Bayesian hierarchical modeling and its applications to clustering and data privacy preservation

by

Peng Yu

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Statistics

Department of Mathematical and Statistical Sciences
University of Alberta

Abstract

The evolution of data acquisition technologies and the exponential growth in computing capabilities have inaugurated an epoch wherein researchers are empowered to procure data of unprecedented dimensionality and complexity. Simultaneously, Bayesian hierarchical models distinguish themselves as promising methodologies, offering versatile frameworks for modeling and addressing intricate data-driven challenges. This thesis harnesses the power of this advanced Bayesian statistical framework to explore innovative solutions in the realms of neuroimaging data interpretation, mental health assessment, and differentially private data analysis. The work is organized into three distinct parts, each dedicated to a specific application of Bayesian hierarchical modeling, reflecting its capacity to tackle diverse analytical problems. The first section of the thesis concentrates on the analysis of electroencephalogram (EEG) data, leveraging Bayesian hierarchical models to uncover latent structures and patterns within the complex signals. This part introduces a groundbreaking approach to EEG data analysis, emphasizing the model's ability to discern intricate neural activity patterns that elude traditional analysis techniques. By applying this novel methodology to EEG datasets, the study not only demonstrates the model's superior analytical prowess but also highlights its potential to revolutionize our understanding of neural dynamics, offering new insights into brain function and disorder diagnostics. In the second segment, attention shifts to the Hamilton Depression Rating Scale (HAMD), a widely recognized metric for assessing depression severity. Here, Bayesian hierarchical models are employed to analyze HAMD data, aiming to identify latent subgroups among patients and predict treatment outcomes more accurately. This section showcases

the application of the model to clinical trial data, revealing its capability to enhance the precision of depression severity assessment and to inform personalized treatment strategies. The use of Bayesian hierarchical models in this context exemplifies the model's adaptability and its potential to contribute meaningfully to the field of mental health. The final part of the thesis addresses the critical issue of data privacy, particularly through the lens of differential privacy (DP). It presents an innovative integration of DP principles within the Bayesian hierarchical modeling framework to safeguard individual privacy in data analysis. This approach not only demonstrates a novel method for achieving privacy-preserving but also improves the efficiency of the statistical inference. By weaving together these diverse applications—ranging from EEG data analysis and mental health assessment to privacy preservation—the thesis underscores the versatility and power of Bayesian hierarchical modeling. Each section, grounded in rigorous theoretical derivations and validated through extensive simulations and real-data applications, contributes to the advancement of statistical analysis in its respective field. Collectively, this thesis not only enriches our understanding of Bayesian hierarchical modeling but also opens new avenues for research and application of Bayesian hierarchical models in neuroscience, mental health, and data privacy.

Preface

The work contained within this thesis has been greatly enriched through collaborative efforts with esteemed colleagues and mentors. This work has benefited from the expertise and support of Dr. Bei Jiang and Dr. Linglong Kong from the University of Alberta, Dr. Zhihua Su from the University of Florida. Assistance is also provided by colleagues like Kaiqiong Zhao from the University of Alberta. The development and execution of the model formulation, as well as the data analysis presented in Chapters 2, 3, and 4, represent my original contributions to the field of statistical analysis.

Chapter 2 of the thesis is submitted and accepted by Statistics and Its Interface as Peng Yu, Kaiqiong Zhao, Bei Jiang, Eva Petkova, Thaddeus Tarpey and R. Todd Ogden, "Associations Between EEG-Defined Subgroups and Antidepressant Response: A Joint Mixture of Probabilistic Multilinear Principle Component Analysis Modeling Approach". Dr. Bei Jiang is the corresponding author. Eva Petkova, Thaddeus Tarpey, and R. Todd Ogden are involved not only as data providers but also in manuscript editing in this study. Kaiqiong Zhao contributed to manuscript edits.

Chapter 3 of the thesis is submitted and under review by the International Conference on Machine Learning (ICML) as Peng Yu and Bei Jiang, "Bayesian Envelope-Based Clustering Model with Non-Ignorable Missingness". Dr. Bei Jiang is the corresponding author.

Chapter 4 of the thesis is submitted and under review by the International Conference on Machine Learning (ICML) too as Peng Yu and Bei Jiang, "Differentially Private Bayesian Envelope Regression". Dr. Bei Jiang is the corresponding author.

Acknowledgements

I wish to express my profound gratitude to my principal supervisor, Dr. Bei Jiang, for her invaluable guidance, patience, and support over the past seven years. Her mentorship was pivotal in shaping my understanding and approach to scientific research. The qualities of a commendable researcher, as demonstrated by Dr. Jiang, have been instrumental in my academic development.

Additionally, I extend my sincere thanks to my academic mentors, Dr. Linglong Kong, Dr. Ivan Mizera, and Dr. Rohana Karunamuni. Their expertise and insights in the field of statistics have significantly contributed to my growth and understanding in this domain, guiding me adeptly through the complexities of my Ph.D. studies. My acknowledgment would be incomplete without mentioning the esteemed members of my dissertation examination committee, whose rigorous evaluation and constructive feedback have been crucial in refining my research.

I am also indebted to my colleagues and friends for their unwavering support and encouragement throughout this journey. Their camaraderie has been a source of strength and motivation. Lastly, heartfelt thanks go to my family—my parents and my sister—for their unconditional love and support. Their enduring belief in my capabilities has been a constant source of encouragement.

This thesis is not only a reflection of my work but also a testament to the collective effort and dedication of all those who have supported me throughout this journey. I am deeply grateful for their contributions.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction and Overview of the Thesis

The relentless advancement of data acquisition technologies and the exponential growth of computing capabilities have ushered in an era where researchers are endowed with the ability to gather data of unprecedented dimensionality and complexity. This evolution has been observed across various fields, encompassing genomic sequencing, neuroimaging techniques such as electroencephalogram (EEG), functional magnetic resonance imaging (fMRI), environmental monitoring, and social media analytics. At the heart of statistical inference lie the critical processes of model estimation and interpretation. It is a principle of statistical science that as the dimensionality of data swells, so does the requisite volume of data needed to sustain a specific level of statistical accuracy. This principle often collides with practical limitations, rendering the accumulation of necessary data volumes prohibitively expensive or logistically unfeasible, as seen in exhaustive clinical trials.

Beyond the challenge of dimensionality, the multifaceted nature of modern datasets introduces a complex array of challenges that extend well beyond the concerns associated with high-dimensional data. As data grows not only in size but in complexity, capturing the nuanced relationships and patterns within it requires advanced analytical techniques. High-dimensional data often contains intertwined layers of information, where variables may exhibit intricate correlations or redundancies, and ob-

servations themselves may be influenced by hierarchical or nested structures. These characteristics complicate the task of discerning the underlying signals amidst the noise, demanding approaches that can untangle these complexities while preserving the essence of the data. Simultaneously, the challenge of ensuring the integrity and utility of analysis in the face of missing or biased data points requires meticulous methodological considerations.

In response to these multifaceted challenges, Bayesian hierarchical models offer a strategic avenue for simplifying complex data into more tractable forms. Their inherent flexibility and the foundational principle of integrating prior knowledge with observed data make them uniquely suited to tackle the complexities presented. The hierarchical construct can be ingeniously divided into three principal echelons: the Data Level, where direct modeling of observations takes place and individual variability is meticulously accounted for; the Parameter Level, which models the parameters that govern the data-level distributions, capturing group or category-specific variability; and the Hyperparameter Level, where hyperparameters controlling the distributions at the parameter level are specified. Such a hierarchical arrangement is instrumental not only in modeling complex data structures with inherent nested or multilevel characteristics but also in enhancing the precision of estimates through the statistical principle of "borrowing strength" across groups or levels. The applicability of Bayesian hierarchical models spans a vast array of fields, including but not limited to ecology, education, psychology, and medical research, thereby underscoring their pivotal role in tackling multifaceted statistical challenges. This dissertation is dedicated to uncovering the profound capabilities of Bayesian hierarchical models in two critical domains: clustering and data privacy preservation. It is through an in-depth exploration that this work aims to shed light on how Bayesian hierarchical models not only proficiently identify inherent groupings within complex datasets but also provide frameworks for gaining efficiency while preserving the privacy of sensitive information.

The proposed Bayesian hierarchical clustering framework application is specifically emphasized within the domain of antidepressant mental health research. It is widely recognized that mental health, constituting a fundamental facet of comprehensive well-being, has garnered escalating attention within the realm of global health research. Among various mental health issues, depression stands out due to its widespread prevalence and profound impact on individuals' lives. The World Health Organization estimates that depression affects over 264 million people worldwide, making it a leading cause of disability. This high prevalence underscores the urgency of developing effective treatments and understanding the underlying mechanisms of depression. Clinical antidepressant studies in the realm of mental health research are laden with a myriad of challenges, primarily stemming from the heterogeneous nature of depression as a disorder. Patients with depression display a vast spectrum of symptoms and exhibit varied responses to treatment, presenting a significant challenge in the analysis and interpretation of clinical data. This heterogeneity necessitates the use of advanced and sophisticated statistical methodologies to ensure accurate results [1]. Another prominent challenge in these studies is the prevalence of missing data, particularly in longitudinal research designs. Such missing data can introduce substantial bias, potentially compromising the validity and reliability of the study findings [2]. This issue is exacerbated in clinical trials involving antidepressants, where patient dropout rates are often high due to side effects or lack of efficacy.

The first project involving clustering in this thesis pivots towards the inherent heterogeneity in patient responses to antidepressants. Major depressive disorder, as a biologically diverse and etiologically complex syndrome, manifests in varied symptoms and treatment responses. This diversity necessitates the use of advanced statistical models to accurately capture and analyze the multifaceted nature of depression. Latent mixture models are valuable tools in identifying latent subgroups within a population based on observable characteristics. This project extends these models to incorporate matrix-variate data, enhancing their capability to handle complex,

multi-dimensional datasets commonly encountered in clinical studies. The extension is inspired by a antidepressant study aimed at examining patient heterogeneity based on baseline (pre-treatment) electroencephalograph (EEG) data and its association with antidepressant response. A distinctive feature of this project is the development of a three-level structure model. The first level deals with the uncertainty of latent class membership in matrix-variate EEG data, using a multinomial logistic model. The second level assumes that class-specific EEG data follow a probabilistic multilinear principal component analysis model. The third level establishes the association between baseline EEG and antidepressant response under the conditional independence given latent class membership. This comprehensive approach allows for a deeper understanding of the relationship between EEG patterns and treatment outcomes, offering potential insights into personalized treatment strategies for depression. The application of this model in the motivating study led to the identification of distinct patient subpopulations, differentiated by their baseline EEG patterns and varied responses to antidepressant treatment. This level of granularity in understanding patient subgroups is a significant advancement over existing clustering methods, offering a more targeted approach to treating depression. The second project involving clustering shifts focus to a pivotal tool in antidepressant study, the 17-item Hamilton Depression Rating Scale (HAMD_{17}), a clinician-administered scale used to assess depression severity. Understanding the progression of depression over time is essential in evaluating the efficacy of antidepressants, and the HAMD_{17} trajectories offer valuable insights into this aspect. This project employs our unique Bayesian clustering approach which assumes that the clustering information of the trajectory only depends on a dimension reduced subspace. A significant challenge in analyzing HAMD_{17} trajectories is the presence of missing data, a common issue in longitudinal studies. The nature of depressive symptoms influences both the likelihood of patients' responses and the incidence of missing data. For instance, severely depressed patients might be less inclined to attend follow-up sessions, resulting in missing HAMD_{17}

scores. This project acknowledges the critical impact of such missingness on study outcomes, advocating for sophisticated methods to handle it effectively. The Bayesian clustering approach introduced in this project is not only designed to identify latent classes of depression severity and treatment response, but also taking into account the missing data mechanism. This approach is particularly important for ensuring accurate and unbiased estimates of depression severity and the effectiveness of interventions. The method goes beyond standard analysis techniques by incorporating multiple imputation and joint modeling approaches, which leverage observed data to infer missing HAMD_{17} scores accurately. The application of the model in the motivating HAMD_{17} trajectories also led to the identification of two distinct subpopulations, one corresponds to the responders to the treatment and the other one yet not.

Apart from Bayesian hierarchical clustering, the application encompasses the utilization of Bayesian hierarchical models for data privacy as well. In an era where data is increasingly digitized and easily accessible, the privacy of sensitive information has become a paramount concern. Project 3 contributes a foundational statistical methodology that can be pivotal in the context of Differential Privacy (DP). The efficient estimation of coefficients under the constraints of DP, as explored in this project, lays the groundwork for handling sensitive data responsibly and effectively. Theoretically speaking, the third project introduces a Bayesian hierarchical modeling framework in the context of DP and linear regression to improve the coefficient estimation efficiency. Considered there is no available dataset at hand, this project only provide simulation study with pseudo unobserved confidential predictors and responses and the only information available is the deferentially private sufficient statistics. The project centers on the concept of DP, a well-established framework that provides robust algorithmic protections to individual privacy. This concept involves introducing calibrated random fluctuations into algorithmic calculations, limiting the probability of revealing individual-specific information through the output. The innovation of this project lies in its unique approach that combines DP with Bayesian

envelope-based hierarchical model which is motivated by the observation that certain variations in predictors may not significantly affect the response variable in a regression model. By building a bridge between the concepts of envelopes, introduced for efficient coefficient estimation, and differential privacy, the project establishes a novel framework. This framework is implemented within a Markov Chain Monte Carlo (MCMC) data augmentation setup, allowing for Bayesian inference in linear regression while adhering to privacy constraints. The comparative analysis between this framework and traditional methods highlights the advantages of utilizing the envelope technique in maintaining privacy without sacrificing statistical rigor.

The rest of the thesis is organized as follows. Chapter 2 introduces the first project modeling and interpreting the associations between EEG-defined subgroups and antidepressant response. Chapter 3 introduces the second project discussing Bayesian envelope-based clustering Model with non-ignorable missingness and its application on HAMD trajectory. Chapter 4 introduces the third project involving the differentially private Bayesian envelope regression.

Chapter 2

Associations Between EEG-Defined Subgroups and Antidepressant Response: A Joint Mixture of Probabilistic Multilinear Principle Component Analysis Modeling Approach

2.1 Introduction

Despite the growing availability of antidepressant medications, major depressive disorder remains a leading cause of disability worldwide [3–5]. The morbidity persists, in part because, as a biologically heterogeneous and etiologically complex syndrome, depression encompasses various symptoms and exhibits divergent treatment responses [6–8]. [9] reported that about 60% of patients respond poorly to their first antidepressant trial. Moreover, for patients who already have tremendous concerns of hopelessness and discouragement, any treatment failure can lead to significant delays in alleviation of depression [10, 11]. Therefore, if patients who will likely respond to a specific antidepressant can be identified in advance of treatment, it would be of great clinical benefit.

Electroencephalography (EEG), which measures the brain's electrical activity acquired from each electrode placed on the scalp, is found to be a promising source of

non-invasive neuroimaging biomarkers of response to antidepressant treatment among patients with major depressive disorder (MDD) (e.g., [7, 12–14]). Specific attention has been paid to the utility of EEG power spectra, a transformation of EEG measures observed over time to the frequency domain (e.g., [15]). EEG spectral domains can be typically divided into the delta (<4 Hz), theta (4-7 Hz), alpha (7-15 Hz) and beta (15-30 Hz)-frequency bands. For example, [16] reported increased EEG alpha power at the posterior region of the brain in MDD patients who responded to a specific antidepressant. Similarly, [17] found that antidepressant responders in their study had higher EEG alpha power compared to non-responders. These findings also suggest that (1) there may exist different subgroups of MDD patients, characterized by different pre-treatment EEG powers, and (2) such EEG-defined subgroups might be of value as predictors of antidepressant response. While many EEG studies have demonstrated the existence of distinct EEG patterns within their study participants (e.g., [18, 19]), EEG-defined heterogeneity among MDD patients is not well understood.

In our current study, we aim to investigate whether pre-treatment EEG powers cluster into meaningful subgroups and how these EEG-defined subgroups are related to the antidepressant response, using the data drawn from the Establishing Moderators and Biosignatures of Antidepressant Response in Clinic Care (EMBARC) study [20, 21]. The pre-treatment EEG data were obtained for 83 MDD patients who underwent 8-week treatment with an antidepressant, and transformed to EEG power spectra using the current source density (CSD) methods [22]. Specifically, we will focus on the EEG powers spanning the theta and alpha frequency bands (4-15 Hz) with a resolution of 0.25 Hz, collected from the total 72 electrode locations on the scalp, yielding EEG powers at 72 electrode locations across 45 unique frequencies; that is, for each MDD patient's EEG data takes the form a 72×45 matrix. Patients' depressive symptom severity was assessed using the 17-item Hamilton Depression Rating Scale (HAMD_{17}) and a patient is considered to be a responder, i.e., who respond favorably to the treatment, if there is a 50% or more reduction of HAMD_{17} after 8

weeks of treatment compared with baseline. Our statistical task is then to cluster these matrix-variate EEG data into subgroups and evaluate the association between the associated subgroup memberships and the binary antidepressant response.

With the increased availability of matrix-variate data in modern scientific studies, the literature on clustering methods for matrix-variate data is growing (e.g., [23–26]). These methods extend multivariate Gaussian mixture models to accommodate matrix-variate data by imposing various low-rank and sparsity constraints to achieve dimension reduction and parsimonious modeling. However, the focus of these works is on clustering performance and as a result, they would not allow joint inference of the latent EEG subgroup memberships and their associated correlations with the antidepressant response of interest. In other words, applying these methods in our current setting would require a two-stage approach. In the first stage, one needs to carry out a clustering task for our EEG data to assign patients to different subgroups and then relate the known subgroup memberships to the antidepressant response in the second stage analysis. Without accounting for the estimation errors in the first stage clustering step, such a two-stage approach can result in attenuation bias when studying the subgroup specific effects on the response of interest (e.g., [27, 28]).

To overcome these shortcomings, we propose a Bayesian joint modeling approach to simultaneously model both the EEG data and the antidepressant response. We summarize our contributions as follows:

- We propose a mixture of probabilistic multilinear principal component analysis (mixture-PMPCA) model for our EEG data, in order to identify "homogeneous" subgroups of MDD patients who share similar pre-treatment EEG patterns. Our mixture-PMPCA model inherits the strength of multilinear principle component analysis (e.g., [29, 30]) by performing low-rank decomposition simultaneously to both the row and column spaces of the matrix-variate data, thus allowing better preservation of their spatial structures, and is a matrix-variate extension

of the classical mixture of probabilistic component analysis model ([31]).

- In contrast to a two-stage approach, our joint modeling approach properly accounts for the uncertainty in the estimation of the latent EEG subgroup memberships (avoiding attenuation bias) and relate the unknown subgroup memberships to the response of interest in one modeling step.

This chapter is structured as follows. Section 2.2 describes the proposed model along with its estimation and inference procedures. In Section 2.3, we describe the detailed data analysis for our motivating study. The paper concludes with a discussion in Section 2.4.

2.2 Model Specification

The task of identifying unobserved EEG subgroups can be naturally formulated as a latent class model (e.g., [32]), which is also referred to as a mixture of experts models (e.g., [33, 34]), with each latent class representing a homogeneous subgroup that has its own EEG pattern and probability of a favorable antidepressant response. For each observation $i \in \{1, \cdots, n\}$ ($n = 83$ in our EMBARC dataset), let $\boldsymbol{x}_i \in \mathbb{R}^{p \times q}$ be the $p \times q$ matrix-variate EEG data with $p = 72$ and $q = 45$, $(f_i, C_i^h)_{i=1}^n$ be the binary scalar covariates denoting gender ($f_i = 0$ as male and 1 as female) and chronicity ($C_i^h = 0$ as no chronic disease and 1 otherwise) respectively, and o_i be the binary response variable with $o_i = 1$ indicating an antidepressant responder.

2.2.1 Model for the Baseline Matrix-Variate Covariates

Suppose that there are G latent subgroups. We let $\boldsymbol{c}_i \doteq (c_{i1}, \ldots, c_{iG})$ with $c_{ig} = 1$ if subject i belongs to subgroup $g \in \{1, \cdots, G\}$, and 0 otherwise, and the probability that subject i is a member of subgroup g, denoted by $\pi_g = \mathbb{P}(c_{ig} = 1)$ follow the multinomial distribution: Multi$(1; \boldsymbol{\pi})$ with event probabilities $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_G)$ ($\sum_{g=1}^G \pi_g = 1$). To accomplish simultaneous dimension reduction and clustering task,

we establish the following mixture-PMPCA model under the previous notation

$$\begin{aligned}
\mathbb{P}(\boldsymbol{x}_i|\boldsymbol{u}_i) &= \mathcal{MN}_{p\times q}\left(\mathbf{A}\boldsymbol{u}_i\mathbf{B}^\top, \phi^{-1}\mathbf{I}_{p\times p}, \mathbf{I}_{q\times q}\right), \\
\mathbb{P}(\boldsymbol{u}_i|c_{ig}=1) &= \mathcal{MN}_{p_0\times q_0}\left(\boldsymbol{\eta}_g, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}\right), \\
(c_{i1},\cdots,c_{iG}) &\sim \text{Multi}\left(1;\boldsymbol{\pi}\right), \text{ with } \boldsymbol{\pi} = (\pi_1,\cdots,\pi_G)^\top, \\
\pi_g &= \mathbb{P}(c_{ig}=1), g=1,\cdots,G.
\end{aligned} \tag{2.1}$$

In model (2.1), $\mathcal{MN}()$ denotes matrix normal distribution. $\mathbf{A}_{p\times p_0}$ and $\mathbf{B}_{q\times q_0}$ are semi-orthogonal matrices, $\boldsymbol{\Gamma}_{p_0\times q_0}$ and $\boldsymbol{\Lambda}_{q_0\times q_0}$ are diagonal, $\boldsymbol{u}_i$ and $\boldsymbol{\eta}_g \in \mathbb{R}^{p_0\times q_0}$. After vectorization, the matrix normal distribution above can be equivalently written as

$$\begin{aligned}
\mathbb{P}(\text{vec}(\boldsymbol{x}_i)|\boldsymbol{u}_i) &= \mathcal{N}_{pq}\left((\mathbf{B}\otimes\mathbf{A})\text{vec}(\boldsymbol{u}_i), \phi^{-1}\mathbf{I}_{pq\times pq}\right), \\
\mathbb{P}(\text{vec}(\boldsymbol{u}_i)|c_{ig}=1) &= \mathcal{N}_{p_0q_0}\left(\text{vec}(\boldsymbol{\eta}_g), \boldsymbol{\Lambda}\otimes\boldsymbol{\Gamma}\right),
\end{aligned} \tag{2.2}$$

In model (2.2), $\otimes$ denotes the Kronecker product. The marginal distribution for $\boldsymbol{x}_i$ then has the following form:

$$\mathbb{P}(\mathbf{x}) = \sum_{g=1}^{G}\int \pi_g \mathbb{P}\left(\boldsymbol{x} \mid \boldsymbol{u}_i\right)\mathbb{P}\left(\boldsymbol{u}_i \mid c_{ig}=1\right)d\boldsymbol{u}_i \tag{2.3}$$

The parameters of the model are $\left\{\mathbf{A},\mathbf{B},\boldsymbol{\Gamma},\boldsymbol{\Lambda},\phi,(\boldsymbol{\eta}_g)_{g=1}^G\right\}$, and the latent variables in this model are the core matrices and the subgroup membership indicators $(\boldsymbol{u}_i,\boldsymbol{c}_i)_{i=1}^n$, where $\boldsymbol{c}_i = (c_{i1},\cdots,c_{iG})$.

2.2.2 Model for the Binary Treatment Response

We use Probit regression for the binary treatment outcome. Instead of regressing o_i on $\boldsymbol{x}_i$ or $\boldsymbol{u}_i$, we relate the likelihood of $o_i = 1$ to the latent class clusters c_{ig}. Specifically, it assumes that the variation inside each latent class, and the random error provides no information in predicting o_i. For simplicity, we denote the indicator function $\text{I}(c_{ig}=1)$ as I_{ig} and let $\boldsymbol{z}_i \doteq \{(\text{I}_{ig}, \text{I}_{ig}f_i, \text{I}_{ig}C_i^h, \text{I}_{ig}f_iC_i^h)_{g=2}^G, f_i, C_i^h, f_iC_i^h\}$ denote all the independent variables. It is worth noting that the independent variables $\boldsymbol{z}_i$ not only

contain the indicator $(\mathrm{I}_{ig})_{g=2}^{G}$, gender f_i and chronicity C_i^h, but also all the higher order interactions $(\mathrm{I}_{ig}f_i, \mathrm{I}_{ig}C_i^h, \mathrm{I}_{ig}f_iC_i^h)_{g=2}^{G}$. The formulation for Probit regression can be expressed as

$$\Phi^{-1}\left(\mathbb{P}(o_i = 1)\right) = \beta_0 + \boldsymbol{z}_i^T\boldsymbol{\beta} \tag{2.4}$$

In (2.4), $\Phi(\cdot)$ denotes the cumulative distribution function for a standard normal distribution. $\boldsymbol{\beta}$ is the corresponding coefficients for $\boldsymbol{z}_i$. Based on model (2.4), the reference group refers to patients in the first latent class with gender as male ($f_i = 0$) and no chronic disease ($C_i^h = 0$). When performing the analysis, we manually select the class with the least average CSD-EEG signals and refer it as the first latent class, i.e. the baseline.

2.2.3 Hierarchical Joint Modeling approach

As mentioned in introduction, in contrast to the two-stage approach, our joint modeling framework estimates the unknown subgroup memberships and relates the subgroup memberships to the response of interest in a unified framework. Our hierarchical joint modeling approach for CSD-EEG matrices and treatment outcome is described in Figure 2.1. In the figure, Diag() means the diagonal elements of a matrix, $\mathcal{MN}()$ denotes matrix normal distribution, $\mathcal{V}_{p_0,p}$ and $\mathcal{V}_{q_0,q}$ denote Stiefel manifold with corresponding dimensions. $\boldsymbol{c}_i$ indicates the latent class of the latent variable $\boldsymbol{u}_i$, on the other hand, $\boldsymbol{c}_i$ is also the independent variable in the Probit regression.

If we use $\boldsymbol{\nu}$ to denote all the model parameters, $\boldsymbol{\nu} \doteq \left\{\mathbf{A}, \mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \phi, (\boldsymbol{\eta}_g)_{g=1}^{G}, \boldsymbol{\pi}, \beta_0, \boldsymbol{\beta}\right\}$, then in our Bayesian joint modeling approach,

$$\mathbb{P}(\boldsymbol{x}_i, o_i, \boldsymbol{c}_i, \boldsymbol{u}_i|\boldsymbol{\nu}) = \mathbb{P}(\boldsymbol{x}_i|\boldsymbol{c}_i, \boldsymbol{u}_i, \boldsymbol{\nu})\mathbb{P}(o_i|\boldsymbol{c}_i, \boldsymbol{u}_i, \boldsymbol{\nu})\mathbb{P}(\boldsymbol{u}_i|c_i, \boldsymbol{\nu})\mathbb{P}(\boldsymbol{c}_i|\boldsymbol{\nu}).$$

The detailed complete data likelihood is given below,

$$
\begin{aligned}
&\mathbb{P}(\boldsymbol{x}, \boldsymbol{o}, \boldsymbol{c}, \boldsymbol{u}|\boldsymbol{\nu}) = \prod_{i=1}^{n} \mathbb{P}(\boldsymbol{x}_i, o_i, \boldsymbol{c}_i, \boldsymbol{u}_i|\boldsymbol{\nu}) \\
&= \prod_{i=1}^{n}\prod_{g=1}^{G} \left[\pi_g \left(\frac{\phi}{2\pi}\right)^{\frac{pq}{2}} \exp\left\{ -\frac{\phi}{2} \left\| \boldsymbol{x}_i - \mathbf{A}\boldsymbol{u}_i\mathbf{B}^\top \right\|_F^2 \right\} \right. \\
&\quad \times \Phi\left(\beta_0 + \boldsymbol{z}_i^T\boldsymbol{\beta}\right)^{\mathrm{I}(o_i=1)} \left[1 - \Phi\left(\beta_0 + \boldsymbol{z}_i^T\boldsymbol{\beta}\right)\right]^{\mathrm{I}(o_i=0)} (2\pi)^{-\frac{p_0 q_0}{2}} \\
&\quad \left. \times \exp\left\{ -\frac{1}{2} \operatorname{vec}\left(\boldsymbol{u}_i - \boldsymbol{\eta}_g\right)^\top \left(\boldsymbol{\Lambda} \otimes \boldsymbol{\Gamma}\right)^{-1} \operatorname{vec}\left(\boldsymbol{u}_i - \boldsymbol{\eta}_g\right) \right\} \right]^{\mathrm{I}(c_{ig}=1)}
\end{aligned}
\tag{2.5}
$$

where $\|\cdot\|_F^2$ represents the Frobenius norm.

Compared with the joint modeling approach, the two-stage approach consists of the following two steps:

- Stage 1: Model CSD-EEG with (2.1) and for each subject i assign

$$\hat{c}_{ig} = 1 \quad \text{where} \quad g = \underset{k\in 1,\ldots,G}{\arg\max} \quad \mathbb{P}(c_{ik} = 1|\boldsymbol{x}_i)$$

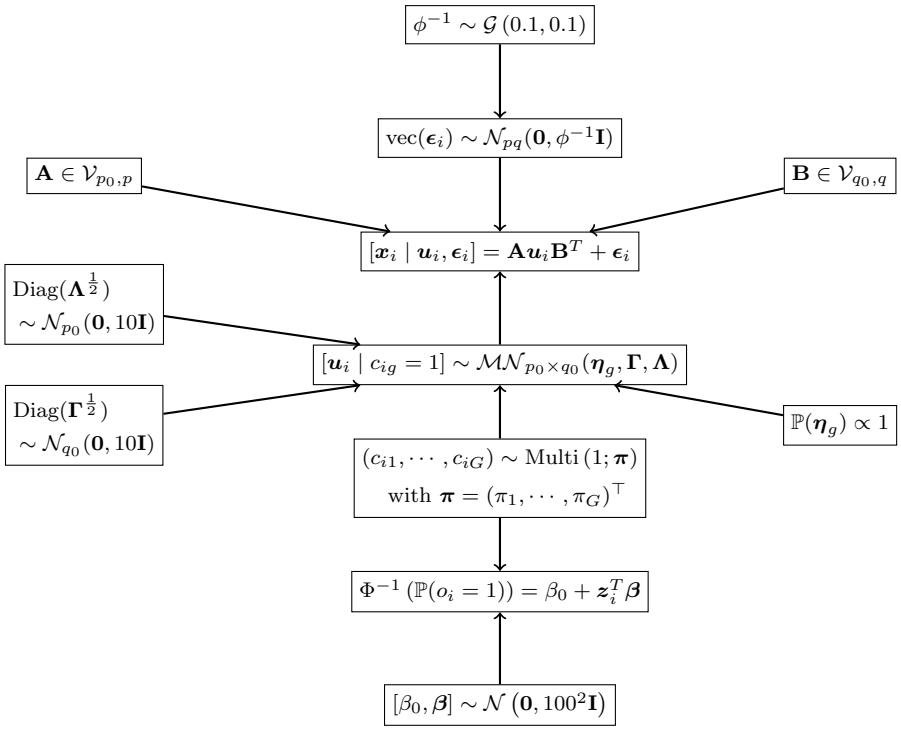
- Stage 2: Predict the binary treatment outcome o_i through the Probit regression model (2.4) given the subject belongs to the g'th class membership.

Remark 1 *From $\widehat{c}_{ig}$, we can define $\widehat{\boldsymbol{c}}_i \in R^G$ with the g'th element equal to 1 and the rest equal to 0s. Basically, $\widehat{\boldsymbol{c}}_i$ stores the subgroup membership of the ith subject. In summary, the two-stage approach separates clustering and outcome prediction steps. They first estimate $\widehat{\boldsymbol{c}}_i$, and then, in the second stage, regress o_i on $\widehat{\boldsymbol{c}}_i$ (along with other clinical covariates). Even though $\widehat{\boldsymbol{c}}_i$ is a random variable, which follows Multi$(1; \widehat{\boldsymbol{\pi}})$ with event probabilities $\widehat{\boldsymbol{\pi}} = (\widehat{\pi}_1, \cdots, \widehat{\pi}_G)$, such a two-stage strategy considers $\widehat{\boldsymbol{c}}_i$ as non-random quantities when predicting the outcome o_i.*

In summary, the two-stage approach fails to take into account the variability in the estimated $\widehat{c}_{ig}$ when fitting the Probit regression model. Indeed, ignoring the variability in $\widehat{c}_{ig}$ can introduce attenuation biases for the regression coefficient estimates, and result in underestimation of their associated uncertainty estimates (such as the width

of credible intervals). In contrast, the joint modeling approach naturally considers the variation in $\widehat{c}_{ig}$ when predicting o_i, which can help reduce the estimation bias and provide correct uncertainty quantification for the regression coefficients in the Probit model.

Figure 2.1: A graphical representation of the hierarchical structure of the joint modeling approach. c_{ig} is the bridge which connects CSD-EEG and the binary treatment outcome. On one hand, c_{ig} indicates the latent class of $\boldsymbol{u}_i$, on the other hand, c_{ig} is the independent variable in the Probit regression.



2.2.4 Prior Distributions

For model (2.1), we need priors for the parameters $\mathbf{A}$, $\mathbf{B}$, $\boldsymbol{\Gamma}$, $\boldsymbol{\Lambda}$, ϕ, $(\boldsymbol{\eta}_g)_{g=1}^G$ and $\boldsymbol{\pi}$. Since $\mathbf{A}_{p\times p_0}$ and $\mathbf{B}_{q\times q_0}$ are semi-orthogonal, we adopt the uniform distribution on Stiefel manifold $\mathcal{V}_{p_0,p}$ as the non-informative prior for $\mathbf{A}_{p\times p_0}$. Similarly, the uniform distribution on Stiefel manifold $\mathcal{V}_{q_0,q}$ is adopted as the prior for $\mathbf{B}_{q\times q_0}$. we assume $\boldsymbol{\Gamma}$ and $\boldsymbol{\Lambda}$ are positive definite diagonal matrices. For such matrices, we can adopt normal distribution as conjugate prior for the diagonal elements [35]. Hence $\mathcal{N}(0, 10)$ as the conjugate diffuse prior for each diagonal element of $\boldsymbol{\Gamma}^{\frac{1}{2}}$ and $\boldsymbol{\Lambda}^{\frac{1}{2}}$ is adopted.

Commonly used noninformative conjugate flat priors is adopted for $\boldsymbol{\eta}_g$, $\mathbb{P}(\boldsymbol{\eta}_g) \propto 1$. Conjugate diffuse Gamma prior is adopted for ϕ^{-1} with $\phi^{-1} \sim \mathcal{G}(0.1, 0.1)$. The latent variable $\boldsymbol{c}_i$ follow a multinomial distribution $\boldsymbol{c_i} \doteq (c_{i1}, \cdots, c_{iG}) \sim \text{Multi}\,(1; \boldsymbol{\pi})$, and we adopt the conjugate prior for $\boldsymbol{\pi}$ as

$$\boldsymbol{\pi} = (\pi_1, \cdots, \pi_G)^\top \sim \text{Dirichlet}(4, \ldots, 4)$$

The choice of our Dirichlet prior is equivalent to assuming a priori of four observations for each class, which can avoid having empty classes [36].

For model (2.4), we adopt conjugate diffuse priors for coefficients $\{\beta_0, \boldsymbol{\beta}\} \sim \mathcal{N}\,(\mathbf{0}, \tau\mathbf{I})$, $\tau = 100^2$

Remark 2 *For all the parameters mentioned above, either conjugate non-informative or conjugate weakly informative priors are selected. The model is not sensitive to the priors as long as the priors are not highly informative. For Gamma distribution* $\mathcal{G}(0.1, 0.1)$*, a decrease of the hyper-parameter* 0.1 *would make the prior less informative. For normal distributions above,* $\mathcal{N}(\mathbf{0}, 10\mathbf{I})$ *and* $\mathcal{N}(\mathbf{0}, 100^2\mathbf{I})$ *a increase of number* 10 *or* 100^2 *would also make the priors less informative.*

Remark 3 *The multilinear transformation matrices* $\mathbf{A} \in \mathbb{R}^{p \times p_0}$ *and* $\mathbf{B} \in \mathbb{R}^{q \times q_0}$*, are semi-orthogonal matrices. The set of such matrices is called the Stiefel manifold and we denote it as* $\mathcal{V}_{p_0,p}$ *and* $\mathcal{V}_{q_0,q}$*, respectively. Uniform distribution is the unique probability measure on Stiefel manifold that is invariant under left and right orthogonal transformations. Let* $\mathbf{A}_{[,k]}$ *denote the* k^{th} *column of* $\mathbf{A}$ *and* $\mathbf{A}_{[,-k]}$ *denote the matrix* $\mathbf{A}$ *with its* k^{th} *column removed. When a uniform prior distribution for* $\mathbf{A}$ *is assumed, the conditional distribution of* $\mathbf{A}_{[,k]}$ *given* $\mathbf{A}_{[,-k]}$ *is equal to the distribution of* $\boldsymbol{N}_{A\{-k\}}\boldsymbol{a}_k$ *where* $\boldsymbol{N}_{A\{-k\}}$ *is a basis for the null space of columns of* $\mathbf{A}_{[,-k]}$ $\left(\left\{\boldsymbol{x} \in \boldsymbol{R}^{p_0-1} : \mathbf{A}_{[,-k]}\boldsymbol{x} = \mathbf{0}\right\}\right)$ *and* $\boldsymbol{a}_k$ *is uniformly distributed on the* $(p - p_0 + 1)$*-dimensional sphere, and* $\mathbf{B}_k$ *is uniformly distributed on the* $(q - q_0 + 1)$*-dimensional sphere. i.e., conditional on* $\mathbf{A}_{[,-k]}, \mathbf{A}_{[,k]} \overset{d}{=} \boldsymbol{N}_{A\{-k\}}\mathbf{A}_k$*. Similar for* $\mathbf{B}$*, conditional on*

$\mathbf{B}_{[,-k]}, \mathbf{B}_{[,k]} \overset{d}{=} \boldsymbol{N}_{B\{-k\}}\boldsymbol{b}_k$, *where* $\boldsymbol{N}_{B\{-k\}}$ *is a basis for the null space of columns of* $\mathbf{B}_{[,-k]}$, *and* $\boldsymbol{b}_k$ *is uniformly distributed on the* $(q - q_0 + 1)$*-dimensional sphere.*

2.2.5 Model Selection with Widely Applicable Information Criterion

For our Bayesian joint modeling framework, the dimension (p_0, q_0) and the number of latent classes G need to be selected. The criterion used in this chapter is called Widely Applicable Information Criterion (WAIC)[37].

Based on [38], if a statistical model is regular and the likelihood can be approximated by Gaussian functions, then AIC [39] and BIC [40] can be applied to such evaluation processes. However, if a statistical model contains hierarchical structure or latent variables, then regularity condition is not satisfied. The information criteria WAIC are devised so as to estimate the generalization loss and the free energy, respectively, even if the posterior distribution is far from any normal distribution and even if the unknown true distribution is not realizable by a statistical model. With our notation, WAIC is with the form WAIC $= log \prod_{i=1}^{n} p_{post}(\boldsymbol{x}_i, o_i) - pwaic$, where $p_{post}(\boldsymbol{x}_i, o_i)$ denotes the posterior probability and $pwaic$ denotes the WAIC penalty term. [41] showed two ways of calculating WAIC during Markov chain Monte Carlo (MCMC) procedures, namely WAIC_1 (2.6) and WAIC_2 (2.7).

$$
\begin{aligned}
& log \prod_{i=1}^{n} \hat{p}_{post}(\boldsymbol{x}_i, o_i) - \hat{p}waic_1 \\
=& \sum_{i=1}^{n} \log \left(\frac{1}{M} \sum_{m=1}^{M} p\left(\boldsymbol{x}_i, o_i \mid \boldsymbol{\nu}^m\right) \right) - \\
& 2 \sum_{i=1}^{n} \left(\log \left(\frac{1}{M} \sum_{m=1}^{M} p\left(\boldsymbol{x}_i, o_i \mid \boldsymbol{\nu}^m\right) \right) - \frac{1}{M} \sum_{m=1}^{M} \log p\left(\boldsymbol{x}_i, o_i \mid \boldsymbol{\nu}^m\right) \right)
\end{aligned}
\tag{2.6}
$$

$$
\begin{aligned}
& log \prod_{i=1}^{n} \hat{p}_{post}(\boldsymbol{x}_i, o_i) - \hat{p}waic_2 \\
=& \sum_{i=1}^{n} \log \left(\frac{1}{M} \sum_{m=1}^{M} p\left(\boldsymbol{x}_i, o_i \mid \boldsymbol{\nu}^m\right) \right) - \sum_{i=1}^{n} \mathrm{V}_{m=1}^{M} \left\{ \log p\left(\boldsymbol{x}_i, o_i \mid \boldsymbol{\nu}^m\right) \right\}
\end{aligned}
\tag{2.7}
$$

In equation (2.6) and (2.7), M denotes the length of total MCMC iterations, $\boldsymbol{\nu}^m$ denotes the m'th posterior MCMC draw of

$$\boldsymbol{\nu} = \left\{\mathbf{A}, \mathbf{B}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \phi, (\boldsymbol{\eta}_g)_{g=1}^G, (\boldsymbol{c}_i, \boldsymbol{u}_i)_{i=1}^n\right\}$$

and $\mathrm{V}_{m=1}^M$ represents the sample variance with $\mathrm{var}_{m=1}^M a_m = \frac{1}{M-1}\sum_{m=1}^M (a_m - \bar{a})$. In this chapter we focused on WAIC_2, because for practical use, WAIC_2 has closer resemblance to the leave one out cross validation (LOO-CV) and also in practice seems to give results closer to LOO-CV [41].

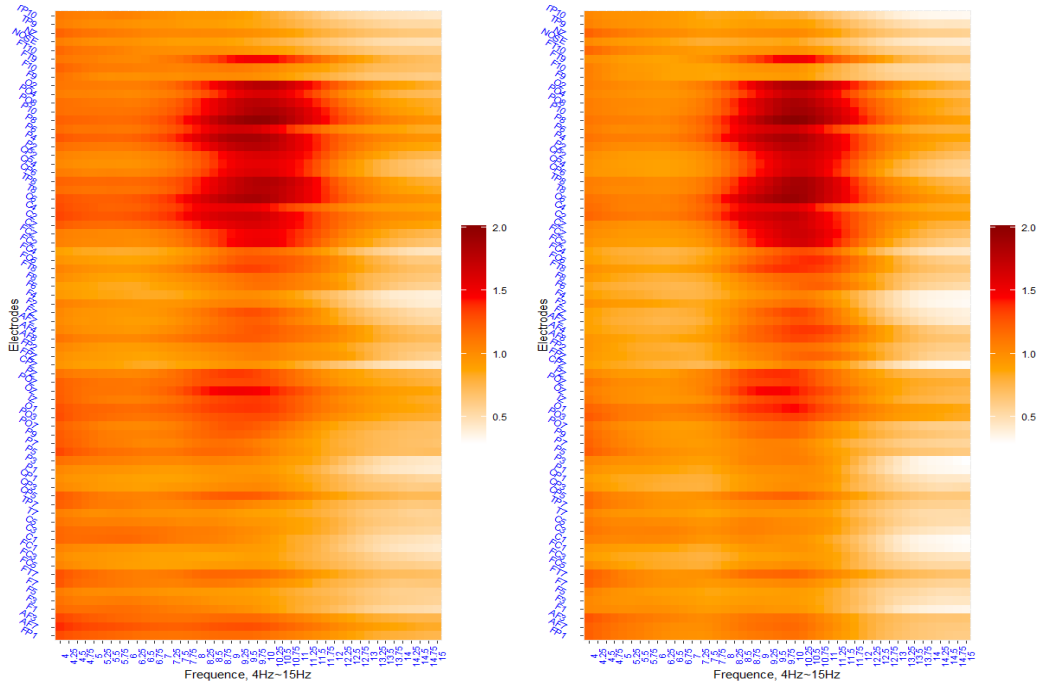
2.3 Numerical Results

The subjects in our study consisted of 83 medication-free MDD patients who were recruited in the EMBARC study and received an antidepressant from the class of selective serotonin reuptake inhibitors (SSRI) for 8 weeks. The depression symptom severity was assessed using the 17-item Hamilton Depression Rating Scale (HAMD_{17}) at baseline, weekly at weeks 1, 2, 3 and 4, and then bi-weekly at weeks 6 and 8 after initiation of the treatment. The outcome o_i is an indicator of a 50% or more reduction of HAMD_{17} scores after 8 weeks of treatment compared with the baseline [42]. The resting state EEG data were collected before treatment at four sites in the United States and transformed to CSD-based EEG power spectra at different frequency bands, leading to 83 72×45 matrices.

Among the 83 subjects, 46 subjects (55%) were classified as treatment responders ($o_i = 1$) and 37(45%) were classified as non-responders ($o_i = 0$). The 72×45 CSD-based EEG matrices, denoted as $\boldsymbol{x}_i, i \in \{1, \ldots, 83\}$, contains the CSD amplitude spectrum values ($\mu V/m^2$) of the total 72 electrodes located in different brain regions, including pre-frontal (Fp), frontal (F), temporal (T), parietal (P), occipital (O), and central (C) lobes, measured at the theta ($4 - 7$Hz) and alpha ($7 - 15$Hz) frequency bands (with the given 0.25Hz frequency resolution, 45 frequencies are recorded). Since the original CSD-EEG data were all positive with large scale, thus they were log-

transformed before the analysis. The scalar covariates contain gender and depression chronicity, which are both clinically relevant covariates for antidepressant response [43], [44], [45].

Figure 2.2 shows the mean CSD-EEG heatmap for antidepressant responders and non-responders. From the figure, we can hardly visualize the difference and a further comprehensive analysis is needed.



(a) Mean CSD-EEG heatmap for non-responders $o = 0$

(b) Mean CSD-EEG heatmap for responders $o = 1$

Figure 2.2: Mean CSD-EEG heatmap based on antidepressant outcome.

Figure 2.3 shows the heatmap of four random selected CSD-EEG subjects. It can be visualized that there is considerable variation. First, we can see that the signals (shown in red) exist especially in the parietal (P) region. In addition, the four heatmaps show significant differences in terms of the signals: case 1 exhibits moderate signal patterns; case 2 has the weakest signal; both case 3 and case 4 are active with strong signals, but clearly the signal in case 3 is stronger. The evidence motivates us to perform latent class analysis and detect the relationship between the

EEG heterogeneity and the treatment outcome.

In the project, we performed WAIC_2 to select (p_0, q_0, G). Based on equation (2.7), the model with the largest WAIC_2 is selected as the best candidate. For all models fitted here, we ran MCMC chains of $25,000$ iterations with the initial $10,000$ iterations discarded as burn-in period, and retained every 3^{rd} sample, leading to $5,000$ posterior samples used in the analysis. The WAIC_2 value under different choices of (p_0, q_0, G) are shown in Table 2.1; the model with $(p_0 = 2, q_0 = 1, G = 4)$ leads to the best $\text{WAIC}_2 = -135692$.

Under the best setup, each $\boldsymbol{x}_i$ $(i \in \{1, \ldots, 83\})$ was assigned to the latent class with the largest posterior probability, and we obtained 4 classes, with sizes of $n_1 = 20$, $n_2 = 47$, $n_3 = 11$ and $n_4 = 5$, respectively. These four latent classes were ordered such that the first latent class had the weakest signal and the fourth had the strongest signal. Figure 2.4 shows the class-specific heatmaps for the $p \times q$ mean structure of each latent class, which is calculated by $\hat{\boldsymbol{\mu}}_g = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{\mu}_g^{(m)}$ where $\boldsymbol{\mu}_g^{(m)} = \mathbf{A}^{(m)} \boldsymbol{\eta}_g^{(m)} \mathbf{B}^{(m)^T}$ with (m) denoting the m'th posterior draw and g denoting the g'th latent class, for $g \in \{1, 2, 3, 4\}$. As in the figure, unique pattern exists in each heatmap: there is almost no signal in the first latent class cluster; a relatively weak signal between 8.25Hz to 12.5Hz for the second latent class; a strong signal between 7.5Hz to 13Hz for the third class; a strong signal not only exists in the alpha bands $(7 - 15)$Hz but also in the theta bands: $(4 - 7)$Hz. Moreover, we found that the majority of the signals are concentrated in the posterior region for all four classes. This observation is consistent with previous studies relating pre-treatment CSD-EEG to treatment outcomes for serotonergic medications [22],[46], which found that the differences between responders and non-responders are most pronounced in the posterior region .

The posterior mean and 95% credible intervals for the Probit regression coefficients are shown in Table 2.2. For illustration, the name of the coefficient is redesigned. The first latent class with gender being male (gen= 0) and no chronicity (chr= 0) is treated as the baseline group.

Table 2.1: WAIC_2 examination different p_0, q_0 and G. Based on experience, the best setup is $(p_0 = 2, q_0 = 1, G = 4)$.

		$q_0 = 1$	$q_0 = 2$	$q_0 = 3$
$G = 2$	$p_0 = 1$	-350048	-298921	-303196
	$p_0 = 2$	-290239	-318353	-393451
	$p_0 = 3$	-267369	-328985	-391602
$G = 3$	$p_0 = 1$	-143244	-144516	-146087
	$p_0 = 2$	-143303	-145614	-165372
	$p_0 = 3$	-143306	-146328	-191468
$G = 4$	$p_0 = 1$	-136690	-214781	-144620
	$p_0 = 2$	$\mathbf{-135692}$	-146566	-156412
	$p_0 = 3$	-137290	-156543	-171317
$G = 5$	$p_0 = 1$	-219218	-143154	-142210
	$p_0 = 2$	-142651	-136745	-171317
	$p_0 = 3$	-145340	-146566	-168129

Table 2.2: The Probit regression coefficient summary under $(p_0 = 2, q_0 = 1, G = 4)$. The name of the variables is redesigned for illustration. The baseline group is the first latent class with gender being male and with no chronicity.

	posterior (mean)	95% C.I.
intercept	-1.03	$(-2.26, 0.01)$
class_2	1.93	$(0.66, 3.33)^*$
class_3	1.22	$(-0.23, 2.77)$
class_4	-12.52	$(-28.20, -0.52)^*$
chr	1.43	$(-0.07, 3.03)$
gen	3.59	$(0.45, 8.26)^*$
gen:chr	-4.19	$(-9.03, -0.65)^*$
class_2:gen	-2.32	$(-4.15, -0.56)^*$
class_3:gen	-1.51	$(-4.08, 1.00)$
class_4:gen	-1.61	$(-18.13, 13.38)$
class_2:chr	-3.67	$(-8.37, -0.16)^*$
class_3:chr	-3.21	$(-8.02, 0.40)$
class_4:chr	-3.81	$(-19.76, 10.48)$
class_2:gen:chr	4.40	$(0.36, 9.42)^*$
class_3:gen:chr	10.57	$(1.94, 22.89)^*$
class_4:gen:chr	-0.04	$(-19.81, 19.26)$

Due to space limits, we focus on interpreting the main effects in our model. Interpretation with the interaction effect can be drawn similarly. Table 2.2 shows that the main effect coefficients for the second and fourth latent classes are significant. The coefficient for the third latent class is almost significant as the lower bound of its 95% credit interval is only -0.23. Our results indicate that, among male patients without chronic disease, patients with Class-2 EEG signals are more likely to be responders than the ones showing Class-1 EEG signals. This conclusion is consistent with previous findings [22], [46] and [17]. They concluded that the patients with greater alpha than expected for control subjects would respond well to the SSRI antidepressant treatment, whereas those with less alpha than expected for control subjects would not. In the forth latent class, we observe a presence of signal not only in the alpha frequency band but also in theta frequency band. Additionally, the coefficient for this class is significantly negative, indicating that patients in this class have a lower likelihood of being responders. However, it is worth noting that the fourth latent class contains less than 10 observations, and thus the effect of this class was estimated with less precision compared to the rest of the latent classes. Aside from the concern of the small sample size, our results suggest that when there is strong signal on both alpha and theta frequency bands, patients are less likely to be responders. This finding highlights the collective impact of theta and alpha frequency on SSRI treatment outcomes, a conclusion that is consistent with previous studies such as [47],[48].

Furthermore, our results show that the main effect of gender is significant, with female patients being more likely to respond to SSRI treatment than male patients. The main effect of chronicity is not shown to be significant.

In addition to the Bayesian joint modeling approach, we also employed two other latent class models to analyze the data. The first approach we tried is the two-stage mixture of common factor analyzers (MCFA) [49], with vectorized CSD-EEG as inputs, using the R function `mcfa()` in the package `EMMIX`. However, due to the high dimensionality of the data ($p \times q = 3240$ and $n = 83$), this approach was not

feasible. The second approach we tried is the two-stage tensor envelope mixture model (TEMM) by [50]. The `temm()` function in the R package `TensorClustering` allows the researcher to select the envelope dimension based on BIC, but not the number of latent classes. In order to compare our approach with the two-stage TEMM, we manually set $G = 4$ when applying function `temm()`. Based on BIC, the selected envelope dimension was $(2, 2)$, which means that both p and q were reduced to 2 through dimension reduction. The mean CSD-EEG structure of each latent class is shown in Figure 2.5. After applying the TEMM, the Probit regression (2.4) was performed to estimate the regression coefficients as shown in Table 2.2. However, among all the coefficients, only interaction effect (class_2:chr) is significant at 0.05 level.

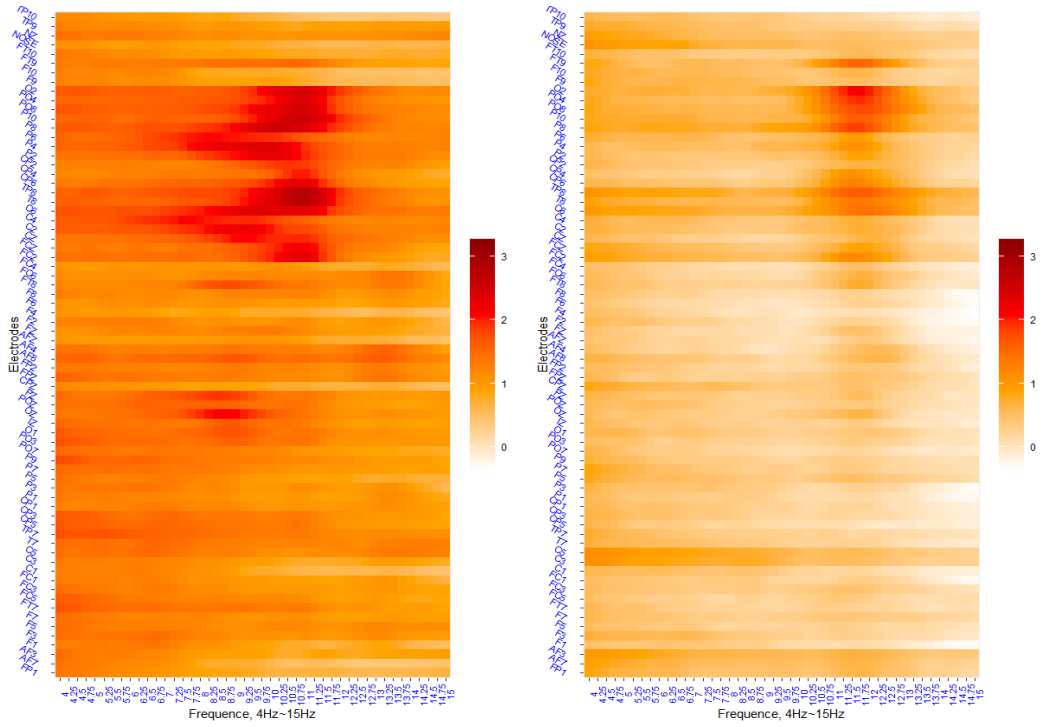
As another alternative method, we applied a non-model based approach using two-stage K-means clustering. Specifically, we first divided the EEG patterns into clusters using K-means, and then used the cluster membership as a predictor in a Probit regression model for treatment outcome. To select the best number of clusters, we used 5-fold cross-validated area under the curve (AUC) of the receiver operating characteristic (ROC). Table 2.3 presents the cross-validated AUC under different cluster numbers. The highest AUC achieved was 0.58 with 4 clusters as the optimal number of clusters.

Table 2.3: Cross-validation AUC for Probit regression with two-stage K-means Clustering

Clusters	2	3	4	5
AUC	0.52	0.53	0.58	0.51

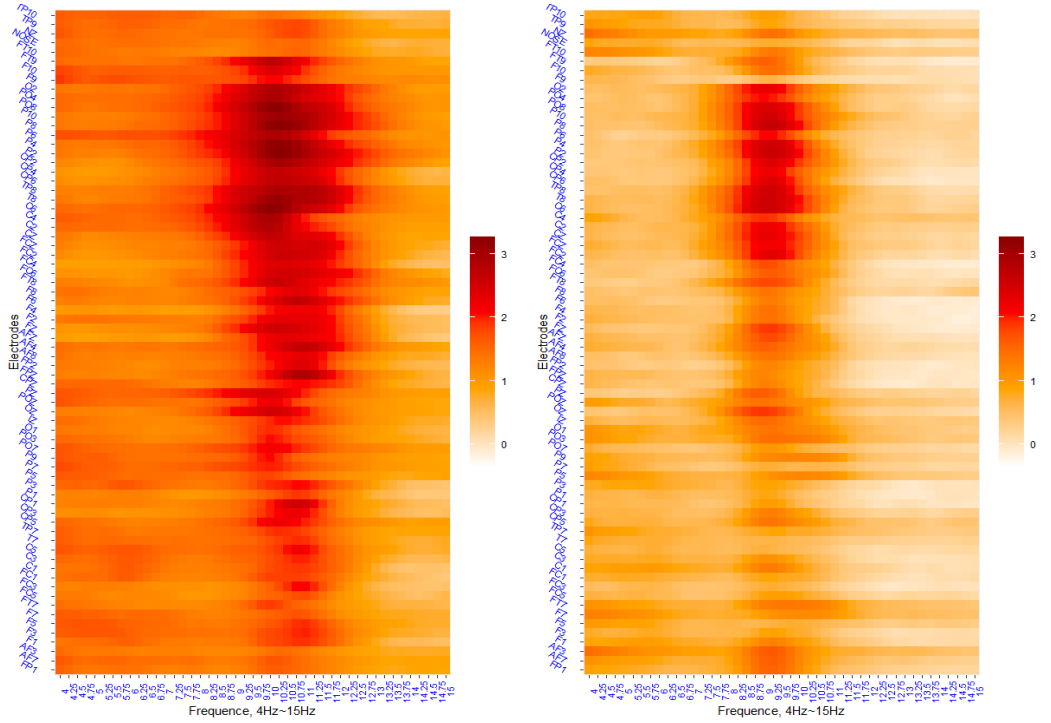
In summary, we employed three different methods to relate the EEG data to SSRI treatment response: our proposed mixture-PMPCA Bayesian joint modeling approach, the two-stage TEMM approach, and the two-stage K-means approach. We evaluated their prediction performance using AUC and presented the results in Table

2.4. The table presents the AUC calculated with all samples for the best candidate models after hyper-parameter selection, which are the mixture-PMPCA joint modeling approach with $(p_0 = 2, q_0 = 1, G = 4)$, the two-stage TEMM approach with $(Enlp_1 = 2, Enlp_2 = 2, G = 4)$, and the two-stage K-means approach with $G = 4$. Note that all the three approaches discussed above use the same Probit regression model, with predictors as outlined in equation (2.4). However, there are two main differences among these approaches. The first distinction is in how they treat the latent EEG subgroup memberships $\widehat{\boldsymbol{c}}_i$ when estimating effects on treatment response. Specifically, our mixture-PMPCA with Probit regression is a joint modeling approach, where the indicator variable $\widehat{\boldsymbol{c}}_i$ is considered as a random variable. In contrast, the two-stage TEMM and K-means treat the indicator variable as a fixed quantity, which ignores the uncertainty in estimating $\widehat{\boldsymbol{c}}_i$, and can thus lead to attenuation bias and worse prediction performance. The second distinction lies in the methods used for uncovering hidden subgroups in the EEG data. Our mixture-PMPCA model performs low-rank decomposition on both the row and column spaces of the matrix-variate EEG data, which results in a better preservation of the EEG data's spatial structure than K-means. Overall, our mixture-PMPCA joint modeling approach shows the best prediction performance, as indicated in Table 2.4.



(a) Sample 1

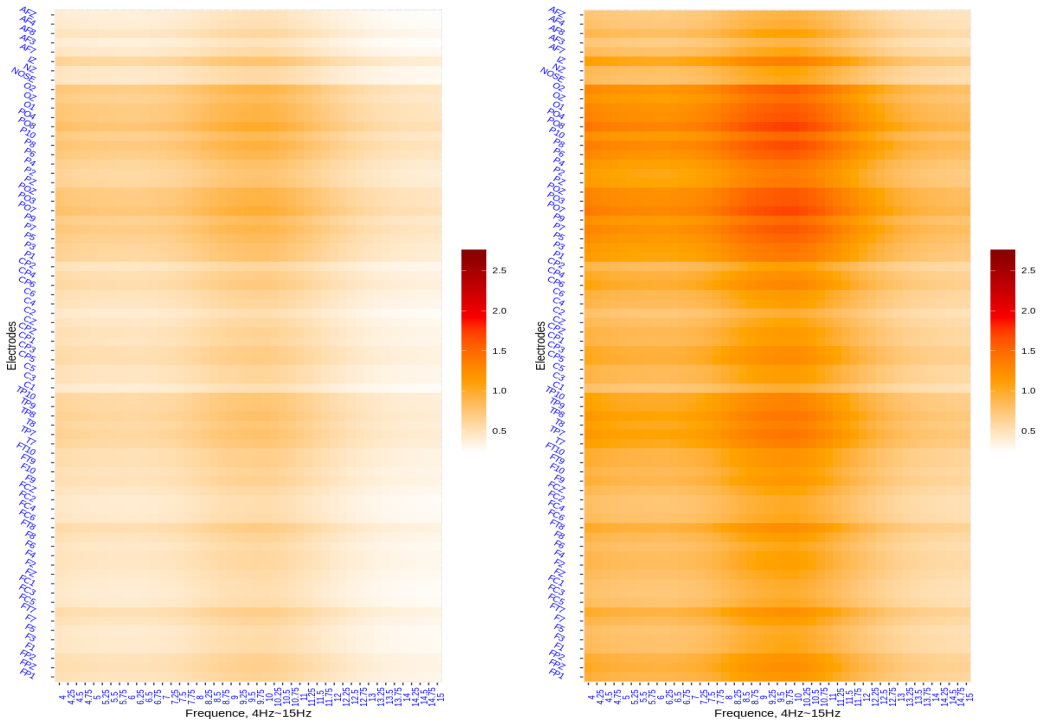
(b) Sample 2



(c) Sample 3

(d) Sample 4

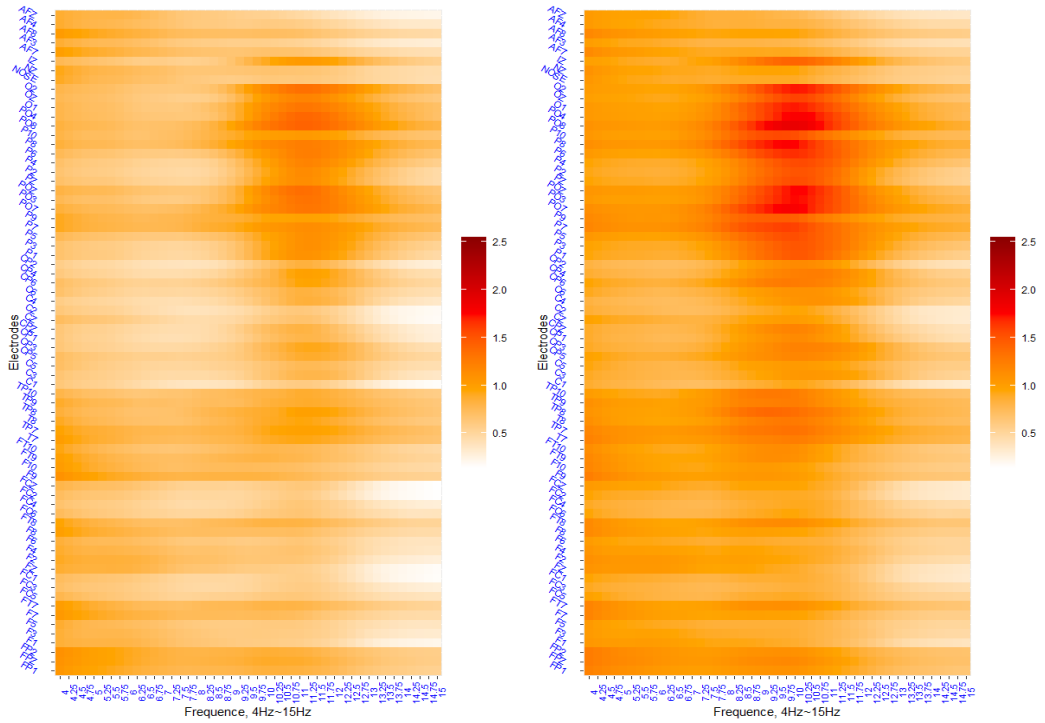
Figure 2.3: Heatmaps of 4 random selected CSD-EEG. The vertical axis denotes the 72 electrodes and the horizontal axis denotes the frequencies between 4Hz to 15Hz.



(a) $\hat{\boldsymbol{\mu}}_1$ (b) $\hat{\boldsymbol{\mu}}_2$

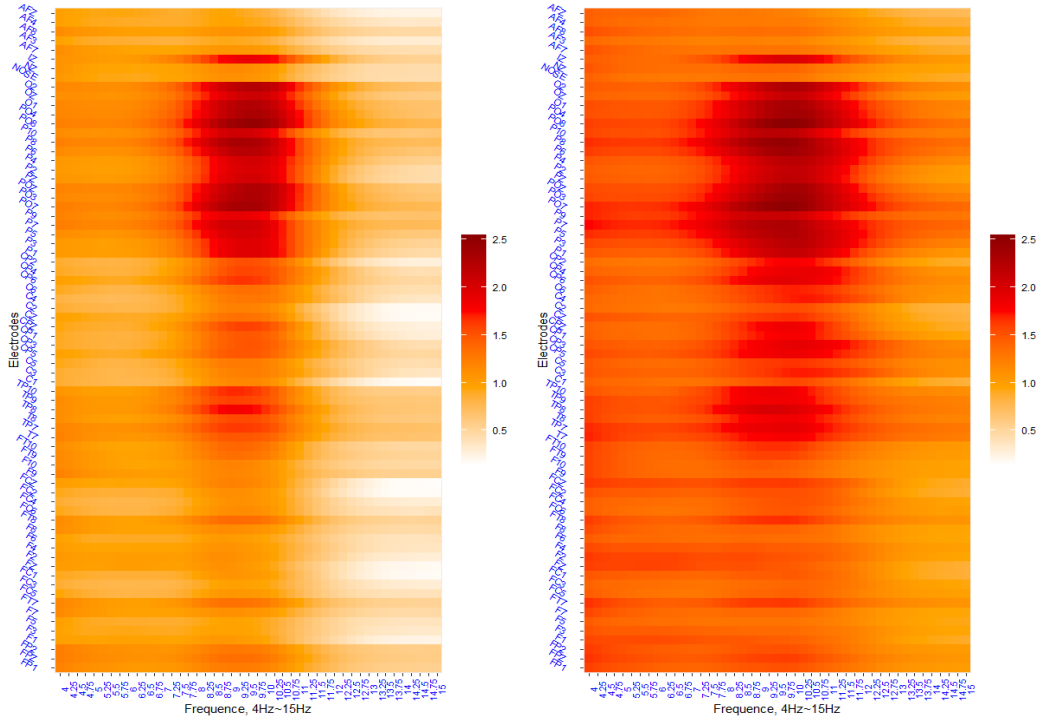
(c) $\hat{\boldsymbol{\mu}}_3$ (d) $\hat{\boldsymbol{\mu}}_4$

Figure 2.4: Heatmaps for the $p \times q$ mean structure of each latent class $\hat{\boldsymbol{\mu}}_g$, ($g \in \{1, \ldots, 4\}$) under our joint modeling framework. The vertical axis denotes the 72 electrodes and the horizontal axis denotes the frequencies between 4Hz to 15Hz.



(a) TEMM latent class$_1$

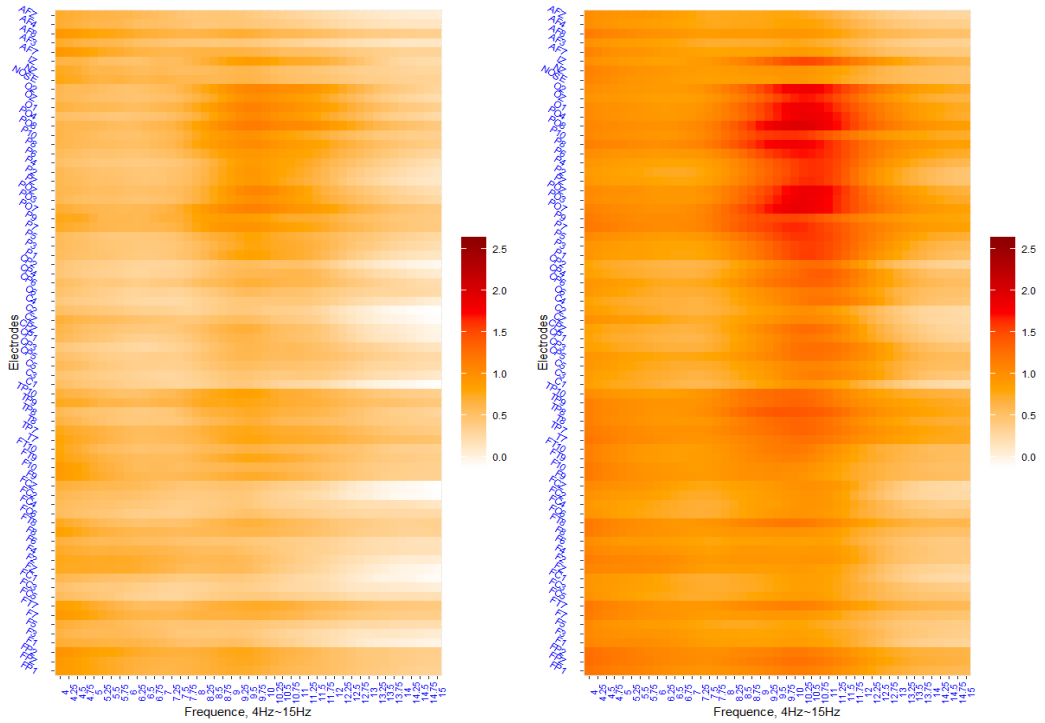
(b) TEMM latent class$_2$



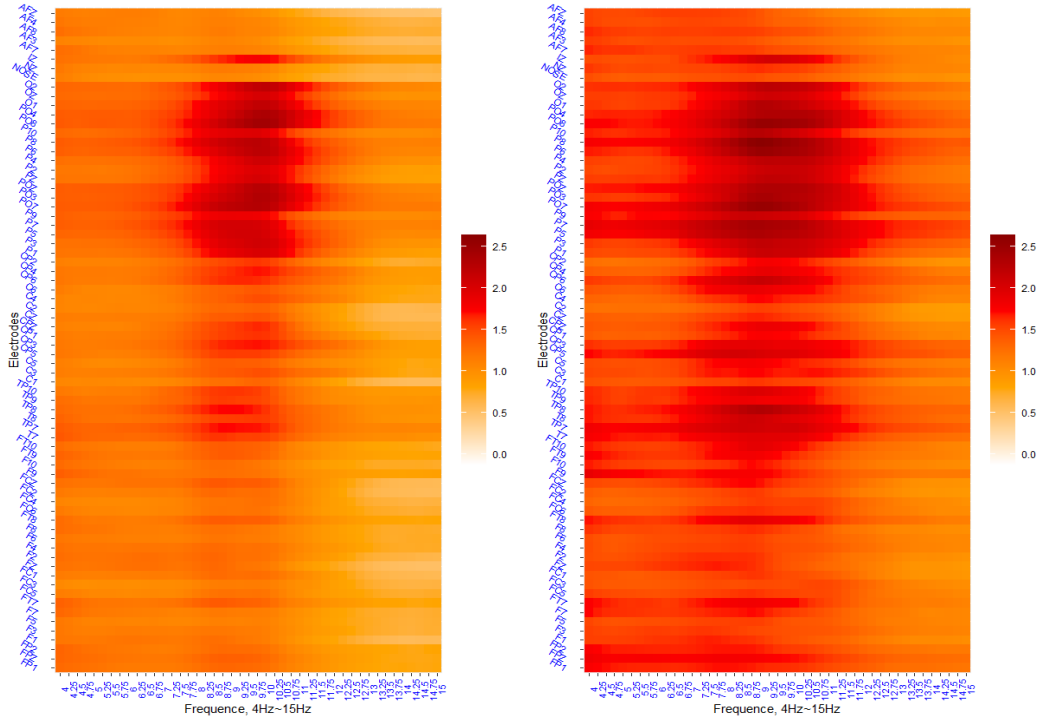
(c) TEMM latent class$_3$

(d) TEMM latent class$_4$

Figure 2.5: Heatmaps for the $p \times q$ mean structure of each latent class under TEMM. The vertical axis denotes the 72 electrodes and the horizontal axis denotes the frequencies between 4Hz to 15Hz.



(a) K-means cluster$_1$ (b) K-means cluster$_2$



(c) K-means cluster$_3$ (d) K-means cluster$_4$

Figure 2.6: Heatmaps for the $p \times q$ mean structure of each K-means cluster. The vertical axis denotes the 72 electrodes and the horizontal axis denotes the frequencies between 4Hz to 15Hz.

We then compare the mean EEG structres in the hidden subgroups derived from our proposed mixture-PMPCA (Figure 2.4) with the ones from TEMM (Figure 2.5) and K-means (Figure 2.6). Overall, the heatmaps generated by the three methods, share some similarities, indicating that all three methods are able to uncover similar structure in the EEG data to some degree. It can be observed that the signal trend increases from Cluster 1 to Cluster 4, and the signal is mainly present in the posterior region.

However, there are still some differences in the heatmaps generated by the three methods. Specifically, the TEMM signal in the second cluster is stronger than that in the mixture-PMPCA method, and the TEMM signal in the third and fourth clusters is more condensed. These differences can be attributed to the fact that mixture-PMPCA utilizes a joint modeling approach that accounts for the uncertainty in the estimation of latent EEG subgroup memberships, whereas the TEMM-based approach treats the cluster membership as fixed quantities. Due to the one-stage nature, our mixture-PMPCA joint modeling approach is able to uncover signal that better reflects the association of EEG and the treatment outcome. This is supported by the improved prediction performance of the mixture-PMPCA method as evidenced in Table 2.4. The same phenomenon can be observed when comparing the signal from the mixture-PMPCA method and the K-means method. Additionally, it can be observed that the heatmap signals generated by the two-stage TEMM and K-means methods are almost identical, further emphasizing the benefits of the joint modeling approach.

2.4 Discussion and Conclusion

Our exploration into the utilization of Bayesian hierarchical models for analyzing electroencephalograph (EEG) data in the context of depression treatment efficacy has yielded compelling insights and methodological advancements. This study, centered on the intricate relationship between baseline EEG patterns and responses to antidepressant medication, demonstrates the profound potential of sophisticated statistical

frameworks in mental health research. The application of Bayesian hierarchical models has not only facilitated a nuanced understanding of patient heterogeneity but also opened avenues for personalized treatment strategies.

The proposed Bayesian hierarchical joint modeling approach successfully summarized the relationship between matrix-variate CSD-EEG and binary antidepressant outcome. On one hand, the proposed mixture-PMPCA model preserves the CSD-EEG matrix structure. On the other hand, Probit regression model successfully summarizes the correlation between CSD-EEG latent classes and treatment outcome even the sample size n is far less than the dimension $p \times q$. Through our model, 4 latent classes with unique signal were found. The latent classes and the corresponding signals we found confirm that there is a collective effect of theta and alpha frequency towards SSRI, which is consistent with the previous studies about CSD-EEG and antidepressant treatment [47]. We also tried the two-stage mixture of common factor analyzers (MCFA) [49] with vectorized CSD-EEG as inputs. However, since the

Table 2.4: AUC for the best candidate models: mixture-PMPCA joint modeling approach with $(p_0 = 2, q_0 = 1, G = 4)$, two-stage TEMM approach with $(Enlp_1 = 2, Enlp_2 = 2, G = 4)$, two-stage Kmeans approach with $G = 4$.

Methods	AUC
Mixture-PMPCA	0.73
K-means	0.67
TEMM	0.69

dimensionality is too large and the R function $mcfa()$ crashed. Aside from MCFA, the two-stage tensor envelope mixture model (TEMM) [50] is also applied. However, there is no existing tool or function to select the number of TEMM clusters G. Another advantage of our Bayesian framework is its joint modeling nature. Unlike

traditional frequentist methods, the Bayesian framework estimates and returns a distribution instead of a point estimator. The joint approach considers the variability of the latent class indicator and helps reducing the bias in the estimator and improving the efficiency. The joint framework is not specific to only using PMPCA and Probit regression. Instead of PMPCA, we can incorporate other clustering strategies for CSD-EEG. For example, the recently proposed Envelope clustering framework [51]. Moreover, instead of the Probit regression with latent class indicators as predictors, researchers can also design other regression frameworks, for example, [52] proposed the partial least square regression; [29] proposed the left and right multiplication of vector when regressing with matrix covariates. Papers such as [26], [53] and [54] have discussed modeling with tensor structured data. As future research directions, our proposed framework can be easily extended to higher mode tensors rather than matrices. With such extension, our framework can be applied to model bio-markers with tensor structures.

Moreover, the introduction of a three-level hierarchical model to analyze matrix-variate EEG data represents a significant methodological innovation. At the core of our approach is the capacity to cluster patients into latent subgroups based on their baseline EEG characteristics, thereby illuminating the complex dynamics between neural activity patterns and treatment outcomes. This clustering approach not only enhances our understanding of the biological underpinnings of depression but also contributes to the broader goal of tailoring treatment plans to individual patient profiles, potentially increasing the efficacy of antidepressant therapies.

Finally, the key finding of our study is the identification of distinct patient subpopulations, differentiated by their EEG signatures and corresponding treatment responses. This granularity in patient categorization underscores the limitations of one-size-fits-all treatment approaches and highlights the critical need for precision medicine in the management of depression. By leveraging Bayesian hierarchical models, we have demonstrated an advanced analytical capability that transcends tradi-

tional clustering methods, offering a more refined lens through which to view the heterogeneity of depressive disorders.

In conclusion, the integration of Bayesian hierarchical models into the analysis of EEG data presents a promising pathway for enhancing our understanding of depression and optimizing treatment interventions. Our research not only contributes to the statistical literature by showcasing the application of advanced modeling techniques but also holds significant implications for clinical practice. As we move forward, it is imperative that future studies continue to build on these methodological foundations, exploring the vast potential of Bayesian hierarchical models in mental health research and beyond.

Chapter 3

Bayesian Envelope-Based Clustering Model with Non-Ignorable Missingness

3.1 Introduction

As the pace of technological innovation surges forward, we find ourselves amidst an ever-expanding ocean of data, brimming with complexity and nuance. In this landscape, the advent of model-based clustering stands as a testament to the sophistication achievable in the domain of data analysis. By positing that the observable data are generated from a finite amalgamation of underlying probability distributions, each indicative of a unique cluster, model-based clustering offers a nuanced lens through which the intrinsic groupings within data can be discerned [55]. This approach is particularly adept at navigating the inherent heterogeneity of real-world data, allowing for the accommodation of clusters of varying shapes and densities. Furthermore, the integration of dimensionality reduction techniques within model-based clustering frameworks serves to illuminate the core characteristics of the data, thereby facilitating a more profound understanding of its inherent structure.

Concurrently, the issue of non-ignorable missingness, or Missing Not at Random (MNAR), underscores the complexities inherent in data analysis. The presence of data missing in a manner contingent upon unobserved values necessitates a departure from

conventional methodologies predicated on simpler missing data mechanisms[56, 57]. The resolution of this challenge through advanced statistical models and sensitivity analyses not only exemplifies the nuanced understanding required for modern data analysis but also highlights the critical importance of accounting for non-ignorable missingness in ensuring the integrity of inferential conclusions. The interplay between model-based clustering and the meticulous handling of non-ignorable missing data structures thus epitomizes the sophisticated analytical strategies requisite for navigating the complexities of contemporary datasets.

In this chapter, we present a Bayesian clustering approach, inspired by the recognition that certain data variations may not significantly influence the clustering result. Essentially, certain data dimensions may be redundant for clustering purposes, as projections onto these dimensions yield no substantive information. This concept resonates with the recent envelope regression methodology, which postulates the existence of an envelope-reducing subspace encapsulating all pertinent information between dependent and independent variables [58, 59]. Due to the motivation, we call our model the Bayesian envelope-based clustering approach. While building on this concept, our methodology extends the envelope notion in regression to clustering context and integrates an additional noise component. This noise is orthogonal to both the envelope-reducing subspace and its complementary, rendering our model more realistic by accounting for inevitable measurement errors present in empirical data. The fundamental premise of our clustering approach assumes the existence of a subspace, denoted as $\mathbf{S}$, encapsulating all clustering-relevant information of the data. Its orthogonal complement, $\mathbf{S}^{\perp}$, is independent of $\mathbf{S}$, and the projection of the data onto this space has no bearing on the clustering outcome. Beyond the scope of clustering, our model is also extended to address the intricacies of the missing data mechanism, with a particular emphasis on the treatment of non-ignorable missingness, a prevalent yet overlooked issue in literature.

The simulation study in the paper explores the performance of our Bayesian

envelope-based clustering approach with missing data, considering various scenarios of sample sizes and noise variances. Results demonstrate the superior performance of our framework across scenarios, highlighting its effectiveness in accurately selecting the correct number of clusters. In our case study, our Bayesian envelope-based clustering framework is applied to analyze heterogeneity within incomplete 17-item Hamilton Depression Rating Scale HAMD_{17} trajectories sourced from the Establishing Moderators and Biosignatures of Antidepressant Response in Clinic Care (EMBARC) study [20]. The results show advantage of of our framework again as the rest of the methods compared in this chapter could not detect the heterogeneity.

In antidepressant trials, the notation HAMD_{17} denotes the 17-item clinician-evaluated Hamilton Depression Rating Scale System. Each corresponding HAMD_{17} trajectory chronicles the temporal variations in a patient's depression severity, with individual data points within the trajectory signifying respective severity scores. Elevated scores on the scale indicate increased severity of depression in the patient. Model-based clustering paradigms pertinent to the HAMD_{17} trajectories are recurrent topics of discussion. Foremost among these methodologies is the Growth Mixture Model, meticulously crafted for longitudinal trajectory examinations. This model, for every latent demographic, delineates a representative progression trajectory and concurrently assesses inter-individual fluctuations. Noteworthy contributions, such as those by [60–62], harness the Growth Mixture Model to elucidate the heterogeneity in HAMD_{17} trajectories. Beyond this archetype, Gaussian Mixture Models have ascended to a prominent stature within the analytical toolkit. A plethora of modern studies, exemplified by [63, 64], champion the merits of Gaussian Mixture Models in unearthing latent classes within the HAMD_{17}, spanning from rudimentary Gaussian architectures to their sophisticated covariance-structured counterparts. Except for the traditional Gaussian Mixture Models, extended mixture models such as the mixture of probabilistic principal components [31], Gaussian components infused with potential contaminants [65], robust multivariate t component mixtures [66], and mixtures

with shared loading matrix [49] can also be considered to expose the multifaceted heterogeneity inherent in the $HAMD_{17}$.

In antidepressant studies utilizing the $HAMD_{17}$, the mechanism behind missing data is critical to address for several reasons. Firstly, the nature of depressive symptoms can influence both the likelihood of response and the missingness of the data. For instance, patients experiencing severe depression may be less likely to attend follow-up appointments or fully engage with the questionnaire, leading to systematically missing $HAMD_{17}$ scores. If such non-ignorable missingness is not accounted for, estimates of depression severity and treatment efficacy could be biased, potentially leading to incorrect conclusions about the effectiveness of interventions. Secondly, in longitudinal studies using the $HAMD_{17}$, the interest often lies in changes in depression severity over time. When data are missing, particularly in a non-random manner related to patients' underlying depression trajectories, standard analysis methods that simply exclude missing data or impute missing values without considering the missing-data mechanism may lead to misleading inferences. Advanced statistical techniques that model the missingness process, such as multiple imputation with chained equations (MICE) or joint modeling approaches, can help mitigate this bias by leveraging the information in the observed data to make plausible inferences about the missing AMD_{17} scores. Research by [57, 67], underscores the importance of using appropriate methods for handling missing data in psychiatric scales to avoid biased parameter estimates and conclusions. Of the three missing mechanisms delineated by [68]—namely, missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR)—only MNAR holds significance for clustering inference. MCAR and MAR's inherent ignorable nature renders them less critical. Conversely, studies by [69, 70] suggest that trajectory data's missing mechanism tends to be non-random. Accordingly, in our Bayesian clustering methodology, we model the missing data mechanism as MNAR, articulated via a Probit regression model, incorporating both $HAMD_{17}$ and its latent class membership as predictors.

The subsequent sections of this chapter are organized as follows. Section 3.2 introduces the Bayesian model-based clustering methodology. It include the Probit regression tailored for addressing the MNAR missing mechanism, the selections of priors, the model's likelihood, prior and posterior distributions. Simulation outcomes are delineated in Section 3.3. Our Bayesian model-based clustering results, juxtaposed with prevailing techniques, are presented in Section 3.4. Concluding remarks and discussions are provided in Section 3.5. Comprehensive derivations, utilizing the Markov Chain Monte Carlo method, can be found in the appendix.

3.2 Model Specification

In this section, details for the Bayesian framework of HAMD_{17} trajectories are presented. For each subject $i \in \{1, \ldots, n\}$, let $\mathbf{Y}_i \in \mathbb{R}^r$ be the full r-dimensional HAMD_{17} trajectory including both the observed and missing data. In this article, the EMBARC dataset of interest is with $r = 7$. Without generalization, suppose the scores $\mathbf{Y}_i$ are centered, with $\text{E}(\mathbf{Y}_i) = \mathbf{0}$. We assume there are K finite latent subgroups and introduce a discrete latent variable $\mathbf{D}_i \doteq (\text{D}_{i1}, \ldots, \text{D}_{iK})$ with $\text{D}_{ik} = 1$ if trajectory i belongs to subgroup $k \in \{1, \cdots, K\}$, and 0 otherwise. $\mathbf{D}_i$ is the latent indicator indicating the latent subgroup that $\mathbf{Y}_i$ belongs to. The probability that trajectory i is a member of subgroup k, denoted by $\pi_k = \mathbb{P}(\text{D}_{ik} = 1)$ follow the multinomial distribution: $\text{Multi}(1; \boldsymbol{\pi})$ with event probabilities $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$ and $\sum_{k=1}^{K} \pi_k = 1$.

3.2.1 Bayesian Envelope-Based Clustering for HAMD_{17} Trajectory

Consider the full r dimensional vector including both the observed and missing data $\mathbf{Y}_i = (\mathbf{Y}_i^{\text{obs}}, \mathbf{Y}_i^{\text{mis}})$, and the latent variable $\mathbf{D}_i$ that indicates the latent classes of $\mathbf{Y}_i$. We assume that there exist a subspace $\mathbf{S} \in \mathbb{R}^u$ that captures all the clustering information of the HAMD_{17}. In this chapter, the orthogonal basis of $\mathbf{S}$ is denoted as $\boldsymbol{\Gamma} \in \mathbb{R}^{r \times u}$, $u \leq r$. Furthermore, we denote $\boldsymbol{\Gamma}_0 \in \mathbb{R}^{r \times (r-u)}$ as the basis of $\mathbf{S}^{\perp}$, the

orthogonal complement of $\mathbf{S}$.

With the previous notation, our model for $\mathbf{Y}_i$ can be expressed as

$$
\begin{aligned}
\mathbf{Y}_i &= \boldsymbol{\Gamma}\mathbf{U}_i + \boldsymbol{\Gamma}_0\mathbf{V}_i + \boldsymbol{\varepsilon}_i \\
[\mathbf{U}_i|\, \mathrm{D}_{ik} = 1] &\sim \mathcal{N}(\boldsymbol{\mu}_k, \boldsymbol{\Omega}) \quad \mathbf{V}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_0) \\
\mathbf{D}_i &\sim \mathrm{Multi}\,(1; \boldsymbol{\pi}), \text{ with } \boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)^\top \\
\boldsymbol{\Gamma} &\in \mathbb{R}^{r\times u} \quad \boldsymbol{\Gamma}_0 \in \mathbb{R}^{r\times(r-u)} \\
\boldsymbol{\Omega} &\in \mathbb{R}^{u\times u} \quad \boldsymbol{\Omega}_0 \in \mathbb{R}^{(r-u)\times(r-u)} \\
\mathbb{P}(\mathrm{D}_{ik} = 1) &= \pi_k, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}) \\
u &\leq r \quad \text{and} \quad k = 1, 2, \ldots, K.
\end{aligned}
\tag{3.1}
$$

Remark 4 *As shown in model (3.1), if we ignore $\boldsymbol{\varepsilon}_i$, $\boldsymbol{\Gamma}_0\mathbf{V}_i$ is invariant under different latent classes and only $\boldsymbol{\Gamma}\mathbf{U}_i$ matters for the clustering of $\mathbf{Y}_i$. We can derive the covariance matrix of the trajectory $\mathbf{Y}_i$, denoted by $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Omega}\boldsymbol{\Gamma}^\top + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}_0^\top$. Furthermore, we can also derive $\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Gamma}\boldsymbol{\Omega}^{-1}\boldsymbol{\Gamma}^\top + \boldsymbol{\Gamma}_0\boldsymbol{\Omega}_0^{-1}\boldsymbol{\Gamma}_0^\top$. If we use $\boldsymbol{\nu}$ to denote all the relevant parameters, then it is straightforward to show that the distribution of $[\mathrm{D}_{ik} = 1|\mathbf{Y}_i]$ is the same as $[\mathrm{D}_{ik} = 1|\boldsymbol{\Gamma}^\top\mathbf{Y}_i]$.*

$$
\begin{aligned}
& \mathbb{P}(\mathrm{D}_{ik} = 1|\mathbf{Y}_i, \boldsymbol{\nu}) \\
= \;& \frac{\pi_k \mathbb{P}(\mathbf{Y}_i|\, \mathrm{D}_{ik} = 1, \boldsymbol{\nu})}{\sum_j \pi_j \mathbb{P}(\mathbf{Y}_i|\mathbf{D}_i = j, \boldsymbol{\nu})} \\
= \;& \frac{\pi_k \cdot e^{(\mathbf{Y}_i - \boldsymbol{\Gamma}\boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\Gamma}\boldsymbol{\mu}_j)}}{\sum_j \pi_j \cdot e^{(\mathbf{Y}_i - \boldsymbol{\Gamma}\boldsymbol{\mu}_j)^\top \boldsymbol{\Sigma}^{-1}(\mathbf{Y}_i - \boldsymbol{\Gamma}\boldsymbol{\mu}_j)}} \\
= \;& \frac{\pi_k \cdot e^{(\boldsymbol{\Gamma}^\top\mathbf{Y}_i - \boldsymbol{\mu}_k)^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{\Gamma}^\top\mathbf{Y}_i - \boldsymbol{\mu}_k)}}{\sum_j \pi_j \cdot e^{(\boldsymbol{\Gamma}^\top\mathbf{Y}_i - \boldsymbol{\mu}_j)^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{\Gamma}^\top\mathbf{Y}_i - \boldsymbol{\mu}_k)}} \\
= \;& \mathbb{P}\left(\mathrm{D}_{ik} = 1|\boldsymbol{\Gamma}^\top\mathbf{Y}_i, \boldsymbol{\nu}\right)
\end{aligned}
\tag{3.2}
$$

Remark 5 *Given the dimension of the subspace $\mathbf{S} \in \mathbf{R}^u$, our goal is to find such space with $\mathbf{S} \doteq span(\boldsymbol{\Gamma})$ for model (3.1). It worth noting that $\boldsymbol{\Gamma}$ is not uniquely defined. In order to make it identifiable, in this chapter we define $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$ as a function of*

an unconstrained matrix $\mathbf{A} \in \mathbb{R}^{(r-u)\times u}$. *Let* $\mathbf{C}_{\mathrm{A}} = (\mathbf{I}_u, \mathbf{A})^\top$ *and* $\mathbf{D}_{\mathbf{A}} = (-\mathbf{A}, \mathbf{I}_{r-u})^\top$, *define*

$$\begin{aligned} \boldsymbol{\Gamma}(\mathbf{A}) &\doteq \mathbf{C}_{\mathbf{A}} \left(\mathbf{C}_{\mathbf{A}}^\top \mathbf{C}_{\mathbf{A}}\right)^{-1/2} \\ \boldsymbol{\Gamma}_0(\mathbf{A}) &\doteq \mathbf{D}_{\mathbf{A}} \left(\mathbf{D}_{\mathbf{A}}^\top \mathbf{D}_{\mathbf{A}}\right)^{-1/2} \end{aligned} \tag{3.3}$$

In equation (3.3), the matrix $\mathbf{A}$ *and span*$(\boldsymbol{\Gamma})$ *are uniquely determined by each other hence* $\boldsymbol{\Gamma}(\mathbf{A})$ *is identifiable. The notation of* $\boldsymbol{\Gamma}(\mathbf{A})$ *as the function of* $\mathbf{A}$ *prevents the prior selection and the posterior updating of* $\boldsymbol{\Gamma}$ *on Stiefel manifolds. As there is no close form posterior for* $\mathbf{A}$, *the update of* $\mathbf{A}$ *is through Metropolis–Hastings algorithm.*

3.2.2 Model for Non-Random Missing Data in HAMD_{17} Trajectory

As highlighted in the introduction, the collection of HAMD_{17} trajectories over several weeks for individual patients frequently results in datasets with missing values. As described by [68], there exist three primary mechanisms for missing data: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). MCAR occurs when the likelihood of missing data is independent of the dataset itself, whereas MAR transpires when the propensity for missing data is solely dependent on the observed values. On the other hand, MNAR is characterized by a dependence on unobserved values, rendering it distinct from MCAR and MAR. While MCAR and MAR are predicated on substantial assumptions that might not hold in practical scenarios, they are also considered ignorable, diminishing their relevance in certain research contexts. Previous studies, such as those conducted by [69, 70], have suggested that missing data in trajectories are often non-random. In response to these findings, the mechanism in the HAMD_{17} prioritizes the MNAR, delving into its nuances and implications in the subsequent sections.

Let $\boldsymbol{m} \in \mathbb{R}^{n\times r}$ denote the missing indicator, n is the sample size and r is the dimension of HAMD_{17} scores. $m_{ij} = 1$ if the data is not missing, $m_{ij} = 0$ if the

data is missing. Let p_{ij} be the probability that the data exists, i.e, $m_{ij} = 1$, then the missing mechanism is modeled through a Probit regression with

$$\Phi^{-1}(p_{ij}) = \theta_0 + \mathbf{Z}_{ij}^{\top}\boldsymbol{\theta} \tag{3.4}$$

where $\mathbf{Z}_{ij} \doteq \{Y_{i(j-1)}, Y_{ij} - Y_{i(j-1)}\}^{\top}$. Note that when θ_1 and θ_2 are not significant, the model becomes $\Phi^{-1}(p_{i,j}) = \theta_0$, which is the case of MCAR. By introducing a latent variable w_{ij} for each m_{ij} with

$$w_{ij} = \theta_0 + \mathbf{Z}_{ij}^{\top}\boldsymbol{\theta} + \epsilon_{ij}^{w} \tag{3.5}$$

where $\epsilon_{ij}^{w} \sim \mathcal{N}(0, 1)$. Then $m_{ij} = 1$ if and only if $w_{ij} > 0$; $m_{ij} = 0$ if and only if $w_{ij} \leq 0$.

3.2.3 Likelihood of the Model-Based Clustering with MNAR

If we still use $\boldsymbol{\nu}$ to represent the model parameters

$$\boldsymbol{\nu} \doteq (\mathbf{A}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\theta}, \boldsymbol{\pi}, \sigma^2)$$

and denote $\mathbf{Y}^{\text{obs}}$ as all the Y_{ij} that is observed, $\mathbf{Y}^{\text{mis}}$ as all the Y_{ij} that is missing, and $\mathbf{D} \doteq \{\mathbf{D}_1, \ldots, \mathbf{D}_n\}$, $i \in \{1, \ldots, n\}$, $j \in \{1, \ldots, r\}$, then the likelihood of $(\mathbf{Y}, \mathbf{D}, \boldsymbol{m})$ given $\boldsymbol{\nu}$ can be expressed as

$$\begin{aligned}
&f(\mathbf{Y}^{\text{obs}}, \mathbf{Y}^{\text{mis}}, \mathbf{D}, \boldsymbol{m}|\boldsymbol{\nu}) \\
&= f\left(\boldsymbol{m}|\mathbf{Y}^{\text{obs}}, \mathbf{Y}^{\text{mis}}, \mathbf{D}, \boldsymbol{\nu}\right) f\left(\mathbf{Y}^{\text{obs}}, \mathbf{Y}^{\text{mis}}|\mathbf{D}, \boldsymbol{\nu}\right) f\left(\mathbf{D}|\boldsymbol{\nu}\right) \\
&\propto \prod_{i=1}^{n}\left(\prod_{k=1}^{K}\pi_k^{\mathrm{I}(\mathrm{D}_{ik}=1)}\right)\left(\prod_{j=1}^{r}\Phi_{ij}^{m_{ij}}\left(1-\Phi_{ij}\right)^{1-m_{ij}}\right) \\
&\qquad \times \left(\prod_{k=1}^{K} e^{-\frac{1}{2}(\tilde{\mathbf{Y}}_i - \boldsymbol{\Gamma}(\mathbf{A})\boldsymbol{\mu}_k)^{\top}\boldsymbol{\Sigma}^{-1}(\tilde{\mathbf{Y}}_i - \boldsymbol{\Gamma}(\mathbf{A})\boldsymbol{\mu}_k)}\right)^{\mathrm{I}(\mathrm{D}_{ik}=1)}
\end{aligned} \tag{3.6}$$

where $\Phi_{ij} = \Phi(\theta_0 + \tilde{\mathbf{Z}}_{ij}^{\top}\boldsymbol{\theta})$ with $\tilde{\mathbf{Z}}_{ij}$ denote $\mathbf{Z}_{ij}$ after imputation, $\tilde{\mathbf{Y}}_i$ denote $\mathbf{Y}_i$ after imputation, and $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}(\mathbf{A})\boldsymbol{\Omega}\boldsymbol{\Gamma}(\mathbf{A})^{\top} + \boldsymbol{\Gamma}(\mathbf{A})_0\boldsymbol{\Omega}_0\boldsymbol{\Gamma}(\mathbf{A})_0^{\top} + \sigma^2\mathbf{I}$

3.2.4 Prior Distribution

$\mathbf{A} \sim \mathcal{MN}_{(r-u)\times u}(\mathbf{A}_0, \mathbf{K}, \mathbf{L})$, where $\mathcal{MN}$ denotes the matrix normal distribution. $\mathbf{A}_0 \in \mathbb{R}^{(r-u)\times u}$, $\mathbf{K} \in \mathbb{R}^{(r-u)\times(r-u)}$, $\mathbf{L} \in \mathbb{R}^{u\times u}$ are positive symmetric matrices. In equation (3.3), $\mathbf{C_A}$ is defined to be $(\mathbf{I}_u, \mathbf{A})^\top$. We set $\mathbf{A}_0 = \mathbf{0}^{(r-u)\times(r-u)}$. It means $\mathbf{\Gamma}$ is assumed centered at $\mathbf{\Gamma}(\mathbf{0}) = (\mathbf{I}_u, \mathbf{0})^\top$, which corresponds to an identity projection to the first u dimensions. $\mathbf{K}$ and $\mathbf{L}$ are chosen to be identity matrices $\mathbf{I}^{(r-u)\times(r-u)}$ and $\mathbf{I}^{u\times u}$. During MCMC, we first perform eigen-decomposition for covariance matrix of the cases without missing data $\mathbf{Y}_{complete}$, $cov(\mathbf{Y}_{complete}) = \mathbf{V\Lambda V}^{-1}$. Suppose the diagnal elements of $\mathbf{\Lambda}$ is decreasing. The first u eigen-vectors $\mathbf{V}_{[,1:u]}$ is set to be the initial value of $\mathbf{\Gamma}$. The corresponding $\mathbf{A}$ such that $\mathbf{\Gamma}(\mathbf{A}) = \mathbf{V}_{[,1:u]}$ is the initial for $\mathbf{A}$.

$\mathbf{\Omega}$ and $\mathbf{\Omega}_0$ follow the Inverse-Wishart distribution: $\mathbf{\Omega} \sim \mathrm{IW}_u(\Psi_Y, \nu_Y)$, $\mathbf{\Omega}_0 \sim \mathrm{IW}_{(r-u)}(\Psi_{0,Y}, \nu_{0,Y})$. In order to determine parameters (Ψ_Y, ν_Y, $\Psi_{0,Y}$, $\nu_{0,Y}$), we first make the projection $\mathbf{V}_{[,1:u]}^\top \mathbf{Y}_{complete}$ and $\mathbf{V}_{[,(u+1):r]}^\top \mathbf{Y}_{complete}$. After the projection, denote $\mathbf{S}_1$ and $\mathbf{S}_2$ as the their corresponding sampling covariance. We set $\Psi_Y = 0.75\mathbf{S}_1$ $\nu_Y = 8.5 + 0.5(u-1)$, $\Psi_{0,Y} = 0.75\mathbf{S}_2$, $\nu_{0,Y} = 8.5 + 0.5(r-u-1)$. By [36], with our choice, the prior expectation of the amount of heterogeneity explained by difference of the group means is 90%: $\mathrm{E}(R_t^2) = 1 - \frac{\Psi(Y)\mathbf{S}_1^{-1}}{\nu_Y - 0.5\cdot(u+1)} = 0.9$. The parameter θ_0, $\boldsymbol{\theta}$ are assumed to be independently normally distributed. θ_0 is assigned with a weakly informative prior $\theta_0 \sim \mathcal{N}(0, 100^2)$. Considered the study from [69], we apply the mildly informative priors for $\boldsymbol{\theta}$ as $\boldsymbol{\theta} \sim \mathcal{N}\left((0, 0.2)^\top, 0.01\mathbf{I}\right)$. σ^2 follow the Inverse-Gamma distribution with $\sigma^2 \sim \mathrm{IG}(0.1, 0.1)$. Finally, $\boldsymbol{\mu}_k \propto \mathbf{1}$ the improper flat prior. $k = 1, \ldots, K$.

3.2.5 Posterior Distribution

- $[\mathbf{D}_i]|\cdot \sim Multinormial(\tilde{\pi}_{i1}, \ldots, \tilde{\pi}_{iK})$ with

$$\tilde{\pi}_{ik} = \mathbb{P}(\mathrm{D}_{ik} = 1|\cdot) = \frac{\Delta_{ik}}{\sum_{j=1}^{K} \Delta_{ij}}$$

  where Δ_{ik} is defined as

$$\Delta_{ik} = \left(\prod_{j=1}^{r} \Phi_{ij}^{m_{ij}} \left(1 - \Phi_{ij}\right)^{1-m_{ij}} \right) \times$$
$$\exp\{-\frac{1}{2}(\tilde{\mathbf{Y}}_i - \mathbf{\Gamma}(\mathbf{A})\boldsymbol{\mu}_k)^\top \mathbf{\Sigma}^{-1}(\tilde{\mathbf{Y}}_i - \mathbf{\Gamma}(\mathbf{A})\boldsymbol{\mu}_k)\}$$

with notation $\Phi_{ij} = \Phi(\theta_0 + \tilde{\mathbf{Z}}_{ij}^\top \boldsymbol{\theta})$ with $\tilde{\mathbf{Z}}_{ij}$ denote $\mathbf{Z}_{ij}$ after imputation, $\tilde{\mathbf{Y}}_i$ denote $\mathbf{Y}_i$ after imputation, and $\mathbf{\Sigma} = \mathbf{\Gamma}(\mathbf{A})\mathbf{\Omega}\mathbf{\Gamma}(\mathbf{A})^\top + \mathbf{\Gamma}(\mathbf{A})_0\mathbf{\Omega}_0\mathbf{\Gamma}(\mathbf{A})_0^\top + \sigma^2\mathbf{I}$

- Since the posterior distribution of $\mathbf{A}$ given other parameters

$$\pi(\mathbf{A}|\cdot) \propto \prod_{i=1}^{N}\prod_{k=1}^{K} \exp\left\{ -\frac{1}{2}\left(\tilde{\mathbf{Y}}_i - \mathbf{\Gamma}\left(\mathbf{A}\right)\boldsymbol{\mu}_k\right)^\top \mathbf{\Sigma}^{-1}\left(\tilde{\mathbf{Y}}_i - \mathbf{\Gamma}\left(\mathbf{A}\right)\boldsymbol{\mu}_k\right)\right\}^{\mathrm{I}(\mathrm{D}_{ik}=1)}$$
$$\times \exp\left\{-\frac{1}{2}\ \mathrm{trace}\ \left[\mathbf{K}^{-1}\left(\mathbf{A}-\mathbf{A}_0\right)\mathbf{L}^{-1}\left(\mathbf{A}-\mathbf{A}_0\right)^\top\right]\right\}$$

has no standard distribution form, thus we need to sample $\mathbf{A}$ using Metropolis-Hasting strategy.

- Since

$$\pi(\boldsymbol{\mu}_k|\cdot) \propto \exp -\frac{1}{2}\sum_{i,\mathbf{D}_i=k}\left\{\left(\tilde{\mathbf{Y}}_i - \mathbf{\Gamma}\left(\mathbf{A}\right)\boldsymbol{\mu}_k\right)^\top \mathbf{\Sigma}^{-1}\left(\tilde{\mathbf{Y}}_i - \mathbf{\Gamma}\left(\mathbf{A}\right)\boldsymbol{\mu}_k\right)\right\}$$

the posterior distribution of $\boldsymbol{\mu}_k, k \in 1\ldots K$, given other parameters

$$\boldsymbol{\mu}_k \sim \mathcal{N}(\frac{\sum_{i,\mathbf{D}_i=k}\mathbf{\Gamma}^\top\tilde{\mathbf{Y}}_i}{n_k}, \mathbf{\Omega} + \sigma^2\mathbf{I})$$

where n_k denotes the number of observations in kth latent class.

- Denote $\mathbf{Y}_{0i} = \mathbf{\Gamma}\mathbf{U}_i + \mathbf{\Gamma}_0\mathbf{V}_i$ and $\tilde{\mathbf{Y}}_{0i}$ as $\mathbf{Y}_{0i}$ after imputation, then

$$\sigma^2 \sim \mathrm{IG}(0.1 + n/2, 0.1 + 1/2\sum_{i=1}^{n}(\tilde{\mathbf{Y}}_{0i} - \tilde{\mathbf{Y}}_i)^2)$$

- $\tilde{\mathbf{Y}}_{0i}|\,\mathrm{D}_{ik} = 1, \tilde{\mathbf{Y}}_i, \cdot \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{\Sigma}_0)$ where

$$\boldsymbol{\mu}_0 = \mathbf{\Sigma}_0(\mathbf{\Sigma}_0\boldsymbol{\mu}_k + \frac{1}{\sigma^2}\tilde{\mathbf{Y}}_i)^{-1}$$

and

$$\mathbf{\Sigma}_0 = \mathbf{\Gamma}(\mathbf{A})\mathbf{\Omega}\mathbf{\Gamma}(\mathbf{A})^\top + \mathbf{\Gamma}(\mathbf{A})_0\mathbf{\Omega}_0\mathbf{\Gamma}(\mathbf{A})_0^\top + \sigma^2\mathbf{I}$$

- $\boldsymbol{\Omega}|\tilde{\mathbf{Y}}_{0i}, \cdot \sim \text{IW}_u(\Psi_Y + \sum_i \sum_k \text{I}(\text{D}_{ik} = 1)(\boldsymbol{\Gamma}^\top \tilde{\mathbf{Y}}_{0i} - \boldsymbol{\mu}_k)(\tilde{\mathbf{Y}}_{0i}^\top \boldsymbol{\Gamma} - \boldsymbol{\mu}_k^\top), \nu_Y + N)$

- $\boldsymbol{\Omega}_0|\tilde{\mathbf{Y}}_{0i}, \cdot \sim \text{IW}_{r-u}(\Psi_{0,Y} + \sum_i \boldsymbol{\Gamma}_0^T \tilde{\mathbf{Y}}_{0i} \tilde{\mathbf{Y}}_{0i}^T \boldsymbol{\Gamma}_0, \nu_{0,Y} + N)$

- $[w_{ij}|\cdot] \sim \mathcal{N}(\theta_0 + \tilde{\mathbf{Z}}_{ij}^\top \boldsymbol{\theta}, 1)$ with $\tilde{\mathbf{Z}}_{ij}$ denoting $\mathbf{Z}_{ij}$ after imputation. The update of wij is as follows:

  If $m_{ij} = 1$

  $$[w_{ij}|\cdot] \sim \mathcal{N}\left(\theta_0 + \tilde{\mathbf{Z}}_{ij}^\top \boldsymbol{\theta}, 1\right) \text{I}_{(0,+\infty)}$$

  If $m_{ij} = 0$,

  $$[w_{ij}|\cdot] \sim \mathcal{N}\left(\theta_0 + \tilde{\mathbf{Z}}_{ij}^\top \boldsymbol{\theta}, 1\right) \text{I}_{(-\infty,0)}$$

- $[\theta_0, \boldsymbol{\theta}|\cdot] \sim \mathcal{N}(M, E)$

  where

  $M = \left(\sum_i [1, \tilde{\mathbf{Z}}_{ij}]^\top [1, \tilde{\mathbf{Z}}_{ij}] + \boldsymbol{\Sigma}_\theta\right)^{-1} \left(\sum_i [1, \tilde{\mathbf{Z}}_{ij}] w_{ij} + \boldsymbol{\Sigma}_\theta \boldsymbol{\mu}_\theta\right)$

  and

  $E = \left(\sum_i [1, \tilde{\mathbf{Z}}_{ij}]^\top [1, \tilde{\mathbf{Z}}_{ij}] + \boldsymbol{\Sigma}_\theta\right)^{-1}$

  with $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ denoting the prior mean and covariance.

- $[\boldsymbol{\pi}|\cdot] \sim \text{Dirichlet}(4 + n_1, \ldots, 4 + n_G)$

  where $n_k = \sum_i \text{D}_{ik}$

The details of the missing data imputation is given in appendix.

3.2.6 Model Selection with Widely Applicable Information Criterion

For our envelope clustering framework, the envelope dimension u and the number of latent classes K need to be selected. The criterion used in this chapter is called Widely Applicable Information Criterion (WAIC) [37].

Similarly [38], if a statistical model is regular and the likelihood can be approximated by Gaussian functions, then AIC [39] and BIC [40] can be applied to such evaluation processes. However, if a statistical model contains hierarchical structure or latent variables, then regularity condition is not satisfied. The information criteria WAIC are devised so as to estimate the generalization loss and the free energy, respectively, even if the posterior distribution is far from any normal distribution and even if the unknown true distribution is not realizable by a statistical model. With our notation, WAIC is with the form WAIC $= log \prod_{i=1}^{n} p_{post}(\mathbf{Y}_i, \boldsymbol{m}_i) - pwaic$, where $p_{post}(\mathbf{Y}_i, \boldsymbol{m}_i)$ denotes the posterior probability and $pwaic$ denotes the WAIC penalty term. [41] showed two ways of calculating WAIC during Markov chain Monte Carlo (MCMC) procedures, namely WAIC_1 (3.7) and WAIC_2 (3.8).

$$
\begin{aligned}
& log \prod_{i=1}^{n} \hat{p}_{post}(\mathbf{Y}_i^{\text{obs}}, \boldsymbol{m}_i) - \hat{p}waic_1 \\
=& \sum_{i=1}^{n} \log \left(\frac{1}{M} \sum_{m=1}^{M} p\left(\boldsymbol{m}_i \mid \tilde{\mathbf{Y}}_i^m, \boldsymbol{\nu}^m\right) p\left(\tilde{\mathbf{Y}}_i^m \mid \boldsymbol{\nu}^m\right) \right) - \\
& 2 \sum_{i=1}^{n} \left\{ \log \left(\frac{1}{M} \sum_{m=1}^{M} p\left(\boldsymbol{m}_i \mid \tilde{\mathbf{Y}}_i^m, \boldsymbol{\nu}^m\right) p\left(\tilde{\mathbf{Y}}_i^m \mid \boldsymbol{\nu}^m\right) \right) - \right. \\
& \left. \frac{1}{M} \sum_{m=1}^{M} \log p\left(\boldsymbol{m}_i \mid \tilde{\mathbf{Y}}_i^m, \boldsymbol{\nu}^m\right) p\left(\tilde{\mathbf{Y}}_i^m \mid \boldsymbol{\nu}^m\right) \right\}
\end{aligned}
\tag{3.7}
$$

$$
\begin{aligned}
& log \prod_{i=1}^{n} \hat{p}_{post}(\mathbf{Y}_i^{\text{obs}}, \boldsymbol{m}_i) - \hat{p}waic_2 \\
=& \sum_{i=1}^{n} \log \left(\frac{1}{M} \sum_{m=1}^{M} p\left(\boldsymbol{m}_i \mid \tilde{\mathbf{Y}}_i^m, \boldsymbol{\nu}^m\right) p\left(\tilde{\mathbf{Y}}_i^m \mid \boldsymbol{\nu}^m\right) \right) - \\
& \sum_{i=1}^{n} \mathrm{V}_{m=1}^{M} \left\{ \log p\left(\boldsymbol{m}_i \mid \tilde{\mathbf{Y}}_i^m, \boldsymbol{\nu}^m\right) p\left(\tilde{\mathbf{Y}}_i^m \mid \boldsymbol{\nu}^m\right) \right\}
\end{aligned}
\tag{3.8}
$$

In equation (3.7) and (3.8), M denotes the length of total MCMC iterations, $\boldsymbol{\nu}^m$ denotes the m'th posterior MCMC draw of all the model parameters and latent variables. $\tilde{\mathbf{Y}}_i^m$ denote the m'th impuation of $\mathbf{Y}_i \doteq (\mathbf{Y}_i^{\text{obs}}, \mathbf{Y}_i^{\text{mis}})$. $\mathrm{V}_{m=1}^{M}$ represents the sample variance with $\text{var}_{m=1}^{M} a_m = \frac{1}{M-1} \sum_{m=1}^{M} (a_m - \bar{a})$. In this chapter we focused

on WAIC_2, because for practical use, WAIC_2 has closer resemblance to the leave one out cross validation (LOO-CV) and also in practice seems to give results closer to LOO-CV [41].

3.3 Simulation Study

In this section, we focus on the simulation about the envelope-based clustering with missing data approach. For all simulation scenarios $n \in \{50, 100\}$, $\sigma^2 \in \{0.1, 1.0, 3.0\}$, the dimension of the simulated data is set to be $r = 4$ and the true dimension of the informative subspace is set to be $u = 2$. The sample size n is set to be either 50 or 100 to emulate the performance of our clustering approach under small sample size. The true classes is set to be $K = 2$ and the probability of individual belonging to each class is assigned as $\pi_1 = 0.4$, and $\pi_2 = 0.6$. For each simulation scenario, the process is replicated 100 times. The detailed information including generating the trajectory and the missing mechanism is listed below.

- generating the full data $\mathbf{Y}$:

  1. given fixed matrix $\mathbf{A} \in \mathbb{R}^{(p-r)\times r}$, define $\mathbf{C_A}$, $\mathbf{D_A}$
     and calculate $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_0$ based on $\mathbf{C_A}$ and $\mathbf{D_A}$ through equation (3.3).
  2. given fixed $\boldsymbol{\mu}_1 = \{0.5, 0.5\}^\top$, $\boldsymbol{\mu}_2 = \{-0.5, -0.5\}^\top$, $\boldsymbol{\Omega} = 5\mathbf{I}_r$ and $\boldsymbol{\Omega}_0 = \mathbf{I}_{p-r}$, generate $(\mathbf{U}_{i1}, \mathbf{U}_{i2}, \mathbf{V}_i)_{i=1}^n$ as $\mathbf{U}_{i1} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Omega})$, $\mathbf{U}_{i2} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\boldsymbol{\mu}_2, \boldsymbol{\Omega})$ and $\mathbf{V}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_0)$
  3. given fixed $\sigma^2 \in \{0.1, 1, 3\}$, for each i, generate $\boldsymbol{\varepsilon}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2\mathbf{I})$.
  4. for each i, assign the class of each individual with the probability π_1 of being in the first class and π_2 in the second class. Then calculate $\mathbf{Y}_i = \boldsymbol{\Gamma}\mathbf{U}_{ik} + \boldsymbol{\Gamma}_0\mathbf{V}_i + \boldsymbol{\varepsilon}_i$, $i \in \{1, \ldots, n\}$.

- generating the missing data indicator $\boldsymbol{m}$:

1. given fixed $\theta_0 = 1.0$, $\theta_1 = 5.0$, generate $w_{ij} = \theta_0 + \theta_1 Y_{ij} + e_{ij}$, where $e_{ij} \sim \mathcal{N}(0,1)$. $i \in \{1,\ldots,n\}$, $j \in \{1,\ldots,r\}$.

2. set $m_{ij} = 1$ (exist) if and only if $w_{ij} > 0$; $m_{ij} = 0$ (not exist) if and only if $w_{ij} \leq 0$. $i \in \{1,\ldots,n\}$, $j \in \{1,\ldots,r\}$.

After the generating process, we keep Y_{ij} with $m_{ij} = 1$ and discard the rest. For simplicity, the dimension of the informative subspace is set to be the true value $u = 2$. Three approaches were compared which includes 1. Bayesian clustering approach with full data; 2. Bayesian clustering approach with the true missing not at random mechanism; 3. Bayesian clustering approach with ignorable missing mechanism; 4. Growth Mixture Model (linear random effect) with ignorable missing mechanism.

Table 3.1: The percentage of selecting the correct number of clusters ($K = 2$).

σ^2	0.1		1.0		3.0	
n	50	100	50	100	50	100
Bys full	96%	99%	94%	99%	91%	95%
Bys nigr	86%	91%	79%	82%	71%	77%
Bys igr	81%	84%	66%	73%	58%	62%
Growth igr	62%	66%	54%	56%	38%	41%

As shown in the table, four methods in selecting the correct number of clusters ($K = 2$) under varying conditions of noise variance σ^2 and sample size n were compared. It is evident from the data that the Bayesian clustering approach with full data consistently outperforms the others across all scenarios, maintaining high accuracy ($91\% - 99\%$) regardless of the increase in noise variance or sample size. Among the rest of the three methods, Bayesian clustering with non-ignorable missing mechanism shows commendable performance, and its accuracy decreases more noticeably as the noise variance increases. The Growth Mixture method, while still effective

to a degree, demonstrates the lowest accuracy among the four methods, particularly struggling in higher noise and lower sample size environments.

3.4 Trajectory Clustering Results of $HAMD_{17}$

The goal of our model-based clustering approach is to find out whether the latent class exists in the $HAMD_{17}$ trajectories during the antidepressant treatment. If the patient reacts to the treatment, its $HAMD_{17}$ trajectory would improve rapidly, otherwise, it is not. Thus we are trying to detect the heterogeneity with one latent class corresponding to the rapid improvement individuals and another corresponding to the slowly improvement individuals. In this section, we apply the Bayesian clustering method to the $HAMD_{17}$ trajectory. First the analyze results with complete trajectories are shown and then followed by the analyze results with all trajectories with missing data. For comparison, growth mixture model [71] and Gaussian mixture model with different covariance structures are also applied and the results are summarized.

3.4.1 Study of Cases Without Missing Data

Among the $n = 94$ $HAMD_{17}$ trajectories, $n_c = 60$ are cases without missing data. Figure (3.1) shows the $HAMD_{17}$ trajectory plot for the 60 complete cases.

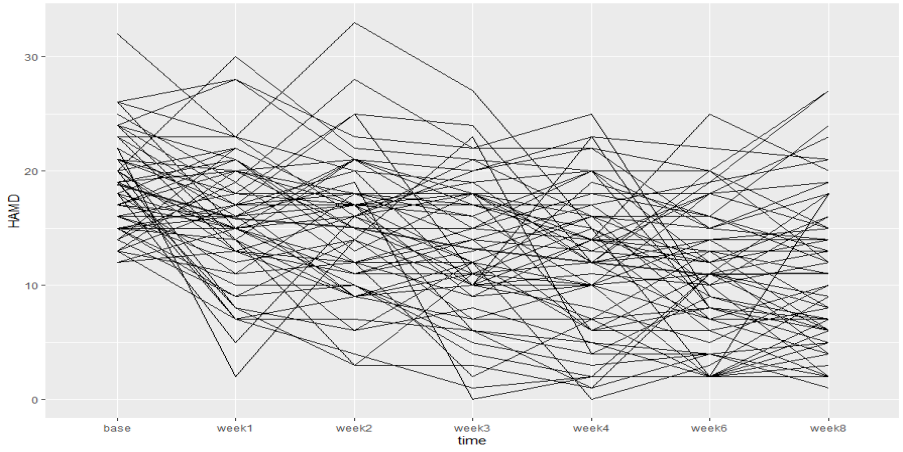


Figure 3.1: $HAMD_{17}$ trajectory plot for the 60 cases without missing data.

The Gaussian Mixture Model (GMM) assumes that the trajectory follows a multi-dimensional Gaussian distribution. We use the R package **mclust** [72] when applying GMM with different within-class covariance structures. The R package **mclust** is used for model-based clustering and classification using Gaussian Mixture Model, where different covariance structures are denoted by abbreviations like 'EII', 'VVI', 'EEE', 'VVI', 'VVV'. These abbreviations represent assumptions about cluster shapes and orientations: 'EEI' implies equal spherical shapes for all clusters; 'VVI' allows for different sizes but maintains spherical shapes; 'EEE' suggests equal ellipsoidal shapes with possible variations in orientation; 'VVV' is the most flexible, allowing each cluster to have its own ellipsoidal shape, size, and orientation. These models provide a range of options for capturing the underlying structure in the data, from simple spherical clusters ('EEI') to more complex, varied ellipsoidal shapes ('VVV'). The choice of model impacts how data is partitioned into clusters, making mclust a versatile tool for data analysis.

Table 3.2 lists the within-class covariance matrix $\mathbf{\Sigma}_k$, $k = 1, \dots K$. In table 3.2, λ and λ_k are real numbers. $\mathbf{T}$ and $\mathbf{T}_k$ are diagonal matrices. $\mathbf{Q}$ and $\mathbf{Q}_k$ are orthogonal matrices.

Table 3.2: Some examples of the within-group covariance matrix $\mathbf{\Sigma}_k$ in Gaussian Mixture Model from simple to complex.

Model	$\mathbf{\Sigma}_k$
EEI	$\lambda\mathbf{T}$
VVI	$\lambda_k\mathbf{T}_k$
EEE	$\lambda\mathbf{Q}\mathbf{T}\mathbf{Q}^\top$
VVE	$\lambda_k\mathbf{Q}\mathbf{T}_k\mathbf{Q}^\top$
VVV	$\lambda_k\mathbf{Q}_k\mathbf{T}_k\mathbf{Q}_k^\top$

Table (3.3) shows the plot of BIC value for different Gaussian Mixture Models. As

Table 3.3: BIC summary table for different GMMs, except for EII, BIC selects EEE with one latent component.

Mixtures	EEI	VVI	**EEE**	VVE	VVV
1	−2717	−2717	**−2614**	−2614	−2614
2	−2623	−2649	−2631	−2642	−2703
3	−2626	−2666	−2650	−2675	−2776
4	−2643	−2691	−2671	-	-
5	−2662	−2731	−2692	-	-
6	−2652	−2777	−2700	-	-

we can see, Gaussian Mixture supports only one latent class. Note that the covariance assumption in our Bayesian clustering framework is based on the assumption that there exist a low-dimensional subspace that fully captures the cluster variation. However, the GMMs have no such interpretation. Because $n_c = 60$ and some of the models cannot be estimated, whose results are denoted by '-' in Table (3.3).

Growth models, typically used in longitudinal data analysis, focus on estimating and predicting individual change over time and understanding the factors influencing this change. These models often use repeated measurements to track development, learning, or progress in a subject. The Growth Mixture Model, an extension of these basic growth models, introduces the concept of latent classes to account for unobserved heterogeneity within the population. Unlike traditional growth models that assume a single underlying growth trajectory for the entire population, Growth Mixture Model allows for the existence of multiple latent subgroups within the population, each with its distinct growth trajectory. This approach is particularly valuable in fields like psychology or education, where individuals' development paths can vary significantly, and understanding these different paths can provide deeper insights into

the factors driving change over time. We fit the Growth Mixture Model with the function **hlme** in R package **lcmm** [73]. Table (3.4) shows the BIC for Growth Mixture Models with linear, quadratic and cubic random effects of time. Unfortunately, for all random effects, the BIC criterion selects only one latent class.

Table 3.4: BIC summary table for complete cases with Growth Mixture Model

	$\boldsymbol{K=1}$	$K=2$	$K=3$
Linear	**2550.49**	2560.17	2566.42
Quadratic	**2533.97**	2545.72	2550.74
Cubic	**2551.04**	2565.61	2580.13

Table (3.5) shows the WAIC for our Bayesian clustering approach. The MCMC has $40,000$ iterations and saved every 5th iteration. The first $20,000$ iterations are discarded as the burn-in period. The WAIC criterion selects the model $(K=2, u=3)$

Table 3.5: WAIC_2 for the 60 trajectories without missing data in our Bayesian clustering approach.

	$K=1$	$K=2$	$K=3$
$u=2$	-1281.7	-1289.4	-1294.5
$u=3$	-1282.7	$\mathbf{-1280.8}$	-1296.4
$u=4$	-1286.0	-1299.6	-1286.2

Figure (3.2) shows our Bayesian clustering result on the original space with the mean trajectories and their corresponding 95% Credible Intervals. As shown in the figure, two classes one with rapid improvement and the other with slowly improvement were captured.

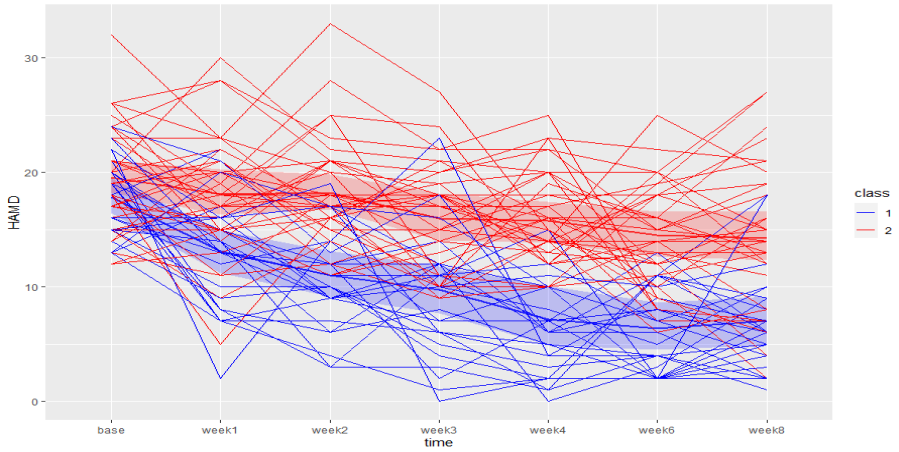


Figure 3.2: Bayesian clustering trajectory result for HAMD_{17} without missing data, $(u = 3, K = 2)$

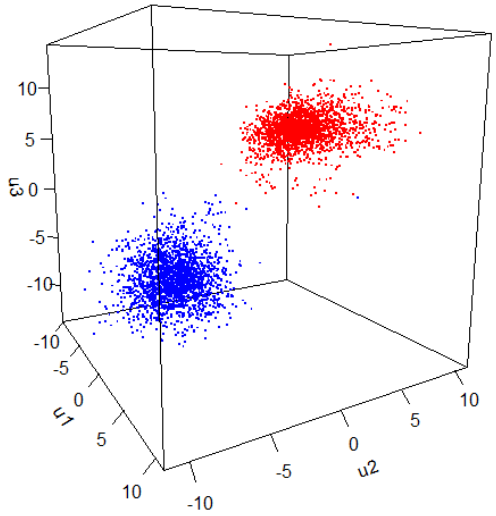


Figure 3.3: Clustering posterior draws of $\boldsymbol{\mu}_k$ on the reduced $span(\boldsymbol{\Gamma})$ 3D-space (material part) for HAMD_{17} trajectories without missing data, $k = 1, 2$. $(u = 3, K = 2)$

Figure (3.3) shows our Bayesian clustering posterior draws of $\boldsymbol{\mu}_k$ on the reduced $u = 3$ dimensional subspace, $k = 1, 2$. Figure (3.4) shows the projected $\boldsymbol{\Gamma}^\top \mathbf{Y}$ on the reduced $u = 3$ dimensional subspace. The red and blue color denotes two classes. By assuming the clustering variation information only depends on the projected space, the $r = 7$ dimensional clustering problem is transferred into $u = 3$ dimensional. It can

be seen that on this 3D space, there almost exist a hyper-plane which can separate the two classes.

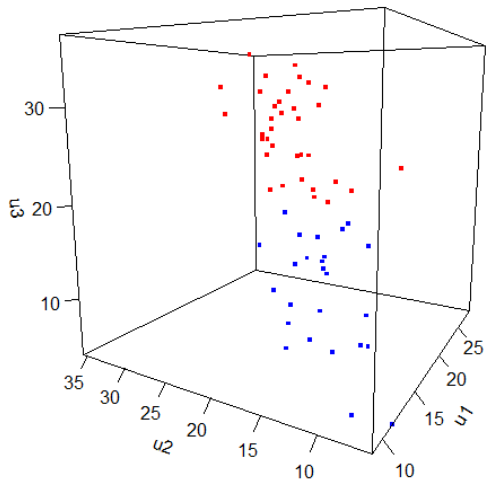
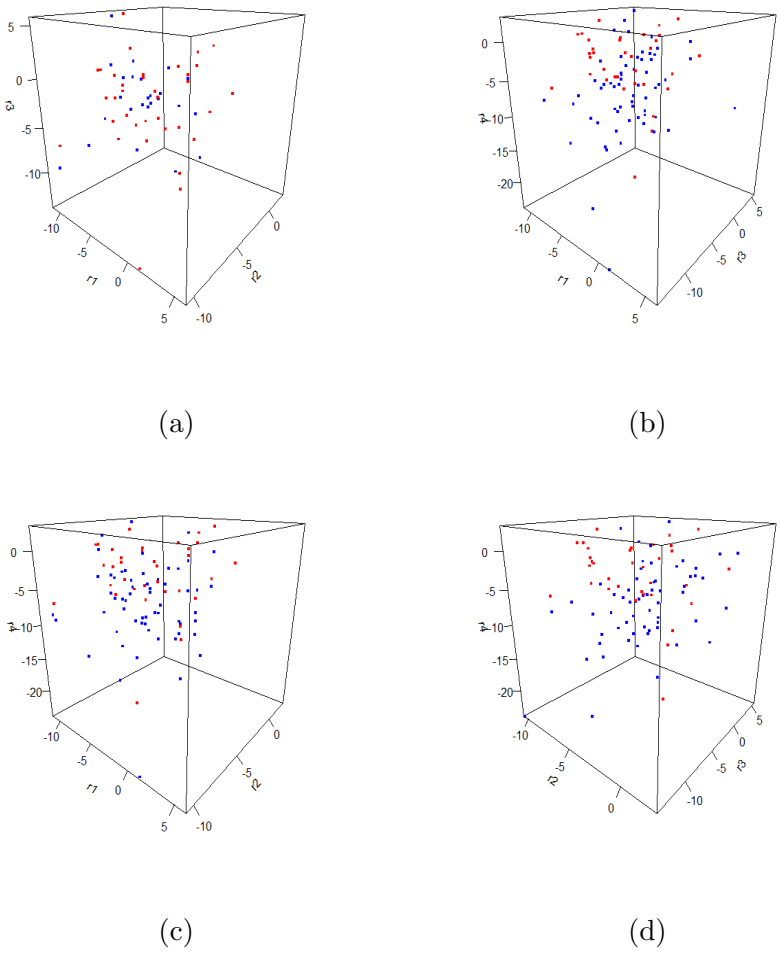


Figure 3.4: Scatter plots on the projected $span(\mathbf{\Gamma})$ 3D-space for HAMD_{17} trajectories without missing data, $(u = 3, K = 2)$.

$\mathbf{\Gamma}_0^\top \mathbf{Y}$ projects the trajectory onto the orthogonal compliment of $span(\mathbf{\Gamma})$ with dimension $r - u = 4$. In order to visualize $\mathbf{\Gamma}_0^\top \mathbf{Y}$, Figure (3.5) shows the 3D scatter plot with all combinations of the four dimensions (r_1, r_2, r_3, r_4): (r_1, r_2, r_3), (r_1, r_2, r_4), (r_1, r_3, r_4), (r_2, r_3, r_4). It can be seen on the scatter plot (3.5), there is no clustering variation information as the dots are random distributed around 0 for both classes.

3.4.2 Study of Cases with Missing Data

The $n = 94$ HAMD_{17} trajectories are analyzed in this section. Table (3.6) shows the BIC for Growth Mixture Models with linear, quadratic and cubic random effects of time for all 94 trajectories. Again, for all random effects, the BIC criterion selects only one latent class. Since there is no R package at hand to apply Gaussian Mixture Model with missing data, the result is not compared in this section.



(a) (b)

(c) (d)

Figure 3.5: Scatter plot on the projected $span(\mathbf{\Gamma}_0)$ space for HAMD_{17} trajectories without missing data, $(u = 3, K = 2)$. Four figures represent four dimension combinations of (r_1, r_2, r_3, r_4).

Table 3.6: BIC using Growth Mixture Model assuming ignorable missingness for all HAMD_{17} trajectories.

	$K = 1$	$K = 2$	$K = 3$
Linear	**−3568.29**	−3577.63	−3586.76
Quadratic	**−3535.31**	−3547.7	−3555.04
Cubic	**−3551.87**	−3564.51	−3579.03

Table (3.7) shows the WAIC for our Bayesian clustering approach. Similarly, the MCMC chain has $40,000$ iterations and saved every 5th iteration. The first $20,000$ iterations are discarded as the burn-in period. The WAIC criterion selects the model $(u = 2, K = 2)$.

Table 3.7: WAIC_2 for the 94 HAMD_{17} trajectories with missing data in our Bayesian clustering approach.

	$K = 1$	$K = 2$	$K = 3$
$u = 1$	-2053.7	-2022.3	-2020.5
$u = 2$	-2058.1	$\mathbf{-2015.8}$	-2021.8
$u = 3$	-2054.4	-2023.0	-2024.1

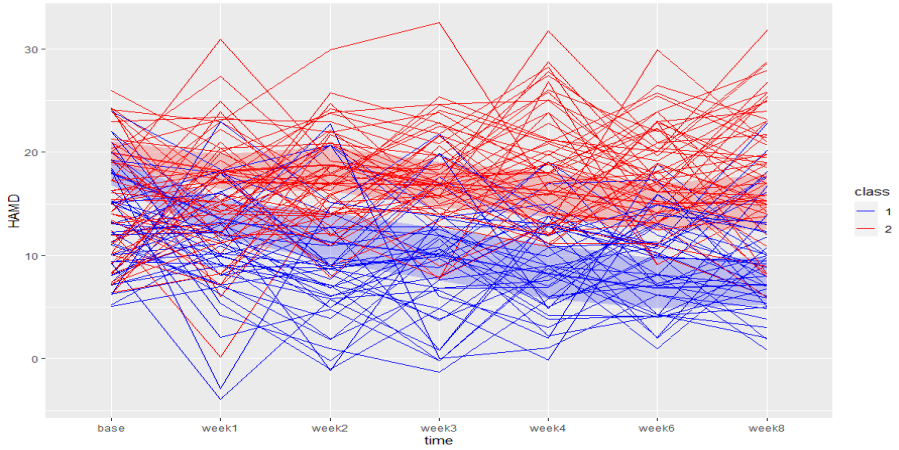


Figure 3.6: Bayesian clustering trajectory for HAMD_{17} with missing data, $(u = 2, K = 2)$ after imputation.

Figure (3.6) shows the Bayesian clustering result on the original space with the mean trajectories and their corresponding 95% Credible Intervals. As shown in the figure, the pattern of the two classes matches the counter part in the previous section: one with rapid improvement and the other with slowly improvement.

Figure (3.7) shows the Bayesian clustering posterior distribution of $\boldsymbol{\mu}_k$ on the reduced $u = 2$ dimensional subspace, $k = 1, 2$. It can be seen that the clustering

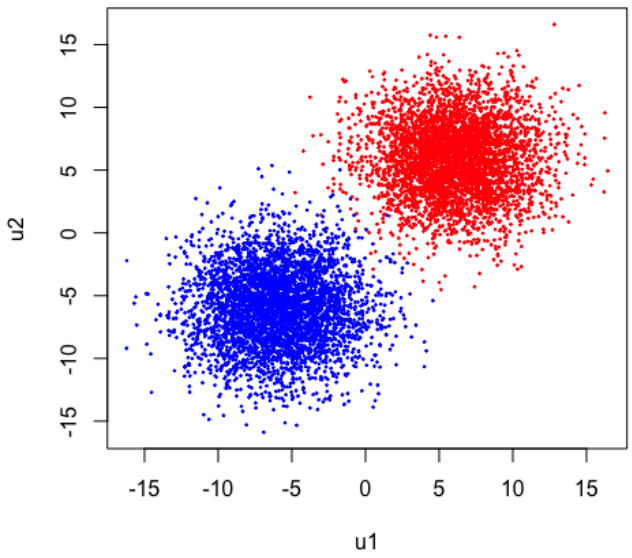


Figure 3.7: Clustering posterior draws of $\boldsymbol{\mu}_k$ on the reduced $span(\boldsymbol{\Gamma})$ 2D-space (material part) for HAMD_{17} trajectories with missing data, $k = 1, 2$. $(u = 2, K = 2)$

variation information depends on $\mathbf{S}$. Figure (3.8) shows the projected $\boldsymbol{\Gamma}^\top \mathbf{Y}$ on the reduced $u = 2$ dimensional subspace. The red and blue color denotes two classes. By assuming the clustering variation information only depends on the projected space, the $r = 7$ dimensional clustering problem is transferred into $u = 2$ dimensional space. It can be seen that on this 2D space, there almost exist a line which can separate the two classes.

$\boldsymbol{\Gamma}_0^\top \mathbf{Y}$ projects the trajectory onto the orthogonal compliment of $span(\boldsymbol{\Gamma})$ with dimension $r - u = 5$. In order to visualize $\boldsymbol{\Gamma}_0^\top \mathbf{Y}$, Figure (3.9) shows the 3D scatter plot with 4 random combinations of the 5 dimensions $(\mathrm{r}_1, r_2, r_3, r_4, r_5)$. It can be seen on the scatter plot (3.5), there is no clustering variation information as the dots are random distributed around 0 for both classes.

Table (3.8) showed the summary statistics for $\boldsymbol{\theta}$. As we can see, θ_2 is significant with the 0.95 threshold. In fact, we should study more MNAR models and perform the robust analysis. However, due to the length of this project we leave it to further

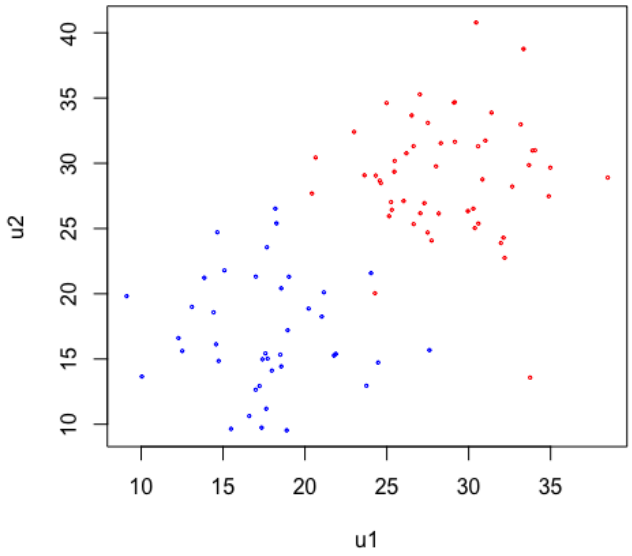


Figure 3.8: Scatter plots on the projected $span(\mathbf{\Gamma})$ 2D-space for HAMD_{17} trajectories with missing data, $(u = 2, K = 2)$.

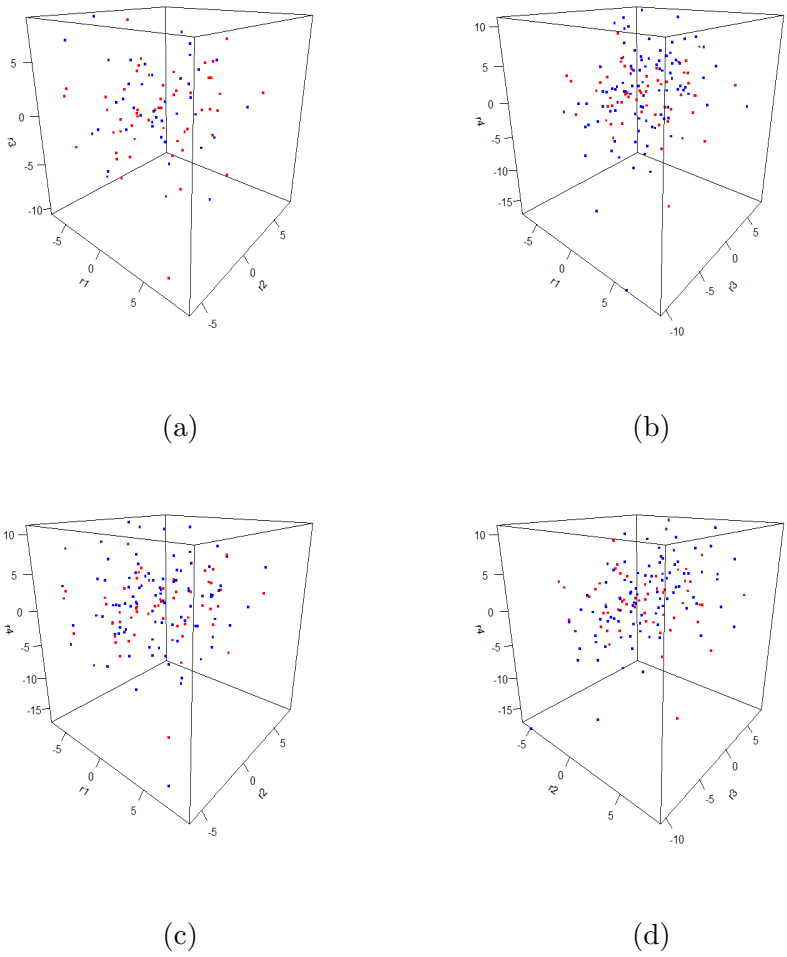
	mean	st.d	95% C.I
θ_0	1.22	0.11	$(1.05, 1.52)$
θ_1	0.03	0.02	$(-0.01, 0.07)$
θ_2	-0.10	0.05	$(-0.20, -0.01)$

Table 3.8: The summary statistics for $\boldsymbol{\theta}$.

studies. Moreover, the results from both section showed the similar trajectory class pattern which showed the consistency of the analysis no matter only with the complete cases or the full dataset.

3.5 Discussion and Conclusion

In this chapter, we have introduced a pioneering Bayesian envelope-based clustering framework designed to navigate the complexities inherent in analyzing HAMD_{17}



(a) (b)

(c) (d)

Figure 3.9: Scatter plot of HAMD_{17} trajectories on the projected $span(\mathbf{\Gamma}_0)$ space, with $(u = 2, K = 2)$. Four figures represent four random dimension combinations.

trajectories, particularly under the challenge of non-ignorable missing data. This framework represents a significant methodological leap, focusing on dimension reduction to identify and capture the quintessence of clustering information within a designated subspace, thereby facilitating a more nuanced and efficient analysis of depression severity over time.

One of the most compelling aspects of our approach is its capacity to discern latent subgroups within the patient data, which are indicative of differential responses to antidepressant treatments. The simulation studies conducted, alongside the empirical analysis using data from the EMBARC study, underscore the robustness and

superior performance of our model, especially when juxtaposed with traditional clustering methods such as Gaussian Mixture Models and Growth Mixture Models. These conventional methods, while valuable, often fall short in addressing the specific challenges posed by mental health data, particularly the nuanced aspects of missing data patterns and their implications on the analysis outcomes.

Our model's innovative integration of a non-ignorable missing data mechanism, modeled through Probit regression, is a testament to the sophisticated understanding required to adequately address the phenomenon of missingness in clinical trials. This aspect is crucial, as the nature of missing data in such studies is seldom random and typically dependent on unobserved variables, such as the severity of depression or the trajectory of response to treatment. By acknowledging and modeling this complexity, our approach not only enhances the reliability of the clustering outcomes but also ensures that the inferences drawn are grounded in a realistic representation of the data.

The development and application of the Bayesian envelope-based clustering model for HAMD_{17} trajectories represent a significant advancement in the field of mental health research. Through this study, we have demonstrated the model's unparalleled ability to identify latent classes within depression studies, thereby illuminating the heterogeneity of patient responses to antidepressant treatments. The detection of distinct subpopulations characterized by rapid and slow improvement trajectories offers critical insights that could revolutionize personalized treatment strategies, emphasizing the potential for targeted interventions tailored to individual patient profiles.

Moreover, the model's adept handling of non-ignorable missing data through a comprehensive missing data mechanism marks a methodological milestone. This feature addresses one of the most pressing challenges in clinical research, ensuring that our findings are not only accurate but also reflective of the complex realities of data collection in mental health studies. The application of our model to the nuanced dataset from the EMBARC study serves as a validation of its utility and efficacy,

highlighting its promise for future applications in both research and clinical settings.

Despite the promising advancements offered by the Bayesian envelope-based clustering model in analyzing HAMD_{17} trajectories, our study is not without limitations. A notable constraint is the assumption of non-ignorable missing data mechanism modeled specifically through Probit regression. While this approach is robust and offers significant insights into the patterns of missingness, it may not encapsulate all possible mechanisms of missing data in clinical studies. Future research could explore alternative models for missing data that accommodate a wider range of missingness patterns, potentially enhancing the model's applicability and accuracy across diverse clinical datasets. Additionally, the current study's focus on HAMD_{17} trajectories within a specific antidepressant treatment context may limit the generalizability of our findings. Expanding the model's application to other treatments and mental health conditions could provide a more comprehensive understanding of patient heterogeneity and treatment efficacy. Moreover, integrating other types of data, such as genetic information or biomarkers, into the Bayesian hierarchical framework could offer deeper insights into the biological underpinnings of depression and its treatment, paving the way for truly personalized medicine. Finally, the implementation of our model in real-world clinical settings presents an exciting avenue for future research. Developing user-friendly software tools and guidelines for applying the Bayesian envelope-based clustering model could facilitate its adoption by clinicians and researchers, ultimately benefiting patient care and advancing the field of mental health research. The journey ahead is filled with opportunities to refine, expand, and apply the innovations presented in this study, promising significant contributions to our understanding and treatment of depression.

Chapter 4

Differentially Private Bayesian Envelope Regression

4.1 Introduction

As technological advancements continue to accelerate, we are faced with the challenge of managing and understanding increasingly complex data. Dimensionality reduction is a fundamental tool for understanding such complex data. Despite each data point often consists a large number of features, the underlying subject of interest is typically lower-dimensional. Reducing the data's "extrinsic" dimension to its "intrinsic" dimension enables analysts to unveil critical structural relationships among features. This dimensionality reduction not only facilitates more efficient utilization of the data for learning tasks, such as classification and regression but also significantly diminishes the storage space required for the data. On one hand, it streamlines these datasets into a more tractable form, retaining crucial information which facilitates simpler analysis and interpretation. On the other hand, dimensionality reduction techniques are pivotal in diminishing the number of variables under consideration. This reduction is essential for addressing challenges such as the curse of dimensionality and the risk of overfitting in statistical models. By lowering the complexity of the data, these techniques contribute to more robust and generalizable model construction.

As data complexity grows, so does the imperative to safeguard data privacy. Differential Privacy (DP) [74] has gained recognition as a prominent mathematical frame-

work for quantifying privacy protection, and several privacy mechanisms [75] have been devised to achieve DP. It entails the introduction of calibrated random fluctuations into algorithmic calculations to demonstrably constrain the probability of individual-specific information being revealed through the algorithm's output. Such a guarantee protects the privacy of individuals while still allowing valuable insights to be extracted from the data. DP is widely used in various applications, especially in scenarios where sensitive data needs to be analyzed, such as healthcare research and census data analysis. Building upon this foundation, our paper introduces an innovative amalgamation of linear regression with DP by leveraging the envelope concept. This framework markedly improves the efficiency of coefficient estimation by adeptly filtering out extraneous information among predictors, showcasing a pioneering application of DP in enhancing statistical analysis precision while concurrently securing sensitive data.

Related Work In the past decade, there has been a significant development of methodologies adapting traditional linear regression to ensure DP. These methodologies are broadly categorized into frequentist and Bayesian approaches. Frequentist approaches include sufficient statistic perturbation and subsample aggregation. Sufficient statistic perturbation involves adding noise to the statistics computed from the data. This method is exemplified in the works of [76] and [77], where noise is added to the sufficient statistics of Ordinary Least Squares (OLS) computations to achieve DP. On the other hand, subsample aggregation, as outlined by [78], involves aggregating results from multiple subsamples, each treated with DP mechanisms. Bayesian approaches, such as the MCMC data augmentation using sufficient statistics, are another significant category. This approach, seen in the works of [79] and [80], employs MCMC methods in conjunction with DP to perform Bayesian linear regression. The focal point of the prior work in the DP linear regression has predominantly centered on enhancing the accuracy of estimation. This trend is evident in the literature, where

the precision of estimation under differential privacy constraints has been the primary objective. Nevertheless, a relatively limited subset of recent research has ventured beyond mere utility aspects to explore the underlying regression structure. A few notable studies in this area include the works of [81–84]. Specifically, these studies incorporate methods such as PCA and regularization techniques like Lasso, Ridge, or Sparse Regression. PCA is employed for dimensionality reduction, whereas regularization methods are used to address issues like overfitting and to enhance model interpretability in the context of high-dimensional data. The incorporation of these conventional techniques in DP linear regression signifies an attempt to strike a balance between maintaining privacy and preserving the integrity of the underlying regression structure. However, the exploration of regression structures in differentially private settings remains an area with ample room for further research and development.

Motivation Motivated by the simple observation that certain variations in the predictors may have no discernible effect on the response variable [59], we focus on a model which assumes that there are "non-important" variations among the predictors when predicting the response and we call it the immaterial component. Our goal is to develop a privacy-preserving linear regression model which can identify the immaterial information among the predictors. In the meantime, such recognition of the material immaterial separation would improve the parameters estimation efficiency. The fundamental principle of our model is to identify a subspace (the envelope) within the predictor variable space that encapsulates the maximum variation related to the response. By focusing only on the material variation, one significant advantage of predictor envelope regression is its potential for efficient estimation of regression coefficients. Moreover, it also offers a clear and interpretable estimation. It segregates the space of the predictor variables into the envelope subspace and its orthogonal complement, simplifying the interpretation of regression results.

Our Contribution In this chapter, we bridge this gap by developing a Differentially private Bayesian envelope linear regression model. Our model innovatively utilizes the predictor envelope methodology within a privacy-preserving framework, leveraging the strengths of envelope methodology to enhance both the efficiency and interpretability of differentially private linear regression. There are two key contributions to our study. Firstly, the introduction of a novel Bayesian hierarchical framework for linear regression that incorporates envelope methodology for identifying immaterial predictor information while ensuring differential privacy. This is the first known application of the envelope idea in privacy preservation, representing a significant interdisciplinary innovation. Secondly, the use of an envelope nucleus within our framework provides a refined approach to dimension reduction and coefficient estimation. By focusing on the essential aspects of the predictor covariance matrix, our model goes beyond traditional regression methods that uniformly treat all components of the response variable's covariance structure. This targeted approach not only improves estimation efficiency but also reduces the risk of overfitting. Through this integration of envelope methodology with differential privacy, our approach marks a significant step forward in developing privacy-preserving statistical methods. It retains the advantages of envelope regression, such as efficiency and clarity in estimation, while introducing a novel dimension of privacy assurance that is increasingly critical in the age with more sensitive data.

The manuscriptis organized as follows: Section 4.2 delineates the foundational principles of differential privacy and envelope methodology, setting the stage for the subsequent discussions. Section 4.3 unveils the novel framework of our differentially private Bayesian envelope analysis. In Section 4.4, simulation experiments are conducted to evaluate the performance of the proposed method against the conventional technique presented by [79]. The results underscore the superiority of our approach in terms of estimation precision and the conciseness of credible intervals, attributing these advancements to the efficient utilization of the underlying low-dimensional data

structure.

4.2 Preliminaries

In this section, we briefly revisit some background materials for differential privacy [74] and envelope methodology [59].

4.2.1 Differential Privacy

Differential privacy is a privacy definition that is tailored to the task of privacy-preserving data analysis. First introduced in [74], it quickly gains popularity as it provides a mathematically rigor framework to quantify the amount of privacy protection. The formal definition of ϵ-Differential Privacy is as follows,

Definition 6 (ϵ-Differential Privacy [74]) *For any $\epsilon > 0$, a mechanism $\mathcal{A}$ is said to be ϵ-differentially private (ϵ-DP) if for all measurable sets $\mathcal{S}$ and for all pairs of neighboring datasets $\mathbf{X}$ and $\mathbf{X}'$, where neighboring datasets refer to two datasets that differ by only one element, the following holds,*

$$\mathbb{P}(\mathcal{A}(\mathbf{X}) \in \mathcal{S}) \leq \exp(\epsilon)\mathbb{P}\left(\mathcal{A}\left(\mathbf{X}'\right) \in \mathcal{S}\right) \tag{4.1}$$

The parameter ϵ quantifies the level of privacy protection. Smaller values of ϵ provide stronger privacy guarantees. As ϵ increases, the privacy protection decreases. A natural relaxation of ϵ-DP is approximate (ϵ, δ)-differential privacy, which allows for a more relaxed level of privacy in some scenarios. Approximate differential privacy has found widespread application in practical settings where a small amount of privacy leakage is acceptable in exchange for improved accuracy or utility of the analysis. In this chapter, we specifically focus on approximate differential privacy and utilize the analytic Gaussian mechanism [85] as our guarantee.

Definition 7 (Approximate Differential Privacy [74, 75]) *For any $\delta \in (0, 1)$ and $\epsilon > 0$, a mechanism $\mathcal{A}$ is said to be approximate (ϵ, δ)- differentially private if for*

all measurable sets $\mathcal{S}$ *and for all pairs of neighboring datasets* X *and* X′*, the following holds true,*

$$\mathbb{P}(\mathcal{A}(\mathbf{X}) \in \mathcal{S}) \leq \exp(\varepsilon)\mathbb{P}\left(\mathcal{A}\left(\mathbf{X}'\right) \in \mathcal{S}\right) + \delta \tag{4.2}$$

To achieve (ϵ, δ)-DP, the analytic Gaussian mechanism [85] is one of the most commonly used mechanisms. It improves the original Gaussian mechanism by calibrating the variance directly using the Gaussian cumulative density function instead of a tail-bound approximation.

Definition 8 (Analytic Gaussian Mechanism [85]) *Let* $f : \mathcal{D} \to \mathbb{R}^d$ *be a function with global* L_2 *sensitivity* Δ_2.[1] *For any* $\varepsilon \geq 0$ *and* $\delta \in [0, 1]$*, the Gaussian output perturbation mechanism* $M(x) = f(x) + Z$ *with* $Z \sim \mathcal{N}\left(0, \sigma_{\text{dp}}^2 I\right)$ *is* (ε, δ)*-differentially private if and only if*

$$\Phi\left(\frac{\Delta}{2\sigma_{\text{dp}}} - \frac{\varepsilon\sigma_{\text{dp}}}{\Delta}\right) - e^{\varepsilon}\Phi\left(-\frac{\Delta}{2\sigma_{\text{dp}}} - \frac{\varepsilon\sigma_{\text{dp}}}{\Delta}\right) \leq \delta \tag{4.3}$$

where $\Phi(\cdot)$ *denotes the cumulative distribution function of standard normal distribution.*

Please refer to [85] for more information and a detailed algorithm for calibrating σ_{dp}^2.

4.2.2 Predictor Envelope Regression

Consider the following univariate linear regression model with p predictor variables,

$$y_i = \boldsymbol{x}_i^{\top}\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n \tag{4.4}$$

$$\varepsilon_i \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2), \quad i = 1, \ldots, n \tag{4.5}$$

Here, y_i represents the scalar response of the ith observation, and $\boldsymbol{x}_i \in \mathbb{R}^p$ represents the predictor vector of the ith observation with a mean of $\mathbf{0}$ and a covariance

[1] $\mathcal{D}$ is the space of the input datasets, and the global L_2 sensitivity of f is defined as $\Delta_2 = \max_{\{\mathbf{X}, \mathbf{X}'\}} \|f(\mathbf{X}) - f\left(\mathbf{X}'\right)\|_2$.

matrix of $\mathbf{\Sigma_x}$. Predictor envelopes for model (4.4) gain efficiency in the estimation of $\boldsymbol{\beta}$ by incorporating the projection $\mathbf{P}_{\mathcal{E}}\boldsymbol{x}$ onto the smallest subspace $\mathcal{E} \in \mathbb{R}^r$ with the following proprieties [86].

1. The distribution of $\mathbf{Q}_{\mathcal{E}}\boldsymbol{x}$ is uncorrelated with $\mathbf{P}_{\mathcal{E}}\boldsymbol{x}$, where $\mathbf{Q}_{\mathcal{E}} = \mathbf{I}_{\mathcal{E}} - \mathbf{P}_{\mathcal{E}}$, and

2. y be uncorrelated with $\mathbf{Q}_{\mathcal{E}}\boldsymbol{x}$ given $\mathbf{P}_{\mathcal{E}}\boldsymbol{x}$.

For any $\mathcal{E}$ with properties (1) and (2), it is said that $\mathbf{Q}_{\mathcal{E}}\boldsymbol{x}$ is **linearly immaterial** to the regression since $\mathbf{Q}_{\mathcal{E}}\boldsymbol{x}$ depends linearly on neither $\mathbf{P}_{\mathcal{E}}\boldsymbol{x}$ nor y. Consequently, $\mathbf{P}_{\mathcal{E}}\boldsymbol{x}$ must carry all of the information that is **linearly material** to the regression, i.e. all of the information that is available about $\boldsymbol{\beta}$ from $\boldsymbol{x}$.

Denote $\mathbb{R}^{m\times n}$ as the collection of all real matrices with size $m \times n$ and $\mathbb{S}^{k\times k}$ as the collection of all real, symmetric matrices with size $k \times k$. For a given matrix $\mathbf{A} \in \mathbb{R}^{m\times n}$, we denote $\text{span}(\mathbf{A}) \subseteq \mathbb{R}^m$ as the subspace spanned by columns of $\mathbf{A}$.

Definition 9 (Reducing Subspace [87]) *A subspace $\mathcal{R} \subseteq \mathbb{R}^p$ is said to be a reducing subspace of $\mathbf{M} \in \mathbb{R}^{p\times p}$ if $\mathcal{R}$ decomposes $\mathbf{M}$ as $\mathbf{M} = \mathbf{P}_{\mathcal{R}}\mathbf{M}\mathbf{P}_{\mathcal{R}} + \mathbf{Q}_{\mathcal{R}}\mathbf{M}\mathbf{Q}_{\mathcal{R}}$. If $\mathcal{R}$ is a reducing subspace of $\mathbf{M}$, we say that $\mathcal{R}$ reduces $\mathbf{M}$.*

Definition 10 (Envelope [59]) *Let $\mathbf{M} \in \mathbb{S}^{p\times p}$ and let $\mathcal{B} \subseteq \text{span}(\mathbf{M})$. Then the M-envelope of $\mathcal{B}$, denoted by $\mathcal{E}_{\mathbf{M}}(\mathcal{B})$, is the intersection of all reducing subspaces of $\mathbf{M}$ that* contain *$\mathcal{B}$.*

With the above definitions, it can be shown that $\mathbf{P}_{\mathcal{E}}$ and $\mathbf{Q}_{\mathcal{E}}$ is the smallest reducing subspace of $\mathbf{\Sigma_x}$, which can be achieved by intersecting all the reducing subspace of $\mathbf{\Sigma_x}$. $\mathbf{P}_{\mathcal{E}}$ is then defined formally as the projection onto $\mathcal{E}_{\Sigma_{\mathbf{x}}}(\text{span}(\boldsymbol{\beta}))$ which "envelopes" $\text{span}(\boldsymbol{\beta})$. Let $\mathbf{B}_1$ be a orthogonal basis of $\mathbf{P}_{\mathcal{E}}$ and $\mathbf{B}_2$ be a orthogonal basis of $\mathbf{Q}_{\mathcal{E}}$, then the envelope parametric version of (4.4) can be expressed as

$$y_i = \boldsymbol{x}_i^\top \mathbf{B}_1\theta + \varepsilon_i, \text{ with } \mathbf{\Sigma_x} = \mathbf{B}_1\mathbf{\Omega}\mathbf{B}_1^\top + \mathbf{B}_2\mathbf{\Omega}_0\mathbf{B}_2^\top, \tag{4.6}$$

where $\boldsymbol{\beta} = \mathbf{B}_1\boldsymbol{\theta}$, $\boldsymbol{\theta} \in \mathbb{R}^r$. r is the dimensionality of $\mathbf{P}_{\mathcal{E}}$, which is also called the **envelope dimension**. $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ are positive definite matrices. The correlation between y and $\boldsymbol{x}$ is only through the projection of $\boldsymbol{x}$ on $\mathbf{P}_{\mathcal{E}}$, which is $\mathbf{B}_1^\top \boldsymbol{x}$. The parameters of interest $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\mathbf{x}}$ depend only on $\mathbf{P}_{\mathcal{E}}$ and not on the basis.

4.3 Data Augmentation MCMC for Envelope Linear Regression via Privatized Sufficient Statistics

In this section, we present our data augmentation framework for Bayesian linear regression. The proposed framework allows the practitioners to perform valid Bayesian inference on linear regression with differentially privatized sufficient statistics generated through the analytic Gaussian mechanism [85]. The key distinction in our approach is our model takes into account the low-dimensional structure of the data by partitioning the predictors into material and immaterial components, based on the observation that certain changes in the predictors might have no impact on the response.

Specifically, let $\boldsymbol{x}_i$ denote a predictor in a p-dimensional space, and let the scalar y_i represent the corresponding response. Let $\boldsymbol{\nu}$ be all the model parameters. Consider the unobserved confidential dataset represented by $\mathbf{x} \doteq (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top \in \mathbb{R}^{n\times p}$ and $\mathbf{y} \doteq (y_1, \ldots, y_n)^\top \in \mathbb{R}^n$. Instead of having direct access to the dataset $(\mathbf{x}, \mathbf{y})$, our observations are limited to privatized sufficient statistics for $\boldsymbol{\nu}$ which is denoted as $\boldsymbol{s}_{\text{dp}}$. Based on the Bayes' rule, we are concerned with the following posterior distribution:

$$p\left(\boldsymbol{\nu} \mid \boldsymbol{s}_{\text{dp}}\right) \propto p(\boldsymbol{\nu}) p\left(\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{\nu}\right) \tag{4.7}$$

As the marginal likelihood $p\left(\boldsymbol{s}_{\text{dp}} \mid \boldsymbol{\nu}\right)$ is often unknown, we augment the MCMC state space with the latent confidential dataset $(\mathbf{x}, \mathbf{y})$,

$$p\left(\boldsymbol{\nu}, \mathbf{x}, \mathbf{y} \mid \boldsymbol{s}_{\mathrm{dp}}\right) \propto p(\boldsymbol{\nu}) f(\mathbf{x}, \mathbf{y} \mid \boldsymbol{\nu}) p\left(\boldsymbol{s}_{\mathrm{dp}} \mid \mathbf{x}, \mathbf{y}\right) \tag{4.8}$$

Marginally, the $\boldsymbol{\nu}$ samples produced by equtaion (4.8) follow the posterior $p\left(\boldsymbol{\nu} \mid \boldsymbol{s}_{\mathrm{dp}}\right)$ in equation (4.7). The joint posterior distribution of $p\left(\boldsymbol{\nu}, \mathbf{x}, \mathbf{y} \mid \boldsymbol{s}_{\mathrm{dp}}\right)$ can be achieved through the following Gibbs sampling procedure: (a) sample the confidential dataset $(\mathbf{x}, \mathbf{y})$ given model parameters $\boldsymbol{\nu}$; (b) sample parameters $\boldsymbol{\nu}$ given latent confidential $(\mathbf{x}, \mathbf{y})$ and $\boldsymbol{s}_{\mathrm{dp}}$ [79]. The following subsections will illustrate the envelope regression structure $[\mathbf{x}, \mathbf{y}|\boldsymbol{\nu}]$, the sufficient statistics $[\boldsymbol{s}_{\mathrm{dp}} \mid \mathbf{x}, \mathbf{y}]$ and their corresponding Gibbs sampling steps.

4.3.1 Hierarchical Envelope Linear Regression

The hierarchical envelope regression structure between $(\mathbf{x}, \mathbf{y})$ given parameters $\boldsymbol{\nu} = (\mathbf{B}_1, \mathbf{B}_2, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\theta}, \sigma)$ can be expressed as follows,

$$\begin{aligned}
\boldsymbol{x}_i &= \mathbf{B}_1 \boldsymbol{\lambda}_i + \mathbf{B}_2 \boldsymbol{\xi}_i \\
[y_i|\boldsymbol{\lambda}_i] &= \boldsymbol{\lambda}_i^{\top} \boldsymbol{\theta} + \varepsilon_i \\
\boldsymbol{\lambda}_i &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}) \\
\boldsymbol{\xi}_i &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_0) \\
\varepsilon_i &\sim \mathcal{N}(0, \sigma^2)
\end{aligned} \tag{4.9}$$

In model (4.9), $\mathbf{B}_1 \in \mathbb{R}^{p \times r}$ and $\mathbf{B}_2 \in \mathbb{R}^{p \times (p-r)}$ are two orthogonal matrices with $\mathbf{B}_1^{\top} \mathbf{B}_2 = \mathbf{0}$. It can be seen that $\boldsymbol{x}_i$ is decomposed into two parts, which is the material $\mathbf{B}_1 \boldsymbol{\lambda}_i$ and the immaterial $\mathbf{B}_2 \boldsymbol{\xi}_i$. $\boldsymbol{\lambda}_i \in \mathbb{R}^r$ and $\boldsymbol{\xi}_i \in \mathbb{R}^{p-r}$ are the corresponding material and immaterial coordinates. The regression dependency between $\boldsymbol{x}_i$ and y_i is only through the material part $\boldsymbol{\lambda}_i$. After the dimension reduction, the regression coefficient $\boldsymbol{\theta}$ is r dimensional.

Remark 11 *As shown in model (4.9), y_i remains the same no matter how $\boldsymbol{x}_i$ varies in the space of span($\mathbf{B}_2$) since the distribution of $[y_i|\boldsymbol{x}_i]$ is the same as $[y_i|\mathbf{B}_1^{\top}\boldsymbol{x}_i]$. It*

is worth noting that for the given dimension r, $\mathbf{B}_1$ and $\mathbf{B}_2$ are not uniquely defined. To make it identifiable, in this chapter we define $\mathbf{B}_1$ and $\mathbf{B}_2$ as a function of an unconstrained matrix $\mathbf{A} \in \mathbb{R}^{(p-r)\times r}$. Let $\mathbf{C_A} = \left(\mathbf{I}_r, \mathbf{A}^\top\right)^\top$ and $\mathbf{D_A} = (-\mathbf{A}, \mathbf{I}_{p-r})^\top$, define

$$\begin{aligned} \mathbf{B}_1(\mathbf{A}) &\doteq \mathbf{C_A}\left(\mathbf{C_A}^T\mathbf{C_A}\right)^{-1/2} \\ \mathbf{B}_2(\mathbf{A}) &\doteq \mathbf{D_A}\left(\mathbf{D_A}^T\mathbf{D_A}\right)^{-1/2} \end{aligned} \tag{4.10}$$

In equation (4.10), the matrix $\mathbf{A}$ and $\text{span}(\mathbf{B}_1)$ is uniquely determined by each other hence $\mathbf{B}_1(\mathbf{A})$ is identifiable. The notation of $\mathbf{B}_1(\mathbf{A})$ as the function of $\mathbf{A}$ prevents the prior selection and the posterior updating of $\mathbf{B}_1$ on Stiefel manifolds. As there is no close form posterior for $\mathbf{A}$, the update of $\mathbf{A}$ is through the Metropolis–Hastings algorithm.

Remark 12 *Given model (4.9), the conditional distribution of $[y_i|\boldsymbol{x}_i]$ is uniquely determined with $[y_i|\boldsymbol{x}_i, \mathbf{B}_1, \boldsymbol{\theta}] \sim \mathcal{N}(\boldsymbol{x}_i^\top\mathbf{B}_1\boldsymbol{\theta}, \sigma^2)$ and the regression coefficient $\boldsymbol{\beta}$ of $[y_i|\boldsymbol{x}_i]$ is given by $\boldsymbol{\beta} = \mathbf{B}_1\boldsymbol{\theta}$. The density function $f(\mathbf{x}, \mathbf{y}|\boldsymbol{\nu})$ is as follows,*

$$\begin{aligned} f(\mathbf{x}, \mathbf{y}|\boldsymbol{\nu}) &= f(\mathbf{y}|\mathbf{x}, \mathbf{B}_1, \boldsymbol{\theta}, \sigma^2) f(\mathbf{x}|\mathbf{B}_1, \mathbf{B}_2, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0) \\ &= \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(y_i - \boldsymbol{x}_i^\top\mathbf{B}_1\boldsymbol{\theta})^2}{2\sigma^2}\right] \times \\ &\quad \prod_{i=1}^{n} \frac{1}{|\boldsymbol{\Sigma}_\mathbf{x}|^{1/2} 2\pi^{p/2}} \exp\left[-\frac{\boldsymbol{x}_i\boldsymbol{\Sigma}_\mathbf{x}^{-1}\boldsymbol{x}_i}{2}\right] \end{aligned} \tag{4.11}$$

where $\boldsymbol{\Sigma}_\mathbf{x} = \mathbf{B}_1\boldsymbol{\Omega}\mathbf{B}_1^\top + \mathbf{B}_2\boldsymbol{\Omega}_0\mathbf{B}_2^\top$.

4.3.2 Differentially Privatized Sufficient Statistic

In this chapter, the analytic Gaussian mechanism [85] is utilized to achieve (ϵ, δ)-DP for sufficient statistics. Observing inequality (4.3), a finite global sensitivity is required to calibrate the DP variance. In literature, the routine procedure entails bounding each predictor and response variable in a manner that is independent of

the data. For simplicity, we set a lower and upper bound (L, U) for all dimensions of $\boldsymbol{x}_i$ and y_i.

Remark 13 *In model (4.9),* $(\mathbf{x}^\top\mathbf{x}, \mathbf{x}^\top\mathbf{y}, \mathbf{y}^\top\mathbf{y})$ *is the sufficient statistics of*

$$\boldsymbol{\nu} = (\mathbf{B}_1, \mathbf{B}_2, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\theta}, \sigma)$$

Denote the L_2 sensitivity for $(\mathbf{x}^\top\mathbf{x}, \mathbf{x}^\top\mathbf{y}, \mathbf{y}^\top\mathbf{y})$ as Δ_2. By analytic Gaussian mechanism, we generate $\boldsymbol{s}_{\text{dp}}$ by adding Gaussian noises $\mathcal{N}(0, \sigma_{\text{dp}}^2)$ independently to each element of $(\mathbf{x}^\top\mathbf{x}, \mathbf{x}^\top\mathbf{y}, \mathbf{y}^\top\mathbf{y})$ where σ_{dp}^2 satisfies (4.3).

When calculating Δ_2, we can reason the worst case influence of an individual on each component of $\mathbf{s}_{\text{dp}} = (\mathbf{x}^\top\mathbf{x}, \mathbf{x}^\top\mathbf{y}, \mathbf{y}^\top\mathbf{y})$. The number of unique elements in $\mathbf{s}_{\text{dp}}$ is $[p(p+1)/2, p, 1]$ and $\Delta_2 = (U-L)^2 p(p+1)/2 + (U-L)^2(p+1)$.

4.3.3 Prior Specification

Denote the parameters in the model (4.9) as $\boldsymbol{\nu} = (\mathbf{B}_1, \mathbf{B}_2, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\theta}, \sigma)$. Note that $\mathbf{B}_1, \mathbf{B}_2$ are set to be functions of an unconstrained matrix $\mathbf{A}$ for identification. Thus, we impose diffuse priors for $(\mathbf{A}, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\theta}, \sigma)$ as follows,

- $\boldsymbol{\theta}$ is normally distributed: $\boldsymbol{\theta} \sim \mathcal{N}(0, 100\mathbf{I}_r)$.
- σ^2 follows Inverse-Gamma distribution: $\sigma^2 \sim \text{IG}(a_0, b_0)$ where $a_0 = b_0 = 0.1$.
- $\boldsymbol{\Omega}$ and $\boldsymbol{\Omega}_0$ follow Inverse-Wishart distribution: $\boldsymbol{\Omega} \sim \text{IW}(\mathbf{S}, \text{s})$, $\boldsymbol{\Omega}_0 \sim \text{IW}(\mathbf{S}_0, \text{s}_0)$, where the parameters are commonly selected as $\mathbf{S} = \mathbf{I}_{p-r}$, $\mathbf{S}_0 = \mathbf{I}_r$ and $\text{s} = p - r + 1$, $\text{s}_0 = r + 1$ [36].
- $\mathbf{A}$ follows matrix normal distribution: $\mathbf{A} \sim \mathcal{MN}(\mathbf{A}_0, \mathbf{K}, \mathbf{L})$. $\mathbf{A}_0 \in \mathbb{R}^{(p-r)\times r}$ is the mean matrix for $\mathbf{A}$. $\mathbf{K} \in \mathbb{S}^{(p-r)\times(p-r)}$, $\mathbf{L} \in \mathbb{S}^{r\times r}$ are positive symmetric covariance matrices. We set $\mathbf{A}_0 = \mathbf{0}$. It means $\mathbf{B}_1$ is assumed to be centered at $\mathbf{B}_1(\mathbf{0}) = (\mathbf{I}_r, \mathbf{0})^\top$, which corresponds to an identity projection to the first r dimensions of the predictors. $\mathbf{K}$ and $\mathbf{L}$ are chosen to be 10 times identity matrix, $10\mathbf{I}_{(p-r)}$ and $10\mathbf{I}_r$.

4.3.4 Privacy-Aware Gibbs Sampler

Let $\lambda \doteq (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_n)^\top$ and $\xi \doteq (\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n)^\top$ be the material and immaterial information for the confidential dataset. Let $(\mathbf{x}^{(t)}, \mathbf{y}^{(t)}, \lambda^{(t)}, \xi^{(t)}, \boldsymbol{\nu}^{(t)})$ denote the state of the Gibbs sampler at the t-th iterations. Based on equation (4.8), the Gibbs sampling procedure can be split into three detailed steps: (**1**) sample $(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)})$ given $\boldsymbol{\nu}^{(t)}$ and $\boldsymbol{s}_{\text{dp}}$; (**2**) calculate $(\lambda^{(t+1)}, \xi^{(t+1)})$ based on $\mathbf{x}^{(t+1)}$ and $\boldsymbol{\nu}^{(t)}$; (**3**) sample $\boldsymbol{\nu}^{(t+1)}$ given $(\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}, \boldsymbol{s}_{\text{dp}})$.

In step (1), the conditional distribution of $(\mathbf{x}, \mathbf{y})$ given $\boldsymbol{s}_{\text{dp}}$ and $\boldsymbol{\nu}^{(t)}$ has no closed form posterior and we employ the algorithm in [79] for the sampler which is Algorithm (1) in appendix. In step (2), $(\lambda^{(t+1)}, \xi^{(t+1)})$ based on $\mathbf{x}^{(t+1)}$ and $\boldsymbol{\nu}^{(t)}$ can be calculated as

$$\lambda^{(t+1)} = \mathbf{x}^{(t+1)} \mathbf{B}_1^{(t)\text{T}}$$

$$\xi^{(t+1)} = \mathbf{x}^{(t+1)} \mathbf{B}_2^{(t)\text{T}}$$

In step (3), all the model parameters are updated and the details are in the appendix.

4.4 Simulation Study

In this section, we conduct a comparative analysis of our data augmentation MCMC framework for Bayesian envelope linear regression and the existing data augmentation MCMC framework by [79]. The comparison is conducted through the utilization of a straightforward, yet informative simulation scenario within the context of (ϵ, δ)-differential privacy. We consider scenarios with different privacy budgets ϵ. δ is considered as $1/n$ which is a common choice for approximate differential privacy. n is the sample size and we choose it to be $n \in \{500, 1000, 5000\}$. The dimension for the predictor is set to be $p = 4$. The dimension for the material part r is set to be $r = 2$.

When comparison, our primary focus is directed towards evaluating the efficiency enhancement of the linear regression coefficients $\boldsymbol{\beta} = \mathbf{B}_1 \boldsymbol{\theta}$, taking into consideration the presence of the material and immaterial separation. When comparing our

approach with the existing data augmentation MCMC [79], it is expected that our framework would lead to a reduction in the uncertainty associated with the estimated coefficients $\hat{\boldsymbol{\beta}}$.

4.4.1 Generating the Differentially Private Sufficient Statistics

Despite the absence of the confidential dataset $(\mathbf{x}, \mathbf{y})$ at our disposal, it is worthwhile to demonstrate the process of how $(\mathbf{x}, \mathbf{y})$ is generated under our framework.

For all simulation scenarios (for each selected n and ϵ), the confidential dataset $(\mathbf{x}, \mathbf{y})$ is generated as the following. Set $\delta = 1/n$, $r = 2$, $\boldsymbol{\Omega} = \mathbf{I}_r$ and $\boldsymbol{\Omega}_0 = 0.1\mathbf{I}_{p-r}$. For each given $n \in \{500, 1000, 5000\}$ and $\epsilon \in \{0.5, 1, 3, 5\}$, generate $(\boldsymbol{x}_i, y_i)_{i=1}^n$ as follow.

1. Given fixed matrix $\mathbf{A} \in \mathbb{R}^{(p-r)\times r}$, compute $\mathbf{B}_1$ and $\mathbf{B}_2$ through (4.10).
2. For each $i \in [n]$:
   (a) generate $\boldsymbol{\lambda}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ and $\boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_0)$,
   (b) compute $\boldsymbol{x}_i = \mathbf{B}_1\boldsymbol{\lambda}_i + \mathbf{B}_2\boldsymbol{\xi}_i$,
   (c) generate $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ given fixed σ^2 and
   (d) compute $y_i = \boldsymbol{\lambda}_i^\top \boldsymbol{\theta} + \varepsilon_i$ given fixed $\boldsymbol{\theta} \in \mathbb{R}^r$.

For simplicity, we set a lower and upper bound (L, U) for all dimensions of $\boldsymbol{x}_i$ and y_i. For a real value z, and $L \leq U$, we define the clamp function $[z]_L^U := \min\{\max\{z, L\}, U\}$. If z is a vector of length d, we use the same notation to apply an entry-wise clamp: $[z]_L^U := \left([z_1]_L^U, [z_2]_L^U, \ldots, [z_d]_L^U\right)^\top$.

After clamping $(\mathbf{x}, \mathbf{y})$ into range (L, U) for each dimension, we denote the clamped dataset as $(\mathbf{x}_C, \mathbf{y}_C)$. Finally, generate $\boldsymbol{s}_{\text{dp}}$ through the analytic Gaussian mechanism [85] by adding Gaussian noises independently to each element of $(\mathbf{x}_C^\top\mathbf{x}_C, \mathbf{x}_C^\top\mathbf{y}_C, \mathbf{y}_C^\top\mathbf{y}_C)$.

For each of the simulation scenarios, we simulate 50 datasets $(\mathbf{x}_C, \mathbf{y}_C)$ and $\boldsymbol{s}_{\text{dp}}$. For each generated $\boldsymbol{s}_{\text{dp}}$, we obtain the posterior samples of all model parameters using the Gibbs sampling algorithm described earlier, retaining 5000 iterations after a burn-in period of 5000 iterations.

For all given parameters in the simulation scenario, $\boldsymbol{\theta}$ and $\mathbf{A}$ are randomly generated for each dataset and $\sigma^2 = 1$. The lower and upper bound is set to be $[-10, 10]$. A more detailed configuration for the simulation is provided in the appendix.

Remark 14 *An essential aspect of MCMC is ergodicity [88], which guarantees the convergence of the MCMC chain to the posterior distribution in terms of total variation. It can be verified that within our model: (a) the chosen priors are proper and* $p(\boldsymbol{\nu}) > 0$ *for all* $\boldsymbol{\nu}$*; (b) the condition* $f(\mathbf{x}, \mathbf{y}|\boldsymbol{\nu}) > 0$ *and* $p(\boldsymbol{s}_{\text{dp}}|\mathbf{x}, \mathbf{y}) > 0$ *is consistently satisfied for all* $(\mathbf{x}, \mathbf{y})$*. Under the two conditions above, it can be proved that the Gibbs sampler for latent confidential dataset* $(\mathbf{x}, \mathbf{y})$ *and parameters* $\boldsymbol{\nu}$ *is ergodic and the limiting distribution is unique [79].*

4.4.2 Evaluation and Comparison

As mentioned previously, the primary focus is directed towards evaluating the efficiency enhancement of the linear regression coefficients $\hat{\boldsymbol{\beta}}$. During Gibbs sampling, the t-th posterior draw of $\mathbf{B}_1$ and $\boldsymbol{\theta}$ is denoted as $\hat{\mathbf{B}}_1^{(t)}$ and $\hat{\boldsymbol{\theta}}^{(t)}$. $\hat{\boldsymbol{\beta}}^{(t)}$ is calculated as $\hat{\boldsymbol{\beta}}^{(t)} = \hat{\mathbf{B}}_1^{(t)} \hat{\boldsymbol{\theta}}^{(t)}$.

To assess the performance of the efficiency gain on the estimation of the coefficient matrix $\hat{\boldsymbol{\beta}}$, we obtain the overall average 95% credible interval width and the overall mean squared error (MSE) based on the 50 simulated confidential dataset $(\mathbf{x}, \mathbf{y})$ and $\boldsymbol{s}_{\text{dp}}$ in each setup. For each $\hat{\beta}_j$ in $\hat{\boldsymbol{\beta}} = \{\hat{\beta}_1, \ldots, \hat{\beta}_p\}$, $j \in \{1, \ldots, p\}$ define MSE as follows,

$$\text{MSE} = \frac{1}{50} \sum_{j=1}^{50} \left\{ \frac{1}{p} \left\| \hat{\boldsymbol{\beta}}^j - \boldsymbol{\beta}^j \right\|_2^2 \right\}$$

where $\|\cdot\|_2$ represents the L_2 norm. $\boldsymbol{\beta}^j$ is the true coefficient from the jth simulation and $\hat{\boldsymbol{\beta}}^j$ is its posterior mean estimate. The overall average 95% credible interval width is then defined as

$$\mathrm{W} = \frac{1}{50}\sum_{j=1}^{50}\left\{\frac{1}{p}\left\|\mathrm{W}_{\hat{\boldsymbol{\beta}}^j}\right\|_2^2\right\}$$

where $\mathrm{W}_{\hat{\boldsymbol{\beta}}^j} \doteq (\mathrm{w}_{\hat{\beta}_1^j}, \ldots, \mathrm{w}_{\hat{\beta}_p^j})$ is the 95% credible interval width vector for all of the p predictors from the jth simulation.

Table (4.1) displays the average MSE and interval width, derived from our framework utilizing the envelope technique, alongside the data augmentation framework outlined in [79] without the envelope approach. It is evident that with limited sample size and a small ϵ value, both methods exhibit pronouncedly elevated MSE and interval width. As the sample size and ϵ increase, there is a reduction in the MSE and interval width for both methods. Upon comparison between the two methods, the observed trend aligns with our expectations—namely, an enhancement in the efficiency of estimating $\hat{\boldsymbol{\beta}}$ across all different combinations of sample size n and privacy budget ϵ.

Figure (4.1) depicts the boxplot illustrating the distribution of the 50 mean squared errors for $\hat{\boldsymbol{\beta}}$, while Figure (4.2) presents the boxplot showcasing the distribution of the 50 average 95% credible intervals. These figures provide a more detailed presentation of the information presented in Table (4.1). Notably, a substantial improvement becomes evident as the sample size increases from $n = 500$ to $n = 1000$. It can be seen that the degree of improvement is less pronounced with further increases in sample size. Similarly, as the privacy budget ϵ ranges from 0.5 to 3, there is a considerable enhancement, but further increases in ϵ do not yield clear improvements.

		MSE		Interval Width	
n	ϵ	Bayes Env	No Env	Bayes Env	No Env
$n = 500$	$\epsilon = 0.5$	6.81	11.20	3.10	6.36
	$\epsilon = 1.0$	3.26	6.26	2.48	5.22
	$\epsilon = 3.0$	1.05	3.01	1.45	2.98
	$\epsilon = 5.0$	0.71	2.98	1.14	2.27
$n = 1000$	$\epsilon = 0.5$	3.30	7.56	2.00	4.04
	$\epsilon = 1.0$	1.93	5.19	1.43	3.01
	$\epsilon = 3.0$	0.67	3.42	0.82	1.65
	$\epsilon = 5.0$	0.65	3.34	0.62	1.28
$n = 5000$	$\epsilon = 0.5$	2.67	3.97	0.65	0.94
	$\epsilon = 1.0$	1.60	2.14	0.39	0.66
	$\epsilon = 3.0$	0.81	1.07	0.20	0.36
	$\epsilon = 5.0$	0.91	0.95	0.16	0.27

Table 4.1: The average MSE and interval width for $\hat{\boldsymbol{\beta}}$, derived from our framework utilizing the envelope technique, alongside the framework without the envelope approach.

4.5 Discussion and Conclusion

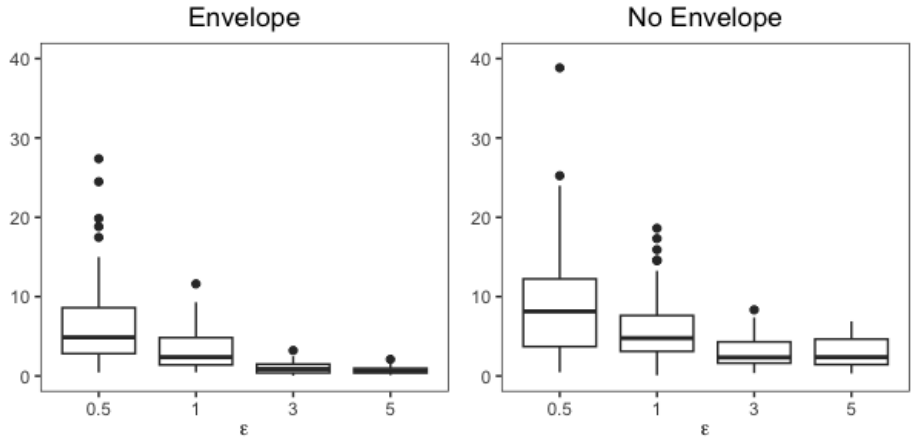
Our work presents a groundbreaking approach to Differentially Private Bayesian Envelope Regression, offering a nuanced solution to the challenges of maintaining privacy while harnessing the utility of data and improving the efficiency of the estimation in regression analysis. By innovatively applying the envelope method within a Bayesian framework, we address the critical need for preserving privacy without sacrificing the richness of data insights, particularly when dealing with datasets that

has low-dimensional structures. This methodological advancement not only enhances efficiency of the parameter estimation in comparison to existing differentially private mechanisms but also facilitates nuanced statistical inference, such as the construction of credible intervals, which is often overlooked in the differential privacy landscape. The simulation studies and theoretical discussions further underscore the superiority of our approach in terms of efficiency improvement over traditional methods.

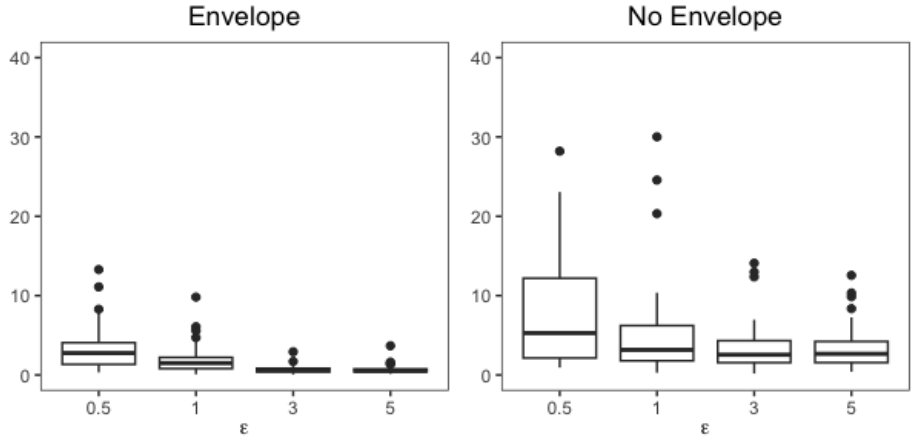
In conclusion, this paper has introduced an innovative Bayesian Envelope Regression model that significantly enhances the efficiency and interpretability of regression analysis in privacy-sensitive contexts. Our approach leverages the strengths of the envelope methodology to focus analysis on the most impactful predictor components, thereby improving parameter estimation efficiency and reducing the risk of overfitting. Through rigorous simulation studies, we have demonstrated the efficiency improvements achieved by our model, highlighting its potential to serve as a valuable tool for researchers and practitioners dealing with sensitive datasets. This pioneering integration of differential privacy with envelope regression methodology not only marks a significant methodological innovation but also paves the way for new research directions in privacy-preserving data analysis. It highlights the potential for sophisticated statistical techniques to enhance data utility while rigorously protecting individual privacy, thereby contributing to the advancement of responsible and ethical data science practices.

Despite the promising results, our approach faces limitations inherent in the assumptions of envelope methods and the specific modeling of privacy mechanisms. Future research could explore extending our framework to accommodate more complex data structures and a wider range of regression models, enhancing its applicability and robustness. Additionally, assessing the performance of our framework with high-dimensional datasets, where predictors substantially outnumber observations, stands as a significant area of interest. Furthermore, examining the framework's robustness to model misspecification and atypical data distributions constitutes another vital

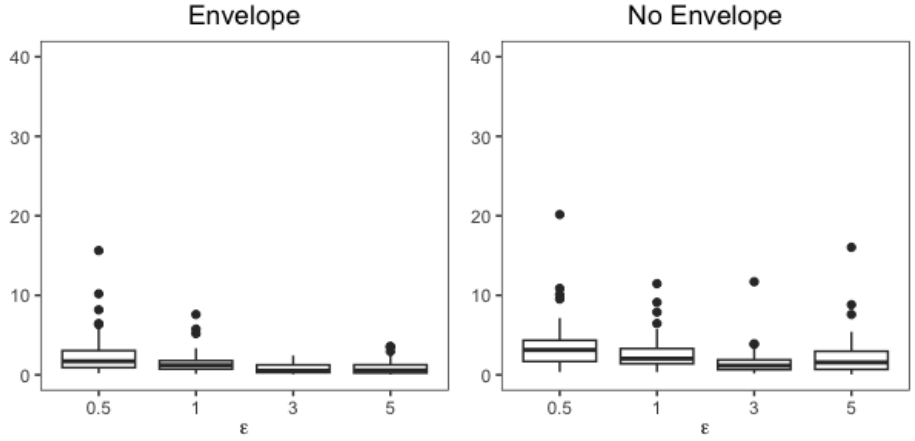
extension. These topics are earmarked for subsequent studies.



(a) $n = 500$

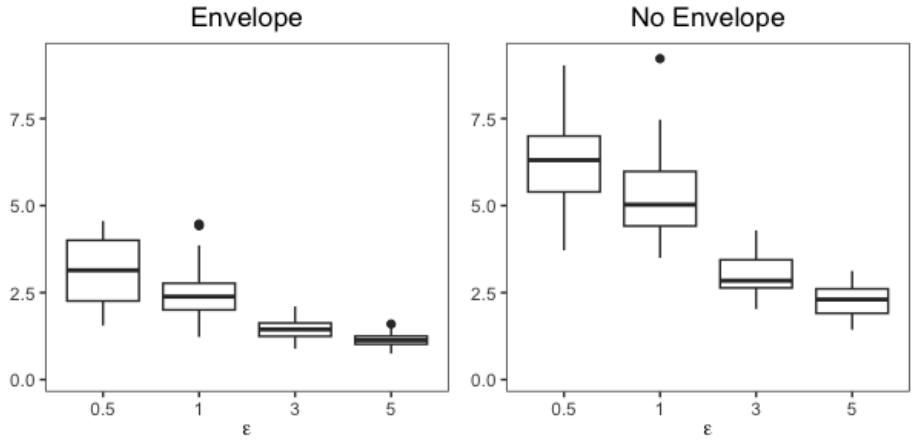


(b) $n = 1000$

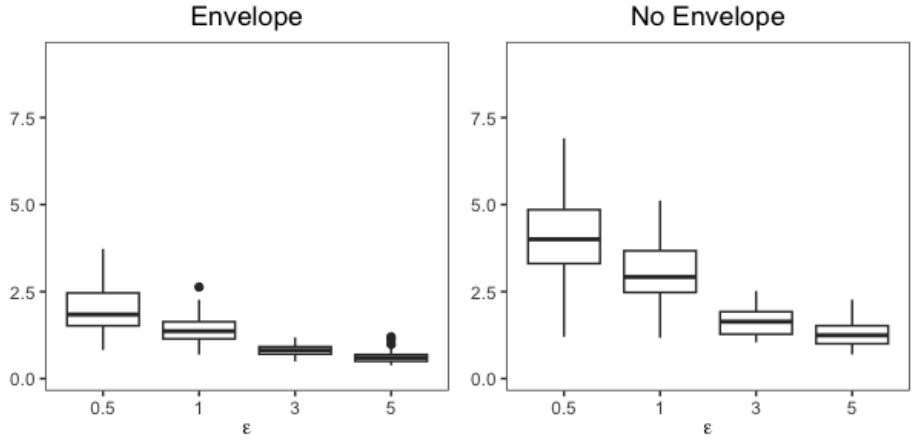


(c) $n = 5000$

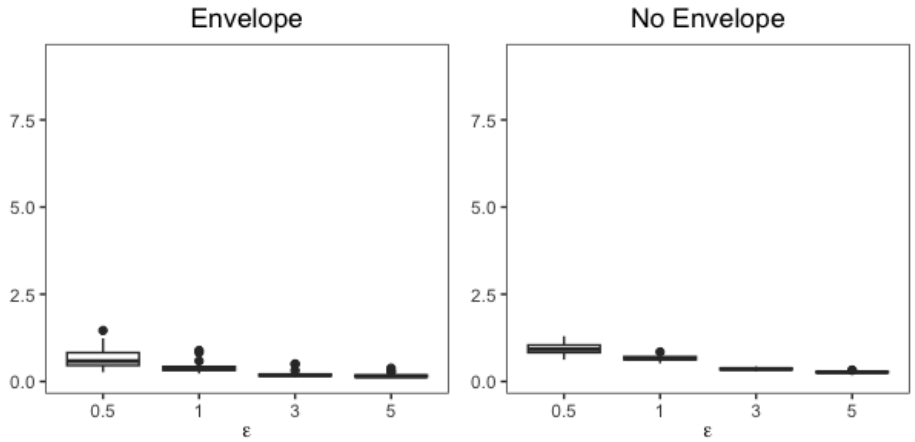
Figure 4.1: The boxplot of average MSE for the coefficient estimation based on our framework and the framework without the envelope approach.



(a) $n = 500$



(b) $n = 1000$



(c) $n = 5000$

Figure 4.2: The boxplot of average 95% credible interval width for the coefficient estimation based on our framework and the framework without the envelope approach.

Bibliography

[1] J. A. Smith, B. C. Jones, and N. Roberts, "Challenges in clinical trials for depression: A comprehensive review," *Journal of Mental Health Research*, vol. 12, no. 3, pp. 234–245, 2018.

[2] P. Jones, "Addressing missing data in longitudinal clinical trials: A practical guide," *Clinical Epidemiology*, vol. 9, pp. 123–130, 2017.

[3] A. J. Ferrari *et al.*, "Burden of depressive disorders by country, sex, age, and year: Findings from the global burden of disease study 2010," *PLoS Medicine*, vol. 10, no. 11, e1001547, 2013.

[4] J. Spijker, R De Graaf, R. Bijl, A. Beekman, J Ormel, and W. Nolen, "Functional disability and depression in the general population. results from the netherlands mental health survey and incidence study (nemesis)," *Acta Psychiatrica Scandinavica*, vol. 110, no. 3, pp. 208–214, 2004.

[5] C. Roehrig, "Mental disorders top the list of the most costly conditions in the united states: $201 billion," *Health Affairs*, vol. 35, no. 6, pp. 1130–1135, 2016.

[6] A. T. Drysdale *et al.*, "Resting-state connectivity biomarkers define neurophysiological subtypes of depression," *Nature Medicine*, vol. 23, no. 1, pp. 28–38, 2017.

[7] W. Wu *et al.*, "An electroencephalographic signature predicts antidepressant response in major depression," *Nature Biotechnology*, vol. 38, no. 4, pp. 439–447, 2020.

[8] A. M. Buch and C. Liston, "Dissecting diagnostic heterogeneity in depression by integrating neuroimaging and genetics," *Neuropsychopharmacology*, vol. 46, no. 1, pp. 156–175, 2021.

[9] M. H. Trivedi *et al.*, "Evaluation of outcomes with citalopram for depression using measurement-based care in star*d: Implications for clinical practice," *American Journal of Psychiatry*, vol. 163, no. 1, pp. 28–40, 2006.

[10] M. Fava, "Diagnosis and definition of treatment-resistant depression," *Biological Psychiatry*, vol. 53, no. 8, pp. 649–659, 2003.

[11] K. S. Al-Harbi, "Treatment-resistant depression: Therapeutic trends, challenges, and future directions," *Patient Preference and Adherence*, vol. 6, p. 369, 2012.

[12] D. A. Pizzagalli *et al.*, "Pretreatment rostral anterior cingulate cortex theta activity in relation to symptom improvement in depression: A randomized clinical trial," *JAMA psychiatry*, vol. 75, no. 6, pp. 547–554, 2018.

[13] A. S. Widge *et al.*, "Electroencephalographic biomarkers for treatment response prediction in major depressive illness: A meta-analysis," *American Journal of Psychiatry*, vol. 176, no. 1, pp. 44–56, 2019.

[14] Y.-C. Tsai *et al.*, "Critical role of rhythms in prefrontal transcranial magnetic stimulation for depression: A randomized sham-controlled study," *Human brain mapping*, vol. 43, no. 5, pp. 1535–1547, 2022.

[15] J. Kayser and C. E. Tenke, "Principal components analysis of laplacian waveforms as a generic method for identifying ERP generator patterns: Ii. adequacy of low-density estimates," *Clinical Neurophysiology*, vol. 117, no. 2, pp. 369–380, 2006.

[16] G Ulrich, E Renfordt, and K Frick, "The topographical distribution of alpha-activity in the resting eeg of endogenous-depressive in-patients with and without clinical response to pharmacotherapy," *Pharmacopsychiatry*, vol. 19, no. 04, pp. 272–273, 1986.

[17] G. E. Bruder, J. P. Sedoruk, J. W. Stewart, P. J. McGrath, F. M. Quitkin, and C. E. Tenke, "Electroencephalographic alpha measures predict therapeutic response to a selective serotonin reuptake inhibitor antidepressant: Pre-and post-treatment findings," *Biological Psychiatry*, vol. 63, no. 12, pp. 1171–1177, 2008.

[18] A. R. Clarke, R. J. Barry, R. McCarthy, and M. Selikowitz, "EEG-defined subtypes of children with attention-deficit/hyperactivity disorder," *Clinical Neurophysiology*, vol. 112, no. 11, pp. 2098–2105, 2001.

[19] A. R. Clarke *et al.*, "Behavioural differences between EEG-defined subgroups of children with attention-deficit/hyperactivity disorder," *Clinical Neurophysiology*, vol. 122, no. 7, pp. 1333–1341, 2011.

[20] M. H. Trivedi *et al.*, "Establishing moderators and biosignatures of antidepressant response in clinical care (embarc): Rationale and design," *Journal of psychiatric research*, vol. 78, pp. 11–23, 2016.

[21] E. Petkova *et al.*, "Statistical analysis plan for stage 1 EMBARC (Establishing Moderators and Biosignatures of Antidepressant Response for Clinical Care) study," *Contemporary clinical trials communications*, vol. 6, pp. 22–30, 2017.

[22] C. E. Tenke *et al.*, "Current source density measures of electroencephalographic alpha predict antidepressant treatment response," *Biological Psychiatry*, vol. 70, no. 4, pp. 388–394, 2011.

[23] C. Viroli, "Finite mixtures of matrix normal distributions for classifying three-way data," *Statistics and Computing*, vol. 21, no. 4, pp. 511–522, 2011.

[24] M. P. Gallaugher and P. D. McNicholas, "Finite mixtures of skewed matrix variate distributions," *Pattern Recognition*, vol. 80, pp. 83–93, 2018.

[25] X. Gao *et al.*, "Regularized matrix data clustering and its application to image analysis," *Biometrics*, vol. 77, no. 3, pp. 890–902, 2021.

[26] Q. Mai, X. Zhang, Y. Pan, and K. Deng, "A doubly enhanced em algorithm for model-based tensor clustering," *Journal of the American Statistical Association*, pp. 1–15, 2021.

[27] R. J. Carroll, D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu, *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC, 2006.

[28] G. Y. Yi, *Statistical analysis with measurement error or misclassification: strategy, method and application*. Springer, 2017.

[29] H. Hung, P. Wu, I. Tu, and S. Huang, "On multilinear principal component analysis of order-two tensors," *Biometrika*, vol. 99, no. 3, pp. 569–583, 2012.

[30] B. Jiang, E. Petkova, T. Tarpey, and R. T. Ogden, "A bayesian approach to joint modeling of matrix-valued imaging data and treatment outcome with applications to depression studies," *Biometrics*, vol. 76, no. 1, pp. 87–97, 2020.

[31] M. E. Tipping and C. M. Bishop, "Mixtures of probabilistic principal component analyzers," *Neural computation*, vol. 11, no. 2, pp. 443–482, 1999.

[32] C. Clogg, *Latent class models in: Handbook of statistical modeling for the social and behavioral sciences. arminger g, clogg cc, sobel me, editor*, 1995.

[33] I. C. Gormley and T. B. Murphy, "Mixture of experts modelling with social science applications," in *Mengersen, K., Robert, C, Titterington, M.(eds.). Mixture: estimation and applications*, Wiley, 2011.

[34] I. C. Gormley and S. Frühwirth-Schnatter, "Mixture of experts models," *Handbook of mixture analysis*, pp. 271–307, 2019.

[35] P. D. Hoff, "Model averaging and dimension selection for the singular value decomposition," *Journal of the American Statistical Association*, vol. 102, no. 478, pp. 674–685, 2007.

[36] S. Frühwirth-Schnatter and S. Frèuhwirth-Schnatter, *Finite mixture and Markov switching models*. Springer, 2006, vol. 425.

[37] S. Watanabe, "A widely applicable bayesian information criterion," *Journal of Machine Learning Research*, vol. 14, no. Mar, pp. 867–897, 2013.

[38] S. Watanabe, "Waic and wbic for mixture models," *Behaviormetrika*, vol. 48, no. 1, pp. 5–21, 2021.

[39] H. Bozdogan, "Model selection and akaike's information criterion (aic): The general theory and its analytical extensions," *Psychometrika*, vol. 52, no. 3, pp. 345–370, 1987.

[40] G. Schwarz *et al.*, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[41] A. Gelman, J. Hwang, and A. Vehtari, "Understanding predictive information criteria for bayesian models," *Statistics and computing*, vol. 24, no. 6, pp. 997–1016, 2014.

[42] J. A. Israel, "Remission in depression: Definition and initial treatment approaches," *Journal of Psychopharmacology*, vol. 20, no. 3_suppl, pp. 5–10, 2006.

[43] A. Pinto-Meza, J. Usall, A. Serrano-Blanco, D. Suárez, and J. M. Haro, "Gender differences in response to antidepressant treatment prescribed in primary care. does menopause make a difference?" *Journal of affective disorders*, vol. 93, no. 1-3, pp. 53–60, 2006.

[44] D. M. Sloan and S. G. Kornstein, "Gender differences in depression and response to antidepressant treatment," *Psychiatric Clinics*, vol. 26, no. 3, pp. 581–594, 2003.

[45] K. Perlman *et al.*, "A systematic meta-review of predictors of antidepressant treatment outcome in major depressive disorder," *Journal of affective disorders*, vol. 243, pp. 503–515, 2019.

[46] C. E. Tenke and J. Kayser, "Reference-free quantification of eeg spectra: Combining current source density (csd) and frequency principal components analysis (fpca)," *Clinical Neurophysiology*, vol. 116, no. 12, pp. 2826–2846, 2005.

[47] V. J. Knott, J. I. Telner, Y. D. Lapierre, M. Browne, and E. R. Horn, "Quantitative eeg in the prediction of antidepressant response to imipramine," *Journal of Affective Disorders*, vol. 39, no. 3, pp. 175–184, 1996.

[48] C. Mulert *et al.*, "Prediction of treatment response in major depression: Integration of concepts," *Journal of Affective Disorders*, vol. 98, no. 3, pp. 215–225, 2007.

[49] J. Baek, G. J. McLachlan, and L. K. Flack, "Mixtures of factor analyzers with common factor loadings: Applications to the clustering and visualization of high-dimensional data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 7, pp. 1298–1309, 2009.

[50] K. Deng and X. Zhang, "Tensor envelope mixture model for simultaneous clustering and multiway dimension reduction," *Biometrics*, vol. 78, no. 3, pp. 1067–1079, 2022.

[51] W. Wang, X. Zhang, and Q. Mai, "Model-based clustering with envelopes," *Electronic Journal of Statistics*, vol. 14, no. 1, pp. 82–109, 2020.

[52] X. Zhang and L. Li, "Tensor envelope partial least-squares regression," *Technometrics*, vol. 59, no. 4, pp. 426–436, 2017.

[53] H. Zhou, L. Li, and H. Zhu, "Tensor regression with applications in neuroimaging data analysis," *Journal of the American Statistical Association*, vol. 108, no. 502, pp. 540–552, 2013.

[54] W. W. Sun and L. Li, "Dynamic tensor clustering," *Journal of the American Statistical Association*, vol. 114, no. 528, pp. 1894–1907, 2019.

[55] C. Fraley and A. E. Raftery, "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, vol. 97, no. 458, pp. 611–631, 2002.

[56] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*, 2nd ed. John Wiley & Sons, 2002.

[57] G. Molenberghs and M. G. Kenward, *Missing Data in Clinical Studies*. John Wiley & Sons, 2007.

[58] R. D. Cook, *Regression graphics: Ideas for studying regressions through graphics*. John Wiley & Sons, 2009, vol. 482.

[59] R. D. Cook, B. Li, and F. Chiaromonte, "Envelope models for parsimonious and efficient multivariate linear regression," *Statistica Sinica*, pp. 927–960, 2010.

[60] B. Muthén, H. Brown, A. Leuchter, and A. Hunter, "General approaches to analysis of course: Applying growth mixture modeling to randomized trials of depression medication," *Causality and psychopathology: Finding the determinants of disorders and their cures. Washington, DC: American Psychiatric Publishing*, pp. 159–78, 2008.

[61] R. Uher *et al.*, "Trajectories of change in depression severity during treatment with antidepressants," *Psychological medicine*, vol. 40, no. 8, p. 1367, 2010.

[62] B. Muthén and H. C. Brown, "Estimating drug effects in the presence of placebo response: Causal inference using growth mixture modeling," *Statistics in medicine*, vol. 28, no. 27, pp. 3363–3385, 2009.

[63] M. N. Kuchibhatla and G. G. Fillenbaum, "Trajectory classes of depression in a randomized depression trial of heart failure patients: A reanalysis of the sadhart-chf trial," *The American journal of geriatric pharmacotherapy*, vol. 9, no. 6, pp. 483–494, 2011.

[64] E. C. Garman, M. Schneider, and C. Lund, "Perinatal depressive symptoms among low-income south african women at risk of depression: Trajectories and predictors," *BMC pregnancy and childbirth*, vol. 19, no. 1, pp. 1–11, 2019.

[65] A. Punzo and P. D. McNicholas, "Parsimonious mixtures of multivariate contaminated normal distributions," *Biometrical Journal*, vol. 58, no. 6, pp. 1506–1537, 2016.

[66] G. J. McLachlan and D. Peel, "Robust cluster analysis via mixtures of multivariate t-distributions," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, Springer, 1998, pp. 658–666.

[67] D. Hedeker and R. D. Gibbons, "Application of random-effects pattern-mixture models for missing data in longitudinal studies," *Psychological Methods*, vol. 2, no. 1, p. 64, 1997.

[68] R. J. Little and D. B. Rubin, *Statistical analysis with missing data*. John Wiley & Sons, 2019, vol. 793.

[69] A. C. Grobler, G. Matthews, and G. Molenberghs, "The impact of missing data on clinical trials: A re-analysis of a placebo controlled trial of hypericum perforatum (st johns wort) and sertraline in major depressive disorder," *Psychopharmacology*, vol. 231, no. 9, pp. 1987–1999, 2014.

[70] A. J. Mason, "Bayesian methods for modelling non-random missing data mechanisms in longitudinal studies," 2010.

[71] B. Muthén and K. Shedden, "Finite mixture modeling with mixture outcomes using the em algorithm," *Biometrics*, vol. 55, no. 2, pp. 463–469, 1999.

[72] L. Scrucca, M. Fop, T. B. Murphy, and A. E. Raftery, "Mclust 5: Clustering, classification and density estimation using gaussian finite mixture models," *The R journal*, vol. 8, no. 1, p. 289, 2016.

[73] C. Proust-Lima, V. Philipps, and B. Liquet, "Estimation of extended mixed models using latent classes and latent processes: The r package lcmm," *Journal of Statistical Software*, vol. 78, no. 2, pp. 1–56, 2017.

[74] C. Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, Springer, 2006, pp. 265–284.

[75] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor, "Our data, ourselves: Privacy via distributed noise generation," in *Advances in Cryptology-EUROCRYPT 2006: 24th Annual International Conference on the Theory and Applications of Cryptographic Techniques, St. Petersburg, Russia, May 28-June 1, 2006. Proceedings 25*, Springer, 2006, pp. 486–503.

[76] Z. Zhang, B. Rubinstein, and C. Dimitrakakis, "On the differential privacy of bayesian inference," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 30, 2016.

[77] F. McSherry and I. Mironov, "Differentially private recommender systems: Building privacy into the netflix prize contenders," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 627–636.

[78] A. Smith, "Efficient, differentially private point estimators," *arXiv preprint arXiv:0809.4794*, 2008.

[79] N. Ju, J. Awan, R. Gong, and V. Rao, "Data augmentation mcmc for bayesian inference from privatized data," *Advances in Neural Information Processing Systems*, vol. 35, pp. 12 732–12 743, 2022.

[80] G. Bernstein and D. R. Sheldon, "Differentially private bayesian linear regression," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[81] A. Dandekar, D. Basu, and S. Bressan, "Differential privacy for regularised linear regression," in *International Conference on Database and Expert Systems Applications*, Springer, 2018, pp. 483–491.

[82] K. Chaudhuri, A. Sarwate, and K. Sinha, "Near-optimal differentially private principal components," *Advances in neural information processing systems*, vol. 25, 2012.

[83] D. Wang and J. Xu, "On sparse linear regression in the local differential privacy model," in *International Conference on Machine Learning*, PMLR, 2019, pp. 6628–6637.

[84] K. Talwar, A. Guha Thakurta, and L. Zhang, "Nearly optimal private lasso," *Advances in Neural Information Processing Systems*, vol. 28, 2015.

[85] B. Balle and Y.-X. Wang, "Improving the gaussian mechanism for differential privacy: Analytical calibration and optimal denoising," in *International Conference on Machine Learning*, PMLR, 2018, pp. 394–403.

[86] R. D. Cook, I. Helland, and Z. Su, "Envelopes and partial least squares regression," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 75, no. 5, pp. 851–877, 2013.

[87] R. D. Cook and X. Zhang, "Foundations for envelope models and methods," *Journal of the American Statistical Association*, pp. 599–611, 2015.

[88] L. Tierney, "Markov chains for exploring posterior distributions," *the Annals of Statistics*, pp. 1701–1728, 1994.

Appendix A: Computation Details in Chapter 2

First, the conditional posterior for semi-orthogonal matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are derived. In the following equations, we denote $\boldsymbol{u}_{ig} \doteq [\boldsymbol{u}_i | c_{ig} = 1]$.

Let $\boldsymbol{A}_{[,k]}$ denote the k^{th} column of $\boldsymbol{A}$ and $\boldsymbol{A}_{[,-k]}$ denote the matrix $\boldsymbol{A}$ with its k^{th} column removed. Based on Proposition 1 from [35], when the distribution of $\boldsymbol{A}$ is the uniform distribution on Stiefel manifold denoted as $\mathcal{V}_{p_0,p}$, then the conditional distribution of $\boldsymbol{A}_{[,k]}$ given $\boldsymbol{A}_{[,-k]}$ is equal to the distribution of $\boldsymbol{N}_{A\{-k\}}\boldsymbol{a}_k$ where $\boldsymbol{N}_{A\{-k\}}$ is a basis for the null space of columns of $\boldsymbol{A}_{[,-k]}$ and $\boldsymbol{a}_k$ is uniformly distributed on the $(p - p_0 + 1)$-dimensional sphere, i.e., conditional on $\boldsymbol{A}_{[,-k]}, \boldsymbol{A}_{[,k]} \overset{d}{=} \boldsymbol{N}_{A\{-k\}}\boldsymbol{a}_k$. When $\boldsymbol{a}_k$ is uniformly distributed on spherical space with parameter $\boldsymbol{\mu}$, we have $\boldsymbol{a}_k \sim \text{vMF}(\boldsymbol{\mu})$, and $log(\boldsymbol{a}_k) \propto \boldsymbol{a}_k^T\boldsymbol{\mu}$. In our setting, we have the following derivation:

$$
\begin{aligned}
&\sum_{i=1}^{n}\sum_{g=1}^{G} I(c_{ig}=1)\phi \left\|\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{u}_i\boldsymbol{B}^\top\right\|_F^2 \\
&\doteq \sum_{\substack{i,g,\\ c_{ig}=1}} \phi \left\|\boldsymbol{x}_i - \boldsymbol{A}\boldsymbol{u}_{ig}\boldsymbol{B}^\top\right\|_F^2 \\
&= \sum_{\substack{i,g,\\ c_{ig}=1}} \phi \left\|\boldsymbol{x}_i - \sum_{t=1}^{p_0}\sum_{s=1}^{q_0} \boldsymbol{A}_{[,t]}\boldsymbol{u}_{ig[t,s]}\boldsymbol{B}_{[,s]}^\top\right\|_F^2 \\
&= \sum_{\substack{i,g,\\ c_{ig}=1}} \phi \left\|\boldsymbol{x}_i - \boldsymbol{A}_{[,k]}\sum_{s=1}^{q_0}\boldsymbol{u}_{ig[k,s]}\boldsymbol{B}_{[,s]}^\top - \sum_{k'\neq k}\sum_{s=1}^{q_0}\boldsymbol{A}_{[,k']}\boldsymbol{u}_{ig[k',s]}\boldsymbol{B}_{[,s]}^\top\right\|_F^2 \\
&\doteq \sum_{\substack{i,g,\\ c_{ig}=1}} \phi \left\|\boldsymbol{x}_i^{-k} - \boldsymbol{A}_{[,k]}\sum_{s=1}^{q_0}\boldsymbol{u}_{ig[k,s]}\boldsymbol{B}_{[,k]}^\top\right\|_F^2 \\
&= \sum_{\substack{i,g,\\ c_{ig}=1}} \phi \left\|\boldsymbol{x}_i^{-k}\right\|_F^2 - 2\boldsymbol{A}_{[,k]}^\top \sum_{\substack{i,g,\\ c_{ig}=1}}\sum_{s=1}^{q_0}\phi\boldsymbol{u}_{ig[k,s]}\boldsymbol{x}_i^{-k}\boldsymbol{B}_{[,k]} + \sum_{\substack{i,g,\\ c_{ig}=1}}\sum_{s=1}^{q_0}\phi\boldsymbol{u}_{ig[k,s]}^2
\end{aligned}
\tag{A.1}
$$

Since conditional on $\boldsymbol{A}_{[,-k]}$, $\boldsymbol{A}_{[,k]} \overset{d}{=} \boldsymbol{N}_{A\{-k\}}\boldsymbol{a}_k$, thus

$$
\boldsymbol{A}_{[,k]}^\top \sum_{\substack{i,g,\\ c_{ig}=1}}\sum_{s=1}^{q_0}\phi\boldsymbol{u}_{ig[k,s]}\boldsymbol{x}_i^{-k}\boldsymbol{B}_{[,k]} =
$$

$$
\boldsymbol{a}_k^\top(\boldsymbol{N}_{A\{-k\}}^\top \sum_{\substack{i,g,\\ c_{ig}=1}}\sum_{s=1}^{q_0}\phi\boldsymbol{u}_{ig[k,s]}\boldsymbol{x}_i^{-k}\boldsymbol{B}_{[,k]})
$$

which means $\boldsymbol{a}_k \sim \text{vMF}(\boldsymbol{\mu})$ with

$$
\boldsymbol{\mu} = \boldsymbol{N}_{A\{-k\}}^\top \sum_{\substack{i,g,\\ c_{ig}=1}}\sum_{s=1}^{q_0}\phi\boldsymbol{u}_{ig[k,s]}\boldsymbol{x}_i^{-k}\boldsymbol{B}_{[,k]}
$$

.

The derivation for $\boldsymbol{B}$ is similar with $\boldsymbol{A}$, and we have the following Gibbs sampling posterior updates

- $\left[\boldsymbol{A}_{[,k]}|\cdot\right] = \boldsymbol{N}_{\boldsymbol{A}_{\{-k\}}}\boldsymbol{a}_k$, for $k \in \{1, \dots p_0\}$ where $\boldsymbol{a}_k \sim \text{vMF}(\boldsymbol{\mu}_a)$ with

$$\boldsymbol{\mu}_a = \boldsymbol{N}^{\top}_{A\{-k\}} \sum_{\substack{i,g,\\ c_{ig}=1}} \sum_{s=1}^{q_0} \phi \boldsymbol{u}_{ig[k,s]} \boldsymbol{x}_i^{-k} \boldsymbol{B}_{[,k]}$$

and

$$\boldsymbol{x}_i^{-k} = \boldsymbol{x}_i - \sum_{k^c \neq k} \sum_{s=1}^{q_0} \boldsymbol{A}_{[,k^c]} \boldsymbol{u}_{ig[k^c,s]} \boldsymbol{B}^{\top}_{[,s]}$$

- $\left[\boldsymbol{B}_{[,d]}|\cdot\right] = \boldsymbol{N}_{\boldsymbol{B}_{\{-d\}}}\boldsymbol{b}_d$, for $d \in \{1, \dots q_0\}$ where $\boldsymbol{b}_d \sim \text{vMF}(\boldsymbol{\mu_b})$ with

$$\boldsymbol{\mu_b} = \boldsymbol{N}^{\top}_{\boldsymbol{B}_{\{-d\}}} \sum_{\substack{i,g,\\ c_{ig}=1}} \sum_{t=1}^{p_0} \boldsymbol{u}_{ig[t,d]} \tilde{\boldsymbol{x}}_i^{-d} \boldsymbol{A}_{[,t]},$$

and

$$\tilde{\boldsymbol{x}}_i^{-d} = \boldsymbol{x}_i - \sum_{d^c \neq d} \sum_{t=1}^{p_0} \boldsymbol{B}_{[,d^c]} \boldsymbol{u}_{ig[t,d^c]} \boldsymbol{A}^{\top}_{[,t]}$$

The update of the rest of the model parameters are standard and the Gibbs sampling posterior update is as follows

- If we denote likelihood equation (2.5) as $L(\boldsymbol{\nu}) \doteq \prod_{i=1}^{n} \prod_{g=1}^{G} [\tilde{p}_{ig}]^{\text{I}(c_{ig}=1)}$, then

  $[\boldsymbol{c}_i|\cdot] \sim Multinomial(1, \tilde{\pi}_1, \dots, \tilde{\pi}_G)$

  where $\tilde{\pi}_m = \frac{\tilde{p}_{im}}{\sum_g \tilde{p}_{ig}}$, $m \in \{1, \dots, G\}$.

- $[\text{vec}(\boldsymbol{\eta}_g)|\cdot] \sim \mathcal{N}(\boldsymbol{M}, \boldsymbol{E})$

  where

  $\boldsymbol{M} = \boldsymbol{E}(\boldsymbol{B} \otimes \boldsymbol{A})^T \sum_{\substack{i,g,\\ c_{ig}=1}} [\text{vec}(\boldsymbol{x}_i) - (\boldsymbol{B}\lambda^{1/2}) \otimes (\boldsymbol{A}\boldsymbol{\Gamma}^{1/2})\, \text{vec}(\tilde{\boldsymbol{u}}_{ig})]$,

  and

  $\boldsymbol{E} = \frac{1}{n_g \phi}\boldsymbol{I}$, $n_g = \sum_i c_{ig}$, $\tilde{\boldsymbol{u}}_{ig} = \boldsymbol{\Gamma}^{-1/2}(\boldsymbol{u}_{ig} - \boldsymbol{\eta}_g)\boldsymbol{\Lambda}^{-1/2}$

- $[\lambda_1^{1/2}, \dots \lambda_{q_0}^{1/2}]$ is updated element-wisely:

  $[\lambda_m^{1/2}|\cdot] \sim \mathcal{N}(M, E)$, $m = 1, \dots, q_0$.

  where

$$M = E \cdot \left(\sum_{\substack{i,g,\\ c_{ig}=1}} \tilde{\boldsymbol{u}}_{ig[,m]}^T \boldsymbol{\Gamma}^{1/2} [(\boldsymbol{A}^T \boldsymbol{x}_i \boldsymbol{B})_{[,m]} - \boldsymbol{\eta}_{j[,m]}] \phi + \mu_\lambda / \sigma_\lambda^2 \right),$$

and

$$E = \left(\sum_{\substack{i,g,\\ c_{ig}=1}} \tilde{\boldsymbol{u}}_{ig[,m]}^T \boldsymbol{\Gamma} \tilde{\boldsymbol{u}}_{ig[,m]} \phi + 1/\sigma_\lambda^2 \right)^{-1},$$

$$\tilde{\boldsymbol{u}}_{ig} = \boldsymbol{\Gamma}^{-1/2} (\boldsymbol{u}_{ig} - \boldsymbol{\eta}_g) \boldsymbol{\Lambda}^{-1/2}$$

- Similarly, $[\psi_1^{1/2}, \dots \psi_{p_0}^{1/2}]$, for $[\psi_m^{1/2}], m = 1, \dots, p_0$:

  $[\psi_m^{1/2} | \cdot] \sim \mathcal{N}(M, E)$

  where

  $$M = E \cdot \left(\sum_{\substack{i,g,\\ c_{ig}=1}} \tilde{\boldsymbol{u}}_{ig[m,]} \boldsymbol{\Lambda}^{1/2} [(\boldsymbol{A}^T \boldsymbol{x}_i \boldsymbol{B})_{[m,]} - \boldsymbol{\eta}_{j[m,]}]^T \phi + \mu_\psi / \sigma_\psi^2 \right),$$

  and

  $$E = \left(\sum_{\substack{i,g,\\ c_{ig}=1}} \tilde{\boldsymbol{u}}_{ig[m,]} \boldsymbol{\Lambda} \tilde{\boldsymbol{u}}_{ig[m,]}^T \phi + 1/\sigma_\psi^2 \right)^{-1}$$

  $$\tilde{\boldsymbol{u}}_{ig} = \boldsymbol{\Gamma}^{-1/2} (\boldsymbol{u}_{ig} - \boldsymbol{\eta}_g) \boldsymbol{\Lambda}^{-1/2}$$

- $[\phi | \cdot] \sim Gamma(a, b)$

  where $a = a_0 + \frac{n_g pq}{2}$, $n_g = \sum_i c_{ig}$,

  $b = b_0 + \frac{1}{2} \sum_{\substack{i,g,\\ c_{ig}=1}} \left\| \boldsymbol{x}_i - \boldsymbol{A}_g (\boldsymbol{\eta}_g + \boldsymbol{\Gamma}^{1/2} \tilde{\boldsymbol{u}}_{ig} \boldsymbol{\Lambda}^{1/2}) \boldsymbol{B}_g^\top \right\|_F^2$

- Here we introduce a latent variable ω_i such that $o_i = I(\omega_i > 0)$,

  and we have $[\omega_i | \cdot] \sim \mathcal{N}(\beta_0 + \mathbb{V}_i^T \boldsymbol{\beta} + \sum_{g=1}^{G-1} \mathbb{X}_{ig}^T \boldsymbol{\delta}_g, 1)$

  The update of ω_i is as follows:

  If $o_i = 1$

  $$[w_i | \cdot] \sim \mathcal{N}\left(\beta_0 + \boldsymbol{z}_i^T \boldsymbol{\beta}, 1\right) \cdot I_{(0, +\infty)}$$

  If $o_i = 0$,

  $$[w_i | \cdot] \sim \mathcal{N}\left(\beta_0 + \boldsymbol{z}_i^T \boldsymbol{\beta}, 1\right) \cdot I_{(-\infty, 0)}$$

- $[\beta_0, \boldsymbol{\beta}|\cdot] \sim \mathcal{N}(M, E)$

  where

  $M = \left(\sum_i [1, \boldsymbol{z}_i]^\top [1, \boldsymbol{z}_i] + \frac{1}{\tau_0}\boldsymbol{I}\right)^{-1} \sum_i [1, \boldsymbol{z}_i] w_i$

  and

  $E = \left(\sum_i [1, \boldsymbol{z}_i]^\top [1, \boldsymbol{z}_i] + \frac{1}{\tau_0}\boldsymbol{I}\right)^{-1}$

- $[\boldsymbol{\pi}|\cdot] \sim Dirichlet(4 + n_1, \ldots, 4 + n_G)$

  where $n_g = \sum_i c_{ig}$

Appendix B: Computation Details in Chapter 3

The missing imputation procedure in chapter 3 is shown below. Let $\boldsymbol{m} \in \boldsymbol{R}^{n,r}$ denote the missing indicator, n is the sample size and r is the dimension of HAMD scores. $m_{i,j} = 1$ if the data is not missing, $m_{i,j} = 0$ if the data is missing. Let $p_{i,j}$ be the probability that the data exists, i.e, $m_{i,j} = 1$, then the model of missing is a probit model with

$$\Phi^{-1}(p_{i,j}) = \theta_0 + \theta_1 \cdot Y_{i,j-1} + \theta_2 \cdot Y_{i,j}$$

Note that by introducing a latent variable $w_{i,j}$ for each $m_{i,j}$ with

$$w_{i,j} = \theta_0 + \theta_1 \cdot Y_{i,j-1} + \theta_2 \cdot Y_{i,j} + \epsilon$$

where $\epsilon \sim \boldsymbol{N}(0, 1)$. Then $m_{i,j} = 1$ if and only if $w_{i,j} > 0$; $m_{i,j} = 0$ if and only if $w_{i,j} \leq 0$.

In the following of this section, we use $\mathbf{Y}$ to denote the full data, i.e, $\mathbf{Y}$ is the combination of the missing and existing $\mathbf{Y}_{mis}$ and Y_{exist}. Similarly, for each patient's score, we still use $\mathbf{Y}_i$ to denote the full, and use $\mathbf{Y}_{i,mis}$, $\mathbf{Y}_{i,exist}$ to denote the missing and existing scores.

Because we need to partition $\mathbf{Y}_i$ into the missing part and the existing part, a further partition for the mean and covariance is needed. In our notation, $[\mathbf{Y}_i|\mathrm{D}_{i,k} = 1] \sim \boldsymbol{N}(\boldsymbol{\mu}_{env,k}, \boldsymbol{\Sigma}_{env})$. After partition, $[\mathbf{Y}_{i,mis}, \mathbf{Y}_{i,exist}|\mathrm{D}_{i,k} = 1] \sim \boldsymbol{N}(\boldsymbol{\mu}_{parti,k}, \boldsymbol{\Sigma}_{parti})$

General derivation:

$$
\begin{aligned}
& f(\mathbf{Y}_{i,mis}|\mathbf{Y}_{i,exist}, \boldsymbol{w}_i, \mathrm{D}_{i,k}=1, \boldsymbol{\mu}_{parti,k}, \boldsymbol{\Sigma}_{parti}) \\
& \propto f(\mathbf{Y}_{i,mis}, \boldsymbol{w}_i|\mathbf{Y}_{i,exist}, \mathrm{D}_{i,k}=1, \boldsymbol{\mu}_{parti,k}, \boldsymbol{\Sigma}_{parti}) \\
& = f(\boldsymbol{w}_i|\mathbf{Y}_{i,mis}, \mathbf{Y}_{i,exist}) \cdot f(\mathbf{Y}_{i,mis}|\mathbf{Y}_{i,exist}, \boldsymbol{\mu}_{parti,k}, \boldsymbol{\Sigma}_{parti}, \mathrm{D}_{i,k}=1) \\
& = [\prod_{j=1}^{r} f(w_{i,j}|\mathbf{Y}_{i,mis}, \mathbf{Y}_{i,exist})] \cdot f(\mathbf{Y}_{i,mis}|\mathbf{Y}_{i,exist}, \boldsymbol{\mu}_{parti,k}, \boldsymbol{\Sigma}_{parti}, \mathrm{D}_{i,k}=1) \\
& \doteq ① \cdot ②
\end{aligned}
\tag{B.1}
$$

For part ②, $f(\mathbf{Y}_{i,mis}|\mathbf{Y}_{i,exist}, \mathrm{D}_{i,k}=1, \cdot)$ is still a multivariate normal distribution. Assume there are r_1 missing values and r_2 existing values with $r_1 + r_2 = r$, i.e,

$$
\mathbf{Y}_{i,partition} = \begin{bmatrix} \mathbf{Y}_{i,mis} \\ \mathbf{Y}_{i,exist} \end{bmatrix} \text{ with sizes } \begin{bmatrix} r_1 \times 1 \\ r_2 \times 1 \end{bmatrix}
$$

The according $\boldsymbol{\mu}_{parti,k}$ and $\boldsymbol{\Sigma}_{parti}$ is

$$
\boldsymbol{\mu}_{parti,k} = \begin{bmatrix} \boldsymbol{\mu}_{1,k} \\ \boldsymbol{\mu}_{2,k} \end{bmatrix} \text{ with sizes } \begin{bmatrix} r_1 \times 1 \\ r_2 \times 1 \end{bmatrix}
$$

$$
\boldsymbol{\Sigma}_{parti} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} \text{ with sizes } \begin{bmatrix} r_1 \times r_1 & r_1 \times r_2 \\ r_2 \times r_1 & r_2 \times r_2 \end{bmatrix}
$$

With the above notation,

$$
f(\mathbf{Y}_{i,mis}|\mathbf{Y}_{i,exist}, \mathrm{D}_{i,k}=1, \cdot) \sim \boldsymbol{N}(\bar{\boldsymbol{\mu}}_k, \bar{\boldsymbol{\Sigma}})
$$

with

$$
\bar{\boldsymbol{\mu}}_k = \boldsymbol{\mu}_{1,k} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\mathbf{Y}_{i,exist} - \boldsymbol{\mu}_{2,k}\right)
$$

$$
\bar{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}
$$

For part ①, the density function is

$$
\prod_{j=2}^{r} \frac{1}{\sqrt{2\pi}} \cdot \exp -\frac{1}{2}(w_{i,j} - \boldsymbol{\theta}^T \tilde{\mathbf{Y}}_i^j)^2
$$

with

$$\boldsymbol{\theta} = \begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \end{bmatrix} \text{ and } \tilde{\mathbf{Y}}_i^j = \begin{bmatrix} 1 \\ Y_{i,j-1} \\ Y_{i,j} \end{bmatrix}$$

It is known that, after considering ① and ②, the conditional distribution of $\mathbf{Y}_{i,mis}$ still follows a multivariate normal distribution. The following material are the steps that finds the kernel.

One loop, four situation:

From ②, the kernel part with second order is $\mathbf{Y}_{i,mis}^T \bar{\boldsymbol{\Sigma}}^{-1} \mathbf{Y}_{i,mis}$. The kernel with first order is $\mathbf{Y}_{i,mis}^T \bar{\boldsymbol{\Sigma}}^{-1} \bar{\boldsymbol{\mu}}_{i,k}$.

Before the loop, set

$$\mathbf{Final.K}_1 = \mathbf{Y}_{i,mis}^T \bar{\boldsymbol{\Sigma}}^{-1} \bar{\boldsymbol{\mu}}_{i,k}$$

$$\mathbf{Final.K}_2 = \mathbf{Y}_{i,mis}^T \bar{\boldsymbol{\Sigma}}^{-1} \mathbf{Y}_{i,mis}$$

From each j, consider the following 4 different situations.

- Both $Y_{i,j-1}$ and $Y_{i,j}$ exist.

  In this case, there is no second order and first order kernel part in ①.

  $$\mathbf{Final.K}_1 = \mathbf{Final.K}_1 + \mathbf{0}$$

  $$\mathbf{Final.K}_2 = \mathbf{Final.K}_2 + \mathbf{0}$$

- $Y_{i,j-1}$ exist but $Y_{i,j}$ miss.

  In this case, the first order contribution from ① is

  $$Y_{i,j}\theta_2 \cdot (w_{i,j} - \theta_0 - \theta_1 Y_{i,j-1})$$

  and the second order contribution is

  $$Y_{i,j}^2 \theta_2^2$$

Suppose $Y_{i,j}$ corresponds to the t'th index in $\mathbf{Y}_{i,mis}$, i.e,

$$\mathbf{Y}_{i,mis} = \begin{bmatrix} \vdots \\ Y_{i,mis,t} = Y_{i,j} \\ \vdots \end{bmatrix}$$

Then

$$\begin{aligned} \mathbf{Final.K}_1 &= \mathbf{Final.K}_1 + \mathbf{Y}_{i,mis}^T \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-1} \\ \theta_2 \cdot (w_{i,j} - \theta_0 - \theta_1 Y_{i,j-1}) \\ 0_{t+1} \\ \vdots \\ 0_{r_1} \end{bmatrix} \\ &= \mathbf{Y}_{i,mis}^T (\bar{\mathbf{\Sigma}}^{-1} \bar{\boldsymbol{\mu}}_{i,k} + \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-1} \\ \theta_2 \cdot (w_{i,j} - \theta_0 - \theta_1 Y_{i,j-1}) \\ 0_{t+1} \\ \vdots \\ 0_{r_1} \end{bmatrix}) \end{aligned} \tag{B.2}$$

$$
\begin{aligned}
\mathbf{Final.K}_2 &= \mathbf{Final.K}_2 + \mathbf{Y}_{i,mis}^T \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-1} \\ \theta_2 \\ 0_{t+1} \\ \vdots \\ 0_{r_1} \end{bmatrix} \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-1} \\ \theta_2 \\ 0_{t+1} \\ \vdots \\ 0_{r_1} \end{bmatrix}^T \mathbf{Y}_{i,mis} \\
&= \mathbf{Y}_{i,mis}^T (\bar{\mathbf{\Sigma}}^{-1} + \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-1} \\ \theta_2 \\ 0_{t+1} \\ \vdots \\ 0_{r_1} \end{bmatrix} \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-1} \\ \theta_2 \\ 0_{t+1} \\ \vdots \\ 0_{r_1} \end{bmatrix}^T) \mathbf{Y}_{i,mis}
\end{aligned} \tag{B.3}
$$

- $Y_{i,j-1}$ miss but $Y_{i,j}$ exist.

  Similarly, the first order contribution from ① is

  $$Y_{i,j-1}\theta_1 \cdot (w_{i,j} - \theta_0 - \theta_2 Y_{i,j})$$

  and the second order contribution is

  $$Y_{i,j-1}^2 \theta_1^2$$

  Again, suppose $Y_{i,j-1}$ corresponds to the $(t-1)$'th index in $\mathbf{Y}_{i,mis}$.

$$
\begin{aligned}
\mathbf{Final.K}_1 &= \mathbf{Final.K}_1 + \mathbf{Y}_{i,mis}^T \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-2} \\ \theta_1 \cdot (w_{i,j} - \theta_0 - \theta_2 Y_{i,j}) \\ 0_t \\ \vdots \\ 0_{r_1} \end{bmatrix} \\
&= \mathbf{Y}_{i,mis}^T (\bar{\mathbf{\Sigma}}^{-1} \bar{\boldsymbol{\mu}}_{i,k} + \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-2} \\ \theta_1 \cdot (w_{i,j} - \theta_0 - \theta_2 Y_{i,j}) \\ 0_t \\ \vdots \\ 0_{r_1} \end{bmatrix})
\end{aligned}
\tag{B.4}
$$

$$
\begin{aligned}
\mathbf{Final.K}_2 &= \mathbf{Final.K}_2 + \mathbf{Y}_{i,mis}^T \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-2} \\ \theta_1 \\ 0_t \\ \vdots \\ 0_{r_1} \end{bmatrix} \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-2} \\ \theta_1 \\ 0_t \\ \vdots \\ 0_{r_1} \end{bmatrix}^T \mathbf{Y}_{i,mis} \\
&= \mathbf{Y}_{i,mis}^T (\bar{\mathbf{\Sigma}}^{-1} + \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-2} \\ \theta_1 \\ 0_t \\ \vdots \\ 0_{r_1} \end{bmatrix} \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-2} \\ \theta_1 \\ 0_t \\ \vdots \\ 0_{r_1} \end{bmatrix}^T) \mathbf{Y}_{i,mis}
\end{aligned}
\tag{B.5}
$$

- Both $Y_{i,j-1}$ and $Y_{i,j}$ miss

  In this case, the first order contribution from ① is

$$\begin{bmatrix} Y_{i,j-1} & Y_{i,j} \end{bmatrix} \cdot (w_{i,j} - \theta_0) \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix}$$

  and the second order contribution is

$$\begin{bmatrix} Y_{i,j-1} & Y_{i,j} \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} \begin{bmatrix} \theta_1 & \theta_2 \end{bmatrix} \begin{bmatrix} Y_{i,j-1} \\ Y_{i,j} \end{bmatrix}$$

  Again, suppose $Y_{i,j-1}$ corresponds to the $(t-1)$'th index in $\mathbf{Y}_{i,mis}$, and $Y_{i,j}$ corresponds to the (t)'th index in $\mathbf{Y}_{i,mis}$ Then

$$\begin{aligned} \mathbf{Final.K}_1 &= \mathbf{Final.K}_1 + \mathbf{Y}_{i,mis}^T \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-2} \\ \theta_1 \cdot (w_{i,j} - \theta_0) \\ \theta_2 \cdot (w_{i,j} - \theta_0) \\ 0_{t+1} \\ \vdots \\ 0_{r_1} \end{bmatrix} \\ &= \mathbf{Y}_{i,mis}^T (\bar{\mathbf{\Sigma}}^{-1} \bar{\boldsymbol{\mu}}_{i,k} + \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-2} \\ \theta_1 \cdot (w_{i,j} - \theta_0) \\ \theta_2 \cdot (w_{i,j} - \theta_0) \\ 0_{t+1} \\ \vdots \\ 0_{r_1} \end{bmatrix} \end{aligned} \tag{B.6}$$

$$\begin{aligned}\mathbf{Final.K}_2 &= \mathbf{Final.K}_2 + \mathbf{Y}_{i,mis}^T \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-2} \\ \theta_1 \\ \theta_2 \\ 0_{t+1} \\ \vdots \\ 0_{r_1} \end{bmatrix} \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-2} \\ \theta_1 \\ \theta_2 \\ 0_{t+1} \\ \vdots \\ 0_{r_1} \end{bmatrix}^T \mathbf{Y}_{i,mis} \\ &= \mathbf{Y}_{i,mis}^T (\bar{\mathbf{\Sigma}}^{-1} + \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-2} \\ \theta_1 \\ \theta_2 \\ 0_{t+1} \\ \vdots \\ 0_{r_1} \end{bmatrix} \begin{bmatrix} 0_1 \\ \vdots \\ 0_{t-2} \\ \theta_1 \\ \theta_2 \\ 0_{t+1} \\ \vdots \\ 0_{r_1} \end{bmatrix}^T) \mathbf{Y}_{i,mis}\end{aligned} \tag{B.7}$$

Loop $r-1$ times and return the final.k1 and final.k2. Calculate the conditional mean and covariance based on it.

Before the running of the algorithm, impute the all the missing values from $\boldsymbol{N}(\boldsymbol{\mu}_{init}, \boldsymbol{\Sigma}_{init})$, where $\boldsymbol{\mu}_{init}$ is the sample mean of the complete cases, and $\boldsymbol{\Sigma}_{init}$ is the covariance matrix of the complete cases.

Appendix C: Computation Details in Chapter 4

The detailed MCMC Gibbs sampling approach in chapter 4 for the parameters and latent variables mentioned in the model is given as below,

- Given $\mathbf{x}$ and $(\mathbf{B}_1, \mathbf{B}_2)$, (λ, ξ) can be calculated as $\lambda = \mathbf{x}\mathbf{B}_1^T$ and $\xi = \mathbf{x}\mathbf{B}_2^T$.

  Thus $(\lambda^{(t+1)}, \xi^{(t+1)})$ based on $\mathbf{x}^{(t+1)}$ and $\boldsymbol{\nu}^{(t)}$

$$\lambda^{(t+1)} = \mathbf{x}^{(t+1)}\mathbf{B}_1^{(t)T}$$

$$\xi^{(t+1)} = \mathbf{x}^{(t+1)}\mathbf{B}_2^{(t)T}$$

- With $\boldsymbol{\theta} \sim \mathcal{N}(0, 100\mathbf{I}_r)$, we have

$$\begin{aligned} & p(\boldsymbol{\theta}|\sigma, \mathbf{x}, \mathbf{y}, \mathbf{B}_1) \\ \propto & f(\mathbf{y}|\mathbf{x}, \mathbf{B}_1, \sigma, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \\ \propto & \exp\{-\frac{\sum_{i=1}^n (y_i - \boldsymbol{x}_i^T \mathbf{B}_1 \boldsymbol{\theta})^2}{2\sigma^2} - \frac{\boldsymbol{\theta}^\top \boldsymbol{\theta}}{200}\} \end{aligned}$$

  Thus $[p(\boldsymbol{\theta}^{(t+1)}|\sigma^{(t)}, \mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}, \mathbf{B}_1^{(t)}] \sim \mathcal{N}(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$ where

$$\boldsymbol{\mu}_\theta = \boldsymbol{\Sigma}_\theta \mathbf{B}_1^{(t)} \mathbf{x}^{(t+1)T} \boldsymbol{y}^{(t+1)}$$

$$\boldsymbol{\Sigma}_\theta = \frac{1}{\sigma^{(t)2}} \mathbf{B}_1^{(t)} \mathbf{x}^{(t+1)T} \mathbf{x}^{(t+1)} \mathbf{B}_1^{(t)T} + \frac{1}{100} \mathbf{I}_r$$

- With $p(\boldsymbol{\Omega}) \sim \mathrm{IW}(\mathbf{S}, s)$, we have

$$\begin{aligned} p(\boldsymbol{\Omega}|\lambda) \propto & \, p(\boldsymbol{\Omega}) f(\lambda|\boldsymbol{\Omega}) \\ \propto & |\boldsymbol{\Omega}|^{-(s+r+1)/2} \exp -\frac{1}{2} \operatorname{tr}\left(S\boldsymbol{\Omega}^{-1}\right) \\ & \times |\boldsymbol{\Omega}|^{-\frac{n}{2}} \exp -\frac{1}{2} \sum_{i=1}^n \boldsymbol{\lambda}_i^T \boldsymbol{\Omega}^{-1} \boldsymbol{\lambda}_i \end{aligned}$$

$[p(\mathbf{\Omega}^{(t+1)}|\lambda^{(t+1)}] \sim \text{IW}(\tilde{\mathbf{S}}, \tilde{s})$ where

$$\tilde{s} = s + n$$

$$\tilde{\mathbf{S}} = \mathbf{S} + \lambda^{(t+1)T}\lambda^{(t+1)}$$

- With $p(\mathbf{\Omega}_0) \sim \text{IW}(\mathbf{S}_0, s_0)$, we have

$$\begin{aligned} p(\mathbf{\Omega}_0|\xi) &\propto p(\mathbf{\Omega}_0) f(\xi|\mathbf{\Omega}_0) \\ &\propto |\mathbf{\Omega}_0|^{-(s_0+p-r+1)/2} \exp -\frac{1}{2}\,\text{tr}\left(S\mathbf{\Omega}_0^{-1}\right) \\ &\quad \times |\mathbf{\Omega}_0|^{-\frac{n}{2}} \exp -\frac{1}{2}\sum_{i=1}^{n} \boldsymbol{\xi}_i^T \mathbf{\Omega}_0^{-1} \boldsymbol{\xi}_i \end{aligned}$$

Thus $[p(\mathbf{\Omega}_0^{(t+1)}|\xi^{(t+1)}] \sim \text{IW}(\tilde{\mathbf{S}}_0, \tilde{s}_0)$ where

$$\tilde{s}_0 = s_0 + n$$

$$\tilde{\mathbf{S}}_0 = \mathbf{S}_0 + \xi^{(t+1)T}\xi^{(t+1)}$$

- With $p(\sigma^2) \sim \text{IG}(a_0, b_0)$, we have

$$\begin{aligned} & p(\sigma^2|\mathbf{x}, \mathbf{y}, \mathbf{B}_1, \boldsymbol{\theta}) \\ \propto & p(\sigma^2) f(\mathbf{y}|\mathbf{x}, \mathbf{B}_1, \boldsymbol{\theta}) \\ \propto & (\sigma^2)^{-(a_0+1)} \exp -\frac{b_0}{\sigma^2} \\ & \times (\sigma^2)^{-\frac{n}{2}} \exp -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T \mathbf{B}_1 \boldsymbol{\theta})^2 \end{aligned}$$

Thus,

$[p(\sigma^{2(t+1)}|\mathbf{x}^{(t+1)}, \mathbf{y}^{(t+1)}, \mathbf{B}_1^{(t)}, \boldsymbol{\theta}^{(t+1)}] \sim \text{IG}(\tilde{a}_0, \tilde{b}_0)$

$$\tilde{a}_0 = a_0 + \frac{n}{2}$$

$$\tilde{b}_0 = b_0 + \frac{1}{2}\sum_{i=1}^{n}(y_i^{(t+1)} - \boldsymbol{x}_i^{(t+1)T} \mathbf{B}_1^{(t)} \boldsymbol{\theta}^{(t+1)})^2$$

- With $\mathbf{B}_1^{(t+1)}$and $\mathbf{B}_2^{(t+1)}$ expressed as functions of a matrix $\mathbf{A}$, we let $p(\mathbf{A}) \sim \mathcal{MN}(\boldsymbol{A}_0, 10\mathbf{I}, 10\mathbf{I})$, we have.

$$
\begin{aligned}
& p(\mathbf{A}|\mathbf{x}, \mathbf{y}, \sigma, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\theta}) \\
\propto & f(\mathbf{y}|\mathbf{x}, \mathbf{B}_1(\mathbf{A}), \sigma, \boldsymbol{\theta}) f(\mathbf{x}|\mathbf{B}_1(\mathbf{A}), \mathbf{B}_2(\mathbf{A}), \boldsymbol{\Omega}, \boldsymbol{\Omega}_0) p(\mathbf{A}) \\
\propto & \left\{ \exp -\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^T \mathbf{B}_1 \boldsymbol{\theta})^2 \right\} \\
& \times \left\{ |\boldsymbol{\Sigma}_{\mathbf{x}}|^{-\frac{n}{2}} \exp -\frac{1}{2} \sum_{i=1}^{n} \boldsymbol{x}_i^T \boldsymbol{\Sigma}_x \boldsymbol{x}_i \right\} p(\mathbf{A})
\end{aligned}
$$

  and $\mathbf{A}$ is updated through the following Metropolis-Hasting algorithm.

  1. Propose $\mathbf{A}^\star$, $\mathbf{A}^\star = \mathbf{A} + \boldsymbol{e}$ with $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\tau}^2)$.
  2. Calculate $p_{\mathbf{A}} = p(\mathbf{A}|\mathbf{x}, \mathbf{y}, \sigma, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\theta})$ and
     $p_{\mathbf{A}}^\star = p(\mathbf{A}^\star|\mathbf{x}, \mathbf{y}, \sigma, \boldsymbol{\Omega}, \boldsymbol{\Omega}_0, \boldsymbol{\theta})$.
  3. Accept the proposed state with probability $\alpha(\mathbf{A}^\star) = min \left\{ \frac{p_{\mathbf{A}}^\star}{p_{\mathbf{A}}}, 1 \right\}$.

  $\boldsymbol{\tau}$ serves as a hyper-parameter that is chosen to ensure that the acceptance rate of $\boldsymbol{A}$ falls within the specified interval of $(0.2, 0.6)$.

To maintain traceable outcomes, we employed distinct seeds for each simulated dataset. For initializing the MCMC process, random values were utilized for the parameters $\boldsymbol{\nu}$ as well as the latent $(\mathbf{x}, \mathbf{y})$.

The conditional distribution of $(\mathbf{x}, \mathbf{y})$ given $\boldsymbol{s}_{\text{dp}}$ and $\boldsymbol{\nu}^{(t)}$ is shown in algorithm (1). The notation $\mathcal{N}(\boldsymbol{t}; \mathbf{0}, \sigma_{\text{dp}}^2 \mathbf{I})$ represents the probability density of $\mathcal{N}(\mathbf{0}, \sigma_{\text{dp}}^2 \mathbf{I})$ in $\boldsymbol{t}$.

Statement on Computing Resources We ran the experiments through software R on Compute Canada high performance cluster. We conducted individual MCMC chains, each comprising 10000 iterations. A standard chain necessitates approximately 9 minutes to complete when considering a sample size of $n = 500$, and around 30 minutes when $n = 1000$. However, with $n = 5000$, the runtime experienced a significant surge, extending to around 12 hours.

Algorithm 1 Update $(\mathbf{x}, \mathbf{y})$ within Gibbs sampler

For each $i = 1, 2, \ldots, n,$ update $(\boldsymbol{x}_i, y_i) \mid \boldsymbol{\nu}, \boldsymbol{s}_{\text{dp}}$.
  (a) Propose $(\boldsymbol{x}_i^\star, y_i^\star)$ as follows:
    $\boldsymbol{x}_i^\star \sim \mathcal{N}(\mathbf{0}, \mathbf{B}_1 \boldsymbol{\Omega} \boldsymbol{B}^\top + \mathbf{B}_2 \boldsymbol{\Omega}_0 \mathbf{B}_2^\top)$
    $y_i^\star | \boldsymbol{x}_i^\star \sim \mathcal{N}(\boldsymbol{x}_i^{\star T} \mathbf{B}_1 \boldsymbol{\theta}, \sigma^2)$
  (b) Set $\boldsymbol{t}_s = \{\mathbf{x}^\top \mathbf{y}, \mathbf{x}^\top \mathbf{x}, \mathbf{y}^\top \mathbf{y}\}$ with current $(\mathbf{x}, \mathbf{y})$.
    Update $\boldsymbol{t}_s$ and set $\boldsymbol{t}_s^+ = \boldsymbol{t}_s - \boldsymbol{t}_i + \boldsymbol{t}_i^\star$.
    $\boldsymbol{t}_i = \{\boldsymbol{x}_i y_i, \boldsymbol{x}_i \boldsymbol{x}_i^\top, y_i^2\}$,
    $\boldsymbol{t}_i^\star = \{\boldsymbol{x}_i^\star y_i^\star, \boldsymbol{x}_i^\star \boldsymbol{x}_i^{\star T}, y_i^{\star 2}\}$
  (c) Accept the proposed state with
    probability $\alpha(\boldsymbol{x}_i^\star, y_i^\star)$ given by:
    $\alpha(\boldsymbol{x}_i^\star, y_i^\star) = min \left\{ \frac{\mathcal{N}(\boldsymbol{t}_s^+; \mathbf{0}, \sigma_{\text{dp}}^2 \mathbf{I})}{\mathcal{N}(\boldsymbol{t}_s; \mathbf{0}, \sigma_{\text{dp}}^2 \mathbf{I})}, 1 \right\}$
  (d) Set $\boldsymbol{t}_s = \boldsymbol{t}_s^+$ if the state is accepted.