j > m$, we have that

$$\begin{aligned}
\epsilon &> \|f_i - f_j\|_{\boldsymbol{\mathcal{S}}} \\
&= \left(\sum_{p=1}^{P}(\alpha_{ip} - \alpha_{jp})^2\int_{\mathcal{T}}\tilde{b}_p(t)^2\mathrm{d}t\right)^{1/2} \\
&= \left(\sum_{p=1}^{P}(\alpha_{ip} - \alpha_{jp})^2\right)^{1/2}.
\end{aligned} \qquad \text{(B.3)}$$

Thus we can see that the sequence α_{ip} is a Cauchy sequence. Since the Euclidean space is a complete metric space, there exists $\alpha_p \in \mathbb{R}$ such that $\alpha_{ip} \to \alpha_p$. Letting $f = \sum_{p=1}^{P} \alpha_p\tilde{b}_p(t)$,

we have

$$\begin{aligned}\|f_i - f\| &= \left(\sum_{p=1}^{P}(\alpha_{ip} - \alpha)^2 \int_{\mathcal{T}} \tilde{b}_p(t)^2 \mathrm{d}t\right)^{1/2} \\ &= \left(\sum_{p=1}^{P}(\alpha_{ip} - \alpha_p)^2\right)^{1/2} \\ &= \sum_{p=1}^{P} ||\alpha_{ip} - \alpha_p||\,, \end{aligned} \tag{B.4}$$

for all $i, j > m$. By definition of $\alpha_{ip} \to \alpha_p$, we know that for $\epsilon_2 = \frac{\epsilon_1}{P}$, there exists a $m_1 \in \mathbb{N}$, such that for all $i > m_1$ we have $||\alpha_{ip} - \alpha_p|| < \epsilon_1$ for $p = 1, \ldots, P$. Thus from equation B.4, we have that for all $i > m_1$, we have

$$\|f_i - f\| < \epsilon_2.$$

Thus by definition, we have that the Cauchy sequence is convergent, and that $\boldsymbol{\mathcal{S}}$ is a closed linear subspace. □

B.1.2 Proof of Lemma 4

Proof. We will start by fixing $\epsilon > 0$. Notice that since b_j are uniformly continuous functions and $\mathcal{T}$ is a closed and bounded domain, we know that b_j is bounded. Thus let R be such that $|b_j(s)| < R$ for $j = 1, \ldots, P$ and any $s \in \mathcal{T}$. Let $\tilde{\epsilon} := \frac{\epsilon}{P^2 RM}$, where M is defined in (b) of lemma 4. Since $b_1, \ldots, b_k$ are uniformly continuous we have that there exists $\delta_i > 0$ such that for all $t, t_* \in \mathcal{T}$, we have

$$\|t - t_*\| < \delta_i \implies |b_i(t) - b_i(t_*)| < \tilde{\epsilon}, \tag{B.5}$$

for $i = 1, \ldots, P$. Define $\tilde{\delta} = \min_i \delta_i$. Thus from equation 3.8, if $\|t - t_*\| < \tilde{\delta}$, then we have

$$\begin{aligned}
\left|C^{(i,j)}(s,t) - C^{(i,j)}(s,t_*)\right| &= \left|\mathbf{B}'(s)\mathrm{Cov}\left(\boldsymbol{\theta}_i, \boldsymbol{\theta}_j\right)\left[\mathbf{B}(t) - \mathbf{B}(t_*)\right]\right| \\
&= \left|\sum_{k=1}^{P}\sum_{l=1}^{P} b_k(s)\mathrm{Cov}\left(\theta_{(i,k)}, \theta_{(j,l)}\right)\left[b_l(t) - b_l(t_*)\right]\right| \\
&\leq \left|\sum_{k=1}^{P}\sum_{l=1}^{P} b_k(s)M\left[b_l(t) - b_l(t_*)\right]\right| \\
&\leq \sum_{k=1}^{P}\sum_{l=1}^{P} \left|b_k(s)M\left[b_l(t) - b_l(t_*)\right]\right| \\
&= \sum_{k=1}^{P}\sum_{l=1}^{P} \left|b_k(s)M\right|\left|b_l(t) - b_l(t_*)\right|.
\end{aligned}$$

From equation B.5, we have that

$$\begin{aligned}
\left|C^{(i,j)}(s,t) - C^{(i,j)}(s,t_*)\right| &< \sum_{k=1}^{P}\sum_{l=1}^{P} \left|b_k(s)M\right|\tilde{\epsilon} \\
&\leq \sum_{k=1}^{P}\sum_{l=1}^{P} RM\tilde{\epsilon} \\
&= \epsilon.
\end{aligned}$$

Thus we have that for any $\epsilon > 0$, there exists a $\tilde{\delta} > 0$, such that for any $t, t_*, s \in \mathcal{T}$ and $1 \leq i \leq j \leq K$, we have

$$\|t - t_*\| < \tilde{\delta} \implies \left|C^{(i,j)}(s,t) - C^{(i,j)}(s,t_*)\right| < \epsilon. \tag{B.6}$$

Consider $\boldsymbol{\mathcal{B}}_Z := \{\mathbf{f} \in \boldsymbol{\mathcal{H}} : \|\mathbf{f}\| < Z\}$ for some $Z \in \mathbb{R}^+$. We will show that the family of functions $\boldsymbol{\mathcal{K}}\mathbf{f}_{\boldsymbol{\mathcal{B}}_Z} := \{\boldsymbol{\mathcal{K}}\mathbf{f} : \mathbf{f} \in \boldsymbol{\mathcal{B}}_Z\}$ is an equicontinuous set of functions. We will fix $\epsilon_1 > 0$.

Letting $\mathbf{f} \in \boldsymbol{\mathcal{B}}_Z$ and $t^{(i)}, t_*^{(i)} \in \mathcal{T}$ such that $\|t^{(i)} - t_*^{(i)}\| < \tilde{\delta}$, we have from equation 3.7 that

$$\begin{aligned}
\left|(\boldsymbol{\mathcal{K}}\mathbf{f})^{(i)}(\mathbf{t}) - (\boldsymbol{\mathcal{K}}\mathbf{f})^{(i)}(\mathbf{t}_*)\right| &= \left|\sum_{k=1}^{K}\int_{\mathcal{T}} C^{(k,i)}\left(s, t^{(i)}\right) f^{(k)}(s) - C^{(k,i))}\left(s, t_*^{(i)}\right) f^{(k)}(s)\mathrm{d}s\right| \\
&\leq \sum_{k=1}^{K}\int_{\mathcal{T}} \left|C^{(k,i)}\left(s, t^{(i)}\right) f^{(k)}(s) - C^{(k,i)}\left(s, t_*^{(i)}\right) f^{(k)}(s)\right| \mathrm{d}s \\
&= \sum_{k=1}^{K}\int_{\mathcal{T}} \left|C^{(k,i)}\left(s, t^{(i)}\right) - C^{(k,i)}\left(s, t_*^{(i)}\right)\right| \left|f^{(k)}(s)\right| \mathrm{d}s. \qquad \text{(B.7)}
\end{aligned}$$

Thus from equation $B.6$ we have that $\left|\left(C^{(k,i)}\left(s, t^{(i)}\right) - C^{(k,i)}\left(s, t_*^{(i)}\right)\right)\right| < \epsilon$. Notice that since $\mathbf{f} \in \mathcal{H}$, we know that $f^{(k)}(s)$ can be written as $f^{(k)}(s) = \sum_{i=1}^{P} a_i b_i(s)$. Since the sum of uniformly continuous functions is also a uniformly continuous function, we know that $f^{(k)}$ is uniformly continuous. Therefore, since $\mathcal{T}$ is a closed and bounded domain, we know that $f^{(k)}$ is bounded. Let M_1 be such that $|f^{(k)}| < M_1$. Thus we can write equation B.7 as

$$\begin{aligned}
\left|(\boldsymbol{\mathcal{K}}\mathbf{f})^{(i)}(\mathbf{t}) - (\boldsymbol{\mathcal{K}}\mathbf{f})^{(i)}(\mathbf{t}_*)\right| &< \sum_{k=1}^{K}\int_{\mathcal{T}} \epsilon M_1 \mathrm{d}s \\
&= \epsilon M_1 \sum_{k=1}^{K}\int_{\mathcal{T}} 1 \mathrm{d}s.
\end{aligned}$$

Since $\mathcal{T}$ is compact subset of $\mathbb{R}^d$, by the Bolzano–Weierstrass theorem, we know that $\mathcal{T}$ is closed and bounded. Therefore, let B be such that $\int_{\mathcal{T}} 1 \mathrm{d}t = B$. Thus, for $i = 1, \ldots, K$, we have

$$\left|(\boldsymbol{\mathcal{K}}\mathbf{f})^{(i)}(\mathbf{t}) - (\boldsymbol{\mathcal{K}}\mathbf{f})^{(i)}(\mathbf{t}_*)\right| < \epsilon M_1 K B.$$

Since

$$||(\boldsymbol{\mathcal{K}}\mathbf{f})(\mathbf{t}), (\boldsymbol{\mathcal{K}}\mathbf{f})(\mathbf{t}_*)|| = \left(\sum_{i=1}^{K}\left|(\boldsymbol{\mathcal{K}}\mathbf{f})^{(i)}(\mathbf{t}) - (\boldsymbol{\mathcal{K}}\mathbf{f})^{(i)}(\mathbf{t}_*)\right|^2\right)^{1/2},$$

we know that

$$||(\boldsymbol{\mathcal{K}}\mathbf{f})(\mathbf{t}), (\boldsymbol{\mathcal{K}}\mathbf{f})(\mathbf{t}_*)|| < \epsilon M_1 K^{3/2} B.$$

If we let $\epsilon = \frac{\epsilon_1}{M_1 K^{3/2} B}$ $\left(\tilde{\epsilon} = \frac{\epsilon_1}{M_1 K^{3/2} B P^2 R M}\right)$, then we have that $||(\boldsymbol{\mathcal{K}}\mathbf{f})(\mathbf{t}), (\boldsymbol{\mathcal{K}}\mathbf{f})(\mathbf{t}_*)|| < \epsilon_1$.Thus, from the assumption that b_j are uniformly continuous (equation B.5), we know there exists a $\tilde{\delta}$ such that for $j = 1, \ldots, K$, we have that

$$\|t - t_*\| < \tilde{\delta} \implies |b_j(t) - b_j(t_*)| < \frac{\epsilon_1}{M_1 K^{3/2} B P^2 R M} \implies ||(\boldsymbol{\mathcal{K}}\mathbf{f})(\mathbf{t}), (\boldsymbol{\mathcal{K}}\mathbf{f})(\mathbf{t}_*)|| < \epsilon_1. \quad \text{(B.8)}$$

Thus by definition, we have proved that $\boldsymbol{\mathcal{K}}\mathbf{f}_{\boldsymbol{\mathcal{B}}_Z}$ is an equicontinuous set of functions. Next, we will show that $\boldsymbol{\mathcal{K}}\mathbf{f}_{\boldsymbol{\mathcal{B}}_Z}$ is a family of uniformly bounded functions. If $t \in \mathcal{T}$, then we have

$$\begin{aligned}
|(\boldsymbol{\mathcal{K}}\mathbf{f})^{(i)}(\mathbf{t})| &= \left| \sum_{k=1}^{K} \int_{\mathcal{T}} \mathbf{B}'(s) \mathrm{Cov}(\boldsymbol{\theta}_k, \boldsymbol{\theta}_i) \mathbf{B}\left(t^{(i)}\right) f^{(k)}(s) \mathrm{d}s \right| \\
&= \left| \sum_{k=1}^{K} \int_{\mathcal{T}} \sum_{p=1}^{P} \sum_{l=1}^{P} b_p(s) \mathrm{Cov}\left(\theta_{(k,l)}, \theta_{(i,p)}\right) b_l\left(t^{(i)}\right) f^{(k)}(s) \mathrm{d}s \right| \\
&= \left| \sum_{k=1}^{K} \sum_{p=1}^{P} \sum_{l=1}^{P} b_l\left(t^{(i)}\right) \mathrm{Cov}\left(\theta_{(k,l)}, \theta_{(i,p)}\right) \int_{\mathcal{T}} b_p(s) f^{(k)}(s) \mathrm{d}s \right| \\
&\leq \left(\sum_{k=1}^{K} \sum_{p=1}^{P} \sum_{l=1}^{P} \left| b_l\left(t^{(i)}\right) \mathrm{Cov}\left(\theta_{(k,l)}, \theta_{(i,p)}\right) \int_{\mathcal{T}} b_p(s) f^{(k)}(s) \mathrm{d}s \right| \right) \\
&= \left(\sum_{k=1}^{K} \sum_{p=1}^{P} \sum_{l=1}^{P} \left| b_l\left(t^{(i)}\right) \right| \left| \mathrm{Cov}\left(\theta_{(k,l)}, \theta_{(i,p)}\right) \right| \left| \int_{\mathcal{T}} b_p(s) f^{(k)}(s) \mathrm{d}s \right| \right).
\end{aligned}$$

Using the R defined such that $|b_j(s)| < R$ for all $s \in \mathcal{T}$, and condition (b), we have that

$$|(\boldsymbol{\mathcal{K}}\mathbf{f})^{(i)}(\mathbf{t})| < \left(\sum_{k=1}^{K} \sum_{p=1}^{P} \sum_{l=1}^{P} RM \left| \int_{\mathcal{T}} b_p(s) f^{(k)}(s) \mathrm{d}s \right| \right).$$

Using Hölder's Inequality, we have

$$\begin{aligned}
\left|\left(\boldsymbol{\mathcal{K}}\mathbf{f}\right)^{(i)}(\mathbf{t})\right| &< \left(\sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{l=1}^{P} RM\left(\int_{\mathcal{T}}\left|b_p(s)\right|^2 \mathrm{d}s\right)^{1/2}\left(\int_{\mathcal{T}}\left|f^{(k)}(s)\right|^2 \mathrm{d}s\right)^{1/2}\right) \\
&< \left(\sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{l=1}^{P} RM\left(\int_{\mathcal{T}} R^2 \mathrm{d}s\right)^{1/2}\left(\int_{\mathcal{T}}\left|f^{(k)}(s)\right|^2 \mathrm{d}s\right)^{1/2}\right).
\end{aligned}$$

Since $\boldsymbol{\mathcal{K}}$ is the direct sum of Hilbert spaces, we know that if $\mathbf{f} \in \boldsymbol{\mathcal{B}}_Z$, then $\|f^{(j)}\| < Z$ for all j, since $\|\mathbf{f}\| = \sum_{j=1}^{K}\|f^{(j)}\|$. Since $\|f^{(j)}\| = \int_{\mathcal{T}} f^{(j)}(s)^2 \mathrm{d}s = \int_{\mathcal{T}}|f^{(j)}(s)|^2 \mathrm{d}s$, we know that $\int_{\mathcal{T}}|f^{(k)}(s)|^2 \mathrm{d}s < Z$. Thus, we have

$$\begin{aligned}
\left|\left(\boldsymbol{\mathcal{K}}\mathbf{f}\right)^{(i)}(\mathbf{t})\right| &< \left(\sum_{k=1}^{K}\sum_{p=1}^{P}\sum_{l=1}^{P} RM\left(R^2 B\right)^{1/2}\left(Z\right)^{1/2}\right) \\
&= KP^2R^2MB^{1/2}Z^{1/2} < \infty. \qquad \text{(B.9)}
\end{aligned}$$

Since $\|\boldsymbol{\mathcal{K}}\mathbf{f}\|^2_{\boldsymbol{\mathcal{H}}} = \sum_{i=1}^{K}\left|\left(\boldsymbol{\mathcal{K}}\mathbf{f}\right)^{(i)}(\mathbf{t})\right|^2$, we have that

$$\|\boldsymbol{\mathcal{K}}\mathbf{f}\|^2_{\boldsymbol{\mathcal{H}}} < K^{3/2}P^2R^2MB^{1/2}Z^{1/2} < \infty.$$

Thus we know that $\boldsymbol{\mathcal{K}}\mathbf{f}_{\boldsymbol{\mathcal{B}}_Z}$ is a bounded equicontinuous set of functions. Therefore, using Ascoli's Theorem (Reed and Simon [1972], page 30), we know that for every sequence $\mathbf{f}_n \in \boldsymbol{\mathcal{B}}_Z$, the set $\boldsymbol{\mathcal{K}}\mathbf{f}_{\boldsymbol{\mathcal{B}}_Z}$ has a subsequence that converges (Reed and Simon [1972], page 199). Therefore, $\boldsymbol{\mathcal{K}}\mathbf{f}_{\boldsymbol{\mathcal{B}}_Z}$ is precompact, which implies that $\boldsymbol{\mathcal{K}}$ is compact.

We will now show that $\boldsymbol{\mathcal{K}}$ is a bounded operator. Let $\mathbf{f} \in \boldsymbol{\mathcal{B}}_Z$. Thus, we have

$$\|\boldsymbol{\mathcal{K}}\mathbf{f}\|^2_{\boldsymbol{\mathcal{H}}} = \sum_{i=1}^{K}\int_{\mathcal{T}}\left|\left(\boldsymbol{\mathcal{K}}\mathbf{f}\right)^{(i)}(\mathbf{t})\right|^2 \mathrm{d}t.$$

From equation B.9, we have that

$$\begin{aligned}\|\boldsymbol{\mathcal{K}}\mathbf{f}\|_{\boldsymbol{\mathcal{H}}}^2 &< \sum_{i=1}^{K} \int_{\mathcal{T}} \left(KP^2R^2MB^{1/2}Z^{1/2}\right)^2 \mathrm{d}t \\ &= K^3P^4R^4M^2BZ \int_{\mathcal{T}} \mathrm{d}t.\end{aligned}$$

Using the B defined above ($\int_{\mathcal{T}} 1\mathrm{d}t < B$), we have that

$$\begin{aligned}\|\boldsymbol{\mathcal{K}}\mathbf{f}\|_{\boldsymbol{\mathcal{H}}}^2 &= K^3P^4R^4M^2BZ \int_{\mathcal{T}} 1\mathrm{d}t \\ &< K^3P^4R^4M^2B^2Z < \infty\end{aligned}$$

Therefore we have that $\boldsymbol{\mathcal{K}}$ is a bounded linear operator. Therefore, if conditions (a) and (b) are met, then $\boldsymbol{\mathcal{K}}$ is a bounded and compact linear operator. □

B.1.3 Proof of Lemma 5

We will start be explicitly defining the functions $\Lambda_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})$, $K_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})$, and $V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})$. Thus we have

$$\begin{aligned}\Lambda_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) =& \log\left(\frac{|(\boldsymbol{\Sigma}_i)_0|^{-1/2} \exp\left\{-\frac{1}{2}\left(\mathbf{Y}_i - (\boldsymbol{\mu}_i)_0\right)' (\boldsymbol{\Sigma}_i)_0^{-1} \left(\mathbf{Y}_i - (\boldsymbol{\mu}_i)_0\right)\right\}}{|\boldsymbol{\Sigma}_i|^{-1/2} \exp\left\{-\frac{1}{2}\left(\mathbf{Y}_i - \boldsymbol{\mu}_i\right)' (\boldsymbol{\Sigma}_i)^{-1} \left(\mathbf{Y}_i - \boldsymbol{\mu}_i\right)\right\}}\right) \\ =& -\frac{1}{2}\left[\log\left(|(\boldsymbol{\Sigma}_i)_0|\right) - \log\left(|\boldsymbol{\Sigma}_i|\right)\right] \\ & -\frac{1}{2}\left[\left(\mathbf{Y}_i - (\boldsymbol{\mu}_i)_0\right)' (\boldsymbol{\Sigma}_i)_0^{-1} \left(\mathbf{Y}_i - (\boldsymbol{\mu}_i)_0\right) - \left(\mathbf{Y}_i - \boldsymbol{\mu}_i\right)' (\boldsymbol{\Sigma}_i)^{-1} \left(\mathbf{Y}_i - \boldsymbol{\mu}_i\right)\right] \\ =& -\frac{1}{2}\left[\sum_{l=1}^{R} \log\left((d_{il})_0 + \sigma_0^2\right) - \log\left(d_{il} + \sigma^2\right)\right] \\ & -\frac{1}{2}\left[\left(\mathbf{Y}_i - (\boldsymbol{\mu}_i)_0\right)' (\boldsymbol{\Sigma}_i)_0^{-1} \left(\mathbf{Y}_i - (\boldsymbol{\mu}_i)_0\right) - \left(\mathbf{Y}_i - \boldsymbol{\mu}_i\right)' (\boldsymbol{\Sigma}_i)^{-1} \left(\mathbf{Y}_i - \boldsymbol{\mu}_i\right)\right] \quad \text{(B.10)}\end{aligned}$$

$$
\begin{aligned}
K_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) =& -\frac{1}{2}\left[\sum_{l=1}^{R} \log\left((d_{il})_0 + \sigma_0^2\right) - \log\left(d_{il} + \sigma^2\right)\right] \\
& -\frac{1}{2}\mathbb{E}_{\boldsymbol{\omega}_0}\left[\left(\mathbf{Y}_i - (\boldsymbol{\mu}_i)_0\right)' \left(\boldsymbol{\Sigma}_i\right)_0^{-1} \left(\mathbf{Y}_i - (\boldsymbol{\mu}_i)_0\right) - \left(\mathbf{Y}_i - \boldsymbol{\mu}_i\right)' \left(\boldsymbol{\Sigma}_i\right)^{-1} \left(\mathbf{Y}_i - \boldsymbol{\mu}_i\right)\right] \\
=& -\frac{1}{2}\left[\sum_{l=1}^{R} \log\left((d_{il})_0 + \sigma_0^2\right) - \log\left(d_{il} + \sigma^2\right)\right] \\
& -\frac{1}{2}\left[R - \left(\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\right) + \left(\left(\boldsymbol{\mu}_i\right)_0 - \boldsymbol{\mu}_i\right)' \left(\boldsymbol{\Sigma}_i\right)^{-1} \left(\left(\boldsymbol{\mu}_i\right)_0 - \boldsymbol{\mu}_i\right)\right)\right] \qquad \text{(B.11)}
\end{aligned}
$$

$$
\begin{aligned}
V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) =& \frac{1}{4}\mathrm{Var}_{\boldsymbol{\omega}_0}\left[\left(\mathbf{Y}_i - (\boldsymbol{\mu}_i)_0\right)' \left(\boldsymbol{\Sigma}_i\right)_0^{-1} \left(\mathbf{Y}_i - (\boldsymbol{\mu}_i)_0\right) - \left(\mathbf{Y}_i - \boldsymbol{\mu}_i\right)' \left(\boldsymbol{\Sigma}_i\right)^{-1} \left(\mathbf{Y}_i - \boldsymbol{\mu}_i\right)\right] \\
=& \frac{1}{4}\mathrm{Var}_{\boldsymbol{\omega}_0}\left[\mathbf{Y}_i' \left(\left(\boldsymbol{\Sigma}_i\right)_0^{-1} + \boldsymbol{\Sigma}_i^{-1}\right) \mathbf{Y}_i - 2\mathbf{Y}_i' \left(\left(\boldsymbol{\Sigma}_i\right)_0^{-1} \left(\boldsymbol{\mu}_i\right)_0 + \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i\right)\right]
\end{aligned}
$$

Letting $\mathbf{M}_v = \left(\boldsymbol{\Sigma}_i\right)_0^{-1} + \boldsymbol{\Sigma}_i^{-1}$, and $\mathbf{m}_v = \left(\boldsymbol{\Sigma}_i\right)_0^{-1} \left(\boldsymbol{\mu}_i\right)_0 + \boldsymbol{\Sigma}_i^{-1}\boldsymbol{\mu}_i$, we have

$$
\begin{aligned}
V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) =& \frac{1}{4}\mathrm{Var}_{\boldsymbol{\omega}_0}\left[\left(\mathbf{Y}_i - \mathbf{M}_v^{-1}\mathbf{m}_v\right)'\mathbf{M}_v\left(\mathbf{Y}_i - \mathbf{M}_v^{-1}\mathbf{m}_v\right)\right] \\
=& \frac{1}{4}\left[2\operatorname{tr}\left(\mathbf{M}_v \left(\boldsymbol{\Sigma}_i\right)_0 \mathbf{M}_v \left(\boldsymbol{\Sigma}_i\right)_0\right) + 4\left(\left(\boldsymbol{\mu}_i\right)_0 - \mathbf{M}_v^{-1}\mathbf{m}_v\right)' \left(\boldsymbol{\Sigma}_i\right)_0 \left(\left(\boldsymbol{\mu}_i\right)_0 - \mathbf{M}_v^{-1}\mathbf{m}_v\right)\right] \\
=& \frac{1}{2}\left[R + 2\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\right) + \operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0 \boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\right)\right] \\
& + \left(\left(\boldsymbol{\mu}_i\right)_0 - \boldsymbol{\mu}_i\right)' \left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0 \boldsymbol{\Sigma}_i^{-1}\right) \left(\left(\boldsymbol{\mu}_i\right)_0 - \boldsymbol{\mu}_i\right). \qquad \text{(B.12)}
\end{aligned}
$$

Let $\boldsymbol{\Omega}_\epsilon(\boldsymbol{\omega}_0) = \{\boldsymbol{\omega} : K_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) < \epsilon \text{ for all i}\}$ for some $\epsilon > 0$. We will assume that $\sigma_0^2 > 0$. Consider the set $\mathcal{B}(\boldsymbol{\omega}_0) = \{\boldsymbol{\omega} : \frac{1}{a}((d_{il})_0 + \sigma_0^2) \leq d_{il} + \sigma^2 \leq a((d_{il})_0 + \sigma_0^2), \| \left(\boldsymbol{\mu}_i\right)_0 - \boldsymbol{\mu}_i\| \leq b\}$ for some $a, b \in \mathbb{R}$ such that $a > 1$ and $b > 0$. Thus for a fixed $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$ and any $\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon) := \mathcal{B}(\boldsymbol{\omega}_0) \cap \boldsymbol{\Omega}_\epsilon(\boldsymbol{\omega}_0)$, we can bound $V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})$. We will let $\lambda_r(\mathbf{A})$ denote the r^{th} eigenvalue of the matrix $\mathbf{A}$, and $\lambda_{max}(\mathbf{A})$ denote the largest eigenvalue of $\mathbf{A}$. Thus we have

$$
\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\right) \leq R\lambda_{max}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\right) \leq \frac{Ra}{\sigma_0^2}\left(\max_l (d_{il} + \sigma_0^2)\right)
$$

$$\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\right)\leq\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\right)^2\leq\left(\frac{Ra}{\sigma_0^2}\left(\max_l(d_{il}+\sigma_0^2)\right)\right)^2$$

$$\begin{aligned}\left(\left(\boldsymbol{\mu}_i\right)_0-\boldsymbol{\mu}_i\right)'\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\boldsymbol{\Sigma}_i^{-1}\right)\left(\left(\boldsymbol{\mu}_i\right)_0-\boldsymbol{\mu}_i\right)&\leq b^2\lambda_{max}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\boldsymbol{\Sigma}_i^{-1}\right)\\&\leq\frac{a^2b^2}{\sigma_0^4}\max_l(d_{il}+\sigma_0^2)\end{aligned}$$

Thus we can see that for any $\boldsymbol{\omega}\in\mathcal{C}(\boldsymbol{\omega}_0,\epsilon)$,

$$\begin{aligned}V_i(\boldsymbol{\omega}_0,\boldsymbol{\omega})\leq&\frac{1}{2}\left[R+2\left(\frac{Ra}{\sigma_0^2}\left(\max_l(d_{il}+\sigma_0^2)\right)\right)+\left(\frac{Ra}{\sigma_0^2}\left(\max_l(d_{il}+\sigma_0^2)\right)\right)^2\right]\\&+\frac{a^2b^2}{\sigma_0^4}\max_l(d_{il}+\sigma_0^2)\\=&M_V.\end{aligned}$$

If we can bound $\lambda_{max}\left((d_{il})_0+\sigma_0\right)$, then we have that $V_i(\boldsymbol{\omega}_0,\boldsymbol{\omega})$ is bounded. Let $\|\cdot\|_F$ be the Frobenius norm. Using the triangle inequality, we have

$$\begin{aligned}\|(\boldsymbol{\Sigma}_i)_0\|_F&\leq\sum_{k=1}^K\sum_{j=1}^K\sum_{p=1}^{KP}Z_{ij}Z_{ik}\|\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{kp})_0(\boldsymbol{\phi}_{jp})_0'\mathbf{S}(\mathbf{t})\|_F+\sigma_0^2\|\mathbf{I}_R\|_F\\&\leq\sum_{k=1}^K\sum_{j=1}^K\sum_{p=1}^{KP}\|\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{kp})_0(\boldsymbol{\phi}_{jp})_0'\mathbf{S}(\mathbf{t})\|_F+\sigma_0^2\|\mathbf{I}_R\|_F\\&=\sum_{k=1}^K\sum_{j=1}^K\sum_{p=1}^{KP}\sqrt{\operatorname{tr}\left(\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{jp})_0(\boldsymbol{\phi}_{kp})_0'\mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{kp})_0(\boldsymbol{\phi}_{jp})_0'\mathbf{S}(\mathbf{t})\right)}+\sqrt{R}\sigma_0^2\\&=\sum_{k=1}^K\sum_{j=1}^K\sum_{p=1}^{KP}\sqrt{\operatorname{tr}\left((\boldsymbol{\phi}_{kp})_0'\mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{kp})_0(\boldsymbol{\phi}_{jp})_0'\mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{jp})_0\right)}+\sqrt{R}\sigma_0^2\\&=\sum_{k=1}^K\sum_{j=1}^K\sum_{p=1}^{KP}\|\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{jp})_0\|_2\|\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{kp})_0\|_2+\sqrt{R}\sigma_0^2\\&=M_{\boldsymbol{\Sigma}_0}<\infty,\end{aligned}$$

for all $i \in \mathbb{N}$. Therefore, we know that $\lambda_{max}\left((d_{il})_0 + \sigma_0\right) \leq M_{\boldsymbol{\Sigma}_0}$, as the Frobenius is the squareroot of the sum of the squared eigenvalues for a square matrix. Therefore, we have for all $i \in \mathbb{N}$ and $\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)$, we have that

$$\frac{V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})}{i^2} \leq \frac{M_V}{i^2}.$$

Since $\sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{\pi^2}{6}$, we have that $\sum_{i=1}^{\infty} \frac{M_V}{i^2} = \frac{M_V \pi^2}{6} < \infty$. Thus we have

$$\sum_{i=1}^{\infty} \frac{V_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})}{i^2} < \infty. \tag{B.13}$$

We will next show that for $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$ and $\epsilon > 0$, $\boldsymbol{\Pi}(\mathcal{C}(\boldsymbol{\omega}_0), \epsilon) > 0$. Fix $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$. While the $(\boldsymbol{\phi}_{jp})_0$ may not be identifiable (for any orthogonal matrix $\mathbf{H}$, $(\boldsymbol{\phi}_{jp})_0 \mathbf{H}\mathbf{H}'(\boldsymbol{\phi}_{kp})_0 = (\boldsymbol{\phi}_{jp})_0'(\boldsymbol{\phi}_{kp})_0$), let $(\boldsymbol{\phi}_{jp})_0$ be such that $\sum_{p=1}^{KP}(\boldsymbol{\phi}_{jp})_0'(\boldsymbol{\phi}_{kp})_0 = (\boldsymbol{\Sigma}_{jk})_0$. Thus we can define the following sets:

$$\begin{aligned}
\boldsymbol{\Omega}_{\boldsymbol{\phi}_{jp}} &= \left\{\boldsymbol{\phi}_{jp} : (\boldsymbol{\phi}_{jp})_0 \leq \boldsymbol{\phi}_{jp} \leq (\boldsymbol{\phi}_{jp})_0 + \epsilon_1 \mathbf{1}\right\} \\
\boldsymbol{\Omega}_{\boldsymbol{\nu}_k} &= \left\{\boldsymbol{\nu}_k : (\boldsymbol{\nu}_k)_0 \leq \boldsymbol{\nu}_k \leq (\boldsymbol{\nu}_k)_0 + \epsilon_2 \mathbf{1}\right\} \\
\boldsymbol{\Omega}_{\sigma^2} &= \left\{\sigma^2 : \sigma_0^2 \leq \sigma^2 \leq (1+\epsilon_1)\sigma_0^2\right\}.
\end{aligned}$$

We define $\boldsymbol{\epsilon}_{1jp}$ and $\boldsymbol{\epsilon}_{2k}$ such that each element of $\boldsymbol{\epsilon}_{1jp}$ is between 0 and ϵ_1, and each element of $\boldsymbol{\epsilon}_{2k}$ is between 0 and ϵ_2. Therefore $(\boldsymbol{\phi}_{jp})_0 + \boldsymbol{\epsilon}_{1jp} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}_{jp}}$ and $(\boldsymbol{\nu}_k)_0 + \boldsymbol{\epsilon}_{2k} \in \boldsymbol{\Omega}_{\boldsymbol{\nu}_k}$. We will define

$$\boldsymbol{\Omega}_{\boldsymbol{\Sigma}_{jk}} := \left\{ \sum_{p=1}^{KP} \boldsymbol{\phi}_{jp}' \boldsymbol{\phi}_{kp} \middle| \boldsymbol{\phi}_{jp} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}_{jp}}, \boldsymbol{\phi}_{kp} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}_{kp}} \right\}.$$

Thus for $\boldsymbol{\Sigma}_i$ such that $\boldsymbol{\phi}_{jp} \in \boldsymbol{\Omega}_{\boldsymbol{\phi}_{jp}}$ and $\sigma^2 \in \boldsymbol{\Omega}_{\sigma^2}$, we have that

$$\begin{aligned}
\boldsymbol{\Sigma}_i =& \sum_{k=1}^{K}\sum_{j=1}^{K} Z_{ik}Z_{ij}\left(\mathbf{S}'(\mathbf{t})\sum_{p=1}^{KP}\left(\left((\boldsymbol{\phi}_{kp})_0+\boldsymbol{\epsilon}_{1kp}\right)\left((\boldsymbol{\phi}_{jp})_0+\boldsymbol{\epsilon}_{1jp}\right)'\right)\mathbf{S}(\mathbf{t})\right)+(1+\epsilon_\sigma)\sigma_0^2\mathbf{I}_R \\
=&(\boldsymbol{\Sigma}_i)_0+\sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{p=1}^{KP} Z_{ik}Z_{ij}\left(\mathbf{S}'(\mathbf{t})\left(\left(\boldsymbol{\epsilon}_{1kp}\right)\left(\boldsymbol{\phi}_{jp}\right)_0'\right)\mathbf{S}(\mathbf{t})\right) \\
&+\sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{p=1}^{KP} Z_{ik}Z_{ij}\left(\mathbf{S}'(\mathbf{t})\left(\left(\boldsymbol{\phi}_{kp}\right)_0\left(\boldsymbol{\epsilon}_{1jp}\right)'\right)\mathbf{S}(\mathbf{t})\right) \\
&+\sum_{k=1}^{K}\sum_{j=1}^{K}\sum_{p=1}^{KP} Z_{ik}Z_{ij}\left(\mathbf{S}'(\mathbf{t})\left(\left(\boldsymbol{\epsilon}_{1kp}\right)\left(\boldsymbol{\epsilon}_{1jp}\right)'\right)\mathbf{S}(\mathbf{t})\right)+\epsilon_\sigma\sigma_0^2\mathbf{I}_R \\
=&(\boldsymbol{\Sigma}_i)_0+\tilde{\boldsymbol{\Sigma}}_i,
\end{aligned}$$

for some ϵ_{kp} and ϵ_σ such that $0 < \epsilon_\sigma \leq \epsilon_1$. Thus, letting $\boldsymbol{\zeta}_{jkp} = \left(\mathbf{S}'(\mathbf{t})\left(\left(\boldsymbol{\epsilon}_{1kp}\right)\left(\boldsymbol{\phi}_{jp}\right)_0'\right)\mathbf{S}(\mathbf{t})\right.$ $\left.+\mathbf{S}'(\mathbf{t})\left(\left(\boldsymbol{\phi}_{kp}\right)_0\left(\boldsymbol{\epsilon}_{1jp}\right)'\right)\mathbf{S}(\mathbf{t})\right)$, we have

$$\begin{aligned}
\left|\left|Z_{ik}Z_{ij}\boldsymbol{\zeta}_{jkp}\right|\right|_F^2 \leq& \left|\left|\boldsymbol{\zeta}_{jkp}\right|\right|_F^2 \\
=&\mathrm{tr}\left(\mathbf{S}'(\mathbf{t})\left(\left(\boldsymbol{\epsilon}_{1kp}\right)\left(\boldsymbol{\phi}_{jp}\right)_0'\right)\mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})\left(\left(\boldsymbol{\phi}_{jp}\right)_0\left(\boldsymbol{\epsilon}_{1kp}\right)'\right)\mathbf{S}(\mathbf{t})\right) \\
&+\mathrm{tr}\left(\mathbf{S}'(\mathbf{t})\left(\left(\boldsymbol{\epsilon}_{1kp}\right)\left(\boldsymbol{\phi}_{jp}\right)_0'\right)\mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})\left(\left(\boldsymbol{\epsilon}_{1jp}\right)\left(\boldsymbol{\phi}_{kp}\right)_0'\right)\mathbf{S}(\mathbf{t})\right) \\
&+\mathrm{tr}\left(\mathbf{S}'(\mathbf{t})\left(\left(\boldsymbol{\phi}_{kp}\right)_0\left(\boldsymbol{\epsilon}_{1jp}\right)'\right)\mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})\left(\left(\boldsymbol{\phi}_{jp}\right)_0\left(\boldsymbol{\epsilon}_{1kp}\right)'\right)\mathbf{S}(\mathbf{t})\right) \\
&+\mathrm{tr}\left(\mathbf{S}'(\mathbf{t})\left(\left(\boldsymbol{\phi}_{kp}\right)_0\left(\boldsymbol{\epsilon}_{1jp}\right)'\right)\mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})\left(\left(\boldsymbol{\epsilon}_{1jp}\right)\left(\boldsymbol{\phi}_{kp}\right)_0'\right)\mathbf{S}(\mathbf{t})\right) \\
\leq&\epsilon_1^2\mathrm{tr}\left(\left(\boldsymbol{\phi}_{jp}\right)_0'\mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})\left(\boldsymbol{\phi}_{jp}\right)_0\left(\mathbf{1}\right)'\mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})\left(\mathbf{1}\right)\right) \\
&+2\mathrm{tr}\left(\left(\boldsymbol{\phi}_{jp}\right)_0'\mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})\left(\boldsymbol{\epsilon}_{1jp}\right)\left(\boldsymbol{\phi}_{kp}\right)_0'\mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})\left(\boldsymbol{\epsilon}_{1kp}\right)\right) \qquad \text{(B.14)} \\
&+\epsilon_1^2\mathrm{tr}\left(\left(\mathbf{1}\right)'\mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})\left(\mathbf{1}\right)\left(\boldsymbol{\phi}_{kp}\right)_0'\mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})\left(\boldsymbol{\phi}_{kp}\right)_0\right).
\end{aligned}$$

Using the Cauchy-Schwarz inequality, we can simplify equation B.14, such that

$$\begin{aligned}(B.14) &= 2\langle \mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{jp})_0, \mathbf{S}'(\mathbf{t})\boldsymbol{\epsilon}_{1jp}\rangle\langle \mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{kp})_0, \mathbf{S}'(\mathbf{t})\boldsymbol{\epsilon}_{1kp}\rangle \\ &\leq 2\|\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{jp})_0\|_2\|\mathbf{S}'(\mathbf{t})\boldsymbol{\epsilon}_{1jp}\|_2\|\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{kp})_0\|_2\|\mathbf{S}'(\mathbf{t})\boldsymbol{\epsilon}_{1kp}\|_2 \\ &\leq 2\epsilon_1^2\|\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{jp})_0\|_2\|\mathbf{S}'(\mathbf{t})\mathbf{1}\|_2\|\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{kp})_0\|_2\|\mathbf{S}'(\mathbf{t})\mathbf{1}\|_2.\end{aligned}$$

Letting

$$\begin{aligned}\tilde{M}_{jkp} = &\|\mathbf{S}'(\mathbf{t})\mathbf{1}\|_2^2\left[\|\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{jp})_0\|_2^2 + \|\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{kp})_0\|_2^2\right] \\ &+ 2\left(\|\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{jp})_0\|_2\|\mathbf{S}'(\mathbf{t})\mathbf{1}\|_2\|\mathbf{S}'(\mathbf{t})(\boldsymbol{\phi}_{kp})_0\|_2\|\mathbf{S}'(\mathbf{t})\mathbf{1}\|_2\right),\end{aligned}$$

we have

$$\left|\left|Z_{ik}Z_{ij}\boldsymbol{\zeta}_{jkp}\right|\right|_F^2 \leq \epsilon_1^2\tilde{M}_{jkp}.$$

In a similar fashion, we can show that

$$\|Z_{ik}Z_{ij}\left(\mathbf{S}'(\mathbf{t})\left((\boldsymbol{\epsilon}_{1kp})\,(\boldsymbol{\epsilon}_{1jp})'\right)\mathbf{S}(\mathbf{t})\right)\|_F^2 \leq \epsilon_1^2\|\mathbf{S}'(\mathbf{t})\mathbf{1}\|_2^4$$

and

$$\|\epsilon_\sigma\sigma_0^2\mathbf{I}_R\|_F^2 \leq \epsilon_1^2\sigma_0^4 R.$$

By using the triangle inequality we have

$$\|\tilde{\boldsymbol{\Sigma}}_i\|_F \leq \epsilon_1\left(\sum_{j=1}^{K}\sum_{k=1}^{K}\sum_{p=1}^{KP}\left(\sqrt{\tilde{M}_{jkp}}\right) + JK^2P\|\mathbf{S}'(\mathbf{t})\mathbf{1}\|_2^2 + \sigma_0^2\sqrt{R}\right) := \epsilon_1 M_{\boldsymbol{\Sigma}} \tag{B.15}$$

for all $i \in \mathbb{N}$. By the Wielandt-Hoffman Theorem (Golub and Van Loan [2013] Theorem

8.1.4), we have that

$$\sum_{r=1}^{R} \left(\lambda_r \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i \right) - \lambda_r \left((\mathbf{\Sigma}_i)_0 \right) \right)^2 \leq \|\tilde{\mathbf{\Sigma}}_i\|_F^2,$$

which implies that

$$\max_r \left| \lambda_r \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i \right) - \lambda_r \left((\mathbf{\Sigma}_i)_0 \right) \right| \leq \|\tilde{\mathbf{\Sigma}}_i\|_F \tag{B.16}$$

where $\lambda_r(\mathbf{A})$ are the eigenvalues of the matrix $\mathbf{A}$. By using equation B.15, we can bound the log-determinant of the ratio of the two covariance matrices as follows

$$\begin{aligned} \log \left(\frac{|\mathbf{\Sigma}_i|}{|(\mathbf{\Sigma}_i)_0|} \right) &= \log \left(\frac{\prod_{r=1}^{R} \lambda_r \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i \right)}{\prod_{r=1}^{R} \lambda_r \left((\mathbf{\Sigma}_i)_0 \right)} \right) \\ &\leq \log \left(\prod_{r=1}^{R} \frac{((d_{ir})_0 + \sigma_0^2) + \epsilon_1 M_{\mathbf{\Sigma}}}{(d_{ir})_0 + \sigma_0^2} \right) \\ &\leq R \log \left(1 + \frac{\epsilon_1 M_{\mathbf{\Sigma}}}{\sigma_0^2} \right). \end{aligned} \tag{B.17}$$

We can also bound $\operatorname{tr}\left(\mathbf{\Sigma}_i^{-1} (\mathbf{\Sigma}_i)_0 \right)$. To do this, we will first consider the spectral norm, defined as $\|\mathbf{A}\|_2 = \sqrt{\mathbf{A}^* \mathbf{A}}$ for some matrix $\mathbf{A}$. In the case where $\mathbf{A}$ is symmetric, we have that $\|\mathbf{A}\|_2 = \max_r |\lambda_r(\mathbf{A})|$. By the submultiplicative property of induced norms, we have that

$$\max_r |\lambda_r(\mathbf{AB})| = \|\mathbf{AB}\|_2 \leq \|\mathbf{A}\|_2 \|\mathbf{B}\|_2 = \max_r |\lambda_r(\mathbf{A})| \max_r |\lambda_r(\mathbf{B})|, \tag{B.18}$$

for two symmetric matrices $\mathbf{A}$ and $\mathbf{B}$. By using the Sherman–Morrison–Woodbury formula, we can see that

$$\begin{aligned} \mathbf{\Sigma}_i^{-1} &= \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i \right)^{-1} \\ &= (\mathbf{\Sigma}_i)_0^{-1} - (\mathbf{\Sigma}_i)_0^{-1} \tilde{\mathbf{\Sigma}}_i \left((\mathbf{\Sigma}_i)_0 + \tilde{\mathbf{\Sigma}}_i \right)^{-1}. \end{aligned}$$

Thus, we have that

$$\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0 = \mathbf{I}_R - \left(\boldsymbol{\Sigma}_i\right)_0^{-1}\tilde{\boldsymbol{\Sigma}}_i\left(\left(\boldsymbol{\Sigma}_i\right)_0 + \tilde{\boldsymbol{\Sigma}}_i\right)^{-1}\left(\boldsymbol{\Sigma}_i\right)_0. \tag{B.19}$$

Using equation B.18, we would like to bound the magnitude of the eigenvalues of $\left(\boldsymbol{\Sigma}_i\right)_0^{-1}\tilde{\boldsymbol{\Sigma}}_i\left(\left(\boldsymbol{\Sigma}_i\right)_0 + \tilde{\boldsymbol{\Sigma}}_i\right)^{-1}\left(\boldsymbol{\Sigma}_i\right)_0$. We know that

$$\max_r \left|\lambda_r\left(\left(\left(\boldsymbol{\Sigma}_i\right)_0 + \tilde{\boldsymbol{\Sigma}}_i\right)^{-1}\right)\right| \leq \frac{1}{\sigma_0^2}$$

and

$$\max_r \left|\lambda_r(\tilde{\boldsymbol{\Sigma}}_i)\right| \leq \epsilon_1 M_{\boldsymbol{\Sigma}},$$

with the second inequality coming from equation B.15. From equation B.19 and basic properties of the trace, we have that

$$\begin{aligned}
\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\right) &= \operatorname{tr}\left(\mathbf{I}_R - \left(\boldsymbol{\Sigma}_i\right)_0^{-1}\tilde{\boldsymbol{\Sigma}}_i\left(\left(\boldsymbol{\Sigma}_i\right)_0 + \tilde{\boldsymbol{\Sigma}}_i\right)^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\right) \\
&= \operatorname{tr}\left(\mathbf{I}_R\right) - \operatorname{tr}\left(\tilde{\boldsymbol{\Sigma}}_i\left(\left(\boldsymbol{\Sigma}_i\right)_0 + \tilde{\boldsymbol{\Sigma}}_i\right)^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\left(\boldsymbol{\Sigma}_i\right)_0^{-1}\right) \\
&= \operatorname{tr}\left(\mathbf{I}_R\right) - \operatorname{tr}\left(\tilde{\boldsymbol{\Sigma}}_i\left(\left(\boldsymbol{\Sigma}_i\right)_0 + \tilde{\boldsymbol{\Sigma}}_i\right)^{-1}\right)
\end{aligned}$$

Thus, using the fact that the trace of a matrix is the sum of its eigenvalues, we have that

$$\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\right) \leq R + R\max_r \left|\lambda_r\left(\tilde{\boldsymbol{\Sigma}}_i\left(\left(\boldsymbol{\Sigma}_i\right)_0 + \tilde{\boldsymbol{\Sigma}}_i\right)^{-1}\right)\right|.$$

Using the submultiplicative property stated in equation B.18, we have

$$\operatorname{tr}\left(\boldsymbol{\Sigma}_i^{-1}\left(\boldsymbol{\Sigma}_i\right)_0\right) \leq R + \frac{R\epsilon_1 M_{\boldsymbol{\Sigma}}}{\sigma_0^2}. \tag{B.20}$$

Lastly, we can bound the quadratic term in $K_i(\boldsymbol{\omega}_0, \boldsymbol{\omega})$ in the following way:

$$\begin{aligned}
\left((\boldsymbol{\mu}_i)_0 - \boldsymbol{\mu}_i\right)' \left(\boldsymbol{\Sigma}_i\right)^{-1} \left((\boldsymbol{\mu}_i)_0 - \boldsymbol{\mu}_i\right) &\leq ||(\boldsymbol{\mu}_i)_0 - \boldsymbol{\mu}_i||_2^2 \max_r \lambda_r(\left(\boldsymbol{\Sigma}_i\right)^{-1}) \\
&\leq \frac{1}{\sigma^2} \sum_{k=1}^{K} \|\mathbf{S}'(\mathbf{t})(\boldsymbol{\nu}_k)_0 - \mathbf{S}'(\mathbf{t})\boldsymbol{\nu}_k\|_2^2 \\
&= \frac{1}{\sigma^2} \sum_{k=1}^{K} \boldsymbol{\epsilon}_{2k}' \mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})\boldsymbol{\epsilon}_{2k} \\
&\leq \frac{KR\epsilon_2^2}{\sigma_0^2} \lambda_{\mathbf{S}(\mathbf{t})}^{max}, \qquad \text{(B.21)}
\end{aligned}$$

where $\lambda_{\mathbf{S}(\mathbf{t})}^{max}$ is the maximum eigenvalue of the matrix $\mathbf{S}(\mathbf{t})\mathbf{S}'(\mathbf{t})$. Thus letting

$$\epsilon_1 < \min\left\{ \frac{\sigma_0^2}{M_{\boldsymbol{\Sigma}}} \left(\exp\left(\frac{2\epsilon}{3R}\right) - 1 \right), \frac{2\epsilon\sigma_0^2}{3RM_{\boldsymbol{\Sigma}}} \right\} \qquad \text{(B.22)}$$

and

$$\epsilon_2 < \sqrt{\frac{2\sigma_0^2\epsilon}{3KR\lambda_{\mathbf{S}(\mathbf{t})}^{max}}}, \qquad \text{(B.23)}$$

we have from equations B.17, B.20, and B.21 that

$$K_i(\boldsymbol{\omega}_0, \boldsymbol{\omega}) < \epsilon \text{ for all } \boldsymbol{\omega} \in \boldsymbol{\Omega}_1$$

where $\boldsymbol{\Omega}_1 := \left(\bigtimes_{j=1}^{K} \bigtimes_{k=1}^{K} \boldsymbol{\Omega}_{\boldsymbol{\Sigma}_{jk}}\right) \times \left(\bigtimes_{k=1}^{K} \boldsymbol{\Omega}_{\boldsymbol{\nu}_k}\right) \times \boldsymbol{\Omega}_{\sigma^2}$. Letting $a > \max\left\{1 + \frac{\epsilon_1 M_{\boldsymbol{\Sigma}}}{\sigma_0^2}, \left(1 - \frac{\epsilon_1 M_{\boldsymbol{\Sigma}}}{\sigma_0^2}\right)^{-1}\right\}$ and $b > \sqrt{KR\epsilon_2^2\lambda_{\mathbf{S}(\mathbf{t})}^{max}}$ in the definition of $\mathcal{C}(\boldsymbol{\omega}_0, \epsilon)$, we have that $\boldsymbol{\Omega}_1 \subset \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)$. Let $H_{\boldsymbol{\phi}}$ be the set of hyper-parameters corresponding to the $\boldsymbol{\phi}$ parameters, and let $\boldsymbol{\Pi}(\boldsymbol{\eta}_{\boldsymbol{\phi}})$ be the prior

distribution on $\boldsymbol{\eta}_{\boldsymbol{\phi}} \in H_{\boldsymbol{\phi}}$. Thus we have that

$$\begin{aligned}\boldsymbol{\Pi}\left(\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)\right) \geq & \int_{H_{\boldsymbol{\phi}}} \prod_{j=1}^{K} \prod_{p=1}^{KP} \prod_{r=1}^{P} \int_{(\phi_{jrp})_0}^{(\phi_{jrp})_0+\epsilon_1} \sqrt{\frac{\gamma_{jrp}\tilde{\tau}_{pj}}{2\pi}} \exp\left\{-\frac{\gamma_{jrp}\tilde{\tau}_{pj}}{2}\phi_{jrp}^2\right\} \mathrm{d}\phi_{jrp} \mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\eta}_{\boldsymbol{\phi}}) \\ & \times \prod_{k=1}^{K} \int_0^{\infty} \int_{(\boldsymbol{\nu}_k)_0}^{(\boldsymbol{\nu}_k)_0+\epsilon_2\mathbf{1}} \left(\frac{\tau_k}{2\pi}\right)^{P/2} |\mathbf{P}|^{-1/2} \exp\left\{\frac{\tau_k}{2}\boldsymbol{\nu}_k'\mathbf{P}\boldsymbol{\nu}_k\right\} \mathrm{d}\boldsymbol{\nu}_k \mathrm{d}\boldsymbol{\Pi}(\tau_k) \\ & \times \int_{\sigma_0^2}^{(1+\epsilon_1)\sigma_0^2} \frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)} (\sigma^2)^{-\alpha_0-1} \exp\left\{-\frac{\beta_0}{\sigma^2}\right\} \mathrm{d}\sigma^2.\end{aligned}$$

Restricting the hyper-parameters of $\boldsymbol{\phi}$ to only a subset of the support, say $\tilde{H}_{\boldsymbol{\phi}}$, where

$$\tilde{H}_{\boldsymbol{\phi}} = \left\{\boldsymbol{\eta}_{\boldsymbol{\phi}} : \frac{1}{10} \leq \gamma_{jrp} \leq 10, 1 \leq \delta_{pj} \leq 2, 1 \leq a_{1j} \leq 10, 1 \leq a_{2j} \leq 10\right\},$$

we can see that there exists a $M_{\phi_{jrp}} > 0$ such that

$$\sqrt{\frac{\gamma_{jrp}\tilde{\tau}_{pj}}{2\pi}} \exp\left\{-\frac{\gamma_{jrp}\tilde{\tau}_{pj}}{2}\phi_{jrp}^2\right\} \geq M_{\phi_{jrp}},$$

for all $\phi_{jrp} \in [(\phi_{jrp})_0, (\phi_{jrp})_0 + \epsilon_1]$. Similarly, we can find a lower bound $M_{\tilde{H}_{\boldsymbol{\phi}}} > 0$, such that

$$\int_{\tilde{H}_{\boldsymbol{\phi}}} \mathrm{d}(\boldsymbol{\eta}_{\boldsymbol{\phi}}) \geq M_{\tilde{H}_{\boldsymbol{\phi}}}.$$

Similarly, if we bound τ_k such that $\frac{1}{10} \leq \tau_k \leq 10$, it is easy to see that there exists constants $M_{\boldsymbol{\nu}_k}, M_{\tau_k}, M_{\sigma^2} > 0$ such that

$$\left(\frac{\tau_k}{2\pi}\right)^{P/2} |\mathbf{P}|^{-1/2} \exp\left\{\frac{\tau_k}{2}\boldsymbol{\nu}_k'\mathbf{P}\boldsymbol{\nu}_k\right\} \geq M_{\boldsymbol{\nu}_k},$$

for all $\boldsymbol{\nu}_k \in [(\boldsymbol{\nu}_k)_0, (\boldsymbol{\nu}_k)_0 + \epsilon_2\mathbf{1}]$,

$$\int_{\frac{1}{10}}^{10} \boldsymbol{\Pi}(\tau_k) \geq M_{\tau_k},$$

and

$$\frac{\beta_0^{\alpha_0}}{\Gamma(\alpha_0)}(\sigma^2)^{-\alpha_0-1}\exp\left\{-\frac{\beta_0}{\sigma^2}\right\} \geq M_{\sigma^2}$$

for all $\sigma^2 \in [\sigma_0^2, (1+\epsilon_1)\sigma_0^2]$. Therefore we have that

$$\begin{aligned}
\mathbf{\Pi}\left(\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)\right) &\geq M_{\tilde{H}_{\boldsymbol{\phi}}} \prod_{j=1}^{K}\prod_{p=1}^{KP}\prod_{r=1}^{P} \epsilon_1 M_{\phi_{jrp}} \\
&\times \prod_{k=1}^{K} M_{\tau_k}\epsilon_2^P M_{\boldsymbol{\nu}_k} \\
&\times \epsilon_1\sigma_0^2 M_{\sigma_0^2} \\
&> 0.
\end{aligned}$$

Therefore, for $\epsilon > 0$, there exists a and b such that $\sum_{i=1}^{\infty}\frac{V_i(\boldsymbol{\omega}_0,\boldsymbol{\omega})}{i^2} < \infty$ for any $\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)$ and $\mathbf{\Pi}\left(\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0, \epsilon)\right) > 0$.

B.1.4 Proof of Lemma 6

Following the notation of Ghosal and Van der Vaart [2017], we will let $P_{\boldsymbol{\omega}_0}^{(N)}$ denote the joint distribution of $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ at $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$. In order to show that the posterior distribution, $\mathbf{\Pi}_N(.|\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$, is weakly consistent at $\boldsymbol{\omega}_0 \in \boldsymbol{\Omega}$, we need to show that $\mathbf{\Pi}_N(\mathcal{U}^c|\mathbf{Y}_1, \ldots, \mathbf{Y}_N) \to 0$ a.s. $[P_{\boldsymbol{\omega}_0}]$ for every weak neighborhood, $\mathcal{U}$ of $\boldsymbol{\omega}_0$. Following a similar notation to Ghosal and Van der Vaart [2017], let ψ_N be measurable mappings, $\psi_N : \boldsymbol{\mathcal{S}}^N \times \boldsymbol{\mathcal{Z}}^N \to [0,1]$, where $\boldsymbol{\mathcal{Z}}$ is the sample space of $\{Z_{i1}, \ldots, Z_{iK}\}$. Let $\psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)$ be the corresponding test function, and $P_{\boldsymbol{\omega}}^N\psi_N = \mathbb{E}_{P_{\boldsymbol{\omega}}^N}\psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) = \int \psi_N \mathrm{d}P_{\boldsymbol{\omega}}^N$, where $P_{\boldsymbol{\omega}}^N$ denotes the joint distribution on $\mathbf{Y}_1, \ldots, \mathbf{Y}_N$ with parameters $\boldsymbol{\omega}$. Suppose there exists tests ψ_N such that $P_{\boldsymbol{\omega}_0}^N\psi_N \to 0$,

and $\sup_{\boldsymbol{\omega}\in\mathcal{U}^c} P_{\boldsymbol{\omega}}^N(1-\psi_N) \to 0$. Since $\psi_N(\mathbf{Y}_1,\ldots,\mathbf{Y}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N) \in [0,1]$, we have that

$$\begin{aligned}\boldsymbol{\Pi}_n(U^c|\mathbf{Y}_1,\ldots,\mathbf{Y}_N) &\leq \boldsymbol{\Pi}_n(U^c|\mathbf{Y}_1,\ldots,\mathbf{Y}_N) + \psi_N(\mathbf{Y}_1,\ldots,\mathbf{Y}_N)\left(1-\boldsymbol{\Pi}_n(U^c|\mathbf{Y}_1,\ldots,\mathbf{Y}_N)\right) \\ &= \psi_N(\mathbf{Y}_1,\ldots,\mathbf{Y}_N) + \frac{\left(1-\psi_N(\mathbf{Y}_1,\ldots,\mathbf{Y}_N)\right)\int_{U^c}\prod_{i=1}^N \frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})}{\int_{\boldsymbol{\Omega}}\prod_{i=1}^N \frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})}.\end{aligned} \tag{B.24}$$

To show that $\boldsymbol{\Pi}_n(U^c|\mathbf{Y}_1,\ldots,\mathbf{Y}_N) \to 0$, it is sufficient to show the following three conditions:

1. $\psi_N(\mathbf{Y}_1,\ldots,\mathbf{Y}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N) \to 0$ a.s. $[P_{\boldsymbol{\omega}_0}]$,
2. $e^{\beta_1 N}\left(1-\psi_N(\mathbf{Y}_1,\ldots,\mathbf{Y}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N)\right)\int_{\mathcal{U}^c}\prod_{i=1}^N \frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega}) \to 0$ a.s. $[P_{\boldsymbol{\omega}_0}]$ for some $\beta_1 > 0$,
3. $e^{\beta N}\left(\int_{\boldsymbol{\Omega}}\prod_{i=1}^N \frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\right) \to \infty$ a.s. $[P_{\boldsymbol{\omega}_0}]$ for all $\beta > 0$.

We will start by proving (c). Fix $\beta > 0$. Thus we have

$$e^{\beta N}\left(\int_{\boldsymbol{\Omega}}\prod_{i=1}^N \frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\right) = e^{\beta N}\left(\int_{\boldsymbol{\Omega}}\exp\left[-\sum_{i=1}^N \log\left(\frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}\right)\right]\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\right).$$

By Fatou's lemma, we have

$$\begin{aligned}&\liminf_{N\to\infty}\int_{\boldsymbol{\Omega}}\exp\left[\beta N - \sum_{i=1}^N \log\left(\frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}\right)\right]\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega}) \\ \geq &\int_{\boldsymbol{\Omega}}\liminf_{N\to\infty}\exp\left[\beta N - \sum_{i=1}^N \log\left(\frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}\right)\right]\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\end{aligned}$$

Let $\beta > \epsilon > 0$ and $a, b > 0$ be defined such that lemma 5 holds. Since $\mathcal{C}(\boldsymbol{\omega}_0,\epsilon) \subset \boldsymbol{\Omega}$, we have

that

$$\int_{\boldsymbol{\Omega}} \liminf_{N\to\infty} \exp\left[\beta N - \sum_{i=1}^{N} \log\left(\frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}\right)\right] \mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})$$
$$\geq \int_{\mathcal{C}(\boldsymbol{\omega}_0,\epsilon)} \liminf_{N\to\infty} \exp\left[\beta N - \sum_{i=1}^{N} \log\left(\frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}\right)\right] \mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})$$

By Kolmogorov's strong law of large numbers for non-identically distributed random variables, we have that

$$\frac{1}{N}\sum_{i=1}^{N} \left(\Lambda_i(\boldsymbol{\omega}_0,\boldsymbol{\omega}) - K_i(\boldsymbol{\omega}_0,\boldsymbol{\omega})\right) \to 0$$

a.s. $[P_{\boldsymbol{\omega}_0}]$. Thus for each $\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0,\epsilon)$, with $P_{\boldsymbol{\omega}_0}$-probability 1,

$$\frac{1}{N}\sum_{i=1}^{N} \Lambda_i(\boldsymbol{\omega}_0,\boldsymbol{\omega}) \to \mathbb{E}(\overline{K_i(\boldsymbol{\omega}_0,\boldsymbol{\omega})}) < \epsilon < B,$$

since $\boldsymbol{\omega} \in \mathcal{C}(\boldsymbol{\omega}_0,\epsilon)$. Therefore, we have that

$$\int_{\mathcal{C}(\boldsymbol{\omega}_0,\epsilon)} \liminf_{N\to\infty} \exp\left[\beta N - \sum_{i=1}^{N} \log\left(\frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}\right)\right] \mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega}) \geq \int_{\mathcal{C}(\boldsymbol{\omega}_0,\epsilon)} \inf_{N\to\infty} \exp\left\{N(\beta-\epsilon)\right\} \mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega}).$$

Since $\beta - \epsilon > 0$, and $\boldsymbol{\Pi}\left(\theta \in \mathcal{C}(\boldsymbol{\omega}_0,\epsilon)\right) > 0$ (lemma 5), we have that

$$e^{\beta N}\left(\int_{\boldsymbol{\Omega}} \prod_{i=1}^{N} \frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)} \mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\right) \to \infty \tag{B.25}$$

a.s. $[P_{\boldsymbol{\omega}_0}]$ for all $\beta > 0$. We will now show that exists measurable mappings such that $P^N_{\boldsymbol{\omega}_0}\psi_N \to 0$ and $\sup_{\boldsymbol{\omega}\in\mathcal{U}^c} P^N_{\boldsymbol{\omega}}(1-\psi_N) \to 0$. Consider weak neighborhoods $\mathcal{U}$ of $\boldsymbol{\omega}_0$ of the form

$$\mathcal{U} = \left\{\boldsymbol{\omega} : \left|\int f_i \mathrm{d}P_{\boldsymbol{\omega}} - \int f_i \mathrm{d}P_{\boldsymbol{\omega}_0}\right| < \epsilon_i, \;\; i = 1,2,\ldots,r\right\}, \tag{B.26}$$

where $r \in \mathbb{N}$, $\epsilon_i > 0$, and f_i are continuous functions such that $f_i : \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{Z}} \to [0,1]$. As shown in Ghosh and Ramamoorthi [2003], for any particular f_i and $\epsilon_i > 0$, $\left|\int f_i \mathrm{d}P_{\boldsymbol{\omega}} - \int f_i \mathrm{d}P_{\boldsymbol{\omega}_0}\right| <$

ϵ_i iff $\int f_i \mathrm{d}P_{\boldsymbol{\omega}} - \int f_i \mathrm{d}P_{\boldsymbol{\omega}_0} < \epsilon_i$ and $\int (1-f_i)\mathrm{d}P_{\boldsymbol{\omega}} - \int (1-f_i)\mathrm{d}P_{\boldsymbol{\omega}_0} < \epsilon$. Since $\tilde{f}_i := (1-f_i)$ is still a continuous function such that $\tilde{f}_i : \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{Z}} \to [0,1]$, we can rewrite equation B.26 as

$$\mathcal{U} = \cap_{i=1}^{2r} \left\{ \boldsymbol{\omega} : \int g_i \mathrm{d}P_{\boldsymbol{\omega}} - \int g_i \mathrm{d}P_{\boldsymbol{\omega}_0} < \epsilon_i \right\}, \tag{B.27}$$

where g_i are continuous functions such that $g_i : \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{Z}} \to [0,1]$ and $\epsilon_i > 0$. Following Ghosal and Van der Vaart [2017], it can be shown by Hoeffding's inequality that using the test function $\tilde{\psi}$, defined as

$$\tilde{\psi}_{iN}(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) := \mathbb{1}\left\{ \frac{1}{N} \sum_{j=1}^{N} g_i(\mathbf{Y}_j, \mathbf{z}_j) > \int g_i \mathrm{d}P_{\boldsymbol{\omega}_0} + \frac{\epsilon_i}{2} \right\}, \tag{B.28}$$

leads to

$$\int \tilde{\psi}_{iN}(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) \mathrm{d}P_{\boldsymbol{\omega}_0} \leq e^{-N\epsilon_i^2/2}$$

and

$$\int \left(1 - \tilde{\psi}_{iN}(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right) \mathrm{d}P_{\boldsymbol{\omega}} \leq e^{-N\epsilon_i^2/2}$$

for any $\boldsymbol{\omega} \in \mathcal{U}^c$. Let $\boldsymbol{\psi}_n = \max_i \tilde{\psi}_{iN}$ be our test function and $\epsilon = \min_i \epsilon_i$. Using the fact that $\mathbb{E}(\max_i \tilde{\psi}_{iN}) \leq \sum_i \mathbb{E}(\tilde{\psi}_{iN})$ and $\mathbb{E}(1 - \max_i \tilde{\psi}_{iN}) \leq \mathbb{E}(1 - \tilde{\psi}_{iN})$, we have

$$\int \psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) \mathrm{d}P_{\boldsymbol{\omega}_0} \leq (2r) e^{-N\epsilon^2/2} \tag{B.29}$$

and

$$\int \left(1 - \psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right) \mathrm{d}P_{\boldsymbol{\omega}} \leq e^{-N\epsilon^2/2}, \tag{B.30}$$

for any $\boldsymbol{\omega} \in \mathcal{U}^c$. Using Markov's inequality on equation B.29, we have that

$$\begin{aligned} P\left(\psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) \geq e^{-nC}\right) &\leq \frac{\mathbb{E}\left(\psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)}{e^{-NC}} \\ &\leq (2r) e^{-N(\epsilon^2/2 - C)} \end{aligned}$$

Thus letting $C < \epsilon^2/2$, we have that $\sum_{N=1}^{\infty} P\left(\psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) \geq e^{-NC}\right) < \infty$. Thus by the Borel-Cantelli lemma, we know that

$$P\left(\limsup_{N\to\infty} P\left(\psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) \geq e^{-NC}\right)\right) = 0$$

Thus we have that $\psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N) \to 0$ a.s. $[P_{\boldsymbol{\omega}_0}]$ (Condition (a)). To prove condition (b), we will first start by taking the expectation with respect to $P_{\boldsymbol{\omega}_0}$:

$$\begin{aligned}
&\mathbb{E}_{P^N_{\boldsymbol{\omega}_0}}\left(e^{\beta N}\left(1-\psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)\int_{\mathcal{U}^c}\prod_{i=1}^{N}\frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\right)\\
=&\int_{\boldsymbol{\mathcal{S}}^N}\left(e^{\beta N}\left(1-\psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)\int_{\mathcal{U}^c}\prod_{i=1}^{N}\frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\right)\mathrm{d}P^N_{\boldsymbol{\omega}_0}\\
=&\int_{\mathcal{U}^c}\left(\prod_{i=1}^{N}\int_{\mathcal{S}}e^{\beta N}\left(1-\psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)f_i(\mathbf{Y}_i;\boldsymbol{\omega})\mathrm{d}\mathbf{Y}_i\right)\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\\
=&e^{\beta N}\int_{\mathcal{U}^c}\mathbb{E}_{P^N_{\boldsymbol{\omega}}}\left(1-\psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\\
\leq&e^{\beta_1 N}e^{-N\epsilon^2/2},
\end{aligned}$$

where the last inequality is from equation B.30. Thus by Markov's inequality and letting $\beta_1 < \epsilon^2/2$, we have that

$$\begin{aligned}
&P\left(e^{\beta N}\left(1-\psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)\int_{\mathcal{U}^c}\prod_{i=1}^{N}\frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega}) \geq e^{-N((\epsilon^2/2-\beta_1)/2)}\right)\\
\leq&\frac{\mathbb{E}_{P^N_{\boldsymbol{\omega}_0}}\left(e^{\beta N}\left(1-\psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)\int_{\mathcal{U}^c}\prod_{i=1}^{N}\frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\right)}{e^{-N((\epsilon^2/2-\beta_1)/2)}}\\
\leq&e^{-N((\epsilon^2/2-\beta_1)/2)}
\end{aligned}$$

Letting E_N be the event that $e^{\beta N}\left(1-\psi_N(\mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{z}_1, \ldots, \mathbf{z}_N)\right)\int_{\mathcal{U}^c}\prod_{i=1}^{N}\frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})$ $\geq e^{-N((\epsilon^2/2-\beta_1)/2)}$, we have that $\sum_{i=1}^{\infty} P(E_N) < \infty$. Thus by the Borel-Cantelli lemma, we

have that

$$e^{\beta N}\left(1-\psi_N(\mathbf{Y}_1,\ldots,\mathbf{Y}_N,\mathbf{z}_1,\ldots,\mathbf{z}_N)\right)\int_{\mathcal{U}^c}\prod_{i=1}^{N}\frac{f_i(\mathbf{Y}_i;\boldsymbol{\omega})}{f_i(\mathbf{Y}_i;\boldsymbol{\omega}_0)}\mathrm{d}\boldsymbol{\Pi}(\boldsymbol{\omega})\to 0$$

a.s. $[P_{\boldsymbol{\omega}_0}]$ for $0<\beta_1<\epsilon^2/2$. Therefore, we have proved conditions (a), (b), and (c). Thus by letting β in condition (c) be such that $\beta=\beta_1$, where $0<\beta_1<\epsilon^2/2$, we can see that $\boldsymbol{\Pi}_N(\mathcal{U}^c|\mathbf{Y}_1,\ldots,\mathbf{Y}_N)\to 0$ a.s. $[P_{\boldsymbol{\omega}_0}]$ for every weak neighborhood, $\mathcal{U}$ of $\boldsymbol{\omega}_0$.

B.2 Case Studies

B.2.1 Simulation Study 1

In this simulation study, we looked at how well we could recover the mean, covariance, and cross-covariance functions at different numbers of functional observations. For this simulation, we used 3 different number of functional observations ($N=40,80,160$), and ran 50 MCMC chains for $500,000$ iterations. To help the chain converge, we used the Multiple Start Algorithm (Algorithm 3) with $n\_try1=50$, $n\_try2=10$, $n\_MCMC1=2000$, and $n\_MCMC2=20000$. Due to our allocated computation budget, we did not use tempered transitions to help move around the space of parameters. In order to save on memory, we only saved every 100 iterations. We used 8 functions to form the basis of the observed functions, such that the observed smooth functions lie in a space spanned by cubic b-spline basis functions with 4 equally spaced internal nodes ($P=8$), and that 3 eigenfunctions can capture the entire covariance process ($M=3$). For this simulation, we used the two feature model ($K=2$). For each simulation, we used the same $\boldsymbol{\nu}$, $\boldsymbol{\Phi}$, and $\boldsymbol{\sigma}^2$ parameters for each simulation. We specified that $\sigma^2=0.001$, while the $\boldsymbol{\nu}$ parameters were drawn according to the following distributions:

$$\boldsymbol{\nu}_1\sim\mathcal{N}\left((6,4,\ldots,-6,-8)',4\mathbf{P}\right),$$

$$\boldsymbol{\nu}_2 \sim \mathcal{N}\left((-8, -6, \ldots, 4, 6)', 4\mathbf{P}\right),$$

where $\mathbf{P}$ is the matrix corresponding with the first order random walk penalty. Due to the non-identifiability described in Section 3.1.2, we drew the $\boldsymbol{\Phi}$ parameters from the subspace orthogonal to the space spanned by the $\boldsymbol{\nu}$ parameters. Thus let $\text{colsp}(\mathbf{B}^{\perp}) := \text{span}\{b_1^{\perp}, \ldots, b_6^{\perp}\} \subset \mathbb{R}^8$ be the subspace orthogonal to the $\boldsymbol{\nu}$ parameters, which can be described as the span of 6 vectors in $\mathbb{R}^8$. The $\boldsymbol{\Phi}$ parameters were drawn according to the following distributions:

$$\boldsymbol{\phi}_{km} = \mathbf{q}_{km}\mathbf{B}^{\perp} \quad k = 1, 2 \quad m = 1, 2, 3,$$

where $\mathbf{q}_{k1} \sim \mathcal{N}(\mathbf{0}_6, 2.25\mathbf{I}_6)$, $\mathbf{q}_{k2} \sim \mathcal{N}(\mathbf{0}_6, \mathbf{I}_6)$, $\mathbf{q}_{k3} \sim \mathcal{N}(\mathbf{0}_6, 0.49\mathbf{I}_6)$. While this may not completely remove the effect of the non-identifiability mentioned in Section 3.1.2, it should help minimize its impact on our recovery of the mean and covariance structures.

For the $\mathbf{z}_i$ and $\boldsymbol{\chi}_{im}$ parameters, we would draw 3 different sets of parameters (corresponding to the various number of functional observations). The χ_{im} parameters were drawn from a standard normal distribution. The $\mathbf{z}_i$ parameters were drawn from a mixture of Dirichlet distributions. Roughly 30% of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution with $\alpha_1 = 10$ and $\alpha_2 = 1$. Another roughly 30% of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution where $\alpha_1 = 1$ and $\alpha_2 = 10$. The rest of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution with $\alpha_1 = \alpha_2 = 1$. For each simulation, we used these parameters to simulate observed functions from our model.

Before getting the posterior median estimates of the functions of interest, we used the Membership Rescale Algorithm (algorithm 4) to help with interpretability and identifiability. From figure B.1, we can see that that we do a good job in recovering the covariance and cross-covariance functions. The estimated functions are slightly conservative, as they tend to slightly underestimate the magnitude of the covariance functions. Figure B.2 show the median posterior mean recovered from each of the 10 MCMC chains when we have 250 functional observations. As we can see from the figure, there is very little variation in the

estimates of the mean function between 10 MCMC chains.

B.2.2 Simulation Study 2

Picking the number of features can be a challenging task for many practitioners, especially when there is little scientific knowledge on the data. Practitioners often rely on information criterion to help aid in picking the number of features. In this simulation, we simulate 10 different "true" data-sets from a 3 feature model to see if information criterion can help pick the correct number of features. For this simulation study, we considered testing the information criterion under the model when $K = 2, 3, 4$, and 5 (where K is the number of features in our model). For each K and each data-set, we ran a MCMC chain for 100,000 iterations each. To help the chain converge, we used the Multiple Start Algorithm (Algorithm 3) with $n\_try1 = 50$, $n\_try2 = 5$, $n\_MCMC1 = 2000$, and $n\_MCMC2 = 4000$. To save on memory, we only saved every 10 iterations.

For the 10 "true" data-sets with 3 functional features ($K = 3$) and 200 functional observations ($N = 200$, $n_i = 100$), we assumed that the observed smooth functions lie in a space spanned by cubic b-spline basis functions with 4 equally spaced internal nodes ($P = 8$), and that 3 eigenfunctions can capture the entire covariance process ($M = 3$). In this simulation, we assumed that $\sigma^2 = 0.001$, and randomly drew the $\boldsymbol{\nu}$ and $\boldsymbol{\Phi}$ parameters for each data-set according to the following distributions:

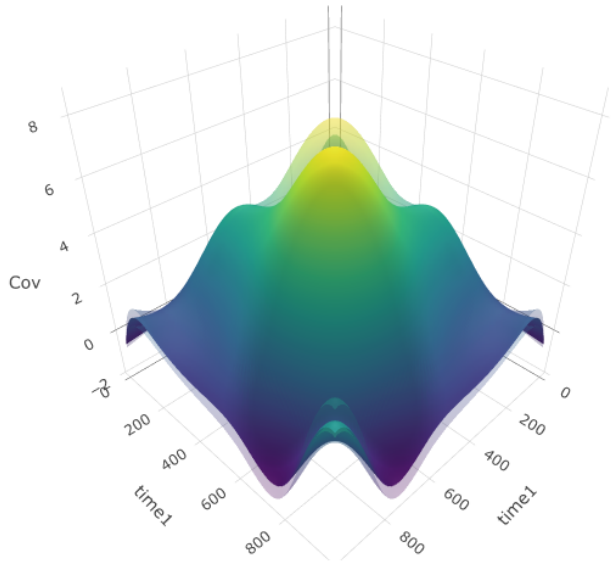
$$\boldsymbol{\nu}_1 \sim \mathcal{N}\left((6, 4, \ldots, -6, -8)', 4\mathbf{P}\right),$$

$$\boldsymbol{\nu}_2 \sim \mathcal{N}\left((-8, -6, \ldots, 4, 6)', 4\mathbf{P}\right),$$

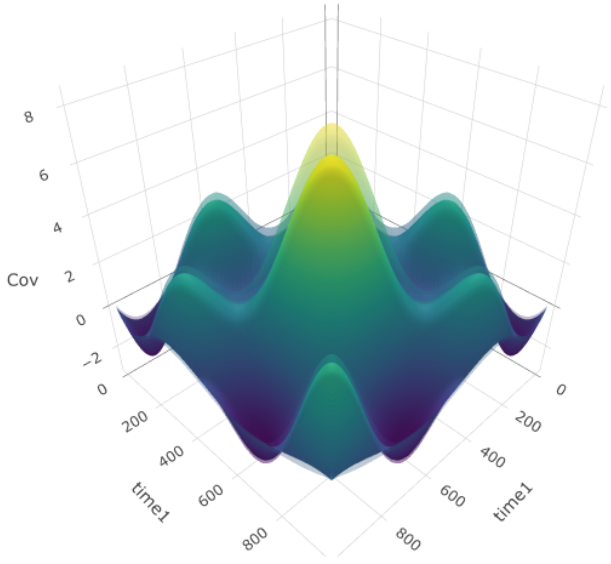
$$\boldsymbol{\nu}_1 \sim \mathcal{N}\left(\mathbf{0}, 4\mathbf{P}\right),$$

$$\boldsymbol{\phi}_{k1} \sim \mathcal{N}\left(\mathbf{0}, \mathbf{I}_8\right),$$

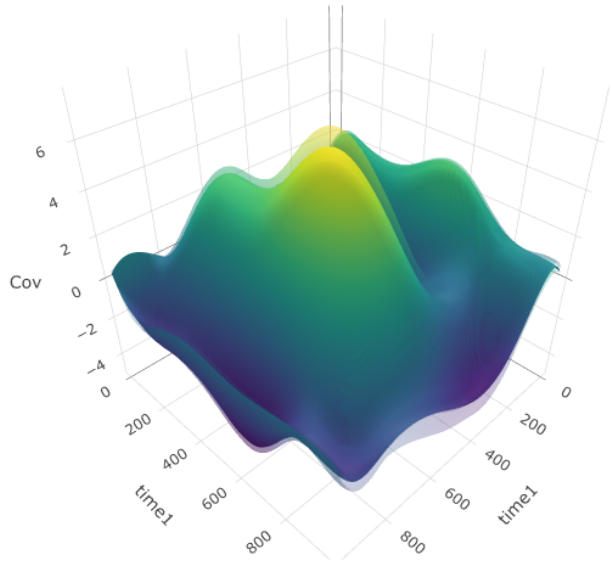
$$\boldsymbol{\phi}_{k2} \sim \mathcal{N}\left(\mathbf{0}, 0.5\mathbf{I}_8\right),$$



(a) $C^{(1,1)}$



(b) $C^{(2,2)}$



(c) $C^{(1,2)}$

Figure B.1: Posterior median estimates of the covariance and cross-covariance functions (opaque) along with the true functions (transparent) for a simulated data set with 160 functional observations.

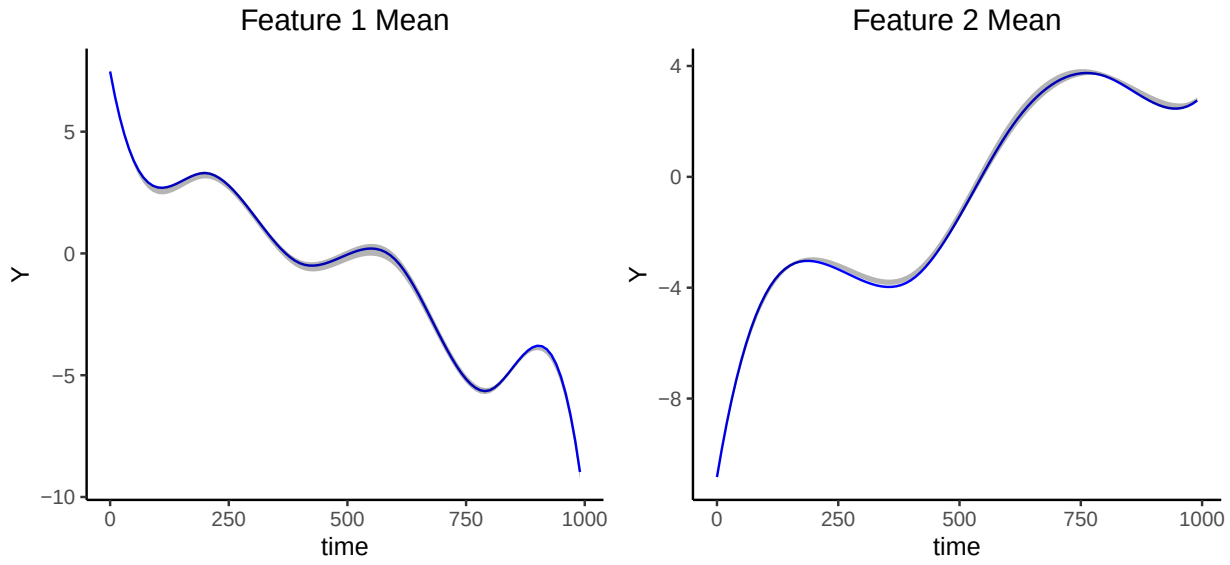


Figure B.2: 95% credible interval of the posterior mean functions for the case when we have 160 functional observations.

$$\boldsymbol{\phi}_{k3} \sim \mathcal{N}\left(\mathbf{0}, 0.2\mathbf{I}_8\right).$$

The χ_{im} parameters were drawn from a standard normal distribution, while the $\mathbf{Z}$ parameters were drawn from a mixture of Dirichlet distributions. 20% of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution with $\alpha_1 = 10$, $\alpha_2 = 1$, and $\alpha_3 = 1$. Another 20% of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution where $\alpha_1 = 1$, $\alpha_2 = 10$, and $\alpha_3 = 1$. Another 20% of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution where $\alpha_1 = 1$, $\alpha_2 = 1$, and $\alpha_3 = 10$. The rest of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution with $\alpha_1 = \alpha_2 = 1$. Once all of the parameters for the "true" data-set were specified, the observed data points were generated according to the model. MCMC was then conducted with various values of K, but with the correct number of eigenfunctions, M, and the correct basis functions.

B.2.3 A Case Study of EEG in ASD

In this study, we grouped patients based off of their T8 electrode signal. We used a two functional feature model, and found that patients in the first feature could be interpreted as $1/f$ noise, while the second feature could be interpreted as a distinct PAF. We also fit a three functional feature mixed membership model, but found that the two feature mixed membership model (AIC = -13091.31, BIC = 9369.66, DIC = -13783.5) seemed to be optimal compared to the three feature paratial membership model (AIC = -12831.17, BIC = 8138.12, DIC= -13815.34). Figure B.3 shows the median of the posterior posterior distribution of the covariance and cross-covariance functions. We can see that the covariance function associated with feature 1 ($C^{(1,1)}$) has high covariance around 6 Hz, which is where we have the highest power of $1/f$ noise.

When looking at the mean function for feature 2 (figure 3.5), we can see that the peak power occurs at around 9 Hz. However, for people who have a distinct PAF pattern, it is common for their peak power to occur anywhere between 9 Hz and 11 Hz. When looking at the covariance function associated with feature 2 ($C^{(2,2)}$), we can see that this is being modeled by the high variance at 9 Hz and at 11 Hz. We can also see that people who only have high Alpha power typically tend to only have one peak in the alpha band, which is also accounted for in our model by the negative covariance between 9 Hz and 11 Hz. When looking at the cross-covariance function, we can see that there is high cross-covariance between 9 Hz in feature 1 and 6 Hz in feature 2, and negative cross-covariance between 11 Hz in feature 1 and 6 Hz in feature 2. This means that patients who are simultaneously in feature 1 and 2 that have moderate $1/f$ noise are likely to have moderate alpha power around 9 Hz and are less likely to have a peak around 11 HZ. According to the scientific literature, this is likely to occur in younger TD individuals.

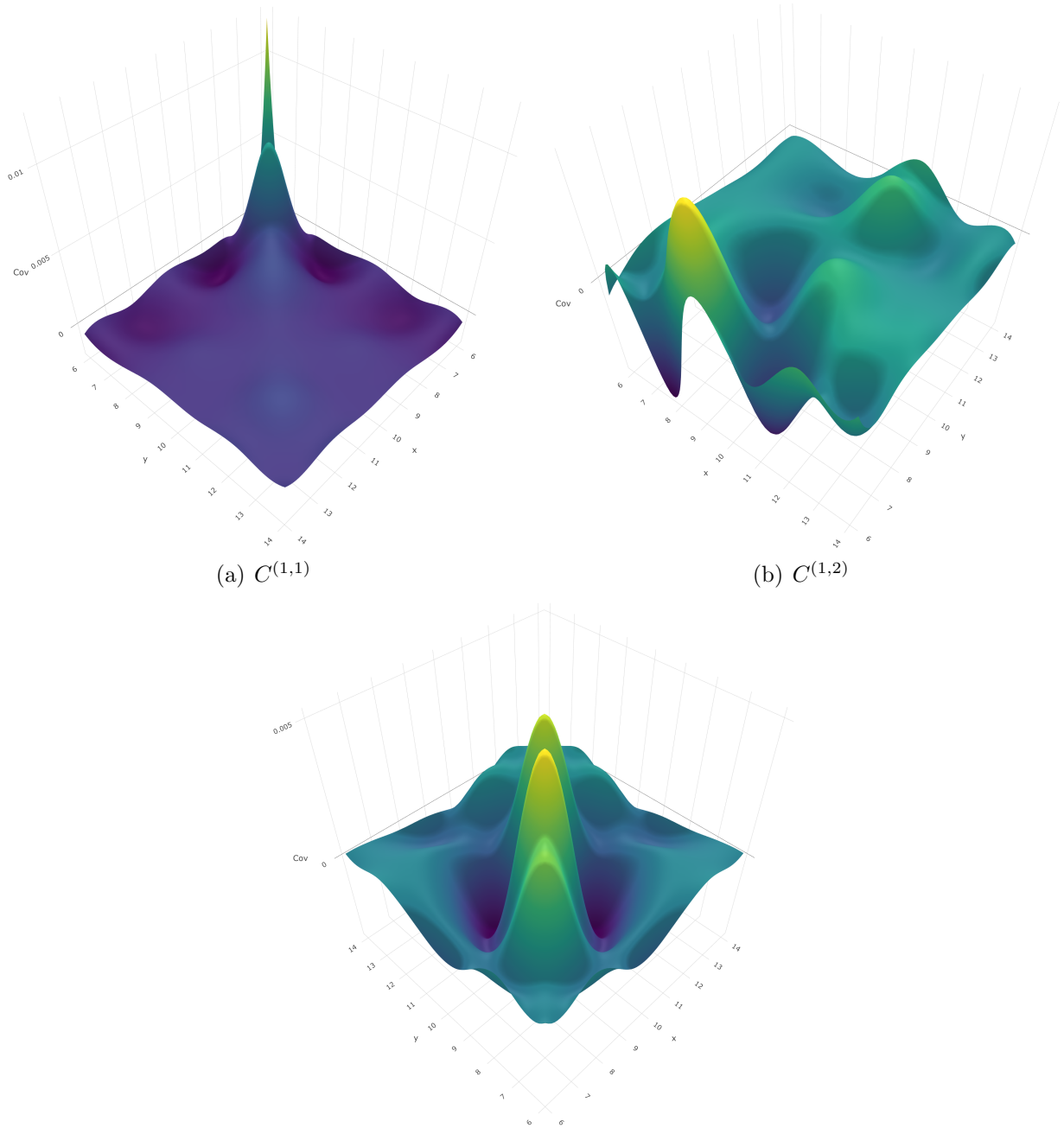
From figure 3.6, we can see that on average ASD children were tended to belong to feature 1 more than feature 2. Thus on average, ASD children tended to have a less distinct PAF

when compared to TD children.

B.2.4 Analysis of Multi-Channel EEG Data

The proposed modeling framework is suitable for the analysis of functional data evaluated over $\mathcal{T} \subset R^d$. Therefore, we extend our analysis in the main manuscript to include EEG data measured on the entire cortex. Specifically, we will use a model with 2 latent functional features ($K = 2$) where $\mathcal{T} \subset \mathbb{R}^3$. Two of the three indices denote the spatial position on the scalp, while the third index contains information on the frequency observed. Similarly, the value of the function at some point $t \in \mathcal{T}$ represents the spectral power of the observed signal. For computational purposes, we project the true three-dimensional coordinates of the electrodes to a two-dimensional bird's eye view of electrodes using the 'eegkit' package developed by Helwig [2018]. In this section, we used two eigenfunctions to capture the covariance process ($M = 2$). We used a tensor product of B-splines to create a basis for our space of functions. For each dimension we used quadratic B-splines, with 3 internal nodes for each spatial index and 2 internal nodes for the frequency index ($P = 180$). Since we are using functional data analysis techniques to model the EEG data, we assume that the smoothness over the spatial and frequency domains. Since EEG data has poor spatial resolution [Grinvald and Hildesheim, 2004] and we have relatively sparse sampling across the spatial domain (25 channels), the smoothness assumption can be thought of as a type of regularization over the domain of our function. Due to computational limitations, we ran the Multiple Start Algorithm (algorithm 3) with $n\_try1 = 6$, $n\_try2 = 1$, $n\_MCMC1 = 3000$, and $n\_MCMC2 = 4000$. We then ran the chain for 19,000 iterations, saving only every 10 iterations.

Figure B.4 reports posterior mean estimates for the feature means over a sample of electrodes. Our findings are similar to our results on electrode T8, analyzed in the main manuscript; one latent feature corresponding to $1/f$ noise, and the other exhibiting well defined PAF across electrodes. Figure B.5, reports the electrode-specific variance at frequency



(a) $C^{(1,1)}$

(b) $C^{(1,2)}$

(c) $C^{(2,2)}$

Figure B.3: Posterior estimates of the covariance and cross-covariance functions

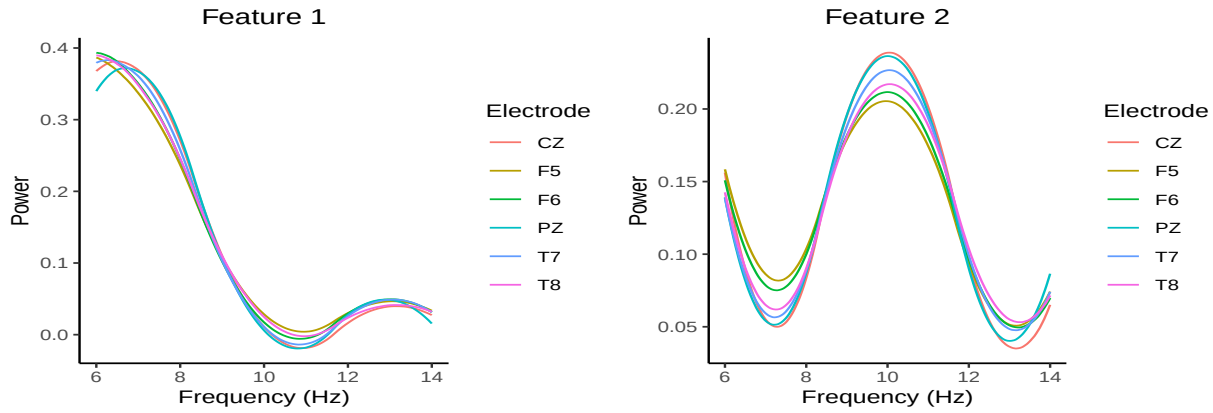
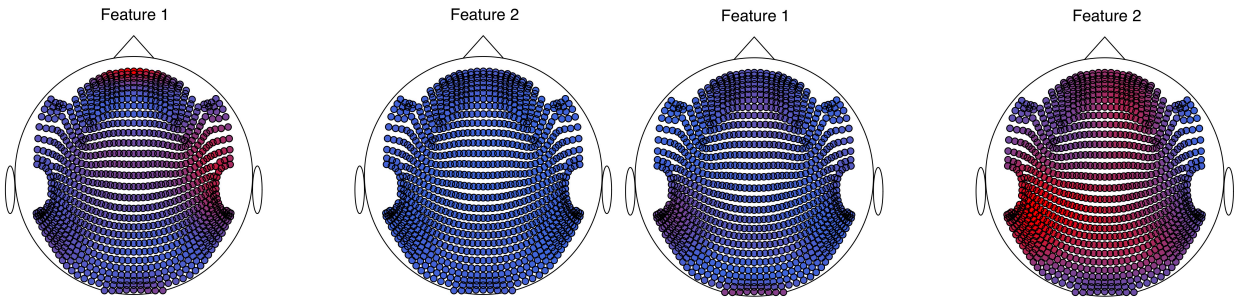


Figure B.4: Posterior estimates of the means of the two functional features viewed at specific electrodes.

6 Hz and 10 Hz, corresponding respectively to the highest relative power in the first latent feature and the average peak alpha frequency in second latent feature.

The right-temporal region around electrode T8, is found to exhibit a high level of heterogeneity (high relative variance) at 6 Hz, within latent feature 1 (poorly defined PAF), which relates to the findings of Scheffler et al. [2019], who identified patterns of variation in the right-temporal region as the highest contributor to log-odds of ASD vs. TD discrimination. Contrastingly, feature 2 (well defined PAF) exhibits high levels of heterogeneity (relative variance) throughout the cortex at frequency 10 Hz, corresponding to the location of the PAF in feature 2. Overall, results for our analysis on the whole set of electrodes agree with our findings for electrode T8 in the main manuscript.

Figure B.6 shows the posterior median estimates of the membership allocations for each individual. We can see from both the mean functions and membership allocations that these results seem to match the univariate results in Section 3.3.3.



(a) 6 Hz (b) 10 Hz

Figure B.5: Variance of the electrodes at 6 and 10 Hz for each functional feature. The relative magnitude of the variance of each electrode is indicated by the color of the electrode (red is relatively high variance, while blue is relatively low variance).

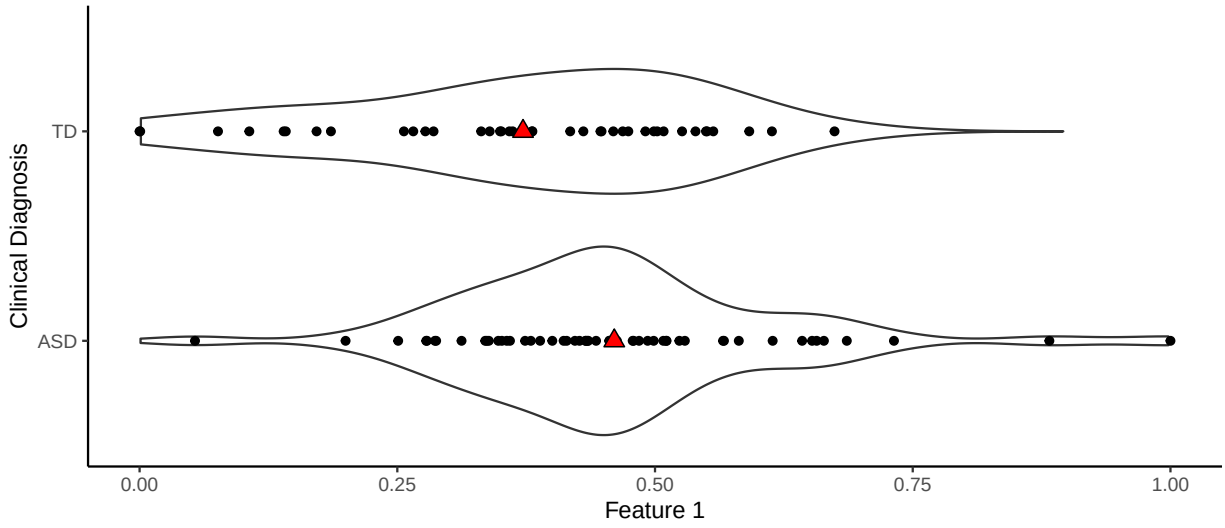


Figure B.6: Posterior estimates of the median membership to the first functional feature.

B.3 Computation

B.3.1 Posterior Distributions and Computation

In this section, we will discuss the computational strategy used to perform Bayesian inference. In cases where the posterior distribution is a known distribution, a Gibbs update will be performed. We will let $\boldsymbol{\Theta}$ be the collection of all parameters, and $\boldsymbol{\Theta}_{-\zeta}$ be the collection of all parameters, excluding the ζ parameter. We will first start with the $\boldsymbol{\phi}_{km}$ parameters, for $j = 1, \ldots, K$ and $m = 1, \ldots, M$. Let $\mathbf{D}_{jm} = \tilde{\tau}_{mj}^{-1} diag\left(\gamma_{j1m}^{-1}, \ldots, \gamma_{jPm}^{-1}\right)$. By letting

$$\begin{aligned}\mathbf{m}_{jm} =& \frac{1}{\sigma^2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} \left(B(t_{il}) \chi_{im} \left(y_i(t_{il}) Z_{ij} - Z_{ij}^2 \boldsymbol{\nu}_j' B(t_{il}) - Z_{ij}^2 \sum_{n \neq m} \left[\chi_{in} \boldsymbol{\phi}_{jn}' B(t_{il}) \right] - \right. \right. \\ & \left. \left. \sum_{k \neq j} Z_{ij} Z_{ik} \left[\boldsymbol{\nu}_k' B(t_{il}) + \sum_{n=1}^{M} \chi_{in} \boldsymbol{\phi}_{kn}' B(t_{il}) \right] \right) \right),\end{aligned}$$

and

$$\mathbf{M}_{jm}^{-1} = \frac{1}{\sigma^2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} \left(Z_{ij}^2 \chi_{im}^2 B(t_{il}) B'(t_{il}) \right) + \mathbf{D}_{jm}^{-1},$$

we have that

$$\boldsymbol{\phi}_{jm} | \boldsymbol{\Theta}_{-\boldsymbol{\phi}_{jm}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N \sim \mathcal{N}(\mathbf{M}_{jm} \mathbf{m}_{jm}, \mathbf{M}_{jm}).$$

The posterior distribution of δ_{1k}, for $k = 1, \ldots, K$, is

$$\begin{aligned}\delta_{1k} | \boldsymbol{\Theta}_{-\delta_{1k}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N \sim & \Gamma \left(a_{1k} + (PM/2), 1 + \frac{1}{2} \sum_{r=1}^{P} \gamma_{k,r,1} \phi_{k,r,1}^2 \right. \\ & \left. + \frac{1}{2} \sum_{m=2}^{M} \sum_{r=1}^{P} \gamma_{k,r,m} \phi_{k,r,m}^2 \left(\prod_{j=2}^{m} \delta_{jk} \right) \right).\end{aligned}$$

The posterior distribution for δ_{ik}, for $i = 2, \ldots, M$ and $k = 1, \ldots, K$, is

$$\delta_{ik}|\boldsymbol{\Theta}_{-\delta_{ik}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N \sim \Gamma\Bigg(a_{2k} + (P(M-i+1)/2), 1 + \frac{1}{2}\sum_{m=i}^{M}\sum_{r=1}^{P}\gamma_{k,r,m}\phi_{k,r,m}^2\left(\prod_{j=1;j\neq i}^{m}\delta_j\right)\Bigg).$$

The posterior distribution for a_{1k} is not a commonly known distribution, however we have that

$$P(a_{1k}|\boldsymbol{\Theta}_{-a_{1k}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N) \propto \frac{1}{\Gamma(a_{1k})}\delta_{1k}^{a_{1k}-1}a_{1k}^{\alpha_1-1}exp\left\{-a_{1k}\beta_1\right\}.$$

Since this is not a known kernel of a distribution, we will have to use Metropolis-Hastings algorithm. Consider the proposal distribution $Q(a'_{1k}|a_{1k}) = \mathcal{N}\left(a_{1k}, \epsilon_1\beta_1^{-1}, 0, +\infty\right)$ (Truncated Normal) for some small $\epsilon_1 > 0$. Thus the probability of accepting any step is

$$A(a'_{1k}, a_{1k}) = \min\left\{1, \frac{P\left(a'_{1k}|\boldsymbol{\Theta}_{-a'_{1k}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N\right)}{P\left(a_{1k}|\boldsymbol{\Theta}_{-a_{1k}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N\right)}\frac{Q\left(a_{1k}|a'_{1k}\right)}{Q\left(a'_{1k}|a_{1k}\right)}\right\}.$$

Similarly for a_{2k}, we have

$$P(a_{2k}|\boldsymbol{\Theta}_{-a_{2k}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N) \propto \frac{1}{\Gamma(a_{2k})^{M-1}}\left(\prod_{i=2}^{M}\delta_{ik}^{a_{2k}-1}\right)a_{2k}^{\alpha_{2k}-1}exp\left\{-a_{2k}\beta_2\right\}.$$

We will use a similar proposal distribution, such that $Q(a'_{2k}|a_{2k}) = \mathcal{N}\left(a_{2k}, \epsilon_2\beta_2^{-1}, 0, +\infty\right)$ for some small $\epsilon_2 > 0$. Thus the probability of accepting any step is

$$A(a'_{2k}, a_{2k}) = \min\left\{1, \frac{P\left(a'_{2k}|\boldsymbol{\Theta}_{-a'_{2k}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N\right)}{P\left(a_{2k}|\boldsymbol{\Theta}_{-a_{2k}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N\right)}\frac{Q\left(a_{2k}|a'_{2k}\right)}{Q\left(a'_{2k}|a_{2k}\right)}\right\}.$$

For the $\gamma_{j,r,m}$ parameters, for $j = 1, \ldots K$, $r = 1, \ldots, P$, and $m = 1, \ldots, M$, we have

$$\gamma_{j,r,m}|\boldsymbol{\Theta}_{-\gamma_{j,r,m}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N \sim \Gamma\left(\frac{\nu_\gamma + 1}{2}, \frac{\phi_{j,r,m}^2\tilde{\tau}_{mj} + \nu_\gamma}{2}\right).$$

The posterior distribution for the $\mathbf{z}_i$ parameters are not a commonly known distribution, so we will have to use the Metropolis-Hastings algorithm. We know that

$$p(\mathbf{z}_i|\boldsymbol{\Theta}_{-\mathbf{z}_i}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N) \propto \prod_{k=1}^{K} Z_{ik}^{\alpha_3 \pi_k - 1} \times \prod_{l=1}^{n_i} exp\left\{ -\frac{1}{2\sigma^2} \left(y_i(t_{il}) - \sum_{k=1}^{K} Z_{ik} \left(\boldsymbol{\nu}_k' B(t_{il}) + \sum_{n=1}^{M} \chi_{in} \boldsymbol{\phi}_{kn}' B(t_{il}) \right) \right)^2 \right\}.$$

We will use $Q(\mathbf{z}_i'|\mathbf{z}_i) = Dir(a_{\mathbf{z}}\mathbf{z}_i)$ for some large $a_{\mathbf{z}} \in \mathbb{R}^+$ as the proposal distribution. Thus the probability of accepting a proposed step is

$$A(\mathbf{z}_i', \mathbf{z}_i) = \min\left\{ 1, \frac{P\left(\mathbf{z}_i'|\boldsymbol{\Theta}_{-\mathbf{z}_i'}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N\right)}{P\left(\mathbf{z}_i|\boldsymbol{\Theta}_{-\mathbf{z}_i}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N\right)} \frac{Q\left(\mathbf{z}_i|\mathbf{z}_i'\right)}{Q\left(\mathbf{z}_i'|\mathbf{z}_i\right)} \right\}.$$

Similarly, a Gibbs update is not available for an update of the $\boldsymbol{\pi}$ parameters. We have that

$$p(\boldsymbol{\pi}|\boldsymbol{\Theta}_{-\boldsymbol{\pi}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N) \propto \prod_{k=1}^{K} \pi_k^{c_k - 1} \times \prod_{i=1}^{N} \frac{1}{B(\alpha_3 \boldsymbol{\pi})} \prod_{k=1}^{K} Z_{ik}^{\alpha_3 \pi_k - 1}.$$

Letting out proposal distribution be such that $Q(\boldsymbol{\pi}'|\boldsymbol{\pi}) = Dir(a_{\boldsymbol{\pi}}\boldsymbol{\pi})$, for some large $a_{\boldsymbol{\pi}} \in \mathbb{R}^+$, we have that our probability of accepting any proposal is

$$A(\boldsymbol{\pi}', \boldsymbol{\pi}) = \min\left\{ 1, \frac{P\left(\boldsymbol{\pi}'|\boldsymbol{\Theta}_{-\boldsymbol{\pi}'}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N\right)}{P\left(\boldsymbol{\pi}|\boldsymbol{\Theta}_{-\boldsymbol{\pi}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N\right)} \frac{Q\left(\boldsymbol{\pi}|\boldsymbol{\pi}'\right)}{Q\left(\boldsymbol{\pi}'|\boldsymbol{\pi}\right)} \right\}.$$

The posterior distribution of α_3 is also not a commonly known distribution, so we will use the Metropolis-Hastings algorithm to sample from the posterior distribution. We have that

$$p(\alpha_3|\boldsymbol{\Theta}_{-\alpha_3}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N) \propto e^{-b\alpha_3} \times \prod_{i=1}^{N} \frac{1}{B(\alpha_3 \boldsymbol{\pi})} \prod_{k=1}^{K} Z_{ik}^{\alpha_3 \pi_k - 1}.$$

Using a proposal distribution such that $Q(\alpha_3'|\alpha_3) = \mathcal{N}(\alpha_3, \sigma^2_{\alpha_3}, 0, +\infty)$ (Truncated Normal), we are left with the probability of accepting a proposed state as

$$A(\alpha_3', \alpha_3) = \min\left\{1, \frac{P\left(\alpha_3'|\boldsymbol{\Theta}_{-\alpha_3'}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N\right)}{P\left(\alpha_3|\boldsymbol{\Theta}_{-\alpha_3}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N\right)} \frac{Q\left(\alpha_3|\alpha_3'\right)}{Q\left(\alpha_3'|\alpha_3\right)}\right\}.$$

Let $\mathbf{P}$ be the following tridiagonal matrix:

$$\mathbf{P} = \begin{bmatrix} 1 & -1 & 0 & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & 0 & -1 & 1 \end{bmatrix}.$$

Thus, letting

$$\mathbf{B}_j = \left(\tau_j \mathbf{P} + \frac{1}{\sigma^2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} Z_{ij}^2 B(t_{il}) B'(t_{il})\right)^{-1}$$

and

$$\mathbf{b}_j = \frac{1}{\sigma^2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} Z_{ij} B(t_{il}) \left(y_i(t_{il}) - \left(\sum_{k \neq j} Z_{ik} \boldsymbol{\nu}_k' B(t_{il})\right) - \left(\sum_{k=1}^{K} \sum_{m=1}^{M} Z_{ik} \chi_{im} \boldsymbol{\phi}_{km}' B(t_{il})\right)\right),$$

we have that

$$\boldsymbol{\nu}_j|\boldsymbol{\Theta}_{-\boldsymbol{\nu}_j}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N \sim \mathcal{N}(\mathbf{B}_j \mathbf{b}_j, \mathbf{B}_j),$$

for $j = 1, \ldots, K$. Thus we can perform a Gibbs update to update our $\boldsymbol{\nu}$ parameters. The τ_l parameters, for $l = 1, \ldots K$, can also be updated by using a Gibbs update since the posterior distribution is:

$$\tau_l|\boldsymbol{\Theta}_{-\tau_l}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N \sim \Gamma\left(\alpha + P/2, \beta + \frac{1}{2} \boldsymbol{\nu}_l' \mathbf{P} \boldsymbol{\nu}_l\right).$$

The parameter σ^2 can be updated by using a Gibbs update. If we let

$$\beta_\sigma = \frac{1}{2}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\left(y_i(t_{il}) - \sum_{k=1}^{K} Z_{ik}\left(\boldsymbol{\nu}_k' B(t_{il}) + \sum_{n=1}^{M}\chi_{in}\boldsymbol{\phi}_{kn}' B(t_{il})\right)\right)^2,$$

then we have

$$\sigma^2|\boldsymbol{\Theta}_{-\sigma^2}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N \sim IG\left(\alpha_0 + \frac{\sum_{i=1}^{N} n_i}{2}, \beta_0 + \beta_\sigma\right),$$

where n_i are the number of time points observed for the i^{th} observed function. Lastly, we can update the χ_{im} parameters, for $i = 1, \ldots, N$ and $m = 1, \ldots, M$, using a Gibbs update. If we let

$$\mathbf{w}_{im} = \frac{1}{\sigma^2}\left(\sum_{l=1}^{n_i}\left(\sum_{k=1}^{K} Z_{ik}\boldsymbol{\phi}_{km}' B(t_{il})\right)\left(y_i(t_{il}) - \sum_{k=1}^{K} Z_{ik}\left(\boldsymbol{\nu}_k' B(t_{il}) + \sum_{n\neq m}\chi_{in}\boldsymbol{\phi}_{kn}' B(t_{il})\right)\right)\right)$$

and

$$\mathbf{W}_{im}^{-1} = 1 + \frac{1}{\sigma^2}\sum_{l=1}^{n_i}\left(\sum_{k=1}^{K} Z_{ik}\boldsymbol{\phi}_{km}' B(t_{il})\right)^2,$$

then we have that

$$\chi_{im}|\boldsymbol{\zeta}_{-\chi_{im}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N \sim \mathcal{N}(\mathbf{W}_{im}\mathbf{w}_{im}, \mathbf{W}_{im}).$$

In our model, we relax the assumption that the $\boldsymbol{\Phi}$ parameters are orthogonal. Even though we relaxed the assumption, we proved that many of the desirable properties still hold. However, if users do not want to relax this assumption, Kowal et al. [2017] describes a framework that allows us to sample when orthogonality constraints are imposed. In our model, orthogonality is defined by the inner product in equation 3.6. Therefore, for p such that $1 \leq p \leq KP$, we must have that

$$\langle \boldsymbol{\Phi}_p, \boldsymbol{\Phi}_j \rangle_{\boldsymbol{\mathcal{H}}} = 0 \;\; \forall j \neq p.$$

By rearranging terms, we can see that we have

$$\begin{aligned}\langle \boldsymbol{\Phi}_p, \boldsymbol{\Phi}_j \rangle_{\boldsymbol{\mathcal{H}}} &= \sum_{k=1}^{K} \int_{\mathcal{T}} \boldsymbol{\phi}'_{kp} \mathbf{B}(t) \boldsymbol{\phi}'_{kj} \mathbf{B}(t) \mathrm{d}t \\ &= \sum_{k \neq i} \int_{\mathcal{T}} \boldsymbol{\phi}'_{kp} \mathbf{B}(t) \boldsymbol{\phi}'_{kj} \mathbf{B}(t) \mathrm{d}t + \boldsymbol{\phi}'_{ip} \int_{\mathcal{T}} \mathbf{B}(t) \mathbf{B}(t)' \mathrm{d}t \boldsymbol{\phi}_{ij},\end{aligned}$$

where $\int_{\mathcal{T}} \mathbf{B}(t)\mathbf{B}(t)'\mathrm{d}t$ is the element-wise integration of the $P \times P$ matrix. Letting

$$\mathbf{L}_{-ip} = \begin{bmatrix} \int_{\mathcal{T}} \mathbf{B}(t)\mathbf{B}(t)'\mathrm{d}t\boldsymbol{\phi}_{i1} \\ \vdots \\ \int_{\mathcal{T}} \mathbf{B}(t)\mathbf{B}(t)'\mathrm{d}t\boldsymbol{\phi}_{i(p-1)} \\ \int_{\mathcal{T}} \mathbf{B}(t)\mathbf{B}(t)'\mathrm{d}t\boldsymbol{\phi}_{i(p+1)} \\ \vdots \\ \int_{\mathcal{T}} \mathbf{B}(t)\mathbf{B}(t)'\mathrm{d}t\boldsymbol{\phi}_{i(KP)} \end{bmatrix} \quad \text{and} \quad \mathbf{c}_{-ip} = \begin{bmatrix} \sum_{k \neq i} \int_{\mathcal{T}} \boldsymbol{\phi}'_{kp} \mathbf{B}(t) \boldsymbol{\phi}'_{k1} \mathbf{B}(t) \mathrm{d}t \\ \vdots \\ \sum_{k \neq i} \int_{\mathcal{T}} \boldsymbol{\phi}'_{kp} \mathbf{B}(t) \boldsymbol{\phi}'_{k(p-1)} \mathbf{B}(t) \mathrm{d}t \\ \sum_{k \neq i} \int_{\mathcal{T}} \boldsymbol{\phi}'_{kp} \mathbf{B}(t) \boldsymbol{\phi}'_{k(p+1)} \mathbf{B}(t) \mathrm{d}t \\ \vdots \\ \sum_{k \neq i} \int_{\mathcal{T}} \boldsymbol{\phi}'_{kp} \mathbf{B}(t) \boldsymbol{\phi}'_{k(KP)} \mathbf{B}(t) \mathrm{d}t \end{bmatrix},$$

we can write our orthogonality constraint for $\boldsymbol{\phi}_{ip}$ given the other $\boldsymbol{\phi}$ parameters as

$$\boldsymbol{\phi}'_{ip} \mathbf{L}_{-ip} = -\mathbf{c}_{-ip}.$$

Thus using the results in Kowal et al. [2017], we have that $\boldsymbol{\phi}_{ip} \sim \mathcal{N}(\tilde{\mathbf{M}}_{ip}\mathbf{m}_{ip}, \tilde{\mathbf{M}}_{ip})$, where

$$\tilde{\mathbf{M}}_{ip} = \mathbf{M}_{ip} - \mathbf{M}_{ip}\mathbf{L}_{-ip} \left(\mathbf{L}'_{-ip}\mathbf{M}_{ip}\mathbf{L}_{-ip}\right)^{-1} \left(\mathbf{L}'_{-ip}\mathbf{M}_{ip} + \mathbf{c}_{-ip}\right).$$

Like in Kowal et al. [2017], $\mathbf{M}_{ip}$ and $\mathbf{m}_{ip}$ are such that when we relax the orthogonal constraints, we have $\boldsymbol{\phi}_{ip} \sim \mathcal{N}(\mathbf{M}_{ip}\mathbf{m}_{ip}, \mathbf{M}_{ip})$ (defined in Section B.3.1). Thus one can use the modified Gibbs update to ensure orthogonality. However, by using this alternative update, the mixing of the Markov chain will likely suffer.

B.3.2 Multiple Start Algorithm

One of the main computational challenges that we encounter in this model is the multi-modal posterior distribution. Often times, the MCMC chain can get stuck in a mode, and it can have trouble moving through areas of low posterior density. One way to traverse through areas of low posterior density is to use tempered transitions. However, tempered transitions are computationally intensive and the hyperparameters can be somewhat difficult to tune. Thus, one of the best ways to converge to the correct mode is to have a good starting point. The Multiple Start Algorithm, found in algorithm 3, is a way to pick an optimal starting point. To get optimal performance our of this algorithm, we recommend that the initial data is standardized before running this algorithm.

The function calls two other functions, `BFPMM_Nu_Z(Y, time, K, n_MCMC1, ...)` and `BFPMM_Theta(P, Y, time, K, n_MCMC2, ...)`. The first function, `BFPMM_Nu_Z(Y, time, K, n_MCMC1, ...)`, starts with random parameter values for $\boldsymbol{\nu}$, $\mathbf{Z}$, σ^2, and other hyperparameters relating to these parameters. We then run an MCMC chain with the values of the χ and $\boldsymbol{\phi}$ variables fixed as 0 (or as the matrix $\mathbf{O}$). The function returns the mean likelihood for the last 20% of the MCMC chain as well as the entire MCMC chain. The variable `n_MCMC1` is assumed to be picked such that the chain converges in the first 80% of the MCMC iterations. Since the starting points are random, the MCMC chains are likely to explore different modes. Once we have a good initial starting point, we estimate the χ, $\boldsymbol{\phi}$, and other parameters that have not already been estimated using the function `BFPMM_Theta(P, Y, time, K, n_MCMC2, ...)`. In this function, we run an MCMC chain while fixing the values of $\boldsymbol{\nu}$ and $\mathbf{Z}$ to their optimal values found previously. We will use the outputs of algorithm 3 as a starting point for our final MCMC chain.

Algorithm 3 Multiple Start Algorithm

Require: n_try1, n_try2, Y, time, K, n_MCMC1, n_MCMC2, ...

$P \leftarrow$ `BFPMM_Nu_Z(Y, time, K, n_MCMC1, ...)` ▷ Returns the likelihood and estimates for $\boldsymbol{\nu}$ and $\mathbf{Z}$

max_likelihood $\leftarrow P$["likelihood"]

$i \leftarrow 1$

while $i \leq$ n_try1 **do**

  $P_i \leftarrow$ `BFPMM_Nu_Z(Y, time, K, n_MCMC1, ...)`

  if max_likelihood $< P_i$["likelihood"] **then**

    max_likelihood $\leftarrow P$["likelihood"]

    $P \leftarrow P_i$

  end if

  $i \leftarrow i + 1$

end while

$\theta \leftarrow$ `BFPMM_Theta(P, Y, time, K, n_MCMC2, ...)` ▷ Returns estimates for the rest of the parameters

max_likelihood $\leftarrow \theta$["likelihood"]

$i \leftarrow 1$

while $i \leq$ n_try2 **do**

  $\theta_i \leftarrow$ `BFPMM_Theta(P, Y, time, K, n_MCMC2, ...)`

  if max_likelihood $< \theta_i$["likelihood"] **then**

    max_likelihood $\leftarrow \theta_i$["likelihood"]

    $\theta \leftarrow \theta_i$

  end if

  $i \leftarrow i + 1$

end while

return (θ, P) ▷ Returns estimates for all model parameters

B.3.3 Tempered Transitions

Tempered transitions are used to help traverse areas of low posterior probability density when running MCMC chains. In problems that have multi-modal posterior distributions, traditional methods often have difficulty moving from one mode to another, which can cause the chain to not explore the entire state-space and therefore not converge to the true posterior distribution. Thus by using tempered transitions, we are potentially able to traverse the state-space to explore multiple modes. In simulations, we found that the tuning parameters can be difficult to tune to get acceptable acceptance probabilities, however in this section we will outline a way to use tempered transitions with our model.

We will be following the works of Behrens et al. [2012] and Pritchard et al. [2000] and only temper the likelihood. The target distribution that we want to temper is usually assumed to be written as

$$p(x) \propto \pi(x) exp\left(-\beta_h h(x)\right),$$

where β_h controls how much the distribution is tempered. We will assume $1 = \beta_0 < \cdots < \beta_h < \cdots < \beta_{N_t}$. The hyperparameters N_t and β_{N_t} are user specified, and will depend on the complexity of the model. For more complex models, we will most likely need a larger N_t. We will also assume that the parameters β_h follow a geometric scheme. We can rewrite our likelihood to fit the above form:

$$\begin{aligned} p_h(y_i(t)|\boldsymbol{\Theta}) &\propto exp\left\{-\beta_h\left(\frac{1}{2}log(\sigma^2) + \frac{1}{2\sigma^2}\left(y_i(t) - \sum_{k=1}^{K} Z_{ik}\left(\boldsymbol{\nu}_k' B(t) + \sum_{n=1}^{M}\chi_{in}\boldsymbol{\phi}_{kn}' B(t)\right)\right)^2\right)\right\} \\ &= \left(\sigma^2\right)^{-\beta_h/2} exp\left\{-\frac{\beta_h}{2\sigma^2}\left(y_i(t) - \sum_{k=1}^{K} Z_{ik}\left(\boldsymbol{\nu}_k' B(t) + \sum_{n=1}^{M}\chi_{in}\boldsymbol{\phi}_{kn}' B(t)\right)\right)^2\right\}. \end{aligned}$$

Let $\boldsymbol{\Theta}_h$ be the set of parameters generated from the model using the tempered likelihood associated with β_h. The tempered transition algorithm can be summarized by the following steps:

1. Start with initial state $\boldsymbol{\Theta}_0$.
2. Transition from $\boldsymbol{\Theta}_0$ to $\boldsymbol{\Theta}_1$ using the tempered likelihood associated with β_1.
3. Continue in this manner until we transition from $\boldsymbol{\Theta}_{N_t-1}$ to $\boldsymbol{\Theta}_{N_t}$ using the tempered likelihood associated with β_{N_t}.
4. Transition from $\boldsymbol{\Theta}_{N_t}$ to $\boldsymbol{\Theta}_{N_t+1}$ using the tempered likelihood associated with β_{N_t}.
5. Continue in this manner until we transition from $\boldsymbol{\Theta}_{2N_t-1}$ to $\boldsymbol{\Theta}_{2N_t}$ using β_1.
6. Accept transition from $\boldsymbol{\Theta}_0$ to $\boldsymbol{\Theta}_{2N_t}$ with probability

$$\min\left\{1, \prod_{h=0}^{N_t-1} \frac{\prod_{i=1}^{N}\prod_{l=1}^{n_i} p_{h+1}(y_i(t_{il})|\boldsymbol{\Theta}_h)}{\prod_{i=1}^{N}\prod_{l=1}^{n_i} p_h(y_i(t_{il})|\boldsymbol{\Theta}_h)} \prod_{h=N_t+1}^{2N_t} \frac{\prod_{i=1}^{N}\prod_{l=1}^{n_i} p_h(y_i(t_{il})|\boldsymbol{\Theta}_h)}{\prod_{i=1}^{N}\prod_{l=1}^{n_i} p_{h+1}(y_i(t_{il})|\boldsymbol{\Theta}_h)}\right\}.$$

Since we only temper the likelihood, we can use many of updates in Section B.3.1. However, we will have to modify how we update the $\boldsymbol{\nu}$, $\boldsymbol{\phi}$, σ^2, $\mathbf{Z}$, and χ parameters. By letting

$$\begin{aligned}(\mathbf{m}_{jm})_h =& \frac{\beta_h}{(\sigma^2)_h}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\Bigg(B(t_{il})(\chi_{im})_h\Bigg(y_i(t_{il})(Z_{ij})_h - (Z_{ij})_h^2(\boldsymbol{\nu}_j)_h'B(t_{il}) \\ &- (Z_{ij})_h^2\sum_{n\neq m}\left[(\chi_{in})_h(\boldsymbol{\phi}_{jn})_h'B(t_{il})\right] \\ &- \sum_{k\neq j}(Z_{ij})_h(Z_{ik})_h\left[(\boldsymbol{\nu}_k)_h'B(t_{il}) + \sum_{n=1}^{M}(\chi_{in})_h(\boldsymbol{\phi}_{kn})_h'B(t_{il})\right]\Bigg)\Bigg),\end{aligned}$$

and

$$(\mathbf{M}_{jm})_h^{-1} = \frac{\beta_h}{(\sigma^2)_h}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\left((Z_{ij})_h^2(\chi_{im})_h^2B(t_{il})B'(t_{il})\right) + (\mathbf{D}_{jm})_h^{-1},$$

we have that

$$(\boldsymbol{\phi}_{jm})_h|\boldsymbol{\Theta}_{-(\boldsymbol{\phi}_{jm})_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N \sim \mathcal{N}((\mathbf{M}_{jm})_h(\mathbf{m}_{jm})_h, (\mathbf{M}_{jm})_h).$$

The posterior distribution for $(\mathbf{z}_i)_h$ is still not a commonly known distribution, so we will

still have to use the Metropolis-Hastings algorithm. The new posterior distribution when using tempered transitions changes into

$$p((\mathbf{z}_i)_h|(\boldsymbol{\Theta}_{-(\mathbf{z}_i)_h})_h, \mathbf{Y}_1, \ldots, \mathbf{Y}_N) \propto \prod_{k=1}^{K} (Z_{ik})_h^{(\alpha_3)_h(\pi_k)_h - 1} \times \prod_{l=1}^{n_i} exp\left\{ -\frac{\beta_h}{2(\sigma^2)_h} \left(y_i(t_{il}) - \sum_{k=1}^{K} (Z_{ik})_h \left((\boldsymbol{\nu}_k)_h' B(t_{il}) + \sum_{n=1}^{M} (\chi_{in})_h (\boldsymbol{\phi}_{kn})_h' B(t_{il}) \right) \right)^2 \right\}.$$

We will use $Q((\mathbf{z}_i)_h'|(\mathbf{z}_i)_{h-1}) = Dir(a_{\mathbf{z}}(\mathbf{z}_i)_{h-1})$ for some large $a_{\mathbf{z}} \in \mathbb{R}^+$ as the proposal distribution. Thus the probability of accepting a proposed step is

$$A((\mathbf{z}_i)_h', (\mathbf{z}_i)_{h-1}) = \min\left\{ 1, \frac{P\left((\mathbf{z}_i)_h'|(\boldsymbol{\Theta}_{-(\mathbf{z}_i)_h'})_h, \mathbf{Y}_1, \ldots, \mathbf{Y}_N\right)}{P\left((\mathbf{z}_i)_{h-1}|\boldsymbol{\Theta}_{-(\mathbf{z}_i)_{h-1}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N\right)} \frac{Q\left((\mathbf{z}_i)_{h-1}|(\mathbf{z}_i)_h'\right)}{Q\left((\mathbf{z}_i)_h'|(\mathbf{z}_i)_{h-1}\right)} \right\}.$$

Letting

$$(\mathbf{B}_j)_h = \left((\tau_j)_h \mathbf{P} + \frac{\beta_h}{(\sigma^2)_h} \sum_{i=1}^{N} \sum_{l=1}^{n_i} (Z_{ij})_h^2 B(t_{il}) B'(t_{il}) \right)^{-1}$$

and

$$(\mathbf{b}_j)_h = \frac{\beta_h}{(\sigma^2)_h} \sum_{i=1}^{N} \sum_{l=1}^{n_i} (Z_{ij})_h B(t_{il}) \left(y_i(t_{il}) - \left(\sum_{k \neq j} (Z_{ik})_h (\boldsymbol{\nu}_k')_h B(t_{il}) \right) - \left(\sum_{k=1}^{K} \sum_{n=1}^{M} (Z_{ik})_h (\chi_{in})_h (\boldsymbol{\phi}_{kn})_h' B(t_{il}) \right) \right),$$

we have that

$$(\boldsymbol{\nu}_j)_h|\boldsymbol{\Theta}_{-(\boldsymbol{\nu}_j)_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N \sim \mathcal{N}((\mathbf{B}_j)_h(\mathbf{b}_j)_h, (\mathbf{B}_j)_h).$$

The parameter $(\sigma^2)_h$ can be updated by using a Gibbs update. If we let

$$(\beta_\sigma)_h = \frac{\beta_h}{2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} \left(y_i(t_{il}) - \sum_{k=1}^{K} (Z_{ik})_h \left((\boldsymbol{\nu}_k)_h' B(t_{il}) + \sum_{n=1}^{M} (\chi_{in})_h (\boldsymbol{\phi}_{kn})_h' B(t_{il}) \right) \right)^2,$$

then we have

$$(\sigma^2)_h|\boldsymbol{\Theta}_{-(\sigma^2)_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N \sim IG\left(\alpha_0 + \frac{\beta_h \sum_{i=1}^N n_i}{2}, \beta_0 + (\beta_\sigma)_h\right).$$

Lastly, letting let

$$\begin{aligned}(\mathbf{w}_{im})_h =& \frac{\beta_h}{(\sigma^2)_h}\left(\sum_{l=1}^{n_i}\left(\sum_{k=1}^{K}(Z_{ik})_h(\boldsymbol{\phi}_{km})_h'B(t_{il})\right)\left(y_i(t_{il})\right.\right.\\ &\left.\left.-\sum_{k=1}^{K}(Z_{ik})_h\left((\boldsymbol{\nu}_k)_h'B(t_{il}) + \sum_{n\neq m}(\chi_{in})_h(\boldsymbol{\phi}_{kn})_h'B(t_{il})\right)\right)\right)\end{aligned}$$

and

$$(\mathbf{W}_{im}^{-1})_h = 1 + \frac{\beta_h}{(\sigma^2)_h}\sum_{l=1}^{n_i}\left(\sum_{k=1}^{K}(Z_{ik})_h(\boldsymbol{\phi}_{km})_h'B(t_{il})\right)^2,$$

then we have that

$$(\chi_{im})_h|\boldsymbol{\zeta}_{-(\chi_{im})_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N \sim \mathcal{N}((\mathbf{W}_{im})_h(\mathbf{w}_{im})_h, (\mathbf{W}_{im})_h).$$

One of the biggest drawbacks to using tempered transition is the computational cost of just one iteration. It is common for N_t to be in the thousands, especially when dealing with a complex model, so each tempered transition will take thousands of times longer than an untempered transition. Thus we recommend using a mixture of tempered transition and untempered transitions to speed up computation. From proposition 1 in Roberts and Rosenthal [2007], we know that an independent mixture of tempered transitions and untempered transitions will still preserve our stationary distribution of our Markov chain.

B.3.4 Membership Rescale Algorithm

As discussed in Section 3.1.2, our model is unidentifiable. To help with interpretability, we will apply a linear transformation to the $\mathbf{Z}$ matrix to ensure that we use as much of the unit

simplex as possible. In the case when $K = 2$, this will correspond to rescaling the observations such that at least one observation is entirely in each feature. This specific assumption that one observation belongs entirely in each feature is known as the *seperability* condition [Papadimitriou et al., 1998, McSherry, 2001, Azar et al., 2001, Chen et al., 2022]. Thus in order to ensure identifiability, algorithm 4 can be used when we only have two features. In the case of a two feature model, the seperability condition is a very weak assumption, however as we move to models with more features, it can be a relatively strong assumption. Weaker geometric assumptions such as the *sufficiently scattered* condition [Huang et al., 2016, Jang and Hero, 2019, Chen et al., 2022]. In cases when we have more than two features, we can use the idea of the sufficiently scattered condition to help with identifiability. From Chen et al. [2022], we have that an allocation matrix $\mathbf{Z}$ is sufficiently scattered if:

1. $\text{cone}(\mathbf{Z}')^* \subseteq \mathcal{K}$
2. $\text{cone}(\mathbf{Z}')^* \cap bd\mathcal{K} \subseteq \{\lambda \mathbf{e}_f, f = 1, \ldots, k, \lambda \geq 0\}$

where $\mathcal{K} := \left\{\mathbf{x} \in \mathbb{R}^K | \|\mathbf{x}\|_2 \leq \mathbf{x}'\mathbf{1}_K\right\}$, $bd\mathcal{K} := \left\{\mathbf{x} \in \mathbb{R}^K | \|\mathbf{x}\|_2 = \mathbf{x}'\mathbf{1}_K\right\}$,
$\text{cone}(\mathbf{Z}')^* := \left\{\mathbf{x} \in \mathbb{R}^K | \mathbf{x}\mathbf{Z}' \geq 0\right\}$, and $\mathbf{e}_f$ is a vector with the i^{th} element equal to 1 and zero elsewhere. The first condition can be interpreted as the allocation parameters should from a convex polytope that contains the dual cone $\mathcal{K}^*$. Thus we have that

$$\text{Conv}(\mathbf{Z}') \subseteq \mathcal{K}^*,$$

where $\mathcal{K}^* := \left\{\mathbf{x} \in \mathbb{R}^K | \mathbf{x}'\mathbf{1}_K \geq \sqrt{k-1}\|\mathbf{x}\|_2\right\}$ and $\text{Conv}(\mathbf{Z}') := \{\mathbf{x} \in \mathbb{R}^K | \mathbf{x} = \mathbf{Z}'\lambda, \lambda \in \Delta^N\}$, where Δ^k denotes the N-dimensional simplex. Ensuring that these two conditions are met is not trivial in our setting. Therefore, we will focus on trying to promote allocation structures such that the first condition is satisfied. Similarly to the case of two functional features, we aim to find a linear transformation such that the convex polytope of our transformed allocation parameters covers the most area. Thus letting $\mathbf{T} \in \mathbb{R}^K \times \mathbb{R}^K$ be our transformation

matrix, we aim to solve the following optimization problem:

$$\max_{\mathbf{T}} \ |\text{Conv}(\mathbf{TZ}')|$$
$$s.t. \ \ \mathbf{z}_i\mathbf{T} \in \mathcal{C} \ \ \forall \ i,$$

where $|\text{Conv}(\mathbf{TZ}')|$ denotes the volume of the convex polytope constructed by the allocation parameters. While this will not ensure that the first condition is met, it will promote an allocation structure that uses the entire simplex. While this algorithm does not ensure identifiability in the case when we have more than 2 functional features, it does make inference on the model more interpretable. Once the memberships are rescaled, we can conduct posterior inference on the mean function and covariance functions using the rescaled parameters.

Algorithm 4 Membership Rescale Algorithm

Require: $\mathbf{Z}, \boldsymbol{\nu}, \boldsymbol{\Phi}, M$
  $T \leftarrow \texttt{matrix}(0, 2, 2)$ ▷ Initialize inverse transformation matrix (2 x 2)
  $i \leftarrow 1$
  while $i \leq 2$ **do**
    max_ind $\leftarrow \texttt{max\_ind}(\mathbf{Z}[, i])$ ▷ Find index of max entry in i^{th} column
    $T[i,] \leftarrow (\mathbf{Z}[\text{max\_ind},])$
    $i \leftarrow i + 1$
  end while
  $\mathbf{Z}\_t \leftarrow \mathbf{Z} * \texttt{inv}(T)$ ▷ Transform the $\mathbf{Z}$ parameters
  $\boldsymbol{\nu}\_t \leftarrow T * \boldsymbol{\nu}$ ▷ Transform the $\boldsymbol{\nu}$ parameters
  $i \leftarrow 1$
  while $i \leq M$ **do**
    $\boldsymbol{\Phi}\_t[, , i] \leftarrow T * \boldsymbol{\Phi}[, , i]$ ▷ Transform the $\boldsymbol{\Phi}$ parameters
    $i \leftarrow i + 1$
  end while
  return $(\mathbf{Z}\_t, \boldsymbol{\nu}\_t, \boldsymbol{\Phi}\_t)$

B.4 Simulation-Based Posterior Inference

Statistical inference is based on Markov chain Monte Carlo samples from the posterior distribution. To achieve this we used the Metropolis-within-Gibbs algorithm. By introducing the latent χ_{im} variables, many of the posterior distributions related to the covariance process were easily sampled through Gibbs updates. More details on the sampling scheme can be found in section C.1 of the web-based supporting materials. The sampling scheme is relatively simple, and was implemented using the RcppArmadillo package created by Eddelbuettel and Sanderson [2014] to speed up computation.

While the naïve sampling scheme is relatively simple, ensuring good exploration of the posterior target can be challenging due to the potentially multimodal nature of the posterior distribution. Specifically, some sensitivity of results to the starting values of the chain can be observed for some data. Section B.3.2 outlines an algorithm for the selection of informed starting values. Furthermore, to mitigate sensitivity to chain initialization, we also implemented a tempered transition scheme, which improves the mixing of the Markov chain by allowing for transitions between modal configuration of the target. Implementation details for the proposed tempered transition scheme are reported in section B.3.3.

Given Monte Carlo samples from the posterior distribution of all parameters of interest, posterior inference is implemented descriptively; either directly on the Monte Carlo samples for parameters of interest, such as the mixed membership proportions $\mathbf{z}_i$, or indirectly through the evaluation of relevant functions of the parameters of interest, e.g. the mean and cross-covariance functions of the latent features.

In this setting, to calculate the simultaneous credible intervals, we will use the simultaneous credible intervals proposed by Crainiceanu et al. [2007]. Let $\mathbf{g}_n$ be simulated realizations using the MCMC samples of the function of interest, and let $\{t_1, \ldots, t_R\}$ be a fine grid of time points in $\mathcal{T}$. Let $\mathbb{E}(\mathbf{g}(t_i))$ be the expected value of the function evaluated at time point $t_i \in \mathcal{T}$, and $\text{SD}(\mathbf{g}(t_i))$ be the standard deviation of the function evaluated at time point

$t_i \in \mathcal{T}$. Let M_α be the $(1-\alpha)$ quantile of $\max_{1\leq i\leq R}\left|\frac{\mathbf{g}_n(t_i)-\mathbb{E}(\mathbf{g}(t_i))}{\mathrm{SD}(\mathbf{g}(t_i))}\right|$, for $1 \leq n \leq N_{MC}$, where N_{MC} are the number of MCMC samples of the converged MCMC chain. Thus the simultaneous credible intervals can be constructed as

$$\mathcal{I}(t_i) = \left[\mathbb{E}(\mathbf{g}(t_i)) - M_\alpha \mathrm{SD}(\mathbf{g}(t_i)), \mathbb{E}(\mathbf{g}(t_i)) + M_\alpha \mathrm{SD}(\mathbf{g}(t_i))\right].$$

Thus we estimate simultaneous credible intervals for all mean functions, $\mu^{(k)}$, and similarly generalize this procedure to define simultaneous credible intervals for the cross-covariance functions, $C^{(k,k')}$. Figure 3.5, illustrates the difference between a simultaneous credible interval and a pointwise credible interval for one of the EEG case studies.

APPENDIX C

Appendix: Covariate Adjusted Mixed Membership Models

C.1 Proof of Lemma 2.1

We will start by defining identifiability and defining some of the notation used in this section. Let $\boldsymbol{\omega} = \{\boldsymbol{\nu}_1, \boldsymbol{\nu}_2, \boldsymbol{\eta}_1, \boldsymbol{\eta}_2, Z_{11}, \ldots, Z_{1N}, \boldsymbol{\Sigma}_{11}, \boldsymbol{\Sigma}_{12}, \boldsymbol{\Sigma}_{22}, \sigma^2\}$, where $\boldsymbol{\Sigma}_{kk'} = \sum_{m=1}^{M} (\boldsymbol{\phi}_{km}\boldsymbol{\phi}'_{k'm})$. We will say that the parameters $\boldsymbol{\omega}$ are unidentifiable if there exists at least one $\boldsymbol{\omega}^* \neq \boldsymbol{\omega}$ such that $\mathcal{L}(\mathbf{Y}_i(\mathbf{t}_i) \mid \boldsymbol{\omega}, \mathbf{x}_i) = \mathcal{L}(\mathbf{Y}_i(\mathbf{t}_i) \mid \boldsymbol{\omega}^*, \mathbf{x}_i)$ for all sets of observations $\{\mathbf{Y}_i(\mathbf{t}_i)\}_{i=1}^N$, that follow assumptions (1)-(3). Otherwise, the parameters $\boldsymbol{\omega}$ are called identifiable. In this case, $\mathcal{L}(\mathbf{Y}_i(\mathbf{t}_i) \mid \boldsymbol{\omega}, \mathbf{x}_i)$ is the likelihood specified in equation 12 in the main text.

From equation 12 in the main text, we have that

$$\mathcal{L}\left(\mathbf{Y}_i(\mathbf{t}_i) \mid \boldsymbol{\omega}, \mathbf{x}_i\right) \propto \exp\left\{-\tfrac{1}{2}\left(\mathbf{Y}_i(\mathbf{t}_i) - \mu_i(\mathbf{x}_i, \mathbf{t}_i)\right)'\left(\mathbf{V}(\mathbf{t}_i, \mathbf{z}_i) + \sigma^2\mathbf{I}_{n_i}\right)^{-1}\left(\mathbf{Y}_i(\mathbf{t}_i) - \mu_i(\mathbf{x}_i, \mathbf{t}_i)\right)\right\}, \quad \text{(C.1)}$$

where

$$\mu_i(\mathbf{x}_i, \mathbf{t}_i) = \sum_{k=1}^{2} Z_{ik}\mathbf{S}'(\mathbf{t}_i)\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k\mathbf{x}'_i\right)$$

and

$$\mathbf{V}(\mathbf{t}_i, \mathbf{z}_i) = \sum_{k=1}^{2}\sum_{k'=1}^{2} Z_{ik}Z_{ik'}\left\{\mathbf{S}'(\mathbf{t}_i)\sum_{m=1}^{M}\left(\boldsymbol{\phi}_{km}\boldsymbol{\phi}'_{k'm}\right)\mathbf{S}(\mathbf{t}_i)\right\}.$$

Assume that $\mathcal{L}\left(\mathbf{Y}_i(\mathbf{t}_i) \mid \boldsymbol{\omega}, \mathbf{x}_i\right) = \mathcal{L}\left(\mathbf{Y}_i(\mathbf{t}_i) \mid \boldsymbol{\omega}^*, \mathbf{x}_i\right)$ for all sets of observations $\{\mathbf{Y}_i(\mathbf{t}_i)\}_{i=1}^N$ that follow assumptions (1)-(3) . Thus we would like to prove that $\boldsymbol{\omega}^* = \boldsymbol{\omega}$ must necessarily

be true. Since $\mathcal{L}\left(\mathbf{Y}_i(\mathbf{t}_i) \mid \boldsymbol{\omega}, \mathbf{x}_i\right)$ is written as a quadratic form in $\mathbf{Y}_i(\mathbf{t}_i)$ and $(\mathbf{V}(\mathbf{t}_i, \mathbf{z}_i) + \sigma^2\mathbf{I}_{n_i})$ is full rank, we have that the following must necessarily be true:

1. $\mu_i^*(\mathbf{x}_i, \mathbf{t}_i) = \mu_i(\mathbf{x}_i, \mathbf{t}_i)$,
2. $\mathbf{V}^*(\mathbf{t}_i, \mathbf{z}_i^*) + (\sigma^2)^*\mathbf{I}_{n_i} = \mathbf{V}(\mathbf{t}_i, \mathbf{z}_i) + \sigma^2\mathbf{I}_{n_i}$.

By (1), we have that

$$\sum_{k=1}^{2} Z_{ik}\mathbf{S}'(\mathbf{t}_i)\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k\mathbf{x}_i'\right) = \sum_{k=1}^{2} Z_{ik}^*\mathbf{S}'(\mathbf{t}_i)\left(\boldsymbol{\nu}_k^* + \boldsymbol{\eta}_k^*\mathbf{x}_i'\right) \quad (i = 1, \ldots, N).$$

Letting $\boldsymbol{\mu}_k = [\boldsymbol{\nu}_k \ \boldsymbol{\eta}_k] \in \mathbb{R}^{P\times(R+1)}$ and $\tilde{\mathbf{x}}_i = [1 \ \mathbf{x}_i]$ ($\tilde{\mathbf{X}} \in \mathbb{R}^{N\times(R+1)}$ is the design matrix with the i^{th} row as $\tilde{\mathbf{x}}$), we have that

$$\begin{aligned} &\sum_{k=1}^{2} Z_{ik}\mathbf{S}'(\mathbf{t}_i)\boldsymbol{\mu}_k\tilde{\mathbf{x}}_i' = \sum_{k=1}^{2} Z_{ik}^*\mathbf{S}'(\mathbf{t}_i)\boldsymbol{\mu}_k^*\tilde{\mathbf{x}}_i' \quad (i = 1, \ldots, N) \\ \iff &\sum_{k=1}^{2} Z_{ik}\boldsymbol{\mu}_k\tilde{\mathbf{x}}_i' = \sum_{k=1}^{2} Z_{ik}^*\boldsymbol{\mu}_k^*\tilde{\mathbf{x}}_i' \quad (i = 1, \ldots, N), \end{aligned}$$

since $n_i \geq P$ by assumption (3). Since $Z_{i1} = (1 - Z_{i1})$ in a two feature mixed membership model, we have

$$\left(Z_{i1}\boldsymbol{\mu}_1 + (1 - Z_{i1})\boldsymbol{\mu}_2\right)\tilde{\mathbf{x}}_i' = \left(Z_{i1}^*\boldsymbol{\mu}_1^* + (1 - Z_{i1}^*)\boldsymbol{\mu}_2^*\right)\tilde{\mathbf{x}}_i' \quad (i = 1, \ldots, N). \tag{C.2}$$

Since $\tilde{\mathbf{X}}$ is full column rank from assumption (1), we know that the solution to the system of equations in equation C.2 takes the following form

$$\begin{aligned} Z_{i1}^* &= aZ_{i1} + b(1 - Z_{i1}), \\ \boldsymbol{\mu}_1^* &= \left(\frac{1-b}{a-b}\right)\boldsymbol{\mu}_1 - \left(\frac{1-a}{a-b}\right)\boldsymbol{\mu}_2, \\ \boldsymbol{\mu}_2^* &= \left(\frac{a}{a-b}\right)\boldsymbol{\mu}_2 - \left(\frac{b}{a-b}\right)\boldsymbol{\mu}_1, \end{aligned}$$

where $a, b \in \mathbb{R}$ such that $a, b > 0$, $a + b = 1$, and $(a, b) \neq (0.5, 0.5)$. From assumption (2), the only way that there can exist $\tilde{i}_1^*, \tilde{i}_2^*$ such that $Z_{\tilde{i}_1^* 1} = 1$ and $Z_{\tilde{i}_2^* 2} = 1$ is that $(a, b) = (1, 0)$ or $(a, b) = (0, 1)$. Since the solution $(a, b) = (0, 1)$ is simply a permutation of the labels (i.e. *label switching*), we can see that $Z_{il} = Z_{il}^*$, $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_1^*$, and $\boldsymbol{\mu}_2 = \boldsymbol{\mu}_2^*$ up to a permutation of the labels. It is important to note that if assumption (2) did not hold, and $\tilde{\mathbf{X}}$ is not full column rank, we could add any vector in the nullspace of $\tilde{\mathbf{X}}$ to each row of $\boldsymbol{\mu}_1^*$ or $\boldsymbol{\mu}_2^*$ and equation C.2 would still hold. Therefore, assuming assumptions (1) - (3) hold, we have that $Z_{ik} = Z_{ik}^*$, $\boldsymbol{\nu}_k = \boldsymbol{\nu}_k^*$, and $\boldsymbol{\eta}_k = \boldsymbol{\eta}_k^*$ up to the permutation of the labels, for $k = 1, 2$ and $i = 1, \ldots, N$.

From (2), we have that

$$\begin{aligned} &\mathbf{V}^*(\mathbf{t}_i, \mathbf{z}_i^*) + (\sigma^2)^* \mathbf{I}_{n_i} = \mathbf{V}(\mathbf{t}_i, \mathbf{z}_i) + \sigma^2 \mathbf{I}_{n_i} \\ \iff &\mathbf{V}^*(\mathbf{t}_i, \mathbf{z}_i^*) - \mathbf{V}(\mathbf{t}_i, \mathbf{z}_i) = ((\sigma^2)^* - \sigma^2)\mathbf{I}_{n_i}. \end{aligned}$$

Suppose that $((\sigma^2)^* - \sigma^2) \neq 0$, then we have that

$$\operatorname{rank}\left(\mathbf{V}^*(\mathbf{t}_i, \mathbf{z}_i^*) - \mathbf{V}(\mathbf{t}_i, \mathbf{z}_i)\right) = \operatorname{rank}\left(((\sigma^2)^* - \sigma^2)\mathbf{I}_{n_i}\right) > 4M,$$

by assumption (3) (there exists i such that $n_i > 4M$) . However, from the definition of $\mathbf{V}(\mathbf{t}_i, \mathbf{z}_i)$, we have that

$$\mathbf{V}(\mathbf{t}_i, \mathbf{z}_i) = \sum_{k=1}^{2} \sum_{k'=1}^{2} Z_{ik} Z_{ik'} \left\{ \mathbf{S}'(\mathbf{t}_i) \sum_{m=1}^{M} (\boldsymbol{\phi}_{km} \boldsymbol{\phi}'_{k'm}) \mathbf{S}(\mathbf{t}_i) \right\}.$$

Thus, we can see from the functional form of $\mathbf{V}(\mathbf{t}_i, \mathbf{z}_i)$, we can see that $\operatorname{rank}(\mathbf{V}(\mathbf{t}_i, \mathbf{z}_i)) \leq KM$, meaning that $\operatorname{rank}\left(\mathbf{V}^*(\mathbf{t}_i, \mathbf{z}_i^*) - \mathbf{V}(\mathbf{t}_i, \mathbf{z}_i)\right) \leq 2KM$, leading to a contradiction. Therefore we have that $(\sigma^2)^* = \sigma^2$ and $\mathbf{V}^*(\mathbf{t}_i, \mathbf{z}_i^*) = \mathbf{V}(\mathbf{t}_i, \mathbf{z}_i)$. From assumptions 2 (there are at least 2 points $\mathbf{z}_i$ in the interior of the simplex) and assumptions 3 ($n_i > P$), as well as the fact that

$Z^*_{ik} = Z_{ik}$ up to a permutation of the labels, we can see that

$$\sum_{m=1}^{M} \left((\boldsymbol{\phi}_{km})^* \left(\boldsymbol{\phi}'_{k'm} \right)^* \right) = \sum_{m=1}^{M} \left(\boldsymbol{\phi}_{km} \boldsymbol{\phi}'_{k'm} \right)$$

up to a permutation of the labels. Therefore, we have that the parameters $\boldsymbol{\nu}_k$, $\boldsymbol{\eta}_k$, Z_{ik}, $\sum_{m=1}^{M} \left(\boldsymbol{\phi}_{km} \boldsymbol{\phi}'_{k'm} \right)$, and σ^2 are identifiable up to a permutation of the labels given assumptions (1)-(3).

C.2 Computation

C.2.1 Posterior Distributions

In this subsection, we will specify the posterior distributions specifically for the functional covariate adjusted mixed membership model proposed in the main manuscript. We will first start with the $\boldsymbol{\phi}_{km}$ parameters, for $j = 1, \ldots, K$ and $m = 1, \ldots, M$. Let $\mathbf{D}_{\boldsymbol{\phi}_{jm}} = \tilde{\tau}^{-1}_{\boldsymbol{\phi}_{mj}} diag\left(\gamma^{-1}_{\boldsymbol{\phi}_{j1m}}, \ldots, \gamma^{-1}_{\boldsymbol{\phi}_{jPm}} \right)$. By letting

$$\begin{aligned} \mathbf{m}_{\boldsymbol{\phi}_{jm}} =& \frac{1}{\sigma^2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} \left(B(t_{il}) \chi_{im} \left(y_i(t_{il}) Z_{ij} - Z^2_{ij} \left(\boldsymbol{\nu}_j + \boldsymbol{\eta}_j \mathbf{x}'_i \right)' B(t_{il}) - Z^2_{ij} \sum_{n \neq m} \chi_{in} \boldsymbol{\phi}'_{jn} B(t_{il}) \right. \right. \\ & \left. \left. - \sum_{k \neq j} Z_{ij} Z_{ik} \left[\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k \mathbf{x}'_i \right)' B(t_{il}) + \sum_{n=1}^{M} \chi_{in} \boldsymbol{\phi}'_{kn} B(t_{il}) \right] \right) \right), \end{aligned}$$

and

$$\mathbf{M}^{-1}_{\boldsymbol{\phi}_{jm}} = \frac{1}{\sigma^2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} \left(Z^2_{ij} \chi^2_{im} B(t_{il}) B'(t_{il}) \right) + \mathbf{D}^{-1}_{\boldsymbol{\phi}_{jm}},$$

we have that

$$\boldsymbol{\phi}_{jm} | \boldsymbol{\Theta}_{-\boldsymbol{\phi}_{jm}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N}\left(\mathbf{M}_{\boldsymbol{\phi}_{jm}} \mathbf{m}_{\boldsymbol{\phi}_{jm}}, \mathbf{M}_{\boldsymbol{\phi}_{jm}} \right).$$

The posterior distribution of δ_{1k}, for $k = 1, \ldots, K$, is

$$\delta_{1k}|\boldsymbol{\Theta}_{-\delta_{1k}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \Gamma\left(a_{1k} + (PM/2), 1 + \frac{1}{2}\sum_{r=1}^{P}\gamma_{k,r,1}\phi_{k,r,1}^2 + \frac{1}{2}\sum_{m=2}^{M}\sum_{r=1}^{P}\gamma_{k,r,m}\phi_{k,r,m}^2\left(\prod_{j=2}^{m}\delta_{jk}\right)\right).$$

The posterior distribution for δ_{ik}, for $i = 2, \ldots, M$ and $k = 1, \ldots, K$, is

$$\delta_{ik}|\boldsymbol{\Theta}_{-\delta_{ik}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \Gamma\left(a_{2k} + (P(M-i+1)/2), 1 + \frac{1}{2}\sum_{m=i}^{M}\sum_{r=1}^{P}\gamma_{\boldsymbol{\xi}_{k,r,m}}\phi_{k,r,m}^2\left(\prod_{j=1;j\neq i}^{m}\delta_{jk}\right)\right).$$

The posterior distribution for a_{1k} $(k = 1, \ldots, K)$ is not a commonly known distribution, however we have that

$$P(a_{1k}|\boldsymbol{\Theta}_{-a_{1k}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}) \propto \frac{1}{\Gamma(a_{1k})}\delta_{1k}^{a_{1k}-1}a_{1k}^{\alpha_1-1}exp\left\{-a_{1k}\beta_1\right\}.$$

Since this is not a known kernel of a distribution, we will have to use Metropolis-Hastings algorithm. Consider the proposal distribution $Q(a'_{1k}|a_{1k}) = \mathcal{N}\left(a_{1k}, \epsilon_1\beta_1^{-1}, 0, +\infty\right)$ (Truncated Normal) for some small $\epsilon_1 > 0$. Thus the probability of accepting any step is

$$A(a'_{1k}, a_{1k}) = \min\left\{1, \frac{P\left(a'_{1k}|\boldsymbol{\Theta}_{-a'_{1k}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)}{P\left(a_{1k}|\boldsymbol{\Theta}_{-a_{1k}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)}\frac{Q\left(a_{1k}|a'_{1k}\right)}{Q\left(a'_{1k}|a_{1k}\right)}\right\}.$$

Similarly for a_{2k} $(k = 1, \ldots, K)$, we have

$$P(a_{2k}|\boldsymbol{\Theta}_{-a_{2k}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}) \propto \frac{1}{\Gamma(a_{2k})^{M-1}}\left(\prod_{i=2}^{M}\delta_{ik}^{a_{2k}-1}\right)a_{2k}^{\alpha_{2k}-1}exp\left\{-a_{2k}\beta_2\right\}.$$

We will use a similar proposal distribution, such that $Q(a'_{2k}|a_{2k}) = \mathcal{N}\left(a_{2k}, \epsilon_2\beta_2^{-1}, 0, +\infty\right)$ for

some small $\epsilon_2 > 0$. Thus the probability of accepting any step is

$$A(a'_{2k}, a_{2k}) = \min\left\{1, \frac{P\left(a'_{2k}|\boldsymbol{\Theta}_{-a'_{2k}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)}{P\left(a_{2k}|\boldsymbol{\Theta}_{-a_{2k}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)} \frac{Q\left(a_{2k}|a'_{2k}\right)}{Q\left(a'_{2k}|a_{2k}\right)}\right\}.$$

The posterior distribution for the $\mathbf{z}_i$ parameters are not a commonly known distribution, so we will use the Metropolis-Hastings algorithm. We know that

$$\begin{aligned} p(\mathbf{z}_i|\boldsymbol{\Theta}_{-\mathbf{z}_i}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}) \propto & \prod_{k=1}^{K} Z_{ik}^{\alpha_3\pi_k - 1} \\ & \times \prod_{l=1}^{n_i} exp\left\{-\frac{1}{2\sigma^2}\left(y_i(t_{il}) - \sum_{k=1}^{K} Z_{ik}\left(\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k \mathbf{x}'_i\right)' B(t_{il})\right.\right.\right. \\ & \left.\left.\left. + \sum_{m=1}^{M} \chi_{im} \boldsymbol{\phi}'_{km} B(t_{il})\right)\right)^2\right\}. \end{aligned}$$

We will use $Q(\mathbf{z}'_i|\mathbf{z}_i) = Dir(a_{\mathbf{z}}\mathbf{z}_i)$ for some large $a_{\mathbf{z}} \in \mathbb{R}^+$ as the proposal distribution. Thus the probability of accepting a proposed step is

$$A(\mathbf{z}'_i, \mathbf{z}_i) = \min\left\{1, \frac{P\left(\mathbf{z}'_i|\boldsymbol{\Theta}_{-\mathbf{z}_i}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)}{P\left(\mathbf{z}_i|\boldsymbol{\Theta}_{-\mathbf{z}_i}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)} \frac{Q\left(\mathbf{z}_i|\mathbf{z}'_i\right)}{Q\left(\mathbf{z}'_i|\mathbf{z}_i\right)}\right\}.$$

Similarly, a Gibbs update is not available for an update of the $\boldsymbol{\pi}$ parameters. We have that

$$\begin{aligned} p(\boldsymbol{\pi}|\boldsymbol{\Theta}_{-\boldsymbol{\pi}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}) \propto & \prod_{k=1}^{K} \pi_k^{c_k - 1} \\ & \times \prod_{i=1}^{N} \frac{1}{B(\alpha_3\boldsymbol{\pi})} \prod_{k=1}^{K} Z_{ik}^{\alpha_3\pi_k - 1}. \end{aligned}$$

Letting out proposal distribution be such that $Q(\boldsymbol{\pi}'|\boldsymbol{\pi}) = Dir(a_{\boldsymbol{\pi}}\boldsymbol{\pi})$, for some large $a_{\boldsymbol{\pi}} \in \mathbb{R}^+$, we have that our probability of accepting any proposal is

$$A(\boldsymbol{\pi}', \boldsymbol{\pi}) = \min\left\{1, \frac{P\left(\boldsymbol{\pi}'|\boldsymbol{\Theta}_{-\boldsymbol{\pi}'}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)}{P\left(\boldsymbol{\pi}|\boldsymbol{\Theta}_{-\boldsymbol{\pi}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)} \frac{Q\left(\boldsymbol{\pi}|\boldsymbol{\pi}'\right)}{Q\left(\boldsymbol{\pi}'|\boldsymbol{\pi}\right)}\right\}.$$

The posterior distribution of α_3 is also not a commonly known distribution, so we will use the Metropolis-Hastings algorithm to sample from the posterior distribution. We have that

$$p(\alpha_3|\boldsymbol{\Theta}_{-\alpha_3}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}) \propto e^{-b\alpha_3} \times \prod_{i=1}^{N} \frac{1}{B(\alpha_3\boldsymbol{\pi})} \prod_{k=1}^{K} Z_{ik}^{\alpha_3\pi_k - 1}.$$

Using a proposal distribution such that $Q(\alpha_3'|\alpha_3) = \mathcal{N}(\alpha_3, \sigma_{\alpha_3}^2, 0, +\infty)$ (Truncated Normal), we are left with the probability of accepting a proposed state as

$$A(\alpha_3', \alpha_3) = \min\left\{1, \frac{P\left(\alpha_3'|\boldsymbol{\Theta}_{-\alpha_3'}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)}{P\left(\alpha_3|\boldsymbol{\Theta}_{-\alpha_3}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)} \frac{Q\left(\alpha_3|\alpha_3'\right)}{Q\left(\alpha_3'|\alpha_3\right)}\right\}.$$

Let $\mathbf{P}$ be the following tridiagonal matrix:

$$\mathbf{P} = \begin{bmatrix} 1 & -1 & 0 & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & 0 & -1 & 1 \end{bmatrix}.$$

Thus, letting

$$\mathbf{B}_{\boldsymbol{\nu}_j} = \left(\tau_{\boldsymbol{\nu}_j}\mathbf{P} + \frac{1}{\sigma^2}\sum_{i=1}^{N}\sum_{l=1}^{n_i} Z_{ij}^2 B(t_{il})B'(t_{il})\right)^{-1}$$

and

$$\mathbf{b}_{\boldsymbol{\nu}_j} = \frac{1}{\sigma^2}\sum_{i=1}^{N}\sum_{l=1}^{n_i} Z_{ij}B(t_{il})\left[y_i(t_{il}) - \left(\sum_{k\neq j} Z_{ik}\boldsymbol{\nu}_k' B(t_{il})\right) - \left(\sum_{k=1}^{K} Z_{ik}\left[\mathbf{x}_i\boldsymbol{\eta}_k' B(t_{il}) + \sum_{m=1}^{M}\chi_{im}\boldsymbol{\phi}_{kn}' B(t_{il})\right]\right)\right],$$

we have that

$$\boldsymbol{\nu}_j|\boldsymbol{\Theta}_{-\boldsymbol{\nu}_j}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N}\left(\mathbf{B}_{\boldsymbol{\nu}_j}\mathbf{b}_{\boldsymbol{\nu}_j}, \mathbf{B}_{\boldsymbol{\nu}_j}\right).$$

Let $\boldsymbol{\eta}_{jd}$ denote the d^{th} column of the matrix $\boldsymbol{\eta}_j$. Thus, letting

$$\mathbf{B}_{\boldsymbol{\eta}_{jd}} = \left(\tau_{\boldsymbol{\eta}_{jd}} \mathbf{P} + \frac{1}{\sigma^2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} Z_{ij}^2 x_{id}^2 B(t_{il}) B'(t_{il}) \right)^{-1}$$

and

$$\begin{aligned} \mathbf{b}_{\boldsymbol{\eta}_{jd}} =& \frac{1}{\sigma^2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} Z_{ij} x_{id} B(t_{il}) \left[y_i(t_{il}) - \left(\sum_{r \neq d} Z_{ij} x_{ir} \boldsymbol{\eta}'_{jr} B(t_{il}) \right) - \left(\sum_{k \neq j} Z_{ik} \mathbf{x}_i \boldsymbol{\eta}'_k B(t_{il}) \right) \right. \\ & \left. - \left(\sum_{k=1}^{K} Z_{ik} \left[\boldsymbol{\nu}'_k B(t_{il}) + \sum_{m=1}^{M} \chi_{im} \boldsymbol{\phi}'_{kn} B(t_{il}) \right] \right) \right], \end{aligned}$$

we have that

$$\boldsymbol{\eta}_{jd} | \boldsymbol{\Theta}_{-\boldsymbol{\eta}_{jd}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N} \left(\mathbf{B}_{\boldsymbol{\eta}_{jd}} \mathbf{b}_{\boldsymbol{\eta}_{jd}}, \mathbf{B}_{\boldsymbol{\eta}_{jd}} \right).$$

Thus we can see that we can draw samples from the posterior of the parameters controlling the mean structure using a Gibbs sampler. Similarly, we can use a Gibbs sampler to draw samples from the posterior distribution of $\tau_{\boldsymbol{\eta}_{jd}}$ and $\tau_{\boldsymbol{\nu}_j}$. We have that the posterior distributions are

$$\tau_{\boldsymbol{\nu}_j} | \boldsymbol{\Theta}_{-\tau_{\boldsymbol{\nu}_j}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \Gamma \left(\alpha_{\boldsymbol{\nu}} + P/2, \beta_{\boldsymbol{\nu}} + \frac{1}{2} \boldsymbol{\nu}'_j \mathbf{P} \boldsymbol{\nu}_j \right)$$

and

$$\tau_{\boldsymbol{\eta}_{jd}} | \boldsymbol{\Theta}_{-\tau_{\boldsymbol{\eta}_{jd}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \Gamma \left(\alpha_{\boldsymbol{\eta}} + P/2, \beta_{\boldsymbol{\eta}} + \frac{1}{2} \boldsymbol{\eta}'_{jd} \mathbf{P} \boldsymbol{\eta}_{jd} \right),$$

for $j = 1, \ldots, K$ and $d = 1, \ldots, R$. The parameter σ^2 can be updated by using a Gibbs update. If we let

$$\beta_\sigma = \frac{1}{2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} \left(y_i(t_{il}) - \sum_{k=1}^{K} Z_{ik} \left(\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k \mathbf{x}'_i \right)' B(t_{il}) + \sum_{n=1}^{M} \chi_{in} \boldsymbol{\phi}'_{kn} B(t_{il}) \right) \right)^2,$$

then we have

$$\sigma^2|\boldsymbol{\Theta}_{-\sigma^2}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim IG\left(\alpha_0 + \frac{\sum_{i=1}^N n_i}{2}, \beta_0 + \beta_\sigma\right).$$

Lastly, we can update the χ_{im} parameters, for $i = 1, \ldots, N$ and $m = 1, \ldots, M$, using a Gibbs update. If we let

$$\begin{aligned}\mathbf{w}_{im} =& \frac{1}{\sigma^2}\left[\sum_{l=1}^{n_i}\left(\sum_{k=1}^{K} Z_{ik}\boldsymbol{\phi}'_{km}B(t_{il})\right)\right. \\ & \left.\left(y_i(t_{il}) - \sum_{k=1}^{K} Z_{ik}\left(\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k\mathbf{x}'_i\right)' B(t_{il}) + \sum_{n\neq m}\chi_{in}\boldsymbol{\phi}'_{kn}B(t_{il})\right)\right)\right]\end{aligned}$$

and

$$\mathbf{W}_{im}^{-1} = 1 + \frac{1}{\sigma^2}\sum_{l=1}^{n_i}\left(\sum_{k=1}^{K} Z_{ik}\boldsymbol{\phi}'_{km}B(t_{il})\right)^2,$$

then we have that

$$\chi_{im}|\boldsymbol{\zeta}_{-\chi_{im}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N}(\mathbf{W}_{im}\mathbf{w}_{im}, \mathbf{W}_{im}).$$

C.2.2 Tempered Transitions

One of the main computational problems we face in these flexible, unsupervised models is a multi-modal posterior distribution. In order to help the Markov chain move across modes, or traverse areas of low posterior probability, we can utilize tempered transitions.

In this paper, we will be following the works of Behrens et al. [2012] and Pritchard et al. [2000] and only temper the likelihood. The target distribution that we want to temper is usually assumed to be written as

$$p(x) \propto \pi(x)exp\left(-\beta_h h(x)\right),$$

where β_h controls how much the distribution is tempered ($1 = \beta_0 < \cdots < \beta_h < \cdots < \beta_{N_t}$). In this setting, we will assume that the hyperparameters N_t and β_{N_t} are user specified, and will depend on the complexity of the model. For more complex or larger models, we will need to set N_t relatively high. In this implementation, we assume the β_h parameters to follow a geometric scheme, but in more complex models, β_{N_t} may need to be relatively small.

We can rewrite our likelihood for the functional covariate adjusted model to fit the above form:

$$\begin{aligned} p_h(y_i(t)|\boldsymbol{\Theta}, \mathbf{X}) \propto & exp\left\{-\beta_h\left(\frac{1}{2}log(\sigma^2) + \frac{1}{2\sigma^2}\left(y_i(t) - \sum_{k=1}^{K} Z_{ik}\left(\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k \mathbf{x}_i'\right)' B(t)\right.\right.\right.\right. \\ & \left.\left.\left.\left. + \sum_{n=1}^{M} \chi_{in} \boldsymbol{\phi}_{k'n}' B(t)\right)\right)^2\right)\right\} \\ = & \left(\sigma^2\right)^{-\beta_h/2} exp\left\{-\frac{\beta_h}{2\sigma^2}\left(y_i(t) - \sum_{k=1}^{K} Z_{ik}\left(\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k \mathbf{x}_i'\right)' B(t)\right.\right.\right. \\ & \left.\left.\left. + \sum_{n=1}^{M} \chi_{in} \boldsymbol{\phi}_{k'n}' B(t)\right)\right)^2\right\}. \end{aligned}$$

Let $\boldsymbol{\Theta}_h$ be the set of parameters generated from the model using the tempered likelihood associated with β_h. The tempered transition algorithm can be summarized by the following steps:

1. Start with initial state $\boldsymbol{\Theta}_0$.
2. Transition from $\boldsymbol{\Theta}_0$ to $\boldsymbol{\Theta}_1$ using the tempered likelihood associated with β_1.
3. Continue in this manner until we transition from $\boldsymbol{\Theta}_{N_t-1}$ to $\boldsymbol{\Theta}_{N_t}$ using the tempered likelihood associated with β_{N_t}.
4. Transition from $\boldsymbol{\Theta}_{N_t}$ to $\boldsymbol{\Theta}_{N_t+1}$ using the tempered likelihood associated with β_{N_t}.
5. Continue in this manner until we transition from $\boldsymbol{\Theta}_{2N_t-1}$ to $\boldsymbol{\Theta}_{2N_t}$ using β_1.

6. Accept transition from $\boldsymbol{\Theta}_0$ to $\boldsymbol{\Theta}_{2N_t}$ with probability

$$\min\left\{1, \prod_{h=0}^{N_t-1} \frac{\prod_{i=1}^{N}\prod_{l=1}^{n_i} p_{h+1}(y_i(t_{il})|\boldsymbol{\Theta}_h, \mathbf{X}_i)}{\prod_{i=1}^{N}\prod_{l=1}^{n_i} p_h(y_i(t_{il})|\boldsymbol{\Theta}_h, \mathbf{X}_i)} \prod_{h=N_t+1}^{2N_t} \frac{\prod_{i=1}^{N}\prod_{l=1}^{n_i} p_h(y_i(t_{il})|\boldsymbol{\Theta}_h, \mathbf{X}_i)}{\prod_{i=1}^{N}\prod_{l=1}^{n_i} p_{h+1}(y_i(t_{il})|\boldsymbol{\Theta}_h, \mathbf{X}_i)}\right\}$$

in the functional case, or

$$\min\left\{1, \prod_{h=0}^{N_t-1} \frac{\prod_{i=1}^{N}\prod_{l=1}^{n_i} p_{h+1}(\mathbf{y}_i|\boldsymbol{\Theta}_h, \mathbf{X}_i)}{\prod_{i=1}^{N}\prod_{l=1}^{n_i} p_h(\mathbf{y}_i)|\boldsymbol{\Theta}_h, \mathbf{X}_i)} \prod_{h=N_t+1}^{2N_t} \frac{\prod_{i=1}^{N}\prod_{l=1}^{n_i} p_h(\mathbf{y}_i|\boldsymbol{\Theta}_h, \mathbf{X}_i)}{\prod_{i=1}^{N}\prod_{l=1}^{n_i} p_{h+1}(\mathbf{y}_i|\boldsymbol{\Theta}_h, \mathbf{X}_i)}\right\}$$

in the multivariate case.

Since we only temper the likelihood, many of the posterior distributions derived in Section C.2.1 can be utilized. Thus the following posteriors are the only ones that change due to the tempering of the likelihood. Starting with the $\boldsymbol{\Phi}$ parameters, we have

$$\begin{aligned}\left(\mathbf{m}_{\boldsymbol{\phi}_{jm}}\right)_h =& \frac{\beta_h}{(\sigma^2)_h}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\Big(B(t_{il})(\chi_{im})_h\Big(y_i(t_{il})(Z_{ij})_h - (Z_{ij})_h^2\left((\boldsymbol{\nu}_j)_h + (\boldsymbol{\eta}_j)_h\mathbf{x}_i'\right)' B(t_{il}) \\ &-(Z_{ij})_h^2\sum_{n\neq m}(\chi_{in})_h(\boldsymbol{\phi}_{jn})_h' B(t_{il}) \\ &-\sum_{k\neq j} Z_{ij}Z_{ik}\left[\left((\boldsymbol{\nu}_k)_h + (\boldsymbol{\eta}_k)_h\mathbf{x}_i'\right)' B(t_{il}) + \sum_{n=1}^{M}\chi_{in}(\boldsymbol{\phi}_{kn})_h' B(t_{il})\right]\Big)\Big),\end{aligned}$$

and

$$\left(\mathbf{M}_{\boldsymbol{\phi}_{jm}}\right)_h^{-1} = \frac{\beta_h}{(\sigma^2)_h}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\left((Z_{ij})_h^2(\chi_{im})_h^2 B(t_{il})B'(t_{il})\right) + \left(\mathbf{D}_{\boldsymbol{\phi}_{jm}}\right)_h^{-1},$$

we have that

$$\left(\boldsymbol{\phi}_{jm}\right)_h|\boldsymbol{\Theta}_{-\left(\boldsymbol{\phi}_{jm}\right)_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N}\left(\left(\mathbf{M}_{\boldsymbol{\phi}_{jm}}\right)_h\left(\mathbf{m}_{\boldsymbol{\phi}_{jm}}\right)_h, \left(\mathbf{M}_{\boldsymbol{\phi}_{jm}}\right)_h\right).$$

As in the untempered case, we have that the posterior distribution $\mathbf{Z}$ parameters under the tempered likelihood is not a commonly known distribution. Therefore, we will use the

Metropolis-Hastings algorithm. We have that

$$p((\mathbf{z}_i)_h|\mathbf{\Theta}_{-(\mathbf{z}_i)_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}) \propto \prod_{k=1}^{K} (Z_{ik})_h^{(\alpha_3)_h(\pi_k)_h - 1} \times \prod_{l=1}^{n_i} exp\left\{ -\frac{\beta_h}{2(\sigma^2)_h} \left(y_i(t_{il}) - \sum_{k=1}^{K} (Z_{ik})_h \left(\left((\boldsymbol{\nu}_k)_h + (\boldsymbol{\eta}_k)_h \mathbf{x}_i' \right)' B(t_{il}) + \sum_{m=1}^{M} (\chi_{im})_h (\boldsymbol{\phi}_{km})_h' B(t_{il}) \right) \right)^2 \right\}.$$

We will use $Q((\mathbf{z}_i)_h'|(\mathbf{z}_i)_h) = Dir(a_{\mathbf{z}}(\mathbf{z}_i)_h)$ for some large $a_{\mathbf{z}} \in \mathbb{R}^+$ as the proposal distribution. Thus the probability of accepting a proposed step is

$$A((\mathbf{z}_i)_h', (\mathbf{z}_i)_h) = \min\left\{ 1, \frac{P\left((\mathbf{z}_i)_h'|\mathbf{\Theta}_{-(\mathbf{z}_i)_h'}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)}{P\left((\mathbf{z}_i)_h|\mathbf{\Theta}_{-(\mathbf{z}_i)_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)} \frac{Q\left((\mathbf{z}_i)_h|(\mathbf{z}_i)_h'\right)}{Q\left((\mathbf{z}_i)_h'|(\mathbf{z}_i)_h\right)} \right\}.$$

Letting

$$\left(\mathbf{B}_{\boldsymbol{\nu}_j}\right)_h = \left(\left(\tau_{\boldsymbol{\nu}_j}\right)_h \mathbf{P} + \frac{\beta_h}{(\sigma^2)_h} \sum_{i=1}^{N} \sum_{l=1}^{n_i} (Z_{ij})_h^2 B(t_{il}) B'(t_{il}) \right)^{-1}$$

and

$$\left(\mathbf{b}_{\boldsymbol{\nu}_j}\right)_h = \frac{\beta_h}{(\sigma^2)_h} \sum_{i=1}^{N} \sum_{l=1}^{n_i} (Z_{ij})_h B(t_{il}) \left[y_i(t_{il}) - \left(\sum_{k \neq j} (Z_{ik})_h (\boldsymbol{\nu}_k')_h B(t_{il}) \right) - \left(\sum_{k=1}^{K} (Z_{ik}) \left[\mathbf{x}_i (\boldsymbol{\eta}_k)_h' B(t_{il}) + \sum_{m=1}^{M} (\chi_{im})_h (\boldsymbol{\phi}_{kn})_h' B(t_{il}) \right] \right) \right],$$

we have that

$$(\boldsymbol{\nu}_j)_h | \mathbf{\Theta}_{-(\boldsymbol{\nu}_j)_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N}\left(\left(\mathbf{B}_{\boldsymbol{\nu}_j}\right)_h \left(\mathbf{b}_{\boldsymbol{\nu}_j}\right)_h, \left(\mathbf{B}_{\boldsymbol{\nu}_j}\right)_h \right).$$

Let $\left(\boldsymbol{\eta}_{jd}\right)_h$ denote the d^{th} column of the matrix $(\boldsymbol{\eta}_j)_h$. Thus, letting

$$\left(\mathbf{B}_{\boldsymbol{\eta}_{jd}}\right)_h = \left(\left(\tau_{\boldsymbol{\eta}_{jd}}\right)_h \mathbf{P} + \frac{\beta_h}{(\sigma^2)_h} \sum_{i=1}^{N} \sum_{l=1}^{n_i} (Z_{ij})_h^2 x_{id}^2 B(t_{il}) B'(t_{il}) \right)^{-1}$$

and

$$\begin{aligned}\left(\mathbf{b}_{\boldsymbol{\eta}_{jd}}\right)_h =& \frac{\beta_h}{(\sigma^2)_h}\sum_{i=1}^{N}\sum_{l=1}^{n_i}(Z_{ij})_h x_{id}B(t_{il})\left[y_i(t_{il}) - \left(\sum_{r\neq d}(Z_{ij})_h x_{ir}(\boldsymbol{\eta}_{jr})'_h B(t_{il})\right)\right. \\ &- \left(\sum_{k\neq j}(Z_{ik})_h \mathbf{x}_i(\boldsymbol{\eta}_k)'_h B(t_{il})\right) \\ &\left. - \left(\sum_{k=1}^{K}(Z_{ik})_h\left[(\boldsymbol{\nu}_k)'_h B(t_{il}) + \sum_{m=1}^{M}(\chi_{im})_h(\boldsymbol{\phi}_{kn})'_h B(t_{il})\right]\right)\right],\end{aligned}$$

we have that

$$\left(\boldsymbol{\eta}_{jd}\right)_h | \boldsymbol{\Theta}_{-\left(\boldsymbol{\eta}_{jd}\right)_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N}\left(\left(\mathbf{B}_{\boldsymbol{\eta}_{jd}}\right)_h \left(\mathbf{b}_{\boldsymbol{\eta}_{jd}}\right)_h, \left(\mathbf{B}_{\boldsymbol{\eta}_{jd}}\right)_h\right).$$

If we let

$$\begin{aligned}(\beta_\sigma)_h = \frac{\beta_h}{2}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\left(y_i(t_{il}) - \sum_{k=1}^{K}(Z_{ik})_h\left(\left((\boldsymbol{\nu}_k)_h + (\boldsymbol{\eta}_k)_h \mathbf{x}_i'\right)' B(t_{il})\right.\right. \\ \left.\left. + \sum_{n=1}^{M}(\chi_{in})_h(\boldsymbol{\phi}_{kn})'_h B(t_{il})\right)\right)^2,\end{aligned}$$

then we have

$$(\sigma^2)_h | \boldsymbol{\Theta}_{-(\sigma^2)_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim IG\left(\alpha_0 + \frac{\beta_h \sum_{i=1}^{N} n_i}{2}, \beta_0 + (\beta_\sigma)_h\right).$$

Lastly, we can update the χ_{im} parameters, for $i = 1, \ldots, N$ and $m = 1, \ldots, M$, using a Gibbs update. If we let

$$\begin{aligned}(\mathbf{w}_{im})_h =& \frac{\beta_h}{(\sigma^2)_h}\left[\sum_{l=1}^{n_i}\left(\sum_{k=1}^{K}(Z_{ik})_h(\boldsymbol{\phi}_{km})'_h B(t_{il})\right)\left(y_i(t_{il})\right.\right. \\ &\left.\left. - \sum_{k=1}^{K}(Z_{ik})_h\left(\left((\boldsymbol{\nu}_k)_h + (\boldsymbol{\eta}_k)_h \mathbf{x}_i'\right)' B(t_{il}) + \sum_{n\neq m}(\chi_{in})_h(\boldsymbol{\phi}_{kn})'_h B(t_{il})\right)\right)\right]\end{aligned}$$

and

$$(\mathbf{W}_{im})_h^{-1} = 1 + \frac{\beta_h}{\sigma^2} \sum_{l=1}^{n_i} \left(\sum_{k=1}^{K} (Z_{ik})_h (\boldsymbol{\phi}_{km})_h' B(t_{il}) \right)^2,$$

then we have that

$$(\chi_{im})_h | \boldsymbol{\zeta}_{-(\chi_{im})_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N}\left((\mathbf{W}_{im})_h (\mathbf{w}_{im})_h, (\mathbf{W}_{im})_h \right).$$

C.3 Simulation Study and Case Studies

C.3.1 Simulation Study

This subsection contains detailed information on how the simulation study in Section 3 of the main text was conducted. This simulation study primarily looked at how well we could recover the true mean structure, covariance structure, and allocation structure. In this simulation study, we simulated datasets from 3 scenarios at 3 different sample sizes for each scenario. Once the datasets were generated, we fit a variety of covariate adjusted functional mixed membership models, as well as unadjusted functional mixed membership models, on the datasets to see how well we could recover the mean, covariance, and allocation structures.

The first scenario we considered was a covariate adjusted functional mixed membership model with 2 true covariates. To generate all of the datasets, we assumed that the observations were in the span of B-spline basis with 8 basis functions. For this scenario, we generated 3 datasets with sample sizes of 60, 120, and 240 functional observations, all observed on a grid of 50 time points. The data was generated by first generating the model parameters (as discussed below) and then generating data from the likelihood specified in Equation 11 of the main text. The model parameters for this dataset were generated as follows:

$$\boldsymbol{\nu}_1 \sim \mathcal{N}\left((6, 4, \ldots, -6, -8)', 4\mathbf{P} \right),$$

$$\boldsymbol{\nu}_2 \sim \mathcal{N}\left((-8, -6, \ldots, 4, 6)', 4\mathbf{P} \right),$$

$$\boldsymbol{\eta}_{k1} \sim \mathcal{N}\left(\mathbf{1}, \mathbf{P}\right) \quad k = 1, 2$$

$$\boldsymbol{\eta}_{k2} \sim \mathcal{N}\left((3, 2, \ldots, -4)', \mathbf{P}\right) \quad k = 1, 2$$

We drew the $\boldsymbol{\Phi}$ parameters from the subspace orthogonal to the space spanned by the $\boldsymbol{\nu}$ parameters. Thus let $\text{colsp}(\mathbf{B}^{\perp}) := \text{span}\{b_1^{\perp}, \ldots, b_6^{\perp}\} \subset \mathbb{R}^8$ be the subspace orthogonal to the $\boldsymbol{\nu}$ parameters, which can be described as the span of 6 vectors in $\mathbb{R}^8$.The $\boldsymbol{\Phi}$ parameters were drawn according to the following distributions:

$$\boldsymbol{\phi}_{km} = \mathbf{q}_{km}\mathbf{B}^{\perp} \quad k = 1, 2 \quad m = 1, 2, 3,$$

where $\mathbf{q}_{k1} \sim \mathcal{N}(\mathbf{0}_6, 2.25\mathbf{I}_6)$, $\mathbf{q}_{k2} \sim \mathcal{N}(\mathbf{0}_6, \mathbf{I}_6)$, $\mathbf{q}_{k3} \sim \mathcal{N}(\mathbf{0}_6, 0.49\mathbf{I}_6)$. The χ_{im} parameters were drawn from a standard normal distribution. The $\mathbf{z}_i$ parameters were drawn from a mixture of Dirichlet distributions. Roughly 30% of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution with $\alpha_1 = 10$ and $\alpha_2 = 1$. Another roughly 30% of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution where $\alpha_1 = 1$ and $\alpha_2 = 10$. The rest of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution with $\alpha_1 = \alpha_2 = 1$. The covariates, $\mathbf{X}$, were drawn from a standard normal distribution. Models in this scenario were run for 500,000 MCMC iterations.

For the second scenario, we considered data drawn from a covariate adjusted functional mixed membership model with one covariate. We considered three sample sizes of 50, 100, and 200 functional samples observed on a grid of 25 time points. The model parameters for this dataset were generated as follows:

$$\boldsymbol{\nu}_1 \sim \mathcal{N}\left((6, 4, \ldots, -6, -8)', 4\mathbf{P}\right),$$

$$\boldsymbol{\nu}_2 \sim \mathcal{N}\left((-8, -6, \ldots, 4, 6)', 4\mathbf{P}\right),$$

$$\boldsymbol{\eta}_{11} \sim \mathcal{N}\left(\mathbf{2}, \mathbf{P}\right)$$

$$\boldsymbol{\eta}_{21} \sim \mathcal{N}\left(-\mathbf{2}, \mathbf{P}\right)$$

We drew the $\boldsymbol{\Phi}$ parameters from the subspace orthogonal to the space spanned by the $\boldsymbol{\nu}$ and $\boldsymbol{\eta}$ parameters. Thus let $\text{colsp}(\mathbf{B}^{\perp}) := \text{span}\{b_1^{\perp}, \ldots, b_4^{\perp}\} \subset \mathbb{R}^8$ be the subspace orthogonal to the $\boldsymbol{\nu}$ and $\boldsymbol{\eta}$ parameters, which can be described as the span of 4 vectors in $\mathbb{R}^8$.The $\boldsymbol{\Phi}$ parameters were drawn according to the following distributions:

$$\boldsymbol{\phi}_{km} = \mathbf{q}_{km}\mathbf{B}^{\perp} \quad k = 1, 2 \quad m = 1, 2, 3,$$

where $\mathbf{q}_{k1} \sim \mathcal{N}(\mathbf{0}_6, 4\mathbf{I}_6)$, $\mathbf{q}_{k2} \sim \mathcal{N}(\mathbf{0}_6, 2.25\mathbf{I}_6)$, $\mathbf{q}_{k3} \sim \mathcal{N}(\mathbf{0}_6, \mathbf{I}_6)$. The χ_{im} parameters were drawn from a standard normal distribution. The $\mathbf{z}_i$ parameters were drawn from a mixture of Dirichlet distributions. Roughly 30% of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution with $\alpha_1 = 10$ and $\alpha_2 = 1$. Another roughly 30% of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution where $\alpha_1 = 1$ and $\alpha_2 = 10$. The rest of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution with $\alpha_1 = \alpha_2 = 1$. The covariates, $\mathbf{X}$, were drawn from a normal distribution with variance of nine and mean of zero. Models in this scenario were run for 300,000 MCMC iterations.

For the third scenario, we generated data from an unadjusted functional mixed membership model. We considered three sample sizes of 40, 80, and 160 functional samples observed on a grid of 25 time points. The model parameters for this dataset were generated as follows:

$$\boldsymbol{\nu}_1 \sim \mathcal{N}\left((6, 4, \ldots, -6, -8)', 4\mathbf{P}\right),$$

$$\boldsymbol{\nu}_2 \sim \mathcal{N}\left((-8, -6, \ldots, 4, 6)', 4\mathbf{P}\right),$$

$$\boldsymbol{\eta}_{11} \sim \mathcal{N}\left(\mathbf{2}, \mathbf{P}\right)$$

$$\boldsymbol{\eta}_{21} \sim \mathcal{N}\left(-\mathbf{2}, \mathbf{P}\right)$$

We drew the $\boldsymbol{\Phi}$ parameters from the subspace orthogonal to the space spanned by the $\boldsymbol{\nu}$

parameters. Thus let $\text{colsp}(\mathbf{B}^{\perp}) := \text{span}\{b_1^{\perp}, \ldots, b_4^{\perp}\} \subset \mathbb{R}^8$ be the subspace orthogonal to the $\boldsymbol{\nu}$ parameters, which can be described as the span of 4 vectors in $\mathbb{R}^8$.The $\boldsymbol{\Phi}$ parameters were drawn according to the following distributions:

$$\boldsymbol{\phi}_{km} = \mathbf{q}_{km}\mathbf{B}^{\perp} \quad k = 1, 2 \quad m = 1, 2,$$

where $\mathbf{q}_{k1} \sim \mathcal{N}(\mathbf{0}_6, 2.25\mathbf{I}_6)$ and $\mathbf{q}_{k2} \sim \mathcal{N}(\mathbf{0}_6, \mathbf{I}_6)$. The χ_{im} parameters were drawn from a standard normal distribution. The $\mathbf{z}_i$ parameters were drawn from a mixture of Dirichlet distributions. Roughly 30% of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution with $\alpha_1 = 10$ and $\alpha_2 = 1$. Another roughly 30% of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution where $\alpha_1 = 1$ and $\alpha_2 = 10$. The rest of the $\mathbf{z}_i$ parameters were drawn from a Dirichlet distribution with $\alpha_1 = \alpha_2 = 1$. Models in this scenario were run for 500,000 MCMC iterations.

The code for running this simulation study can be found on Github.

C.4 Mean and Covariance Covariate-dependent Mixed Membership Model

C.4.1 Model Specification

In this section, we completely specify a mixed membership model where the mean and covariance structures are dependent on the covariates of interest. As in the main text of this manuscript, we will let $\{\mathbf{Y}_i(.)\}_{i=1}^N$ be the observed sample paths and $\mathbf{t}_i = [t_{i1}, \ldots, t_{in_i}]'$ denote the time points at which the i^{th} function was observed over. We will also let $\mathbf{X} \in \mathbb{R}^{N \times R}$ denote the design matrix and $\mathbf{x}_i = [X_{i1} \ldots X_{iR}]$ denote the i^{th} row of the design matrix (or the covariates associated with the i^{th} observation). By introducing covariate-dependent pseudo-eigenfunctions, we arrive at the likelihood of our mixed membership model where

the mean and covariance structures are dependent on the covariates of interest:

$$\mathbf{Y}_i(\mathbf{t}_i) \mid \boldsymbol{\Theta}, \mathbf{X} \sim \mathcal{N}\left\{\sum_{k=1}^{K} Z_{ik}\left(\mathbf{S}'(\mathbf{t}_i)\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k \mathbf{x}_i'\right) + \sum_{m=1}^{M} \chi_{im}\mathbf{S}'(\mathbf{t}_i)\left(\boldsymbol{\phi}_{km} + \boldsymbol{\xi}_{km}\mathbf{x}_i'\right)\right),\ \sigma^2\mathbf{I}_{n_i}\right\}. \quad \text{(C.3)}$$

From equation C.3, we can see that $\boldsymbol{\xi}_{km} \in \mathbb{R}^{P\times R}$, directly controls the effect that the covariates have on the pseudo-eigenfunctions for $k = 1, \ldots, K$ and $m = 1, \ldots, M$. By integrating out the χ_{im} parameters ($i = 1, \ldots, N$ and $m = 1, \ldots, M$), we get a model of the following form:

$$\mathbf{Y}_i(\mathbf{t}_i) \mid \boldsymbol{\Theta}_{-\chi}, \mathbf{X} \sim \mathcal{N}\left\{\sum_{k=1}^{K} Z_{ik}\mathbf{S}'(\mathbf{t}_i)\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k \mathbf{x}_i'\right),\ \mathbf{V}(\mathbf{t}_i, \mathbf{z}_i) + \sigma^2\mathbf{I}_{n_i}\right\}, \quad \text{(C.4)}$$

where $\boldsymbol{\Theta}_{-\chi}$ is the collection of our model parameters excluding the χ_{im} variables, and the error-free mixed membership covariance is

$$\mathbf{V}(\mathbf{t}_i, \mathbf{z}_i) = \sum_{k=1}^{K}\sum_{k'=1}^{K} Z_{ik}Z_{ik'}\left\{\mathbf{S}'(\mathbf{t}_i)\sum_{m=1}^{M}\left[\left(\boldsymbol{\phi}_{km} + \boldsymbol{\xi}_{km}\mathbf{x}_i'\right)\left(\boldsymbol{\phi}_{k'm} + \boldsymbol{\xi}_{k'm}\mathbf{x}_i'\right)'\right]\mathbf{S}(\mathbf{t}_i)\right\}. \quad \text{(C.5)}$$

As with the pseudo-eigenfunctions in the unadjusted model, we will utilize the multiplicative gamma process prior as our prior on the $\boldsymbol{\xi}_{km}$ variables. Letting $\xi_{(krm)_p}$ denote the element in the p^{th} row and r^{th} column of $\boldsymbol{\xi}_{km}$. Thus we have:

$$\xi_{(krm)_p} \mid \gamma_{\boldsymbol{\xi}_{krmp}}, \tilde{\tau}_{\boldsymbol{\xi}_{mkr}} \sim \mathcal{N}\left(0, \gamma_{\boldsymbol{\xi}_{krmp}}^{-1}\tilde{\tau}_{\boldsymbol{\xi}_{mkr}}^{-1}\right), \quad \gamma_{\xi_{krmp}} \sim \Gamma\left(\nu_\gamma/2, \nu_\gamma/2\right), \quad \tilde{\tau}_{\boldsymbol{\xi}_{mkr}} = \prod_{n=1}^{m}\delta_{\boldsymbol{\xi}_{nkr}},$$

$$\delta_{\boldsymbol{\xi}_{1kr}} \mid a_{\boldsymbol{\xi}_{1kr}} \sim \Gamma(a_{\boldsymbol{\xi}_{1kr}}, 1), \quad \delta_{\boldsymbol{\xi}_{jkr}} \mid a_{\boldsymbol{\xi}_{2kr}} \sim \Gamma(a_{\boldsymbol{\xi}_{2kr}}, 1), \quad a_{\boldsymbol{\xi}_{1kr}} \sim \Gamma(\alpha_1, \beta_1), \quad a_{\boldsymbol{\xi}_{2kr}} \sim \Gamma(\alpha_2, \beta_2),$$

for $k = 1, \ldots, K$, $r = 1, \ldots, R$, $m = 1, \ldots, M$, and $p = 1, \ldots P$. The rest of the parameters in the model have the same prior distributions as the model with the covariate-dependence

on the mean structure only in the main text. Specifically, we have

$$\phi_{kpm}|\gamma_{kpm},\tilde{\tau}_{mk}\sim\mathcal{N}\left(0,\gamma_{kpm}^{-1}\tilde{\tau}_{mk}^{-1}\right),\quad \gamma_{kpm}\sim\Gamma\left(\nu_\gamma/2,\nu_\gamma/2\right),\quad \tilde{\tau}_{mk}=\prod_{n=1}^{m}\delta_{nk},$$

$$\delta_{1k}|a_{1k}\sim\Gamma(a_{1k},1),\quad \delta_{jk}|a_{2k}\sim\Gamma(a_{2k},1),\quad a_{1k}\sim\Gamma(\alpha_1,\beta_1),\quad a_{2k}\sim\Gamma(\alpha_2,\beta_2),$$

for $k=1,\ldots,K$, $m=1,\ldots,M$, and $p=1,\ldots P$. Similarly, we have

$$P(\boldsymbol{\nu}_k|\tau_{\boldsymbol{\nu}_k})\propto exp\left(-\frac{\tau_{\boldsymbol{\nu}_k}}{2}\sum_{p=1}^{P-1}\left(\nu'_{pk}-\nu_{(p+1)k}\right)^2\right),$$

for $k=1,\ldots,K$, where $\tau_{\boldsymbol{\nu}_k}\sim\Gamma(\alpha_{\boldsymbol{\nu}},\beta_{\boldsymbol{\nu}})$ and ν_{pk} is the p^{th} element of $\boldsymbol{\nu}_k$. Likewise, we have that

$$P(\{\eta_{prk}\}_{p=1}^{P}|\tau_{\boldsymbol{\eta}_{rk}})\propto exp\left(-\frac{\tau_{\boldsymbol{\eta}_{rk}}}{2}\sum_{p=1}^{P-1}\left(\eta'_{prk}-\eta_{(p+1)rk}\right)^2\right),$$

for $k=1,\ldots,K$ and $r=1,\ldots,R$, where $\tau_{\boldsymbol{\eta}_{rk}}\sim\Gamma(\alpha_{\boldsymbol{\eta}},\beta_{\boldsymbol{\eta}})$ and η_{prk} is the p^{th} row and r^{th} column of $\boldsymbol{\eta}_k$. Lastly, we assume that $\mathbf{z}_i\mid\boldsymbol{\pi},\alpha_3\sim_{iid}Dir(\alpha_3\boldsymbol{\pi})$, $\boldsymbol{\pi}\sim Dir(\mathbf{c})$, $\alpha_3\sim Exp(b)$, and $\sigma^2\sim IG(\alpha_0,\beta_0)$.

C.4.2 Posterior Distributions

In this subsection, we will specify the posterior distributions specifically for the functional covariate adjusted mixed membership model where the covariance is covariate-dependent. We will first start with the $\boldsymbol{\phi}_{km}$ parameters, for $j=1,\ldots,K$ and $m=1,\ldots,M$. Let $\mathbf{D}_{\boldsymbol{\phi}_{jm}}=\tilde{\tau}_{\boldsymbol{\phi}_{mj}}^{-1}diag\left(\gamma_{\boldsymbol{\phi}_{j1m}}^{-1},\ldots,\gamma_{\boldsymbol{\phi}_{jPm}}^{-1}\right)$. By letting

$$\begin{aligned}\mathbf{m}_{\boldsymbol{\phi}_{jm}}=&\frac{1}{\sigma^2}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\Bigg(B(t_{il})\chi_{im}\Bigg(y_i(t_{il})Z_{ij}-Z_{ij}^2\left(\boldsymbol{\nu}_j+\boldsymbol{\eta}_j\mathbf{x}_i'\right)'B(t_{il})-Z_{ij}^2\sum_{n\neq m}\chi_{in}\boldsymbol{\phi}'_{jn}B(t_{il})\\&-Z_{ij}^2\sum_{n=1}^{M}\chi_{in}\mathbf{x}_i\boldsymbol{\xi}'_{jn}B(t_{il})-\sum_{k\neq j}Z_{ij}Z_{ik}\left[\left(\boldsymbol{\nu}_k+\boldsymbol{\eta}_k\mathbf{x}_i'\right)'B(t_{il})+\sum_{n=1}^{M}\chi_{in}\left(\boldsymbol{\phi}_{kn}+\boldsymbol{\xi}_{kn}\mathbf{x}_i'\right)'B(t_{il})\right]\Bigg)\Bigg),\end{aligned}$$

and

$$\mathbf{M}_{\boldsymbol{\phi}_{jm}}^{-1} = \frac{1}{\sigma^2}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\left(Z_{ij}^2\chi_{im}^2 B(t_{il})B'(t_{il})\right) + \mathbf{D}_{\boldsymbol{\phi}_{jm}}^{-1},$$

we have that

$$\boldsymbol{\phi}_{jm}|\boldsymbol{\Theta}_{-\boldsymbol{\phi}_{jm}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N}\left(\mathbf{M}_{\boldsymbol{\phi}_{jm}}\mathbf{m}_{\boldsymbol{\phi}_{jm}}, \mathbf{M}_{\boldsymbol{\phi}_{jm}}\right).$$

Let $\boldsymbol{\xi}_{krm}$ be the r^{th} column of the matrix $\boldsymbol{\xi}_{km}$. We will let $\mathbf{D}_{\boldsymbol{\xi}_{krm}} = \tilde{\tau}_{\boldsymbol{\xi}_{mjr}}^{-1} diag\left(\gamma_{\boldsymbol{\xi}_{jrm1}}^{-1}, \ldots, \gamma_{\boldsymbol{\phi}_{\boldsymbol{\xi}_{jrmP}}}^{-1}\right)$. We will also let x_{ir} denote the r^{th} element of $\mathbf{x}_i$. Thus, letting

$$\begin{aligned}\mathbf{m}_{\boldsymbol{\xi}_{kdm}} =& \frac{1}{\sigma^2}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\left(B(t_{il})\chi_{im}x_{id}Z_{ik}\left(y_i(t_{il}) - \sum_{j=1}^{K}Z_{ij}\left[\left(\boldsymbol{\nu}_j + \boldsymbol{\eta}_j\mathbf{x}_i'\right)'B(t_{il}) + \sum_{n=1}^{M}\chi_{in}\boldsymbol{\phi}_{jn}'B(t_{il})\right]\right.\right.\\ &\left.\left. - \sum_{(j,n,r)\neq(k,m,d)} Z_{ij}\chi_{in}x_{ir}\boldsymbol{\xi}_{krn}'B(t_{il})\right)\right)\end{aligned}$$

$$\mathbf{M}_{\boldsymbol{\xi}_{kdm}}^{-1} = \frac{1}{\sigma^2}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\left(Z_{ik}^2\chi_{im}^2x_{id}^2 B(t_{il})B'(t_{il})\right) + \mathbf{D}_{\boldsymbol{\xi}_{kdm}}^{-1},$$

we have that

$$\boldsymbol{\xi}_{kdm}|\boldsymbol{\Theta}_{-\boldsymbol{\xi}_{kdm}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N}\left(\mathbf{M}_{\boldsymbol{\xi}_{kdm}}\mathbf{m}_{\boldsymbol{\xi}_{kdm}}, \mathbf{M}_{\boldsymbol{\xi}_{kdm}}\right).$$

The posterior distribution of $\delta_{\boldsymbol{\phi}_{1k}}$, for $k = 1, \ldots, K$, is

$$\begin{aligned}\delta_{\boldsymbol{\phi}_{1k}}|\boldsymbol{\Theta}_{-\delta_{\boldsymbol{\phi}_{1k}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim& \Gamma\left(a_{\boldsymbol{\phi}_{1k}} + (PM/2), 1 + \frac{1}{2}\sum_{r=1}^{P}\gamma_{\boldsymbol{\phi}_{k,r,1}}\phi_{k,r,1}^2\right.\\ &\left. + \frac{1}{2}\sum_{m=2}^{M}\sum_{r=1}^{P}\gamma_{\boldsymbol{\phi}_{k,r,m}}\phi_{k,r,m}^2\left(\prod_{j=2}^{m}\delta_{\boldsymbol{\phi}_{jk}}\right)\right).\end{aligned}$$

The posterior distribution for $\delta_{\boldsymbol{\phi}_{ik}}$, for $i = 2, \ldots, M$ and $k = 1, \ldots, K$, is

$$\delta_{\boldsymbol{\phi}_{ik}}|\boldsymbol{\Theta}_{-\delta_{\boldsymbol{\phi}_{ik}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \Gamma\Bigg(a_{\boldsymbol{\phi}_{2k}} + (P(M-i+1)/2), 1 + \frac{1}{2}\sum_{m=i}^{M}\sum_{r=1}^{P}\gamma_{\boldsymbol{\xi}_{k,r,m}}\phi^2_{k,r,m}\left(\prod_{j=1;j\neq i}^{m}\delta_{\boldsymbol{\phi}_{jk}}\right)\Bigg).$$

The posterior distribution of $\delta_{\boldsymbol{\xi}_{1kd}}$, for $k = 1, \ldots, K$ and $d = 1, \ldots, R$, is

$$\delta_{\boldsymbol{\xi}_{1kd}}|\boldsymbol{\Theta}_{-\delta_{\boldsymbol{\xi}_{1kd}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \Gamma\left(a_{\boldsymbol{\xi}_{1kd}} + (PM/2), 1 + \frac{1}{2}\sum_{r=1}^{P}\gamma_{\boldsymbol{\xi}_{kdr1}}\xi^2_{kdr1} + \frac{1}{2}\sum_{m=2}^{M}\sum_{r=1}^{P}\gamma_{\boldsymbol{\xi}_{kdrm}}\xi^2_{kdrm}\left(\prod_{j=2}^{m}\delta_{\boldsymbol{\xi}_{jkd}}\right)\right).$$

The posterior distribution for $\delta_{\boldsymbol{\xi}_{ikd}}$, for $i = 2, \ldots, M$, $k = 1, \ldots, K$, and $d = 1, \ldots, D$ is

$$\delta_{\boldsymbol{\xi}_{ikd}}|\boldsymbol{\Theta}_{-\delta_{\boldsymbol{\xi}_{ikd}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \Gamma\Bigg(a_{\boldsymbol{\xi}_{2kd}} + (P(M-i+1)/2), 1 + \frac{1}{2}\sum_{m=i}^{M}\sum_{r=1}^{P}\gamma_{\boldsymbol{\xi}_{kdrm}}\phi^2_{kdrm}\left(\prod_{j=1;j\neq i}^{m}\delta_{\boldsymbol{\xi}_{jkd}}\right)\Bigg).$$

The posterior distribution for $a_{\boldsymbol{\phi}_{1k}}$ $(k = 1, \ldots, K)$ is not a commonly known distribution, however we have that

$$P(a_{\boldsymbol{\phi}_{1k}}|\boldsymbol{\Theta}_{-a_{\boldsymbol{\phi}_{1k}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}) \propto \frac{1}{\Gamma(a_{\boldsymbol{\phi}_{1k}})}\delta_{\boldsymbol{\phi}_{1k}}^{a_{\boldsymbol{\phi}_{1k}}-1}a_{\boldsymbol{\phi}_{1k}}^{\alpha_1-1}exp\left\{-a_{\boldsymbol{\phi}_{1k}}\beta_1\right\}.$$

Since this is not a known kernel of a distribution, we will have to use Metropolis-Hastings algorithm. Consider the proposal distribution $Q(a'_{\boldsymbol{\phi}_{1k}}|a_{\boldsymbol{\phi}_{1k}}) = \mathcal{N}\left(a_{\boldsymbol{\phi}_{1k}}, \epsilon_1\beta_1^{-1}, 0, +\infty\right)$ (Trun-

cated Normal) for some small $\epsilon_1 > 0$. Thus the probability of accepting any step is

$$A(a'_{\boldsymbol{\phi}_{1k}}, a_{\boldsymbol{\phi}_{1k}}) = \min\left\{1, \frac{P\left(a'_{\boldsymbol{\phi}_{1k}}|\boldsymbol{\Theta}_{-a'_{\boldsymbol{\phi}_{1k}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)}{P\left(a_{\boldsymbol{\phi}_{1k}}|\boldsymbol{\Theta}_{-a_{\boldsymbol{\phi}_{1k}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)} \frac{Q\left(a_{\boldsymbol{\phi}_{1k}}|a'_{\boldsymbol{\phi}_{1k}}\right)}{Q\left(a'_{\boldsymbol{\phi}_{1k}}|a_{\boldsymbol{\phi}_{1k}}\right)}\right\}.$$

Similarly for $a_{\boldsymbol{\phi}_{2k}}$ $(k = 1, \ldots, K)$, we have

$$P(a_{\boldsymbol{\phi}_{2k}}|\boldsymbol{\Theta}_{-a_{\boldsymbol{\phi}_{2k}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}) \propto \frac{1}{\Gamma(a_{\boldsymbol{\phi}_{2k}})^{M-1}} \left(\prod_{i=2}^{M} \delta_{\boldsymbol{\phi}_{ik}}^{a_{\boldsymbol{\phi}_{2k}}-1}\right) a_{\boldsymbol{\phi}_{2k}}^{\alpha_{\boldsymbol{\phi}_{2k}}-1} exp\left\{-a_{\boldsymbol{\phi}_{2k}}\beta_2\right\}.$$

We will use a similar proposal distribution, such that $Q(a'_{\boldsymbol{\phi}_{2k}}|a_{\boldsymbol{\phi}_{2k}}) = \mathcal{N}\left(a_{\boldsymbol{\phi}_{2k}}, \epsilon_2\beta_2^{-1}, 0, +\infty\right)$ for some small $\epsilon_2 > 0$. Thus the probability of accepting any step is

$$A(a'_{\boldsymbol{\phi}_{2k}}, a_{\boldsymbol{\phi}_{2k}}) = \min\left\{1, \frac{P\left(a'_{\boldsymbol{\phi}_{2k}}|\boldsymbol{\Theta}_{-a'_{\boldsymbol{\phi}_{2k}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)}{P\left(a_{\boldsymbol{\phi}_{2k}}|\boldsymbol{\Theta}_{-a_{\boldsymbol{\phi}_{2k}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)} \frac{Q\left(a_{\boldsymbol{\phi}_{2k}}|a'_{\boldsymbol{\phi}_{2k}}\right)}{Q\left(a'_{\boldsymbol{\phi}_{2k}}|a_{\boldsymbol{\phi}_{2k}}\right)}\right\}.$$

Similarly, the posterior distribution for $a_{\boldsymbol{\xi}_{1kd}}$ $(k = 1, \ldots, K$ and $d = 1, \ldots, R)$ is not a commonly known distribution, however we have that

$$P(a_{\boldsymbol{\xi}_{1kd}}|\boldsymbol{\Theta}_{-a_{\boldsymbol{\xi}_{1kd}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}) \propto \frac{1}{\Gamma(a_{\boldsymbol{\xi}_{1kd}})} \delta_{\boldsymbol{\xi}_{1kd}}^{a_{\boldsymbol{\xi}_{1kd}}-1} a_{\boldsymbol{\xi}_{1kd}}^{\alpha_1-1} exp\left\{-a_{\boldsymbol{\xi}_{1kd}}\beta_1\right\}.$$

We will use a similar proposal distribution, such that $Q(a'_{\boldsymbol{\xi}_{1kd}}|a_{\boldsymbol{\xi}_{1kd}}) = \mathcal{N}\left(a_{\boldsymbol{\xi}_{1kd}}, \epsilon_1\beta_1^{-1}, 0, +\infty\right)$ for some small $\epsilon_1 > 0$. Thus the probability of accepting any step is

$$A(a'_{\boldsymbol{\xi}_{1kd}}, a_{\boldsymbol{\xi}_{1kd}}) = \min\left\{1, \frac{P\left(a'_{\boldsymbol{\xi}_{1kd}}|\boldsymbol{\Theta}_{-a'_{\boldsymbol{\xi}_{1kd}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)}{P\left(a_{\boldsymbol{\xi}_{1kd}}|\boldsymbol{\Theta}_{-a_{\boldsymbol{\phi}_{1k}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)} \frac{Q\left(a_{\boldsymbol{\xi}_{1kd}}|a'_{\boldsymbol{\xi}_{1kd}}\right)}{Q\left(a'_{\boldsymbol{\xi}_{1kd}}|a_{\boldsymbol{\xi}_{1kd}}\right)}\right\}.$$

Similarly for $a_{\boldsymbol{\xi}_{2kd}}$ ($k = 1, \ldots, K$ and $d = 1, \ldots, R$), we have

$$P(a_{\boldsymbol{\xi}_{2kd}}|\boldsymbol{\Theta}_{-a_{\boldsymbol{\xi}_{2kd}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}) \propto \frac{1}{\Gamma(a_{\boldsymbol{\xi}_{2kd}})^{M-1}} \left(\prod_{i=2}^{M} \delta_{\boldsymbol{\xi}_{ikd}}^{a_{\boldsymbol{\xi}_{2kd}}-1} \right) a_{\boldsymbol{\xi}_{2kd}}^{\alpha_{\boldsymbol{\xi}_{2kd}}-1} exp\left\{ -a_{\boldsymbol{\xi}_{2kd}} \beta_2 \right\}.$$

We will use a similar proposal distribution, such that $Q(a'_{\boldsymbol{\xi}_{2kd}}|a_{\boldsymbol{\xi}_{2kd}}) = \mathcal{N}\left(a_{\boldsymbol{\xi}_{2kd}}, \epsilon_2 \beta_2^{-1}, 0, +\infty\right)$ for some small $\epsilon_2 > 0$. Thus the probability of accepting any step is

$$A(a'_{\boldsymbol{\xi}_{2kd}}, a_{\boldsymbol{\xi}_{2kd}}) = \min \left\{ 1, \frac{P\left(a'_{\boldsymbol{\xi}_{2kd}}|\boldsymbol{\Theta}_{-a'_{\boldsymbol{\xi}_{2kd}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)}{P\left(a_{\boldsymbol{\xi}_{2kd}}|\boldsymbol{\Theta}_{-a_{\boldsymbol{\xi}_{2kd}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)} \frac{Q\left(a_{\boldsymbol{\xi}_{2kd}}|a'_{\boldsymbol{\xi}_{2kd}}\right)}{Q\left(a'_{\boldsymbol{\xi}_{2kd}}|a_{\boldsymbol{\xi}_{2kd}}\right)} \right\}.$$

For the $\gamma_{\phi_{jrm}}$ parameters, for $j = 1, \ldots K$, $p = 1, \ldots, P$, and $m = 1, \ldots, M$, we have

$$\gamma_{\phi_{jpm}}|\boldsymbol{\Theta}_{-\gamma_{\phi_{jpm}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \Gamma\left(\frac{\nu_\gamma + 1}{2}, \frac{\phi_{jpm}^2 \tilde{\tau}_{\boldsymbol{\phi}_{mj}} + \nu_\gamma}{2} \right).$$

Similarly, for the $\gamma_{\boldsymbol{\xi}_{jdpm}}$ parameters, we have

$$\gamma_{\xi_{jrpm}}|\boldsymbol{\Theta}_{-\gamma_{\xi_{jrpm}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \Gamma\left(\frac{\nu_\gamma + 1}{2}, \frac{\xi_{jrpm}^2 \tilde{\tau}_{\boldsymbol{\xi}_{mjr}} + \nu_\gamma}{2} \right),$$

for $j = 1, \ldots, K$, $r = 1, \ldots, R$, $p = 1, \ldots, P$, and $m = 1, \ldots, M$. The posterior distribution for the $\mathbf{z}_i$ parameters are not a commonly known distribution, so we will use the Metropolis-Hastings algorithm. We know that

$$\begin{aligned} p(\mathbf{z}_i|\boldsymbol{\Theta}_{-\mathbf{z}_i}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}) \propto & \prod_{k=1}^{K} Z_{ik}^{\alpha_3 \pi_k - 1} \\ & \times \prod_{l=1}^{n_i} exp\left\{ -\frac{1}{2\sigma^2} \left(y_i(t_{il}) - \sum_{k=1}^{K} Z_{ik} \left((\boldsymbol{\nu}_k + \boldsymbol{\eta}_k \mathbf{x}'_i)' B(t_{il}) \right. \right. \right. \\ & \left. \left. \left. + \sum_{m=1}^{M} \chi_{im} \left(\boldsymbol{\phi}_{km} + \boldsymbol{\xi}_{km} \mathbf{x}'_i \right)' B(t_{il}) \right) \right)^2 \right\}. \end{aligned}$$

We will use $Q(\mathbf{z}_i'|\mathbf{z}_i) = Dir(a_{\mathbf{z}}\mathbf{z}_i)$ for some large $a_{\mathbf{z}} \in \mathbb{R}^+$ as the proposal distribution. Thus the probability of accepting a proposed step is

$$A(\mathbf{z}_i', \mathbf{z}_i) = \min\left\{1, \frac{P\left(\mathbf{z}_i'|\boldsymbol{\Theta}_{-\mathbf{z}_i}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)}{P\left(\mathbf{z}_i|\boldsymbol{\Theta}_{-\mathbf{z}_i}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)} \frac{Q\left(\mathbf{z}_i|\mathbf{z}_i'\right)}{Q\left(\mathbf{z}_i'|\mathbf{z}_i\right)}\right\}.$$

Similarly, a Gibbs update is not available for an update of the $\boldsymbol{\pi}$ parameters. We have that

$$\begin{aligned} p(\boldsymbol{\pi}|\boldsymbol{\Theta}_{-\boldsymbol{\pi}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}) \propto & \prod_{k=1}^{K} \pi_k^{c_k - 1} \\ & \times \prod_{i=1}^{N} \frac{1}{B(\alpha_3 \boldsymbol{\pi})} \prod_{k=1}^{K} Z_{ik}^{\alpha_3 \pi_k - 1}. \end{aligned}$$

Letting out proposal distribution be such that $Q(\boldsymbol{\pi}'|\boldsymbol{\pi}) = Dir(a_{\boldsymbol{\pi}}\boldsymbol{\pi})$, for some large $a_{\boldsymbol{\pi}} \in \mathbb{R}^+$, we have that our probability of accepting any proposal is

$$A(\boldsymbol{\pi}', \boldsymbol{\pi}) = \min\left\{1, \frac{P\left(\boldsymbol{\pi}'|\boldsymbol{\Theta}_{-\boldsymbol{\pi}'}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)}{P\left(\boldsymbol{\pi}|\boldsymbol{\Theta}_{-\boldsymbol{\pi}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)} \frac{Q\left(\boldsymbol{\pi}|\boldsymbol{\pi}'\right)}{Q\left(\boldsymbol{\pi}'|\boldsymbol{\pi}\right)}\right\}.$$

The posterior distribution of α_3 is also not a commonly known distribution, so we will use the Metropolis-Hastings algorithm to sample from the posterior distribution. We have that

$$\begin{aligned} p(\alpha_3|\boldsymbol{\Theta}_{-\alpha_3}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}) \propto & e^{-b\alpha_3} \\ & \times \prod_{i=1}^{N} \frac{1}{B(\alpha_3 \boldsymbol{\pi})} \prod_{k=1}^{K} Z_{ik}^{\alpha_3 \pi_k - 1}. \end{aligned}$$

Using a proposal distribution such that $Q(\alpha_3'|\alpha_3) = \mathcal{N}(\alpha_3, \sigma_{\alpha_3}^2, 0, +\infty)$ (Truncated Normal), we are left with the probability of accepting a proposed state as

$$A(\alpha_3', \alpha_3) = \min\left\{1, \frac{P\left(\alpha_3'|\boldsymbol{\Theta}_{-\alpha_3'}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)}{P\left(\alpha_3|\boldsymbol{\Theta}_{-\alpha_3}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X}\right)} \frac{Q\left(\alpha_3|\alpha_3'\right)}{Q\left(\alpha_3'|\alpha_3\right)}\right\}.$$

Let $\mathbf{P}$ be the following tridiagonal matrix:

$$\mathbf{P} = \begin{bmatrix} 1 & -1 & 0 & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & 0 & -1 & 1 \end{bmatrix}.$$

Thus, letting

$$\mathbf{B}_{\boldsymbol{\nu}_j} = \left(\tau_{\boldsymbol{\nu}_j} \mathbf{P} + \frac{1}{\sigma^2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} Z_{ij}^2 B(t_{il}) B'(t_{il}) \right)^{-1}$$

and

$$\begin{aligned} \mathbf{b}_{\boldsymbol{\nu}_j} = \frac{1}{\sigma^2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} Z_{ij} B(t_{il}) \Bigg[y_i(t_{il}) - \Bigg(\sum_{k \neq j} Z_{ik} \boldsymbol{\nu}_k' B(t_{il}) \Bigg) \\ - \left(\sum_{k=1}^{K} Z_{ik} \left[\mathbf{x}_i \boldsymbol{\eta}_k' B(t_{il}) + \sum_{m=1}^{M} \chi_{im} \left(\boldsymbol{\phi}_{kn} + \boldsymbol{\xi}_{kn} \mathbf{x}_i' \right)' B(t_{il}) \right] \right) \Bigg], \end{aligned}$$

we have that

$$\boldsymbol{\nu}_j | \boldsymbol{\Theta}_{-\boldsymbol{\nu}_j}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N} \left(\mathbf{B}_{\boldsymbol{\nu}_j} \mathbf{b}_{\boldsymbol{\nu}_j}, \mathbf{B}_{\boldsymbol{\nu}_j} \right).$$

Let $\boldsymbol{\eta}_{jd}$ denote the d^{th} column of the matrix $\boldsymbol{\eta}_j$. Thus, letting

$$\mathbf{B}_{\boldsymbol{\eta}_{jd}} = \left(\tau_{\boldsymbol{\eta}_{jd}} \mathbf{P} + \frac{1}{\sigma^2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} Z_{ij}^2 x_{id}^2 B(t_{il}) B'(t_{il}) \right)^{-1}$$

and

$$\begin{aligned} \mathbf{b}_{\boldsymbol{\eta}_{jd}} = & \frac{1}{\sigma^2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} Z_{ij} x_{id} B(t_{il}) \Bigg[y_i(t_{il}) - \Bigg(\sum_{r \neq d} Z_{ij} x_{ir} \boldsymbol{\eta}_{jr}' B(t_{il}) \Bigg) - \Bigg(\sum_{k \neq j} Z_{ik} \mathbf{x}_i \boldsymbol{\eta}_k' B(t_{il}) \Bigg) \\ & - \left(\sum_{k=1}^{K} Z_{ik} \left[\boldsymbol{\nu}_k' B(t_{il}) + \sum_{m=1}^{M} \chi_{im} \left(\boldsymbol{\phi}_{kn} + \boldsymbol{\xi}_{kn} \mathbf{x}_i' \right)' B(t_{il}) \right] \right) \Bigg], \end{aligned}$$

we have that

$$\boldsymbol{\eta}_{jd}|\boldsymbol{\Theta}_{-\boldsymbol{\eta}_{jd}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N}\left(\mathbf{B}_{\boldsymbol{\eta}_{jd}}\mathbf{b}_{\boldsymbol{\eta}_{jd}}, \mathbf{B}_{\boldsymbol{\eta}_{jd}}\right).$$

Thus we can see that we can draw samples from the posterior of the parameters controlling the mean structure using a Gibbs sampler. Similarly, we can use a Gibbs sampler to draw samples from the posterior distribution of $\tau_{\boldsymbol{\eta}_{jd}}$ and $\tau_{\boldsymbol{\nu}_j}$. We have that the posterior distributions are

$$\tau_{\boldsymbol{\nu}_j}|\boldsymbol{\Theta}_{-\tau_{\boldsymbol{\nu}_j}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \Gamma\left(\alpha_{\boldsymbol{\nu}} + P/2, \beta_{\boldsymbol{\nu}} + \frac{1}{2}\boldsymbol{\nu}_j'\mathbf{P}\boldsymbol{\nu}_j\right)$$

and

$$\tau_{\boldsymbol{\eta}_{jd}}|\boldsymbol{\Theta}_{-\tau_{\boldsymbol{\eta}_{jd}}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \Gamma\left(\alpha_{\boldsymbol{\eta}} + P/2, \beta_{\boldsymbol{\eta}} + \frac{1}{2}\boldsymbol{\eta}_{jd}'\mathbf{P}\boldsymbol{\eta}_{jd}\right),$$

for $j = 1, \ldots, K$ and $d = 1, \ldots, R$. The parameter σ^2 can be updated by using a Gibbs update. If we let

$$\beta_\sigma = \frac{1}{2}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\left(y_i(t_{il}) - \sum_{k=1}^{K} Z_{ik}\left(\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k\mathbf{x}_i'\right)' B(t_{il}) + \sum_{n=1}^{M}\chi_{in}\left(\boldsymbol{\phi}_{kn} + \boldsymbol{\xi}_{kn}\mathbf{x}_i'\right)' B(t_{il})\right)\right)^2,$$

then we have

$$\sigma^2|\boldsymbol{\Theta}_{-\sigma^2}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim IG\left(\alpha_0 + \frac{\sum_{i=1}^{N} n_i}{2}, \beta_0 + \beta_\sigma\right).$$

Lastly, we can update the χ_{im} parameters, for $i = 1, \ldots, N$ and $m = 1, \ldots, M$, using a Gibbs update. If we let

$$\begin{aligned}\mathbf{w}_{im} =& \frac{1}{\sigma^2}\left[\sum_{l=1}^{n_i}\left(\sum_{k=1}^{K} Z_{ik}\left(\boldsymbol{\phi}_{km} + \boldsymbol{\xi}_{km}\mathbf{x}_i'\right)' B(t_{il})\right)\right. \\ &\left.\left(y_i(t_{il}) - \sum_{k=1}^{K} Z_{ik}\left(\left(\boldsymbol{\nu}_k + \boldsymbol{\eta}_k\mathbf{x}_i'\right)' B(t_{il}) + \sum_{n\neq m}\chi_{in}\left(\boldsymbol{\phi}_{kn} + \boldsymbol{\xi}_{kn}\mathbf{x}_i'\right)' B(t_{il})\right)\right)\right]\end{aligned}$$

and

$$\mathbf{W}_{im}^{-1} = 1 + \frac{1}{\sigma^2} \sum_{l=1}^{n_i} \left(\sum_{k=1}^{K} Z_{ik} \left(\boldsymbol{\phi}_{km} + \boldsymbol{\xi}_{km} \mathbf{x}_i' \right)' B(t_{il}) \right)^2,$$

then we have that

$$\chi_{im} | \boldsymbol{\zeta}_{-\chi_{im}}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N}(\mathbf{W}_{im} \mathbf{w}_{im}, \mathbf{W}_{im}).$$

C.4.3 Tempered Transitions

Since we only temper the likelihood, many of the posterior distributions derived in Section C.4.2 can be utilized. Starting with the $\boldsymbol{\Phi}$ parameters, we have

$$\begin{aligned}
\left(\mathbf{m}_{\boldsymbol{\phi}_{jm}} \right)_h = & \frac{\beta_h}{(\sigma^2)_h} \sum_{i=1}^{N} \sum_{l=1}^{n_i} \Bigg(B(t_{il}) (\chi_{im})_h \Bigg(y_i(t_{il}) (Z_{ij})_h - (Z_{ij})_h^2 \left((\boldsymbol{\nu}_j)_h + (\boldsymbol{\eta}_j)_h \mathbf{x}_i' \right)' B(t_{il}) \\
& - (Z_{ij})_h^2 \sum_{n \neq m} (\chi_{in})_h (\boldsymbol{\phi}_{jn})_h' B(t_{il}) - (Z_{ij})_h^2 \sum_{n=1}^{M} (\chi_{in}) \mathbf{x}_i (\boldsymbol{\xi}_{jn})_h' B(t_{il}) \\
& - \sum_{k \neq j} Z_{ij} Z_{ik} \left[\left((\boldsymbol{\nu}_k)_h + (\boldsymbol{\eta}_k)_h \mathbf{x}_i' \right)' B(t_{il}) + \sum_{n=1}^{M} \chi_{in} \left((\boldsymbol{\phi}_{kn})_h + (\boldsymbol{\xi}_{kn})_h \mathbf{x}_i' \right)' B(t_{il}) \right] \Bigg) \Bigg),
\end{aligned}$$

and

$$\left(\mathbf{M}_{\boldsymbol{\phi}_{jm}} \right)_h^{-1} = \frac{\beta_h}{(\sigma^2)_h} \sum_{i=1}^{N} \sum_{l=1}^{n_i} \left((Z_{ij})_h^2 (\chi_{im})_h^2 B(t_{il}) B'(t_{il}) \right) + \left(\mathbf{D}_{\boldsymbol{\phi}_{jm}} \right)_h^{-1},$$

we have that

$$\left(\boldsymbol{\phi}_{jm} \right)_h | \boldsymbol{\Theta}_{-\left(\boldsymbol{\phi}_{jm} \right)_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N} \left(\left(\mathbf{M}_{\boldsymbol{\phi}_{jm}} \right)_h \left(\mathbf{m}_{\boldsymbol{\phi}_{jm}} \right)_h, \left(\mathbf{M}_{\boldsymbol{\phi}_{jm}} \right)_h \right).$$

Letting

$$\begin{aligned}\left(\mathbf{m}_{\boldsymbol{\xi}_{kdm}}\right)_h =& \frac{\beta_h}{(\sigma^2)_h}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\Bigg(B(t_{il})(\chi_{im})_h x_{id}(Z_{ik})_h\Bigg(y_i(t_{il})- \\ & \sum_{j=1}^{K}(Z_{ij})_h\left[\left((\boldsymbol{\nu}_j)_h+(\boldsymbol{\eta}_j)_h\mathbf{x}_i'\right)' B(t_{il})+\sum_{n=1}^{M}(\chi_{in})_h(\boldsymbol{\phi}_{jn})_h' B(t_{il})\right] \\ & -\sum_{(j,n,r)\neq(k,m,d)}(Z_{ij})_h(\chi_{in})_h x_{ir}(\boldsymbol{\xi}_{krn})_h' B(t_{il})\Bigg)\Bigg)\end{aligned}$$

$$\left(\mathbf{M}_{\boldsymbol{\xi}_{kdm}}\right)_h^{-1} = \frac{\beta_h}{(\sigma^2)_h}\sum_{i=1}^{N}\sum_{l=1}^{n_i}\left((Z_{ik})_h^2(\chi_{im})_h^2 x_{id}^2 B(t_{il})B'(t_{il})\right)+\left(\mathbf{D}_{\boldsymbol{\xi}_{kdm}}\right)_h^{-1},$$

we have that

$$(\boldsymbol{\xi}_{kdm})_h\,|\boldsymbol{\Theta}_{-(\boldsymbol{\xi}_{kdm})_h},\mathbf{Y}_1,\ldots,\mathbf{Y}_N,\mathbf{X}\sim\mathcal{N}\left(\left(\mathbf{M}_{\boldsymbol{\xi}_{kdm}}\right)_h\left(\mathbf{m}_{\boldsymbol{\xi}_{kdm}}\right)_h,\left(\mathbf{M}_{\boldsymbol{\xi}_{kdm}}\right)_h\right).$$

As in the untempered case, we have that the posterior distribution $\mathbf{Z}$ parameters under the tempered likelihood is not a commonly known distribution. Therefore, we will use the Metropolis-Hastings algorithm. We have that

$$\begin{aligned}p((\mathbf{z}_i)_h|\boldsymbol{\Theta}_{-(\mathbf{z}_i)_h},\mathbf{Y}_1,\ldots,\mathbf{Y}_N,\mathbf{X}) \propto & \prod_{k=1}^{K}(Z_{ik})_h^{(\alpha_3)_h(\pi_k)_h-1} \\ & \times\prod_{l=1}^{n_i}exp\left\{-\frac{\beta_h}{2(\sigma^2)_h}\left(y_i(t_{il})-\sum_{k=1}^{K}(Z_{ik})_h\left(\left((\boldsymbol{\nu}_k)_h+(\boldsymbol{\eta}_k)_h\mathbf{x}_i'\right)' B(t_{il})\right.\right.\right. \\ & \left.\left.\left.+\sum_{m=1}^{M}(\chi_{im})_h\left((\boldsymbol{\phi}_{km})_h+(\boldsymbol{\xi}_{km})_h\mathbf{x}_i'\right)' B(t_{il})\right)\right)^2\right\}.\end{aligned}$$

We will use $Q((\mathbf{z}_i)_h'|(\mathbf{z}_i)_h)=Dir(a_{\mathbf{z}}(\mathbf{z}_i)_h)$ for some large $a_{\mathbf{z}}\in\mathbb{R}^+$ as the proposal distribution.

Thus the probability of accepting a proposed step is

$$A((\mathbf{z}_i)'_h,(\mathbf{z}_i)_h)=\min\left\{1,\frac{P\left((\mathbf{z}_i)'_h|\boldsymbol{\Theta}_{-(\mathbf{z}_i)'_h},\mathbf{Y}_1,\ldots,\mathbf{Y}_N,\mathbf{X}\right)}{P\left((\mathbf{z}_i)_h|\boldsymbol{\Theta}_{-(\mathbf{z}_i)_h},\mathbf{Y}_1,\ldots,\mathbf{Y}_N,\mathbf{X}\right)}\frac{Q\left((\mathbf{z}_i)_h|(\mathbf{z}_i)'_h\right)}{Q\left((\mathbf{z}_i)'_h|(\mathbf{z}_i)_h\right)}\right\}.$$

Letting

$$\left(\mathbf{B}_{\boldsymbol{\nu}_j}\right)_h=\left(\left(\tau_{\boldsymbol{\nu}_j}\right)_h\mathbf{P}+\frac{\beta_h}{(\sigma^2)_h}\sum_{i=1}^{N}\sum_{l=1}^{n_i}(Z_{ij})_h^2B(t_{il})B'(t_{il})\right)^{-1}$$

and

$$\begin{aligned}\left(\mathbf{b}_{\boldsymbol{\nu}_j}\right)_h=&\frac{\beta_h}{(\sigma^2)_h}\sum_{i=1}^{N}\sum_{l=1}^{n_i}(Z_{ij})_hB(t_{il})\left[y_i(t_{il})-\left(\sum_{k\neq j}(Z_{ik})_h(\boldsymbol{\nu}'_k)_hB(t_{il})\right)\right.\\&\left.-\left(\sum_{k=1}^{K}(Z_{ik})\left[\mathbf{x}_i(\boldsymbol{\eta}_k)'_hB(t_{il})+\sum_{m=1}^{M}(\chi_{im})_h\left((\boldsymbol{\phi}_{kn})_h+(\boldsymbol{\xi}_{kn})_h\mathbf{x}'_i\right)'B(t_{il})\right]\right)\right],\end{aligned}$$

we have that

$$(\boldsymbol{\nu}_j)_h|\boldsymbol{\Theta}_{-(\boldsymbol{\nu}_j)_h},\mathbf{Y}_1,\ldots,\mathbf{Y}_N,\mathbf{X}\sim\mathcal{N}\left(\left(\mathbf{B}_{\boldsymbol{\nu}_j}\right)_h\left(\mathbf{b}_{\boldsymbol{\nu}_j}\right)_h,\left(\mathbf{B}_{\boldsymbol{\nu}_j}\right)_h\right).$$

Let $\left(\boldsymbol{\eta}_{jd}\right)_h$ denote the d^{th} column of the matrix $(\boldsymbol{\eta}_j)_h$. Thus, letting

$$\left(\mathbf{B}_{\boldsymbol{\eta}_{jd}}\right)_h=\left(\left(\tau_{\boldsymbol{\eta}_{jd}}\right)_h\mathbf{P}+\frac{\beta_h}{(\sigma^2)_h}\sum_{i=1}^{N}\sum_{l=1}^{n_i}(Z_{ij})_h^2x_{id}^2B(t_{il})B'(t_{il})\right)^{-1}$$

and

$$\begin{aligned}\left(\mathbf{b}_{\boldsymbol{\eta}_{jd}}\right)_h=&\frac{\beta_h}{(\sigma^2)_h}\sum_{i=1}^{N}\sum_{l=1}^{n_i}(Z_{ij})_hx_{id}B(t_{il})\left[y_i(t_{il})-\left(\sum_{r\neq d}(Z_{ij})_hx_{ir}(\boldsymbol{\eta}_{jr})'_hB(t_{il})\right)\right.\\&-\left(\sum_{k\neq j}(Z_{ik})_h\mathbf{x}_i(\boldsymbol{\eta}_k)'_hB(t_{il})\right)\\&\left.-\left(\sum_{k=1}^{K}(Z_{ik})_h\left[(\boldsymbol{\nu}_k)'_hB(t_{il})+\sum_{m=1}^{M}(\chi_{im})_h\left((\boldsymbol{\phi}_{kn})_h+(\boldsymbol{\xi}_{kn})_h\mathbf{x}'_i\right)'B(t_{il})\right]\right)\right],\end{aligned}$$

we have that

$$\left(\boldsymbol{\eta}_{jd}\right)_h | \boldsymbol{\Theta}_{-\left(\boldsymbol{\eta}_{jd}\right)_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N}\left(\left(\mathbf{B}_{\boldsymbol{\eta}_{jd}}\right)_h \left(\mathbf{b}_{\boldsymbol{\eta}_{jd}}\right)_h, \left(\mathbf{B}_{\boldsymbol{\eta}_{jd}}\right)_h\right).$$

If we let

$$(\beta_\sigma)_h = \frac{\beta_h}{2} \sum_{i=1}^{N} \sum_{l=1}^{n_i} \left(y_i(t_{il}) - \sum_{k=1}^{K} (Z_{ik})_h \left(\left((\boldsymbol{\nu}_k)_h + (\boldsymbol{\eta}_k)_h \mathbf{x}_i' \right)' B(t_{il}) + \sum_{n=1}^{M} (\chi_{in})_h \left((\boldsymbol{\phi}_{kn})_h + (\boldsymbol{\xi}_{kn})_h \mathbf{x}_i' \right)' B(t_{il}) \right) \right)^2,$$

then we have

$$(\sigma^2)_h | \boldsymbol{\Theta}_{-(\sigma^2)_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim IG\left(\alpha_0 + \frac{\beta_h \sum_{i=1}^{N} n_i}{2}, \beta_0 + (\beta_\sigma)_h\right).$$

Lastly, we can update the χ_{im} parameters, for $i = 1, \ldots, N$ and $m = 1, \ldots, M$, using a Gibbs update. If we let

$$(\mathbf{w}_{im})_h = \frac{\beta_h}{(\sigma^2)_h} \left[\sum_{l=1}^{n_i} \left(\sum_{k=1}^{K} (Z_{ik})_h \left((\boldsymbol{\phi}_{km})_h + (\boldsymbol{\xi}_{km})_h \mathbf{x}_i' \right)' B(t_{il}) \right) \left(y_i(t_{il}) - \sum_{k=1}^{K} (Z_{ik})_h \left(\left((\boldsymbol{\nu}_k)_h + (\boldsymbol{\eta}_k)_h \mathbf{x}_i' \right)' B(t_{il}) + \sum_{n \neq m} (\chi_{in})_h \left((\boldsymbol{\phi}_{kn})_h + (\boldsymbol{\xi}_{kn})_h \mathbf{x}_i' \right)' B(t_{il}) \right) \right) \right]$$

and

$$(\mathbf{W}_{im})_h^{-1} = 1 + \frac{\beta_h}{\sigma^2} \sum_{l=1}^{n_i} \left(\sum_{k=1}^{K} (Z_{ik})_h \left((\boldsymbol{\phi}_{km})_h + (\boldsymbol{\xi}_{km})_h \mathbf{x}_i' \right)' B(t_{il}) \right)^2,$$

then we have that

$$(\chi_{im})_h | \boldsymbol{\zeta}_{-(\chi_{im})_h}, \mathbf{Y}_1, \ldots, \mathbf{Y}_N, \mathbf{X} \sim \mathcal{N}\left((\mathbf{W}_{im})_h (\mathbf{w}_{im})_h, (\mathbf{W}_{im})_h \right).$$

Bibliography

Hirotugu Akaike. A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723, 1974.

David H Alexander, John Novembre, and Kenneth Lange. Fast model-based estimation of ancestry in unrelated individuals. *Genome research*, 19(9):1655–1664, 2009.

D American Psychiatric Association, American Psychiatric Association, et al. *Diagnostic and statistical manual of mental disorders: DSM-5*, volume 5. American psychiatric association Washington, DC, 2013.

Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models–going beyond svd. In *2012 IEEE 53rd annual symposium on foundations of computer science*, pages 1–10. IEEE, 2012.

Yossi Azar, Amos Fiat, Anna Karlin, Frank McSherry, and Jared Saia. Spectral analysis of data. In *Proceedings of the thirty-third annual ACM symposium on Theory of computing*, pages 619–626, 2001.

Gundula Behrens, Nial Friel, and Merrilee Hurn. Tuning tempered transitions. *Statistics and computing*, 22(1):65–78, 2012.

JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, AFM Smith, and M West. Bayesian factor regression models in the "large p, small n" paradigm. *Bayesian statistics*, 7:733–742, 2003.

Anirban Bhattacharya and David B Dunson. Sparse bayesian infinite factor models. *Biometrika*, pages 291–306, 2011.

Norbert Binkiewicz, Joshua T Vogelstein, and Karl Rohe. Covariate-assisted spectral clustering. *Biometrika*, 104(2):361–377, 2017.

Christopher M Bishop and Markus Svenskn. Bayesian hierarchical mixtures of experts. In *Proceedings of the Nineteenth conference on Uncertainty in Artificial Intelligence*, pages 57–64, 2002.

D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *JMLR*, 2003.

T Broderick, J Pitman, and MI Jordan. Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, 8(4):801–836, 2013.

Babette A Brumback and John A Rice. Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association*, 93(443): 961–976, 1998.

Carlos M Carvalho, Jeffrey Chang, Joseph E Lucas, Joseph R Nevins, Quanli Wang, and Mike West. High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008.

Gilles Celeux, Florence Forbes, Christian P Robert, and D Mike Titterington. Deviance information criteria for missing data models. *Bayesian analysis*, 1(4):651–673, 2006.

Ming-Hui Chen, Qi-Man Shao, and Joseph G Ibrahim. *Monte Carlo methods in Bayesian computation*. Springer Science & Business Media, 2012.

Yinyin Chen, Shishuang He, Yun Yang, and Feng Liang. Learning topic models: Identifiability and finite-sample analysis. *Journal of the American Statistical Association*, pages 1–16, 2022.

Jeng-Min Chiou and Pai-Ling Li. Functional clustering and identifying substructures of longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(4):679–699, 2007.

Taeryon Choi and Mark J. Schervish. On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, 98(10):1969–1987, 2007. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2007.01.004.

Ciprian M Crainiceanu, David Ruppert, Raymond J Carroll, Adarsh Joshi, and Billy Goodner. Spatially adaptive bayesian penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics*, 16(2):265–288, 2007.

Geraldine Dawson, Laura Grofer Klinger, Heracles Panagiotides, Arthur Lewy, and Paul Castelloe. Subgroups of autistic children based on social behavior display distinct patterns of brain activity. *Journal of abnormal child psychology*, 23(5):569–583, 1995.

Emilie Devijver. Finite mixture regression: a sparse variable selection by model selection for clustering. 2015.

Abigail Dickinson, Charlotte DiStefano, Damla Senturk, and Shafali Spurling Jeste. Peak alpha frequency is a neural marker of cognitive function across the autism spectrum. *European Journal of Neuroscience*, 47(6):643–651, 2018.

David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? *Advances in neural information processing systems*, 16, 2003.

Dirk Eddelbuettel and Conrad Sanderson. Rcpparmadillo: Accelerating r with high-performance c++ linear algebra. *Computational Statistics & Data Analysis*, 71:1054–1063, 2014.

Fifth Edition. Diagnostic and statistical manual of mental disorders. *Am Psychiatric Assoc*, 21(21):591–643, 2013.

E Erosheva, S Fienberg, and J Lafferty. Mixed membership models of scientific publications. *PNAS*, 2004.

Julian J Faraway. Regression analysis for a functional response. *Technometrics*, 39(3):254–261, 1997.

Susana Faria and Gilda Soromenho. Fitting mixtures of linear regressions. *Journal of Statistical Computation and Simulation*, 80(2):201–225, 2010.

Pascal Fries. A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in cognitive sciences*, 9(10):474–480, 2005.

Scott Gaffney and Padhraic Smyth. Trajectory clustering with mixtures of regression models. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 63–72, 1999.

April Galyardt. *Interpreting mixed membership: Implications of Erosheva's representation theorem.*, chapter 3. CRC Press, 2014.

Zoubinm Ghahramani, Shakir Mohamed, and Katherine Heller. *A Simple and General Exponential Family Framework for Partial Membership and Factor Analysis*, chapter 4. CRC Press, 2014.

Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

J. Ghosh and R. Ramamoorthi. Bayesian nonparametrics. *Springer Series in Statistics*, 01 2003.

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

Jeff Goldsmith, Fabian Scheipl, Lei Huang, Julia Wrobel, J Gellar, J Harezlak, MW McLean, B Swihart, L Xiao, C Crainiceanu, et al. Refund: Regression with functional data. *R package version 0.1-16*, 572, 2016.

Gene H Golub and Charles F Van Loan. *Matrix computations*. The Johns Hopkins University Press, 2013.

TL Griffiths and Z Ghahramani. The Indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12:1185–1224, 2011.

Amiram Grinvald and Rina Hildesheim. Vsdi: a new era in functional imaging of cortical dynamics. *Nature Reviews Neuroscience*, 5(11):874–885, 2004.

Jonathan Gruhl and Elena Erosheva. *A Tale of Two (Types of) Memberships: Comparing Mixed and Partial Membership with a Continuous Data Example*, chapter 2. CRC Press, 2014.

Bettina Grün, Friedrich Leisch, et al. Applications of finite mixtures of regression models. *URL: http://cran. r-project. org/web/packages/flexmix/vignettes/regression-examples. pdf*, 2007.

Sharmistha Guha and Rajarshi Guhaniyogi. Bayesian covariate-dependent clustering of undirected networks with brain-imaging data. 2022.

Saskia Haegens, Helena Cousijn, George Wallis, Paul J Harrison, and Anna C Nobre. Inter- and intra-individual variability in alpha peak frequency. *Neuroimage*, 92:46–55, 2014.

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

Katherine A Heller, Sinead Williamson, and Zoubin Ghahramani. Statistical models for partial membership. In *Proceedings of the 25th International Conference on Machine learning*, pages 392–399, 2008.

Nathaniel E. Helwig. *eegkit: Toolkit for Electroencephalography Data*, 2018. URL `https://CRAN.R-project.org/package=eegkit`. R package version 1.0-4.

Christiana Hennig, Marina Meila, Fionn Murtagh, and roberto Rocci. *Handbook of Cluster Analysis*. CRC Press, 2015.

Jason Hou-Liu and Ryan P Browne. Chimeral clustering. *Journal of Classification*, pages 1–20, 2022.

Kejun Huang, Xiao Fu, and Nikolaos D Sidiropoulos. Anchor-free correlated topic modeling: Identifiability and algorithm. *Advances in Neural Information Processing Systems*, 29, 2016.

Sangwon Hyun, Mattias Rolf Cape, Francois Ribalet, and Jacob Bien. Modeling cell populations measured by flow cytometry with covariates using sparse mixture of regressions. *The Annals of Applied Statistics*, 17(1):357–377, 2023.

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

Gareth M James and Catherine A Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462):397–408, 2003.

Byoungwook Jang and Alfred Hero. Minimum volume topic modeling. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3013–3021. PMLR, 2019.

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

Abbas Khalili and Jiahua Chen. Variable selection in finite mixture of regression models. *Journal of the american Statistical association*, 102(479):1025–1038, 2007.

Wolfgang Klimesch, Paul Sauseng, and Simon Hanslmayr. Eeg alpha oscillations: the inhibition–timing hypothesis. *Brain research reviews*, 53(1):63–88, 2007.

Michael R Kosorok and Eric B Laber. Precision medicine. *Annual review of statistics and its application*, 6:263–286, 2019.

Daniel R Kowal and Daniel C Bourgeois. Bayesian function-on-scalars regression for high-dimensional data. *Journal of Computational and Graphical Statistics*, 29(3):629–638, 2020.

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

Robert T Krafty, Phyllis A Gimotty, David Holtz, George Coukos, and Wensheng Guo. Varying coefficient model with unknown within-subject covariance for analysis of tumor growth curves. *Biometrics*, 64(4):1023–1031, 2008.

Richard L Kravitz, Naihua Duan, and Joel Braslow. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *The Milbank Quarterly*, 82(4):661–687, 2004.

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

Paul O Lewis, Wangang Xie, Ming-Hui Chen, Yu Fan, and Lynn Kuo. Posterior predictive bayesian phylogenetic model selection. *Systematic biology*, 63(3):309–321, 2014.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2):129–137, 1982.

Catherine Lord, Mayada Elsabbagh, Gillian Baird, and Jeremy Veenstra-Vanderweele. Autism spectrum disorder. *The Lancet*, 392(10146):508–520, 2018.

Nicholas Marco, Damala Şentürk, Shafali Jeste, Charlotte DiStefano, Abigail Dickinson, and Donatello Telesca. Functional partial membership models. *arXiv preprint arXiv:2206.12084*, 2022a.

Nicholas Marco, Damla Şentürk, Shafali Jeste, Charlotte DiStefano, Abigail Dickinson, and Donatello Telesca. Functional mixed membership models, 2022b. URL `https://arxiv.org/abs/2206.12084`.

Nicholas Marco, Damla Şentürk, Shafali Jeste, Charlotte DiStefano, Abigail Dickinson, and Donatello Telesca. Flexible regularized estimation in high-dimensional mixed membership models, 2022c. URL `https://arxiv.org/abs/2212.06906`.

Kevin McEvoy, Kyle Hasenstab, Damla Senturk, Andrew Sanders, and Shafali S Jeste. Physiologic artifacts in resting state oscillations in young children: methodological considerations for noisy data. *Brain imaging and behavior*, 9(1):104–114, 2015.

Geoffrey J McLachlan and Kaye E Basford. *Mixture models: Inference and applications to clustering*, volume 38. M. Dekker New York, 1988.

Geoffrey J McLachlan and David Peel. *Finite mixture models*. John Wiley & Sons, 2004.

Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual review of statistics and its application*, 6:355–378, 2019.

Frank McSherry. Spectral partitioning of random graphs. In *Proceedings 42nd IEEE Symposium on Foundations of Computer Science*, pages 529–537. IEEE, 2001.

Volodymyr Melnykov and Ranjan Maitra. Finite mixture models and model-based clustering. *Statistics Surveys*, 4:80–116, 2010.

Jeffrey S Morris. Functional regression. *Annual Review of Statistics and Its Application*, 2: 321–359, 2015.

Jeffrey S Morris and Raymond J Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(2):179–199, 2006.

Peter Müller, Fernando Quintana, and Gary L Rosner. A product partition model with regression on covariates. *Journal of Computational and Graphical Statistics*, 20(1):260–278, 2011.

XuanLong Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. *The Annals of Statistics*, 41(1):370–400, 2013.

Garritt L Page and Fernando A Quintana. Spatial product partition models. 2016.

Christos H Papadimitriou, Hisao Tamaki, Prabhakar Raghavan, and Santosh Vempala. Latent semantic indexing: A probabilistic analysis. In *Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems*, pages 159–168, 1998.

Ju-Hyun Park and David B Dunson. Bayesian generalized product partition model. *Statistica Sinica*, pages 1203–1226, 2010.

Joel S Parker, Michael Mullins, Maggie CU Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology*, 27(8): 1160, 2009.

Sonia Petrone, Michele Guindani, and Alan E Gelfand. Hybrid dirichlet mixture models for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(4):755–782, 2009.

LI Pettit. The conditional predictive ordinate for the normal distribution. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(1):175–184, 1990.

Aleix Prat, Estela Pineda, Barbara Adamo, Patricia Galván, Aranzazu Fernández, Lydia Gaba, Marc Díez, Margarita Viladot, Ana Arance, and Montserrat Muñoz. Clinical im-

plications of the intrinsic molecular subtypes of breast cancer. *The Breast*, 24:S26–S35, 2015.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Natalya Pya Arnqvist, Per Arnqvist, and Sara Sjöstedt de Luna. fdamocca: Model-based clustering for functional data with covariates. r package version 0.1-0. 2021.

Li-Xuan Qin and Steven G Self. The clustering of regression models method with applications in gene expression data. *Biometrics*, 62(2):526–533, 2006.

J. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer, 2005a. ISBN 9780387400808. URL `https://books.google.com/books?id=mU3dop5wY\_4C`.

JO Ramsay and BW Silverman. Principal components analysis for functional data. *Functional data analysis*, pages 147–172, 2005b.

Michael Reed and Barry Simon. *Methods of modern mathematical physics*, volume 1. Elsevier, 1972.

Philip T Reiss, Lei Huang, and Maarten Mennes. Fast function-on-scalar regression with penalized basis expansions. *The international journal of biostatistics*, 6(1), 2010.

John A Rice and Bernard W Silverman. 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, 1991.

Gareth O Roberts and Jeffrey S Rosenthal. Coupling and ergodicity of adaptive markov chain monte carlo algorithms. *Journal of applied probability*, 44(2):458–475, 2007.

EI Rodríguez-Martínez, FJ Ruiz-Martínez, CI Barriga Paulino, and Carlos M Gómez. Frequency shift in topography of spontaneous brain rhythms from childhood to adulthood. *Cognitive neurodynamics*, 11(1):23–33, 2017.

Kathryn Roeder and Larry Wasserman. Practical bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92(439):894–902, 1997.

Judith Rousseau and Kerrie Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(5):689–710, 2011.

Enrique H Ruspini, James C Bezdek, and James M Keller. Fuzzy clustering: A historical perspective. *IEEE Computational Intelligence Magazine*, 14(1):45–55, 2019.

Aaron W Scheffler, Donatello Telesca, Catherine A Sugar, Shafali Jeste, Abigail Dickinson, Charlotte DiStefano, and Damla Şentürk. Covariate-adjusted region-referenced generalized functional linear model for eeg data. *Statistics in medicine*, 38(30):5587–5602, 2019.

Gideon Schwarz. Estimating the dimension of a model. *The annals of statistics*, pages 461–464, 1978.

J Shamshoian, D Senturk, S Jeste, and D Telesca. Bayesian analysis of longitudinal and multidimensional functional data. *Biostatistics*, 23(2):558–573, 2022.

Han Lin Shang. A survey of functional principal component analysis. *AStA Advances in Statistical Analysis*, 98:121–142, 2014.

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

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

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

Tatiana A Stroganova, Elena V Orekhova, and Irina N Posikera. Eeg alpha rhythm in infants. *Clinical neurophysiology*, 110(6):997–1012, 1999.

Hua Tang, Jie Peng, Pei Wang, and Neil J Risch. Estimation of individual admixture: analytical and study design considerations. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society*, 28(4):289–301, 2005.

Adrienne L Tierney, Laurel Gabard-Durnam, Vanessa Vogel-Farley, Helen Tager-Flusberg, and Charles A Nelson. Developmental trajectories of resting eeg power: an endophenotype of autism spectrum disorder. *PloS one*, 7(6):e39127, 2012.

Stef Van Buuren and Karin Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of statistical software*, 45:1–67, 2011.

Haidong Wang, Mohsen Naghavi, Christine Allen, Ryan M Barber, Zulfiqar A Bhutta, Austin Carter, Daniel C Casey, Fiona J Charlson, Alan Zian Chen, Matthew M Coates, et al. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the global burden of disease study 2015. *The lancet*, 388(10053):1459–1544, 2016.

Yanxun Xu, Peter Müller, and Donatello Telesca. Bayesian inference for latent biologic structure with determinantal point processes (dpp). *Biometrics*, 72(3):955–964, 2016.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American statistical association*, 100(470):577–590, 2005.

Fang Yao, Yuejiao Fu, and Thomas CM Lee. Functional mixture regression. *Biostatistics*, 12(2):341–353, 2011.

LA Zadeth. Fuzzy sets. *Information and control*, 8(3):338–353, 1965.