STATISTICAL MACHINE LEARNING AND DATA FUSION METHODOLOGIES: APPLICATIONS IN HEALTHCARE

BY

YU DING

BS, University of Science and Technology of China, 2017
MS, Wayne State University, 2019

DISSERTATION

Submitted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Industrial and Systems Engineering
in the Graduate School of
Binghamton University
State University of New York
2024

Accepted in partial fulfillment of the requirements for
the degree of Doctor of Philosophy in Industrial and Systems Engineering
in the Graduate School of
Binghamton University
State University of New York
2024

May 31st, 2024

Dr. Bing Si, Chair and Faculty Advisor
Department of Systems Science and Industrial Engineering, Binghamton University - State University of New York

Dr. Shuxia (Susan) Lu, Committee Member
Department of Systems Science and Industrial Engineering, Binghamton University - State University of New York

Dr. Hiroki Sayama, Committee Member
Department of Systems Science and Industrial Engineering, Binghamton University - State University of New York

Dr. Sung Hoon Chung, Committee Member
Department of Systems Science and Industrial Engineering, Binghamton University - State University of New

Dr. Yu Chen, Outside Examiner
Department of Electrical and Computer Engineering, Binghamton University - State University of New York

Abstract

The increasing availability of healthcare data from diverse sources, such as large biobanks, electronic healthcare records, medical tests, and wearable sensors, has paved the way for the development of novel machine learning (ML) models. These models aim to capture the complexity of human health and disease, thereby enhancing healthcare data analysis. This dissertation addresses three major topics within this domain, presenting innovative solutions for analyzing multi-modal mixed-type data, federated learning for functional regression, and privacy-preserving telemedicine.

The first topic introduces a Multi-modal Mixed-type Structural Equation Model (M2-SEM) with structured sparsity for subgroup discovery from heterogeneous healthcare data. This model effectively handles both continuous and categorical data modalities through a novel Gauss-Hermite-enabled Expectation-Majorization-Minimization (GH-EMM) algorithm. Extensive simulation studies and applications to cardiometabolic risk factors demonstrate the model's ability to identify at-risk subgroups, highlighting its potential for enabling targeted health interventions and improving population health management.

The second topic focuses on Federated Function-on-Function Regression with an efficient Gradient Boosting algorithm (fed-GB-LSA). This approach ensures privacy-preserving telemedicine by allowing collaborative model training across multiple data sources without sharing sensitive data. The GB-based algorithm facilitates the sparse selection of functional and non-functional features, providing an efficient estimation

method. Its application to the telemonitoring of Obstructive Sleep Apnea (OSA) showcases the model's capability to maintain performance comparable to global models while preserving patient privacy, thereby supporting remote health monitoring and personalized treatment plans.

The third topic extends the research to Vertical Federated Learning (VFL) with Differential Privacy for function-on-function regression models. By integrating differential privacy into the federated gradient boosting process, we address the critical trade-off between model performance and privacy protection. Empirical results from simulation studies and a case study on OSA validate the method's robustness and practical relevance, demonstrating its applicability in privacy-sensitive healthcare environments where data security and patient confidentiality are paramount.

Overall, this dissertation significantly advances the field of healthcare data analysis by developing innovative machine learning models and algorithms that address the complexities of multi-modal mixed-type and functional health data. These methodologies ensure data privacy and computational efficiency, laying a strong foundation for future research and development. The findings and approaches proposed here contribute to improving health outcomes and advancing personalized medicine, ultimately enhancing healthcare delivery and patient care.

If I have seen further, it is by standing on the shoulders of Giants.
-Isaac Newton

Acknowledgements

I have received a lot of help along my journey to this day. My Ph.D. advisor, Dr. Bing Si, has provided invaluable advice, ranging from research to personal growth. She has given me maximum space and freedom while patiently handling my whimsical thoughts.

I am profoundly grateful to the Faculty of Systems Science and Industrial Engineering (SSIE) for their unwavering support towards every student. I also wish to extend my deep appreciation to the committee members, Dr. Susan Lu, Dr. Hiroki Sayama, and Dr. Sung Hoon Chung, for their invaluable insights that enhanced my research.

I would like to thank my parents, Jixiang Ding and Zheng Wang, for their unwavering support as I pursued my studies abroad over the years.

I am also grateful to my grandparents, who raised me from infancy and provided me with warm childhood memories.

I would like to thank Fan Xing, whose companionship has been a pillar of support through both challenging and joyous times.

Lastly, I would like to thank everyone who has contributed to our understanding of the external and internal worlds. I aspire to add my modest contribution to this vast pool of human knowledge.

Table of Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Background

Advancements in diagnostic imaging, smart sensing, and health information systems have led to a data-rich environment in healthcare. It is now feasible to meticulously track all information related to a patient's diagnosis, treatment, and care. This creates significant opportunities for Personalized Medicine (PM), enabling precise medical decision-making tailored to individuals at optimal times. However, the volume and complexity of this data exceed existing modeling capabilities of statistical methods. Additionally, the extensive use of statistical models raises concerns about privacy breaches, given the highly sensitive nature of healthcare data pertaining to individuals and service providers.

The aim of this research is to develop privacy-preserving statistical machine learning and data fusion methodologies to enhance the quality and performance of healthcare systems, from accurate diagnosis to phenotype discovery.

This dissertation addresses three emerging challenges in healthcare by developing novel statistical models that cater to the unique data structures and objectives of specific problem domains. The first topic focuses on multimodality imaging data fusion and the development of novel latent variable models for phenotype discovery. This involves integrating diverse imaging modalities to uncover latent phenotypes, enhancing our understanding of complex diseases. The second topic explores federated function-on-function regression for privacy-preserving telemedicine. This approach enables the

analysis of functional data from multiple sources without compromising patient privacy, thus facilitating secure and effective remote healthcare services. The third topic delves into vertical federated gradient boosting for functional regression with differential privacy. This method aims to improve the accuracy of functional regression models while ensuring the privacy of sensitive healthcare data through differential privacy techniques. Collectively, these studies aim to advance the field of personalized medicine by providing robust, privacy-preserving statistical tools that enhance the quality and performance of healthcare systems. Through these contributions, this research seeks to address the critical balance between leveraging rich healthcare data for improved medical decision-making and safeguarding the privacy of individuals and healthcare providers.

1.2 Summary of Research Topics and State of the Art

Topic (I): Multi-modal mixed-type structural equation modeling with structured sparsity for subgroup discovery from heterogeneous health data. The increasing availability of health data from resources such as large biobanks, electronic healthcare records, medical tests, and wearable sensors, has set the stage for the development of novel machine learning (ML) models for multi-modal mixed-type data to capture the complexity of human health and disease. Clustering is a type of ML model that aims to identify homogenous subgroups from heterogeneous data, providing a data-driven solution to targeted, subgroup-specific studies and interventions. While such data contain diverse and complementary information to facilitate decision-making and improve population health, clustering of high-dimensional multi-modal mixed-type data poses major challenges to existing ML and statistical models. We propose a novel Multi-modal Mixed-type Structural Equation Model (M2-SEM) with structured sparsity to cluster heterogeneous health data

for precise subgroup discovery. To accommodate a mix of continuous and categorical data modalities, we developed a novel Gauss-Hermite-enabled Expectation-Majorization-Minimization (GH-EMM) algorithm that integrates the GH quadrature and the Majorization Maximization (MM) algorithm within the Expectation Maximization (EM) framework for efficient model estimation. The proposed M2-SEM and GH-EMM are first tested in extensive simulation studies in comparison with benchmarks, and then applied to identify subgroups of individuals with low- and high-risk of developing adverse cardiometabolic (CM) outcomes based on a full spectrum of CM risk factors such as poor nutrition and mental health, physical inactivity, and sleep deprivation. These findings shed light on the promise of using multi-modal mixed-type health data for early identification and targeted intervention of at-risk individuals for health promotion at the population level.

Topic (II): Federated Function-on-Function Regression with an Efficient Gradient Boosting Algorithm for Privacy-Preserving Telemedicine. Federated Learning (FL) is an emerging computing paradigm to collaboratively train Machine Learning (ML) models across multi-source data while preserving privacy. The major challenge of the "meaningful" implementation of FL for any ML model is how to guarantee that the federated ML model can achieve comparable performance compared to the global model trained using the combined data. Moreover, there are very limited studies on FL of the functional regression models that analyze functional data, a commonly encountered type of data in many fields. This study develops the first-of-its-kind federated Gradient Boosting algorithm with the Least Squares Approximation (fed-GB-LSA) for efficient, privacy-preserving federated learning of the function-on-function regression with several distinct merits: (1) The GB-based algorithm allows the sparse selection of multivariate functional and non-functional

features in the function-on-function regression prediction, which is not straightforward in the functional regression; (2) The parameter estimation by the GB algorithm results in separate sub-optimization problems with explicitly analytical solutions for each of the features, providing an efficient estimation algorithm for the function-on-function regression; (3) The LSA-enabled fed-GB provides a "one-shot" approach for FL that is communicationally and statistically efficient, providing theoretical guarantees to the federated model's performance without data sharing across local servers. The proposed fed-GB-LSA is tested in extensive simulation studies and applied in a real-world dataset for privacy-preserving telemonitoring of Obstructive Sleep Apnea (OSA).

Topic (III): Vertical Federated Functional Gradient Boosting with Differential Privacy. Vertical Federated Learning (VFL) has emerged as a significant technique for facilitating data collaboration among multiple organizations while complying with privacy regulations such as the General Data Protection Regulation (GDPR). This chapter presents an innovative approach to VFL by integrating Gradient Boosting for Functional Regression with Differential Privacy, specifically aimed at function-on-function regression models. The study addresses the challenge of preserving model performance while ensuring privacy through differentially private gradient sharing. The proposed method is evaluated for its prediction accuracy and privacy-preserving capability through simulation studies. The findings indicate a trade-off between prediction accuracy and privacy protection, with stricter privacy requirements leading to a decrease in prediction accuracy. However, enhanced privacy protection is confirmed through a membership inference attack. A case study on Obstructive Sleep Apnea (OSA) using data from the Sleep Heart Health Study (SHHS) illustrates the practical application and effectiveness of the proposed method. The

SHHS dataset includes functional features from electrocardiogram (ECG) and electroencephalogram (EEG) signals, in addition to non-functional features. The results demonstrate that the federated model achieves performance comparable to a global model and superior to local models, highlighting its potential in privacy-sensitive healthcare settings. This work advances the field of federated learning by incorporating differential privacy into functional regression, setting the stage for future developments in privacy-preserving predictive modeling.

1.3 Significance and Contribution of Research

Three studies have been conducted to address these issues. In the first study, we introduce a novel Multi-modal Mixed-type Structural Equation Model (M2-SEM) with structured sparsity, designed to cluster heterogeneous health data for precise subgroup discovery. M2-SEM employs a unique double L_{21} penalized likelihood formulation, facilitating the hierarchical selection of informative imaging modes and features. This formulation satisfies a Quadratic Majorization (QM) condition, enabling the development of an efficient Group-wise Majorization Descent (GMD) algorithm for model estimation. To handle a combination of continuous and categorical data modalities, we developed an innovative Gauss-Hermite-enabled Expectation-Majorization-Minimization (GH-EMM) algorithm. This algorithm integrates the Gauss-Hermite quadrature and the Majorization Maximization (MM) algorithm within the Expectation Maximization (EM) framework for efficient model estimation. The proposed M2-SEM and GH-EMM were rigorously tested through extensive simulation studies against benchmark models and subsequently applied to identify subgroups of individuals at low and high risk for adverse cardiometabolic (CM) outcomes. These outcomes were assessed based on a comprehensive spectrum of CM risk

factors, including poor nutrition, mental health issues, physical inactivity, and sleep deprivation.

The second study presents the development of an unprecedented federated Gradient Boosting algorithm with the Least Squares Approximation (fed-GB-LSA), aimed at efficient and privacy-preserving federated learning for function-on-function regression. This algorithm offers several distinct advantages: (1) It allows for the sparse selection of multivariate functional and non-functional features in function-on-function regression prediction, which is typically challenging in functional regression; (2) The parameter estimation by the GB algorithm results in separate sub-optimization problems with explicitly analytical solutions for each feature, ensuring efficient estimation for function-on-function regression; (3) The LSA-enabled fed-GB provides a "one-shot" approach for federated learning that is both communicationally and statistically efficient, with theoretical guarantees for the federated model's performance without necessitating data sharing across local servers. The proposed fed-GB-LSA was subjected to extensive simulation studies and applied to a real-world dataset for privacy-preserving telemonitoring of Obstructive Sleep Apnea (OSA).

In the third study, we introduce an innovative approach to Vertical Federated Learning (VFL) by integrating Gradient Boosting for Functional Regression with Differential Privacy, specifically targeting function-on-function regression models. This study addresses the challenge of preserving model performance while ensuring privacy through differentially private gradient sharing. It demonstrates that with minimal sacrifice in model performance, privacy can be maintained at a high level. The proposed method was evaluated for its prediction accuracy and privacy-preserving capability through

simulation studies, with prediction accuracies assessed under various privacy settings. Additionally, the privacy-preserving capability was tested via membership inference attacks, yielding outstanding results. Moreover, the proposed method was applied to a real-world dataset for the privacy-preserving telemonitoring of Obstructive Sleep Apnea (OSA). The results indicated that the global model under VFL significantly outperforms any of the local models while maintaining the patient's privacy.

Next three chapters provide detailed discussion of challenges, data, methods and results of each of study.

1.4 Dissertation Organization

The remainder of this dissertation is organized as follows. Chapter 2 develops a Multi-modal mixed-type structural equation modeling with structured sparsity for subgroup discovery from heterogeneous health data. Chapter 3 proposes a federated function-on-function regression with an efficient gradient boosting algorithm for privacy-preserving telemedicine. Chapter 4 introduces a vertical federated functional gradient boosting model with differential privacy. Finally, Chapter 5 presents the discussion and future work.

Chapter 2 Multi-modal Mixed-type Structural Equation Modeling with Structured Sparsity for Subgroup Discovery from Heterogeneous Health Data

2.1 Introduction

Subgroup discovery is of critical importance for heterogeneity delineation for many complex systems, particularly for those systems in which domain knowledge of the underlying mechanisms is too scarce to explicitly articulate and quantify the individual-to-individual similarities and dissimilarities. Fortunately, such information may have already been embedded in the data, which makes the data-driven approach a promising solution to delineate the heterogeneity in complex systems. Below we present several case studies across various domains that demand subgroup discovery.

- Subgroup discovery is critical to health management of chronic conditions such as cardiometabolic (CM) diseases. CM diseases are interconnected conditions including hypertension, diabetes, and cardiovascular diseases, known to be the leading cause of preventable death in the United States and worldwide (Shah et al., 2019). There is an estimated 47 million Americans living with CM diseases (American College of Cardiology, 2021) costing US healthcare systems more than $677 billion each year (Fryar et al., 2012; American Diabetes Association, 2018; Kirkland et al., 2018). One major task in CM health management is to identify high- vs low-risk subgroups in a large, vulnerable, and heterogeneous population to guide cost-effective and targeted disease intervention and management, eventually leading to improved population health (Buxton et al., 2018; Liu et al., 2021; Jiang et al., 2022; Alramadeen et al., 2023).

- With the help of advanced metering infrastructure, modern power systems can obtain real-time data with high resolution and large volumes. This enables researchers to discover power consumption patterns by clustering (Si et al., 2021), which can be used to support customer segmentation, enact tariff policies, detect load anomalies, and support load forecasting and demand-side responses. The inherent fluctuations and intermittent and uncertain characteristics of clean energy sources highlight the importance of a clear understanding of demand-side characteristics to ensure the power system is resilient and robust. Subgroups are identified based on the number, magnitude, draft, and lag of the peaks, as well as other characteristics of the demand curves over different time scales, ranging from residential to industrial sectors (Ryu et al., 2019; Lin et al., 2017).
- Clustering users and content on social media can provide critical information on many aspects, such as user behavior, content popularity, sentiment analysis, and target audience segmentation, which can be useful for businesses, marketers, and social media platforms to better understand and engage their users. For example, identifying medium vulnerabilities is often challenging for organizations because it is difficult to balance the cost of solving them with the associated risks. However, with the help of subgroup discovery, medium vulnerabilities can be timely identified by detecting continued discussions on social media (Allen et al., 2017).

Clustering is the natural choice for subgroup discovery from big data in complex systems (Tan et al., 2016, Sutherland et al., 2024, Mueller et al., 2024). However, there are significant challenges in employing conventional clustering methods to discover subgroups from high-dimensional heterogeneous data collected in the complex system of interest.

Specifically, big data has many challenging properties that overwhelm the modeling capacity of many existing machine learning and statistical models. First, it is common to have data of multi-modalities to characterize multi-faceted aspects of the complex system, providing complementary information for more precise differentiation among heterogeneous subjects. Taking CM data as an example, these CM diseases are known to be associated with a spectrum of multi-modal risk factors including but not limited to socioeconomics, sleep-related, nutritional, environmental, and other behavioral factors, upon which more precise determination of CM risk subgroups can be made. Second, due to distinct data collection procedures, multiple data modalities are likely to be mixed-typed, e.g., continuous, nominal, or ordinal. For example, the data modality collected by wearable sensors can be continuous, whereas survey data collection can result in a modality of categorical ordinal features on the Likert scale. Third, pooling together a large number of features in multi-modalities is likely to result in a dataset that has a latent factor structure underlying the original features in each modality. The presence of modality-specific latent factors needs to be considered in the modeling for both dimension reduction and ease of interpretation. Last but not least, the high dimensionality of multi-modal features requires structured variable selection approaches to be integrated within the clustering algorithm. That is, the variable selection strategy should consider the hierarchy of features within multi-modalities by selecting significant features by modality, aiming to reveal the significance of each modality in differentiating CM risk subgroups and facilitating domain knowledge discovery from high-dimensional multi-modal data.

To address these challenges, this study proposes to develop a novel Multi-modal Mixed-type Structural Equation Model (M2-SEM) with structured sparsity for precise

subgroup discovery from multi-modal, mixed-type, high-dimensional data. The proposed M2-SEM results in a complex objective function with both observed data and latent variables that need to be estimated by an Expectation Maximization (EM) framework. However, the traditional EM does not suffice due to the presence of categorical features, which introduce non-analytical formulas that overwhelm the traditional EM algorithm. While Monte Carlo (MC) simulation can be integrated within the EM to obtain a numerical approximation to the non-analytical formulas by repeatedly random sampling, MC EM relies on a large number of simulation runs and is computationally inefficient, especially for high-dimensional big data. Instead, this chapter proposes to use the Gauss-Hermite (GH) Quadrature, a numerical optimization approach, to approximate the non-analytical terms, thus achieving a drastic reduction in the computational cost. Additionally, the GH-enabled analytical terms are proven to have good computational properties such as Lipschitz-continuity and concavity, which can be efficiently solved by a Majorization-Minimization (MM) algorithm. Consequently, this chapter proposes a novel GH-enabled Expectation Majorization Minimization algorithm (GH-EMM) for model estimation. Finally, the proposed M2-SEM and GH-EMM are examined in simulation studies and then applied to a real-world dataset for CM subgroup discovery from multi-modal mixed-type CM data. The findings of the proposed model are consistent with medical intuition and domain knowledge that shed light on precise CM risk stratification and CM health promotion in the population.

The novel contributions of the proposed M2-SEM include the following: 1. development of a novel M2-SEM that can cluster multi-modal mixed-type data while preserving the modal-specific information and providing structured sparse selection; 2.

proposal of a novel GH-enabled Expectation-Majorization-Minimization (GH-EMM) for efficient model estimation; 3. deployment of the proposed machine learning method to identify CM subgroups from the multi-modal mixed-type real-world dataset to enrich the knowledge bank of CM subgroups and facilitate Precise Medicine in CM health for the first time. The remainder of this chapter is structured as follows: Chapter 2.2 discusses the relevant works; Chapter 2.3 introduces the mathematical formulation of the proposed M2-SEM; Chapter 2.4 describes the efficient model estimation algorithm; Chapter 2.5 presents the simulation studies; Chapter 2.6 discusses the application of the proposed method on a real-world dataset; Chapter 2.7 concludes the chapter.

2.2 Literature Review

Clustering is the most important subfield of unsupervised learning. Compared with supervised learning methods that depend on labeled data to "supervise" the model to classify or predict the response variable of interest, clustering handles unlabeled data by grouping heterogeneous samples into relatively homogeneous clusters that aim to maximize the within-cluster similarities and between-cluster dissimilarities. Most of the existing clustering methods are heuristic approaches such as K-mean, hierarchical clustering, and DBSCAN (Tan et al., 2016). K-mean is the most classic center-based clustering method. That is, a cluster that results from K-mean contains a set of observations in which each observation is closer to the centroid of this particular cluster than to the centroid of any other cluster. Hierarchical clustering is an agglomerative clustering method that produces a hierarchical clustering tree by starting with each observation as a singleton cluster and then repeatedly merging the two closest clusters until a single, all-encompassing cluster remains. DBSCAN is a density-based clustering method that defines clusters as

dense regions of observations while observations in low-density regions are classified as noise and omitted. While heuristic approaches are widely used, they face several common challenges such as subjective selection of optimal number of clusters and lack of statistical rigor.

Model-based clustering (MBC) encompasses clustering methods that are based on statistical models and therefore can leverage rigorous statistical criteria for model selection and inference. Gaussian Mixture Model (GMM) is the most common MBC method that models the data as a mixture of several Gaussian distributions. Each distribution corresponds to a cluster and the mean and covariance of each distribution provide a description of the corresponding cluster in terms of its center and spread. However, the conventional GMM can handle only continuous data, and there is a lack of consideration of categorical data. In recent years, increasing attention has been given to extending the classic GMM to cluster mixed-type data including both continuous and categorical data, due to the increasing need for many real-world applications. Lebret et al. (2015) proposed a mixed-MBC method for mixed-type data by modeling continuous data and categorical data with Gaussian distribution and multinominal distribution, respectively. Marbac et al. (2020) proposed a mixed-MBC method for mixed-type data and used BIC with a modified Expectation-Maximization algorithm for variable selection in mixed data clustering. The probability distributions of the mixed-type features are jointly considered in an integrated complete-date likelihood for model estimation. However, multinominal distribution can be used to model only categorical nominal features, and therefore both methods cannot handle categorical ordinal features. Alternatively, McParland and Gormley (2016) developed another mixed-MBC procedure by assuming each observed categorical feature as a

categorical nominal or ordinal manifestation of the latent continuous variable following a Gaussian distribution, while clustering is based on the continuous latent variables. However, none of the mixed-MBC has considered either latent variable or multi-modal data structure.

Structural Equation Models (SEMs) are referred to as a set of statistical models used to analyze the interconnected relationship of both observed variables and latent variables, providing a flexible framework for developing and processing complex relationships among multiple variables (Wang and Wang, 2019). Depending on different applications, SEMs can successfully accommodate various interconnected observed and latent variables through model specification to achieve tasks in both supervised and unsupervised learning. With appropriate specifications, an SEM can achieve clustering of multi-modal data with latent variable structure by assuming a latent categorical variable indicating the cluster membership to be linked with multi-modal data. However, most existing SEMs don't consider sparse learning and variable selection in the model estimation and thus fall short in modeling high-dimensional data. There are limited studies on sparse SEMs. The sparse-aware SEM was developed by employing lasso penalties for variable selection, and the proposed model can be estimated with the Block Coordinate Descent algorithm integrated with the Proximal Alternating Linearized Minimization (PALM) (Cai et al., 2013; Zhou and Cai, 2022). However, the proposed sparse-aware method can only consider SEMs with observed variables. Jacobucci et al. (2016) proposed a general sparse SEM framework with both observed and latent variables by comparing the similarity between the model-implied covariance with a sample covariance matrix but it may result in suboptimal solutions in empirical studies (Huang, 2020). Alternatively,

Huang et al. (2017) and Huang (2018) proposed a penalized likelihood (PL) method of the complete likelihood function in SEMs using smoothly clipped absolute deviation (SCAD) or Minimax Concave Penalty (MCP) approaches that were shown to achieve better convergence in model estimation (Huang, 2020). However, all the existing sparse SEMs assume that latent variables follow Gaussian distributions, and thus cannot be used for clustering-type SEMs in which a categorical latent variable is used to indicate latent cluster membership, not to mention clustering of multi-modal data.

2.3 Model Formulation

2.3.1 Conceptual framework

The objective of this study is to develop a novel generalized SEM with structured sparsity to cluster multi-modal mixed-typed high-dimensional data to reveal the heterogeneity in complex systems with little to no domain knowledge of the underlying mechanisms to explicitly articulate and quantify the heterogeneity. To fix the notations, let $\left\{\left\{\boldsymbol{x}_i^{(m)}\right\}_{m=1}^{M}\right\}_{i=1}^{N}$ denote features nested within M different modalities across N subjects to be clustered. $\boldsymbol{x}_i^{(m)}$ is a P-dimensional feature vector within the m-th modality for subject i for $i = 1, \dots, N$ and $m = 1, \dots, M$. To simplify the notation, we assume each modality contains the same type of features and denote the M_1 numerical modalities and M_2 categorical modalities with $\left\{\left\{\boldsymbol{x}_i^{(m)}\right\}_{m=1}^{M_1}\right\}_{i=1}^{N}$ and $\left\{\left\{\boldsymbol{x}_i^{(m)}\right\}_{m=M_1+1}^{M_1+M_2}\right\}_{i=1}^{N}$, respectively. For example, a numerical modality $\left\{\boldsymbol{x}_i^{(m)}\right\}_{i=1}^{N}$ may consist of sleep measures in a sensor-based sleep study, e.g., measures for sleep events, heart rate, oxygen saturation, or sleep architecture, whereas a categorical modality may consist of a set of questions from a health

survey questionnaire in a Likert scale. Next, we introduce the conceptual framework of the proposed method as shown in Figure 1.

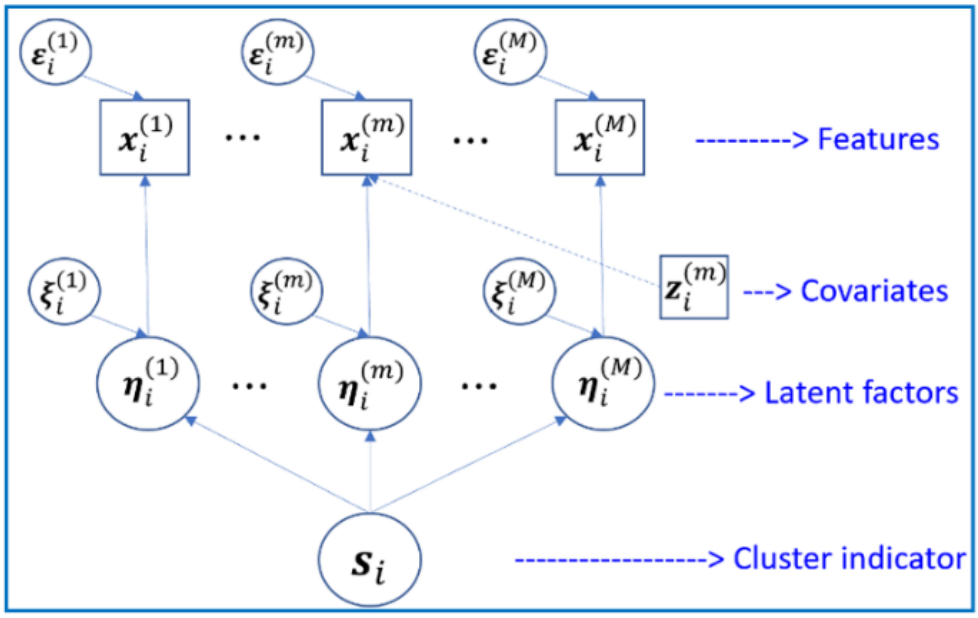

Figure 1: Graphical illustration of the proposed sparse SEM for multi-modal data clustering

The first step is to link $\boldsymbol{x}_i^{(m)}$ with a small number of latent factors $\boldsymbol{\eta}_i^{(m)}$ with Generalized Linear Models (GLMs) (Dobson and Barnett, 2018).

$$g\left(\boldsymbol{x}_i^{(m)}\right) = \boldsymbol{\alpha}^{(m)} + \boldsymbol{L}^{(m)}\boldsymbol{\eta}_i^{(m)} + \mathbf{B}^{(m)}\boldsymbol{z}_i^{(m)} + \boldsymbol{\varepsilon}_i^{(m)} \quad \text{for } i = 1, \dots, N \text{ and } m = 1, \dots, M. \tag{2.1}$$

where N is the total number of samples and M is the total number of modalities; $g(\cdot)$ is the logit function in GLM that takes different forms depending on the type of features $\boldsymbol{x}^{(m)}$, i.e., ordinal, nominal, or continuous; $\boldsymbol{\alpha}^{(m)}$ is the intercept and can be level-specific for categorical variables; $\boldsymbol{\eta}^{(m)}$ are unobserved latent factors while $\boldsymbol{z}^{(m)}$ are known covariates. The next step is to enable probabilistic clustering by linking the latent factors $\{\boldsymbol{\eta}^{(m)}\}_{m=1}^{M}$ with another latent factor $\boldsymbol{s}$ that indicates the cluster membership. We assume the total

number of clusters is K and let $\mathbf{s} = (s_1, \dots, s_K)^T$ follow a multinomial distribution. For example, if the subject belongs to the k-th cluster, then $s_k = 1$. Applying the similar idea of probabilistic clustering (Reynolds, 2009), the proposed method assumes that the distribution of latent factors $\boldsymbol{\eta}^{(m)}$ depends on the subject's cluster membership, e.g., $\boldsymbol{\eta}^{(m)}|s_k = 1 \sim N\left(\boldsymbol{\mu}^{(m,k)}, \boldsymbol{\Sigma}^{(m,k)}\right)$. As a result, the distribution of $\boldsymbol{\eta}^{(m)}$ can be written as the mixture of K Gaussian distributions as follows:

$$\boldsymbol{\eta}^{(m)} \sim \sum_{k=1}^{K} w_k N\left(\boldsymbol{\mu}^{(m,k)}, \boldsymbol{\Sigma}^{(m,k)}\right), \tag{2.2}$$

where $\boldsymbol{w} = (w_1, \dots, w_K)^T$ corresponds to the probabilities of different clusters. Last, we have included many modalities and features in clustering and propose to employ sparse learning techniques to select informative modalities and features. The three steps are not separate but need to be jointly estimated. Compared with existing clustering methods, the proposed M2-SEM has the following distinct advantages.

- To accommodate modalities with **mixed-typed features**, the GLM can be used to link mixed-typed features $\boldsymbol{x}_i^{(m)}$ with the latent factors $\boldsymbol{\eta}_i^{(m)}$ depending on the feature types, i.e., ordinal, nominal, or continuous. For example, surveys are cost-effective strategies to collect data from a large population and a data modality of survey questions often uses Likert scales, resulting in a large number of categorical ordinal features that instead require an ordinal GLM.
- To factor out the impact of **known covariates** $\boldsymbol{z}_i^{(m)}$ on clustering, these covariates can be considered as additional predictors in the model (2.1). For example, blood-based biomarkers or mental health conditions could be significantly associated with

covariates such as age and gender. Such measures should be adjusted for the known covariates when used in clustering.

- To accommodate **multi-modalities**, a modality index m is used to indicate different data modality for $m = 1, \dots, M$. Instead of pooling features across multi-modalities, we propose the modality-specific latent factor modeling as shown in Figure 1 to preserve modality-specific information.
- To address **high-dimensionality**, we propose a structured sparse learning approach to hierarchically select data modalities and features. That is, if a modality is selected as non-informative to clustering, all its features are excluded from clustering.

2.3.2 Sparse Multi-modal Mixed-type Structural Equation Model (M2-SEM)

This subchapter derives the GLM-type mathematical formulation of M2-SEM in detail. For a numerical modality, a regression model is applied to link features $\boldsymbol{x}_i^{(m)} \in \boldsymbol{R}^{P\times 1}$ with latent factors $\boldsymbol{\eta}_i^{(m)} \in \boldsymbol{R}^{Q\times 1}$ ($Q \ll P$) and covariates $\boldsymbol{z}_i^{(m)} \in \boldsymbol{R}^{R\times 1}$, i.e.,

$$\boldsymbol{x}_i^{(m)} = \mathbf{L}^{(m)}\boldsymbol{\eta}_i^{(m)} + \mathbf{B}^{(m)}\boldsymbol{z}_i^{(m)} + \boldsymbol{\varepsilon}_i^{(m)} \text{ for } i = 1, \dots, N \text{ and } m = 1, \dots, M_1. \quad (2.3)$$

Note that the intercept term $\boldsymbol{\alpha}^{(m)}$ in (2.1) can be omitted for continuous feature $\boldsymbol{x}_i^{(m)}$ for simplicity. $\mathbf{L}^{(m)}$ and $\mathbf{B}^{(m)}$ are common loading matrices of size $P \times R$, $P \times Q$, respectively. $\boldsymbol{\varepsilon}_i^{(m)}$ follows a zero-mean Gaussian distribution with a covariance matrix $\boldsymbol{\Psi}^{(m)}$ of size $P \times P$. For continuous feature $\boldsymbol{x}_i^{(m)}$, the GLM-type formulation reduces to the ordinal regression model for which the intercept term can be omitted for simplicity. For a categorical ordinal or nominal modality, we assume it contains P categorical features with C levels. For example, the categorical ordinal modality can be individual survey items in

the Epworth Sleepiness Scale (ESS) to rate daytime sleepiness, in which each feature uses a 4-point Likert scale with 1-4 corresponding to "no chance of dozing", "slight chance of dozing", "moderate chance of dozing", and "high chance of dozing", respectively, i.e., $C = 4$. Similarly, a GLM is applied to the categorial modality resulting in the model as follows:

$$log\left(\frac{\mathrm{P}\left(\boldsymbol{x}_i^{(m)} \leq c \times \mathbf{1}_P\right)}{1-\mathrm{P}\left(\boldsymbol{x}_i^{(m)} \leq c \times \mathbf{1}_P\right)}\right) = \boldsymbol{\alpha}_c^{(m)} + \mathbf{L}^{(m)}\boldsymbol{\eta}_i^{(m)} + \boldsymbol{B}^{(m)}\boldsymbol{z}_i^{(m)} + \boldsymbol{\varepsilon}_i^{(m)},$$

$$\text{for } i = 1, \dots, N \text{ and } m = M_1 + 1, \dots, M_1 + M_2. \quad (2.4)$$

For categorical <u>ordinal</u> modality, we have $\boldsymbol{\alpha}^1 \leq \boldsymbol{\alpha}^2 \leq \cdots \leq \boldsymbol{\alpha}^{C-1}$ as the constraints inherited from GLM for ordinal responses due to the nondecreasing property of a cumulative distribution function (Agresti, 2010), while the same constraints are not required for categorical <u>nominal</u> modalities. The difference between models (2.2-2.3) and classic GLMs is that the predictors $\boldsymbol{\eta}_i^{(m)}$ are unobserved latent factors and need to be estimated.

Next, we enable clustering in the SEM framework by assuming a latent variable $\boldsymbol{s}_i = \left(s_{1,i}, \dots, s_{K,i}\right)^T$ to indicate cluster membership for subject i. For example, if subject i belongs to the k-th cluster, then $s_{k,i} = 1$ and $s_{\tilde{k},i} = 0$ for $\tilde{k} \neq k$. K is the number of clusters. $\boldsymbol{s}_i$ is assumed to follow a multinomial distribution with parameters $\boldsymbol{w} = (w_1, \dots, w_K)^T$ that correspond to the probabilities of different clusters. Then, we assume the distribution of a subject's latent factors depends on its cluster membership and link the latent factor $\boldsymbol{\eta}_i^{(m)}$ with the latent variable $\boldsymbol{s}_i$ for $m = 1, \dots, M$ as follows:

$$\boldsymbol{\eta}_i^{(m)} = \mathbf{U}^{(m)}\boldsymbol{s}_i + \boldsymbol{\xi}_i^{(m)}, \quad (2.5)$$

where $\boldsymbol{\xi}_i^{(m)} \sim N\left(\mathbf{0}, \boldsymbol{\Sigma}^{(m)}\right)$ and $\mathbf{U}^{(m)}$ is a Q $\times K$ coefficient matrix, i.e., $\mathbf{U}^{(m)} = \left[\boldsymbol{\mu}^{(m,1)}, \dots, \boldsymbol{\mu}^{(m,K)}\right]^T$. Given a subject in cluster k, i.e., $s_{k,i} = 1$, the distribution of its latent

factors is a Gaussian distribution with mean $\boldsymbol{\mu}^{(m,k)}$ and covariance matrix $\boldsymbol{\Sigma}^{(m)}$, i.e., $\boldsymbol{\eta}_i^{(m)}|s_{k,i} = 1 \sim N\left(\boldsymbol{\mu}^{(m,k)}, \boldsymbol{\Sigma}^{(m)}\right)$, $m = 1, \dots, M$. In SEMs, model identifiability is a common issue that can be addressed by adding constraints on select parameters to ensure the uniqueness of the model estimate. Specifically, in this study, we add the constraints to the mean and covariance of the latent factors $\boldsymbol{\eta}_i^{(m)}$ for $m = 1, \dots, M$ to ensure that $E\left(\boldsymbol{\eta}_i^{(m)}\right) = \mathbf{0}$ and $Var\left(\boldsymbol{\eta}_i^{(m)}\right) = \mathbf{I}$. The mean constraint can be easily satisfied by standardizing $\boldsymbol{x}_i^{(m)}$. The covariance constraint can be satisfied using a mathematical trick as follows: Assume the original covariance, $Var\left(\boldsymbol{\eta}_i^{(m)}\right)$, is not an identity matrix and let $\mathbf{L}^{(m)}$ be the original loading matrix. Using Cholesky Decomposition we can decompose $Var\left(\boldsymbol{\eta}_i^{(m)}\right) = \boldsymbol{\tau}_{m,i}\boldsymbol{\tau}_{m,i}^T$ and update the loading matrix by $\mathbf{L}^{(m)\prime} = \mathbf{L}^{(m)}\boldsymbol{\tau}_{m,i}$. As $Var\left(\mathbf{L}^{(m)}\boldsymbol{\eta}_i^{(m)}\right) = \mathbf{L}^{(m)}Var\left(\boldsymbol{\eta}_i^{(m)}\right)\mathbf{L}^{(m)^T} = \mathbf{L}^{(m)}\boldsymbol{\tau}_{m,i}\boldsymbol{\tau}_{m,i}^T\mathbf{L}^{(m)^T} = \mathbf{L}^{(m)\prime}\mathbf{L}^{(m)\prime T}$, the updated $\mathbf{L}^{(m)\prime}$ automatically sets the covariance of the latent factors to be an identify matrix, i.e., the covariance constraint is satisfied.

Finally, we impose sparsity on models in (2.3-2.5) to achieve a hierarchical selection of modalities and features with two group lasso-based penalties. Specifically, the cluster differentiation in each modality, e.g., modality m, is based on the difference in cluster-specific mean values, e.g., $\left\{\boldsymbol{\mu}^{(m,1)}, \dots, \boldsymbol{\mu}^{(m,K)}\right\}$. In other words, if the modality m does not contribute to clustering, it means that the mean values are invariant across all the clusters. Considering the constraint that $E\left(\boldsymbol{\eta}_i^{(m)}\right) = \sum_{k=1}^{K} w_k\boldsymbol{\mu}^{(m,k)} = \mathbf{0}_{Q\times 1}$, if $\boldsymbol{\mu}^{(m,k)}$ are invariant with respect to k, it is equivalent to that $\boldsymbol{\mu}^{(m,k)} = \mathbf{0}_{Q\times 1}$ for $k = 1, \dots, K$, i.e., $\mathbf{U}^{(m)} = \mathbf{0}_{Q\times K}$. Therefore, to enable modality selection, a group lasso penalty is applied to

the vectorized $\{\mathbf{U}^{(1)}, \dots, \mathbf{U}^{(M)}\}$ by treating parameters in each modality as a group. Moreover, for feature selection, a second group lasso penalty is employed to the coefficient matrices $\{\mathbf{L}^{(1)}, \dots, \mathbf{L}^{(M)}\}$ in the models (2.3-2.4) by treating each row of the coefficient matrices as a group. Take the continuous features in model (2.3) as an example. If the p-th row of a loading matrix $\mathbf{L}^{(m)}$ are estimated to be zeros, then the linear relationship between the p-th corresponding feature in $\boldsymbol{x}_i^{(m)}$ with all the latent factors $\boldsymbol{\eta}_i^{(m)}$ is eliminated, resulting in the removal of this feature from clustering.

The proposed M2-SEM contains many parameters to be estimated that can be put in a set $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3, \boldsymbol{\Theta}_4\}$, where $\boldsymbol{\Theta}_1 = \{\boldsymbol{\Theta}_{1m}\}_{m=1}^{M_1} = \{\mathbf{L}^{(m)}, \mathbf{B}^{(m)}, \boldsymbol{\Psi}^{(m)}\}_{m=1}^{M_1}$, $\boldsymbol{\Theta}_2 = \{\boldsymbol{\Theta}_{2m}\}_{m=M_1+1}^{M_1+M_2} = \left\{\left\{\boldsymbol{\alpha}_c^{(m)}\right\}_{c=1}^{C-1}, \mathbf{L}^{(m)}, \mathbf{B}^{(m)}, \boldsymbol{\Psi}^{(m)}\right\}_{m=M_1+1}^{M_1+M_2}$, $\boldsymbol{\Theta}_3 = \{\boldsymbol{\Theta}_{3m}\}_{m=1}^{M_1+M_2} = \left\{\left\{\boldsymbol{\mu}^{(m,k)}, \boldsymbol{\Sigma}^{(m,k)}\right\}_{k=1}^{K}\right\}_{m=1}^{M_1+M_2}$, and $\boldsymbol{\Theta}_4 = \{\boldsymbol{w}\}$. Let $\mathbf{X}_m = \left\{\boldsymbol{x}_i^{(m)}\right\}_{i=1}^{N}$, $\mathbf{Z}_m = \left\{\boldsymbol{z}_i^{(m)}\right\}_{i=1}^{N}$, $\mathbf{H}_m = \left\{\boldsymbol{\eta}_i^{(m)}\right\}_{i=1}^{N}$, and $\mathbf{s} = \{\boldsymbol{s}_i\}_{i=1}^{N}$. The complete log-likelihood function can be written as

$$l\big(f(\boldsymbol{\Theta}; \{\mathbf{X}_m, \mathbf{Z}_m, \mathbf{H}_m\}_{m=1}^{M}, \boldsymbol{s})\big) = \sum_{m=1}^{M_1} \log\big(f(\mathbf{X}_m, \mathbf{Z}_m|\mathbf{H}_m; \boldsymbol{\Theta}_1)\big) + \sum_{m=M_1+1}^{M_1+M_2} \log\big(f(\mathbf{X}_m, \mathbf{Z}_m|\mathbf{H}_m; \boldsymbol{\Theta}_2)\big) + \sum_{m=1}^{M} \log\big(f(\mathbf{H}_m|\boldsymbol{s}; \boldsymbol{\Theta}_3)\big) + \log\big(f(\boldsymbol{s}; \boldsymbol{\Theta}_4)\big). \quad (2.6)$$

After adding two group lasso penalties, the model estimation becomes an optimization problem as follows:

$$\min_{\boldsymbol{\Theta}} \tilde{l}(\boldsymbol{\Theta}) = -l\big(f(\boldsymbol{\Theta}; \{\mathbf{X}_m, \mathbf{Z}_m, \mathbf{H}_m\}_{m=1}^{M}, \boldsymbol{s})\big) + \lambda_1 \sum_{m=1}^{M} \sum_{p=1}^{P} \left\| \boldsymbol{l}_p^{(m)} \right\|_2 + \lambda_2 \sum_{m=1}^{M} \left\| \mathbf{u}^{(m)} \right\|_2 . \quad (2.7)$$

where $\boldsymbol{l}_p^{(m)}$ is the transpose of the p-th row of the loading matrix $\mathbf{L}^{(m)}$ and $\mathbf{u}^{(m)}$ is a column vector consisting of all the elements in the matrix $\mathbf{U}^{(m)}$.

2.4 Model Estimation

2.4.1 Expectation-Maximization (EM) framework

An Expectation-Maximization (EM) framework (Dempster et al., 1977) is adopted for model estimation due to the presence of unobserved latent factors, e.g., $\boldsymbol{\eta}_i^{(m)}$, and latent cluster membership, e.g., $\boldsymbol{s}_i$, which is an iterative algorithm between E-steps and M-steps. Specifically, in the j-th iteration of the EM, the E-step needs to evaluate the expectation of $\tilde{l}(\boldsymbol{\Theta})$ in (2.7) with respect to the conditional distribution of latent variables $\{\mathbf{H}_m\}_{m=1}^M$ and $\mathbf{s}$ given observed data $\{\mathbf{X}_m\}_{m=1}^M$ and $\boldsymbol{\Theta}^{(j-1)}$, i.e.,

$$Q\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(j-1)}\right) \triangleq E_{\{\mathbf{H}_m\}_{m=1}^M,\boldsymbol{s}|\{\mathbf{X}_m\}_{m=1}^M;\,\boldsymbol{\Theta}^{(j-1)}}\left\{\tilde{l}(\boldsymbol{\Theta};\,\{\mathbf{X}_m,\mathbf{H}_m\}_{m=1}^M,\boldsymbol{s})\right\}. \tag{2.8}$$

Following Bayes' Theorem, (2.8) can be rewritten as the sum of four separate expectation terms as follows:

$$Q\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(j-1)}\right) =$$

$$E_{\{\mathbf{H}_m\}_{m=1}^M,\boldsymbol{s}|\{\mathbf{X}_m\}_{m=1}^M;\,\boldsymbol{\Theta}^{(j-1)}}\left\{\begin{array}{c} -\sum_{m=1}^{M_1}\log\left(f(\mathbf{X}_m|\mathbf{H}_m;\boldsymbol{\Theta}_{1m})\right)+\lambda_1\sum_{m=1}^{M_1}\sum_{p=1}^{P}\left\|\boldsymbol{l}_p^{(m)}\right\|_2 \\ -\sum_{m=M_1+1}^{M_1+M_2}\log\left(f(\mathbf{X}_m|\mathbf{H}_m;\boldsymbol{\Theta}_{2m})\right)+\lambda_1\sum_{m=M_1+1}^{M_1+M_2}\sum_{p=1}^{P}\left\|\boldsymbol{l}_p^{(m)}\right\|_2 \\ -\sum_{m=1}^{M}\log\left(f(\mathbf{H}_m|\boldsymbol{s};\boldsymbol{\Theta}_{3m})\right)+\lambda_2\sum_{m=1}^{M}\left\|\mathbf{u}^{(m)}\right\|_2 \\ -\log\left(f(\boldsymbol{s};\boldsymbol{\Theta}_4)\right) \end{array}\right\}. \tag{2.9}$$

The M-step aims to maximize the expectation in (2.9) that can be further separated into four convex functions in which each depends on a separate subset of parameters in $\boldsymbol{\Theta}$:

$$\begin{aligned} Q\left(\boldsymbol{\Theta};\boldsymbol{\Theta}^{(j-1)}\right) = {} & \varphi_1\left(\left\{\mathbf{L}^{(m)},\mathbf{B}^{(m)},\boldsymbol{\Psi}^{(m)}\right\}_{m=1}^{M_1}\right) \\ & +\varphi_2\left(\left\{\left\{\boldsymbol{\alpha}_c^{(m)}\right\}_{c=1}^{C-1},\mathbf{L}^{(m)},\mathbf{B}^{(m)},\boldsymbol{\Psi}^{(m)}\right\}_{m=M_1+1}^{M_1+M_2}\right) \\ & +\varphi_3\left(\left\{\left\{\boldsymbol{\mu}^{(m,k)},\boldsymbol{\Sigma}^{(m,k)}\right\}_{k=1}^{K}\right\}_{m=1}^{M}\right)+\varphi_4(\boldsymbol{w}), \end{aligned} \tag{2.10}$$

where $\varphi_1(\cdot)$, $\varphi_2(\cdot)$, $\varphi_3(\cdot)$, and $\varphi_4(\cdot)$ can be minimized as four separate optimization problems.

However, the traditional EM algorithm does not suffice in estimating the proposed M2-SEM with significant challenges in conducting both the E-step and M-Step. In the E-step, the expectation in $\varphi_2(\cdot)$ is intractable and needs to be first approximated with the numerical optimization method. This is because the likelihood function $f(\mathbf{X}_m|\mathbf{H}_m;\boldsymbol{\Theta})$ follows the multinominal distributions for categorical modalities $\mathbf{X}_m$, $m = M_1 + 1, \ldots, M_1 + M_2$, and thus its conditional expectation cannot be explicitly derived as the likelihood function with Gaussian assumption in $\varphi_1(\cdot)$. To provide a traceable solution, this chapter proposes to use the GH quadrature (Ehrich S, 2002) to approximate the implicit integral in (2.10). Compared with other methods such as Monte Carlo EM, GH is more computationally efficient, especially for high-dimensional data. In the M-step, the optimization problem in (2.10) is non-smooth and conventional optimization methods are computationally costly particularly in handing large-scale data. The conventional solvers such as Newton-type algorithm (Dennis et al., 1983; Schnabel et al., 1985), Nesterov's method (Liu et al., 2009), and Block Coordinate Descent algorithm (Hildreth 1957; Warga 1963) are time-consuming. To provide an efficient estimation approach, this chapter proposes to use the Majorization Maximization (MM) algorithm (Heiser, 1995; Sun et al. 2016) known to be 5-10 times faster empirically.

2.4.2 E-step Enabled by Gauss-Hermite Approximation

In the *j*th iteration, the E-step derives the four expectation terms in (2.10), each of which depends on a subset of parameters in $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3, \boldsymbol{\Theta}_4\}$. After dropping terms not consisting of parameters to be estimated, the four terms can be written as

$$\varphi_1(\boldsymbol{\Theta}_1) = \sum_{m=1}^{M_1} \varphi_{1m}\left(\mathbf{L}^{(m)}, \mathbf{B}^{(m)}, \boldsymbol{\Psi}^{(m)}\right); \tag{2.11}$$

$$\varphi_2(\boldsymbol{\Theta}_2) = \sum_{m=M_1+1}^{M_1+M_2} \varphi_{2m}\left(\left\{\boldsymbol{\alpha}_c^{(m)}\right\}_{c=1}^{C-1}, \mathbf{L}^{(m)}, \mathbf{B}^{(m)}, \boldsymbol{\Psi}^{(m)}\right); \tag{2.12}$$

$$\varphi_3(\boldsymbol{\Theta}_3) = \sum_{m=1}^{M} \varphi_{3m}\left(\left\{\boldsymbol{\mu}^{(m,k)}, \boldsymbol{\Sigma}^{(m,k)}\right\}_{k=1}^{K}\right); \tag{2.13}$$

$$\varphi_4(\boldsymbol{\Theta}_4) = -\sum_{i=1}^{N}\sum_{k=1}^{K} \log(w_k)\, \tilde{f}\left(s_{i,k}=1 \middle| \boldsymbol{x}_i^{(1)}, \cdots \boldsymbol{x}_i^{(M)};\, \widehat{\boldsymbol{\Theta}}^{(j-1)}\right). \tag{2.14}$$

We have

$$\varphi_{1m}\left(\mathbf{L}^{(m)}, \mathbf{B}^{(m)}, \boldsymbol{\Psi}^{(m)}\right) =$$

$$-\sum_{i=1}^{N}\left\{\begin{array}{c} \frac{1}{2}\log\left|\boldsymbol{\Psi}^{(m)}\right| \\ +\frac{1}{2}\left(\boldsymbol{x}_i^{(m)} - L^{(m)}E\left(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right) - B^{(m)}\boldsymbol{z}_i^{(m)}\right)^T \boldsymbol{\Psi}^{(m)^{-1}}\left(\boldsymbol{x}_i^{(m)} - L^{(m)}E\left(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right) - B^{(m)}\boldsymbol{z}_i^{(m)}\right) \\ +\frac{1}{2}tr\left(L^{(m)^T}\boldsymbol{\Psi}^{(m)^{-1}}L^{(m)}\left(E\left(\boldsymbol{\eta}_i^{(m)}\boldsymbol{\eta}_i^{(m)^T}|\boldsymbol{x}_i^{(m)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right) - E\left(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)E\left(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)^T\right)\right) \end{array}\right\} +$$

$$\lambda_1 \sum_{p=1}^{P}\left\|\boldsymbol{l}_p^{(m)}\right\|_2; \tag{2.15}$$

$$\varphi_{2m}\left(\left\{\boldsymbol{\alpha}_c^{(m)}\right\}_{c=1}^{C-1}, \mathbf{L}^{(m)}, \mathbf{B}^{(m)}, \boldsymbol{\Psi}^{(m)}\right) =$$

$$-\sum_{i=1}^{N}\sum_{k=1}^{K} f\left(s_{ik}=1|\boldsymbol{x}_i^{(m)};\,\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)\sum_{p=1}^{P}\int\left(log\left(f\left(\boldsymbol{x}_{ip}^{(m)}|\boldsymbol{\eta}_i^{(m)};\,\boldsymbol{\Theta}_{2m}\right)\right)f\left(\boldsymbol{\eta}_i^{(m)}|s_{ik}=1;\,\widehat{\boldsymbol{\Theta}}^{(j)}\right)\right)d\boldsymbol{\eta}_i^{(m)} + \lambda_1\sum_{p=1}^{P}\left\|\boldsymbol{l}_p^{(m)}\right\|_2; \tag{2.16}$$

$$\varphi_{3m}\left(\left\{\boldsymbol{\mu}^{(m,k)}, \boldsymbol{\Sigma}^{(m,k)}\right\}_{k=1}^{K}\right) =$$

$$-\sum_{i=1}^{N}\sum_{m=1}^{M}\sum_{k=1}^{K}\left\{\begin{array}{c} \frac{1}{2}\log\left|\boldsymbol{\Sigma}^{(m,k)}\right| \\ +\frac{1}{2}\left(\widetilde{\boldsymbol{\rho}}_{m,k}\left(\boldsymbol{x}_i^{(m)}\right) - \boldsymbol{\mu}^{(m,k)}\right)^T \boldsymbol{\Sigma}^{(m,k)^{-1}}\left(\widetilde{\boldsymbol{\rho}}_{m,k}\left(\boldsymbol{x}_i^{(m)}\right) - \boldsymbol{\mu}^{(m,k)}\right) \\ +\frac{1}{2}tr\left(\boldsymbol{\Sigma}^{(m,k)^{-1}}\left(\widetilde{\boldsymbol{\Upsilon}}_{m.k} + \widetilde{\boldsymbol{\rho}}_{m,k}\left(\boldsymbol{x}_i^{(m)}\right)^T\widetilde{\boldsymbol{\rho}}_{m,k}\left(\boldsymbol{x}_i^{(m)}\right)\right)\right) \end{array}\right\}\tilde{f}\left(s_{i,k}=1\middle|\boldsymbol{x}_i^{(1)}, \cdots \boldsymbol{x}_i^{(M)};\,\widehat{\boldsymbol{\Theta}}^{(j-1)}\right) +$$

$$\lambda_2 \sum_{m=1}^{M}\left\|\mathbf{u}^{(m)}\right\|_2. \tag{2.17}$$

where $\widetilde{\boldsymbol{\Upsilon}}_{m.k} = \left(\tilde{L}^{(m)^T}\widetilde{\boldsymbol{\Psi}}^{(m)^{-1}}\tilde{L}^{(m)} + \widetilde{\boldsymbol{\Sigma}}^{(m,k)^{-1}}\right)^{-1}$ and $\widetilde{\boldsymbol{\rho}}_{m,k}\left(\boldsymbol{x}_i^{(m)}\right) = \left(\tilde{L}^{(m)^T}\widetilde{\boldsymbol{\Psi}}^{(m)^{-1}}\tilde{L}^{(m)} + \widetilde{\boldsymbol{\Sigma}}^{(m,k)^{-1}}\right)^{-1}\left(\tilde{L}^{(m)^T}\widetilde{\boldsymbol{\Psi}}^{(m)^{-1}}\boldsymbol{x}_i^{(m)} + \widetilde{\boldsymbol{\Sigma}}^{(m,k)^{-1}}\boldsymbol{\mu}^{(m,k)}\right)$. While the derivation of $\varphi_1(\cdot)$, $\varphi_3(\cdot)$, and $\varphi_4(\cdot)$ is relatively straightforward and more details can be found in Appendix A, $\varphi_2(\cdot)$ cannot be analytically derived because it contains a few inexplicit terms, i.e., $f\left(s_{ik}=1|\boldsymbol{x}_i^{(m)};\,\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)$ and $\int\left(log\left(f\left(\boldsymbol{x}_{ip}^{(m)}|\boldsymbol{\eta}_i^{(m)};\,\boldsymbol{\Theta}_{2m}\right)\right)f\left(\boldsymbol{\eta}_i^{(m)}|s_{ik}=\right.\right.$

1; $\widehat{\boldsymbol{\Theta}}^{(j-1)}\Big)\Big) d\boldsymbol{\eta}_i^{(m)}$ in (2.16) that do not have closed-forms. Fortunately, we have derived Propositions 2.1-2.3 below to approximate the inexplicit terms with the GH method to provide a closed-form formulation of (2.12) in the following two steps. First, we introduce the GH approximation in Definition 2.1 followed by a discussion of the approximation error in Proposition 2.1. GH quadrature is a widely used tool in numerical optimization to approximate the integral in the form of $\int_{-\infty}^{\infty} \exp\{-z^2\} g(z) dz$, in which $g(z)$ is a function of z and is infinitely differentiable w.r.t z (Sauer and Xu, 1995; Davis and Rabinowitz, 2007). The univariate GH quadrature can be generated to approximate the multivariate integrals, i.e., $\int_{\mathbb{R}^Q} exp\{-\boldsymbol{z}^T\boldsymbol{z}\} g(\boldsymbol{z}) d\boldsymbol{z}$, where $\boldsymbol{z}$ is a Q-dimensional vector and g is a function of $\boldsymbol{z}$, as shown in Definition 2.1. We also have proved the GH approximation error will reduce to zero with sufficient number of notes in Proposition 2.1. Second, we employ the multivariate GH quadrature to approximate the two non-analytical terms in (2.12). Since this approximation is not trivial, we derived Propositions 2.2 and 2.3 to describe how each of the terms is approximated by GH in detail.

Definition 2.1 (GH approximation). Given vector $\boldsymbol{z}$ with $\text{rank}(\boldsymbol{z}) = Q$, and function $g \in \mathbb{C}^{2T}: \mathbb{R}^Q \to \mathbb{R}$ by applying Hermite interpolation , we have

$$\int_{\boldsymbol{S}} exp\{-\boldsymbol{z}^T\boldsymbol{z}\} g(\boldsymbol{z}) d\boldsymbol{z} \approx \sum_{t_1=1}^{T} \cdots \sum_{t_Q=1}^{T} \omega_{t_1} \cdots \omega_{t_Q} g(\boldsymbol{z}_t) \tag{2.18}$$

where $\boldsymbol{S} \in \mathbb{R}^Q$ is the intergration set, $\boldsymbol{z}_t = (z_{t_1}, \cdots, z_{t_Q})^T$ and z_{t_q} are the roots of Hermite polynomial of order T, $H_T(x) = (-1)^T e^{x^2} \frac{d^T}{dx^T} e^{-x^2}$ for $t_q \in \{1, \cdots, T\}$ and $q \in \{1, \cdots, Q\}$. The weight, ω_{t_q}, is given by $\omega_{t_q} = \frac{2^{T+1} T! \sqrt{\pi}}{[H_{T+1}(z_{t_q})]^2}$.

Proposition 2.1 (The GH approximate error is bounded). If the integration set $\boldsymbol{S}$ in Definition 2.1 is closed, the GH approximation error, determined by $\frac{T!\sqrt{\pi}}{2^T(2T)!}g^{(2T)}(\xi)$, reduces to zero for a sufficiently large *T*.

(*Due to the limitation of space, the detailed proof is listed in Appendix B.*)

Proposition 2.2. The non-analytical term $f\left(s_{ik}=1|\boldsymbol{x}_i^{(m)};\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)$ in (2.16) can be analytically approximated by GH Quadrature, i.e.,

$$\tilde{f}\left(s_{ik}=1|\boldsymbol{x}_i^{(m)};\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)\triangleq\tilde{f}\left(s_{ik}=1|\boldsymbol{x}_i^{(m)};\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)=\frac{\tilde{f}\left(\boldsymbol{x}_i^{(m)}|s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)f\left(s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)}{\sum_{k=1}^{K}\tilde{f}\left(\boldsymbol{x}_i^{(m)}|s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)f\left(s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)},\tag{2.19}$$

where $\tilde{f}\left(\boldsymbol{x}_i^{(m)}|s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)=(\pi)^{-Q/2}\sum_{t_1=1}^{T}\dots\sum_{t_Q=1}^{T}w_{t_1}\dots w_{t_Q}f\left(\boldsymbol{x}_i^{(m)}|\sqrt{2}\boldsymbol{\Sigma}^{(m,k)^{\frac{1}{2}}}\widetilde{\boldsymbol{\eta}}_{i,t}^{(m)}+\boldsymbol{\mu}^{(m,K)};\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)$

with $\left\{w_{t_1},\dots,w_{t_Q}\right\}$ and$\boldsymbol{\eta}_i^{(m)}=\sqrt{2}\boldsymbol{\Sigma}^{(m,k)^{\frac{1}{2}}}\widetilde{\boldsymbol{\eta}}_i^{(m)}+\boldsymbol{\mu}^{(m,K)}$, $\widetilde{\boldsymbol{\eta}}_{i,t}^{(m)}=(\widetilde{\eta}_{i,t_1}^{(m)},\cdots,\widetilde{\eta}_{i,t_Q}^{(m)})^T$ representing the weights and polynomial roots of GH approximation in Definition 2.1.

(*Due to the limitation of space, the detailed proof is listed in Appendix C.*)

Proposition 2.3. The non-analytical term $\int_{\boldsymbol{\eta}}^{\square}\left(log\left(f\left(\boldsymbol{x}_{ip}^{(m)}|\boldsymbol{\eta};\ \boldsymbol{\Theta}_{2m}\right)\right)f\left(\boldsymbol{\eta}|s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)\right)d\boldsymbol{\eta}$ in (2.16) can be analytically approximated by GH Quadrature, i.e.,

$$\begin{aligned}&\int_{\boldsymbol{\eta}}^{\square}\left(log\left(f\left(\boldsymbol{x}_{ip}^{(m)}|\boldsymbol{\eta}_i^{(m)};\ \boldsymbol{\Theta}_{2m}\right)\right)f\left(\boldsymbol{\eta}_i^{(m)}|s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)\right)d\boldsymbol{\eta}\\&\approx(\pi)^{-Q/2}\sum_{t_1=1}^{T}\cdots\sum_{t_Q=1}^{T}\omega_{t_1}\cdots\omega_{t_Q}log\left(f\left(\boldsymbol{x}_{ip}^{(m)}|\sqrt{2}\boldsymbol{\Sigma}^{(m,k)^{\frac{1}{2}}}\widetilde{\boldsymbol{\eta}}_{i,t}^{(m)}+\boldsymbol{\mu}^{(m,K)};\ \boldsymbol{\Theta}_{2m}\right)\right)\end{aligned}\tag{2.20}$$

where $\boldsymbol{\eta}_i^{(m)} = \sqrt{2}\boldsymbol{\Sigma}^{(m,k)\frac{1}{2}}\tilde{\boldsymbol{\eta}}_i^{(m)} + \boldsymbol{\mu}^{(m,K)}$, and $\tilde{\boldsymbol{\eta}}_{i,t}^{(m)} = \left(\tilde{\eta}_{i,t_1}^{(m)}, \cdots, \tilde{\eta}_{i,t_Q}^{(m)}\right)$ are the roots of the Hermite polynomial of order T, T is the number of quadrature points of $\boldsymbol{\eta}_{i,t_q}^{(m)}$, and the weights are given by $\omega_{t_q} = \frac{2^{T+1}\mathrm{T}!\sqrt{\pi}}{[H_{T+1}(\tilde{\eta}_{i,t_q}^{(m)})]^2}$.

(*Due to the limitation of space, the detailed proof is listed in Appendix D.*)

Based on Propositions 2.2 and 2.3, the implicit terms in (2.16) can be replaced by their GH Quadrature and the objective function in (2.12) can be rewritten as

$$\varphi_2(\boldsymbol{\Theta}_2) \approx \sum_{m=M_1+1}^{M_1+M_2} \tilde{\varphi}_{2m}\left(\left\{\boldsymbol{\alpha}_c^{(m)}\right\}_{c=1}^{C-1}, \mathbf{L}^{(m)}, \mathbf{B}^{(m)}, \boldsymbol{\Psi}^{(m)}\right), \tag{2.21}$$

where

$$\tilde{\varphi}_{2m}\left(\left\{\boldsymbol{\alpha}_c^{(m)}\right\}_{c=1}^{C-1}, \mathbf{L}^{(m)}, \mathbf{B}^{(m)}, \boldsymbol{\Psi}^{(m)}\right) =$$

$$\sum_{p=1}^{P}\left\{-\sum_{i=1}^{N}\sum_{k=1}^{K}\tilde{f}\left(s_{ik}=1|\boldsymbol{x}_i^{(m)};\hat{\boldsymbol{\Theta}}^{(j-1)}\right)(\pi)^{-\frac{Q}{2}}\sum_{t_1=1}^{T}\cdots\sum_{t_Q=1}^{T}\omega_{t_1}\cdots\omega_{t_Q}log\left(f\left(\boldsymbol{x}_{ip}^{(m)}|\sqrt{2}\boldsymbol{\Sigma}^{(m,k)\frac{1}{2}}\tilde{\boldsymbol{\eta}}_{i,t}^{(m)}+\boldsymbol{\mu}^{(m,K)};\boldsymbol{\Theta}_{2m}\right)\right)\right\}. \tag{2.22}$$
$$+\lambda_2\left\|\boldsymbol{\iota}_p^{(m)}\right\|_2$$

2.4.3 M-step Integrated with Majorization-Minimization Algorithm

In the *j* th iteration, the subsequent M-step minimizes the expected objective function derived in the E-step by solving four separate optimization problems each of which depends on a subset of parameters in $\boldsymbol{\Theta} = \{\boldsymbol{\Theta}_1, \boldsymbol{\Theta}_2, \boldsymbol{\Theta}_3, \boldsymbol{\Theta}_4\}$. After dropping terms not consisting of parameters to be estimated, the four optimization problems can be written as

$$\left\{\widehat{\boldsymbol{\Theta}}_{1m}{}^{(j)} = \left\{\hat{\mathbf{L}}^{(m)(j)}, \widehat{\mathbf{B}}^{(m)(j)}, \widehat{\boldsymbol{\Psi}}^{(m)(j)}\right\}\right\}_{m=1}^{M_1} = \underset{\{\boldsymbol{\Theta}_{1m}\}_{m=1}^{M_1}}{\operatorname{argmin}}\ \varphi_{1m}\left(\mathbf{L}^{(m)}, \mathbf{B}^{(m)}, \boldsymbol{\Psi}^{(m)}\right); \tag{2.23}$$

$$\left\{\widehat{\boldsymbol{\Theta}}_{2m}{}^{(j)} = \left\{\left\{\widehat{\boldsymbol{\alpha}}_c^{(m)}\right\}_{c=1}^{C-1}, \hat{\mathbf{L}}^{(m)(j)}, \widehat{\mathbf{B}}^{(m)(j)}, \widehat{\boldsymbol{\Psi}}^{(m)(j)}\right\}\right\}_{m=M_1+1}^{M_1+M_2} = \underset{\boldsymbol{\Theta}_{2m}}{\operatorname{argmin}}\sum_{m=M_1+1}^{M_1+M_2}\tilde{\varphi}_{2m}\left(\left\{\widehat{\boldsymbol{\alpha}}_c^{(m)}\right\}_{c=1}^{C-1}, \mathbf{L}^{(m)}, \mathbf{B}^{(m)}, \boldsymbol{\Psi}^{(m)}\right); \tag{2.24}$$

$$\left\{\widehat{\boldsymbol{\Theta}}_{3m}{}^{(j)} = \left\{\widehat{\boldsymbol{\mu}}^{(m,k)(j)}, \widehat{\boldsymbol{\Sigma}}^{(m,k)(j)}\right\}_{k=1}^{K}\right\}_{m=1}^{M} = \underset{\boldsymbol{\Theta}_3}{\operatorname{argmin}} \sum_{m=1}^{M} \varphi_{3m}\left(\left\{\boldsymbol{\mu}^{(m,k)}, \boldsymbol{\Sigma}^{(m,k)}\right\}_{k=1}^{K}\right); \tag{2.25}$$

$$\widehat{\boldsymbol{w}}^{(j)} = \underset{\boldsymbol{\Theta}_4}{\operatorname{argmin}}\, \tilde{\varphi}_4(\boldsymbol{w}). \tag{2.26}$$

The term in (2.26) can be analytically optimized while the three terms in (2.23-2.25) are non-smooth and require to be optimized by iterative algorithms. All the three objectives in (2.23-2.25) are shown to be jointly convex and thus can be minimized by the Block Coordinate Descent (BCD) algorithm. BCD sequentially minimizes the objective function in each block coordinate while the other coordinates are held fixed. Hereinafter, we take the term in (2.24) as an example to describe how each BCD iteration is conducted, and similar procedures are applicable to terms in (2.23) and (2.25).

The objective function in (2.24) is jointly convex with respect to its parameters, i.e., $\boldsymbol{\Theta}_2 = \{\boldsymbol{\Theta}_{2m}\}_{m=M_1+1}^{M_1+M_2} = \left\{\left\{\boldsymbol{\alpha}_c^{(m)}\right\}_{c=1}^{C-1}, \mathbf{L}^{(m)}, \mathbf{B}^{(m)}, \boldsymbol{\Psi}^{(m)}\right\}_{m=M_1+1}^{M_1+M_2}$, and thus can be optimized by the BCD algorithm. Since (2.24) has a constraint that $\boldsymbol{\alpha}_1^{(m)} \leq \boldsymbol{\alpha}_2^{(m)} \leq \cdots \leq \boldsymbol{\alpha}_{C-1}^{(m)}$ for $m \in \{M_1+1, \cdots, M_1+M_2\}$, to get rid of its influence, we modify the objective function of (2.24) as

$$\tilde{\varphi}_{2m}\left(\{\boldsymbol{\alpha}_c^{(m)}\}_{c=1}^{C-1}, \mathbf{L}^{(m)}, \mathbf{B}^{(m)}, \boldsymbol{\Psi}^{(m)}\right) =$$

$$\sum_{p=1}^{P}\left\{-\sum_{i=1}^{N}\sum_{k=1}^{K}\tilde{f}\left(s_{ik}=1|\boldsymbol{x}_i^{(m)};\, \widehat{\boldsymbol{\Theta}}^{(j-1)}\right)(\pi)^{-\frac{Q}{2}}\sum_{t_1=1}^{T}\cdots\sum_{t_Q=1}^{T}\omega_{t_1}\cdots\omega_{t_Q} log\left(f\left(\boldsymbol{x}_{ip}^{(m)}|\sqrt{2}\boldsymbol{\Sigma}^{(m,k)\frac{1}{2}}\widetilde{\boldsymbol{\eta}}_{i,t}^{(m)} + \boldsymbol{\mu}^{(m,K)};\, \boldsymbol{\Theta}_{2m}\right)\right)\right\} +$$

$$+\lambda_2\left\|\boldsymbol{l}_p^{(m)}\right\|_2$$

$$\mathbb{I}\left(\{\boldsymbol{\alpha}_c^{(m)}\}_{c=1}^{C-1}\right), \text{where } \mathbb{I}\left(\left\{\boldsymbol{\alpha}_c^{(m)}\right\}_{c=1}^{C-1}\right) = \begin{cases} 0, & \text{if } \boldsymbol{\alpha}_1^{(m)} \leq \boldsymbol{\alpha}_2^{(m)} \leq \cdots \leq \boldsymbol{\alpha}_{C-1}^{(m)} \\ \infty, & \text{else} \end{cases}.$$

The new objective function represents an extended-value extension of the original objective function, as described in Chapter 3.1.2 of Boyd et al., 2004. This approach allows us to relax the convex constraint while ensuring that the new objective function remains convex. The most challenging coordinate block to optimize is the one involving non-smooth penalty terms, i.e., features' loading matrix on latent factors $\mathbf{L}^{(m)}$. Note that

although covariates' coefficient matrix $\mathbf{B}^{(m)}$ is not include in the group lasso penalty, it can be assigned into the same block as $\mathbf{L}^{(m)}$ for simplicity. Instead of traditional solvers of non-smooth convex optimization, this study proposes to adopt an efficient Majorization Maximization (MM) algorithm (Hunter & Lange, 2004; Mairal, 2015) to optimize the non-smooth convex objective function in (2.24) with respect to $\left\{\mathbf{L}^{(m)}, \mathbf{B}^{(m)}\right\}_{m=M_1+1}^{M_1+M_2}$. The principle of MM is to successively find and minimize an upper bound of the complex objective function $\varphi(\cdot)$ as a majorizing surrogate, in which each upper bound is simple, locally tight and each minimization step of the upper bound can result in decrease of the objective function's value. Depending on the property of the objective function, different surrogates can be adopted. Next, we present Theorem 2.1 (Hunter & Lange, 2004; Mairal, 2015) and briefly discuss the convergence of the MM algorithm in estimating M2-SEM in this chapter. Based on Theorem 2.1, the objective function in (2.24) is differentiable and convex with respect to $\left\{\mathbf{L}^{(m)}, \mathbf{B}^{(m)}\right\}_{m=M_1+1}^{M_1+M_2}$ and its first-order derivative is Lipschitz continuous, the first-order surrogate function is sufficient to "majorize" the objective function in (2.24). Moreover, all the three non-smooth objectives in (2.23-2.25) are proven to satisfy the assumptions on differentiability, convexity, and Lipschitz continuity and therefore can be solved effectively by the MM algorithm with the first-order surrogate model. More details can be found in Proposition 2.4 below.

Theorem 2.1 (Convergence of the MM algorithm) (Hunter & Lange, 2004; Mairal 2015). Assume the following optimization problem where $\boldsymbol{l}$ denotes the parameters to be estimated and $\mathbf{D}$ denotes the data:

$$\min_{\boldsymbol{l}} \varphi(\boldsymbol{l}|\mathbf{D}) + \lambda\|\boldsymbol{l}\|_2$$

If the objective function $\varphi(\boldsymbol{l}|\mathbf{D})$ is differentiable and convex with respect to $\boldsymbol{l}$ and its first-order derivative $\nabla\varphi(\boldsymbol{l}|\mathbf{D})$ is Lipschitz continuous, the MM algorithm with the first-order surrogate function is guaranteed to achieve the Karush–Kuhn–Tucker (KKT) conditions upon convergence.

(*Due to the limitation of space, the detailed proof is listed in Appendix E.*)

Proposition 2.4. The smooth objective functions in (2.23-2.25) are jointly convex and have Lipschitz continuous gradients with respect to $\{\mathbf{L}^{(m)}, \mathbf{B}^{(m)}\}$ for $m = 1, \ldots, M_1$, $\{\mathbf{L}^{(m)}, \mathbf{B}^{(m)}\}$ for $m = M_1 + 1, \ldots, M_1 + M_2$, and $\{\boldsymbol{\mu}^{(m,k)}\}_{k=1}^{K}$ for $m = 1, \ldots, M_1 + M_2$, respectively.

(*Due to the limitation of space, the detailed proof is listed in Appendix F.*)

2.4.4 Gauss-Hermite Expectation-Majorization-Minimization (GH-EMM) algorithm

This chapter develops a novel GH-EMM algorithm for efficient estimation of the proposed M2-SEM. The GH-EMM iterates over E-step enabled by Gauss-Hermite Approximation to accommodate mixed-type data modalities and M-step integrated with the efficient MM algorithm as described in Chapter 2.4.2 and Chapter 2.4.3, respectively. This algorithm is summarized in Table 1 below. The performance of GH-EMM will be tested in simulation studies and then used in a real-world application in Chapters 2.5 and 2.6, respectively.

Table 1: Major steps of the proposed GH-EMM algorithm for parameter estimation in M2-SEM

At j-th iteration of GH-EMM: given $\{\mathbf{X}_m, \mathbf{Z}_m\}_{m=1}^{M}$ and $\boldsymbol{\Theta}^{(j-1)}$
E-step (Chapter 2.4.2): Derive the expectation term $Q\left(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(j-1)}\right)$ in (2.10) • Derive the explicit terms $\varphi_1(\boldsymbol{\Theta}_1)$, $\varphi_3(\boldsymbol{\Theta}_3)$, and $\varphi_4(\boldsymbol{\Theta}_4)$ in (2.11), (2.13), and (2.14) • Derive the non-explicit term $\varphi_2(\boldsymbol{\Theta}_2)$ in (2.12) with GH approximation (see **Props 2.1-2.3**)
M-step (Chapter 2.4.3): Update parameters $\boldsymbol{\Theta}^{(j)}$ by maximizing the derived expectation term $Q\left(\boldsymbol{\Theta}; \boldsymbol{\Theta}^{(j-1)}\right)$ • Update $\widehat{\boldsymbol{\Theta}}_1^{(j)}$, $\widehat{\boldsymbol{\Theta}}_2^{(j)}$, and $\widehat{\boldsymbol{\Theta}}_3^{(j)}$ by solving optimization problems in (2.23-2.25) using MM (see **Theorem 2.1** & **Prop 2.4**) • Update $\widehat{\boldsymbol{\Theta}}_4^{(j)}$ by solving the optimization problem in (2.26) resulting in an analytical estimate
Go to the next iteration until it reaches convergence

The GH-EMM algorithm includes four hyperparameters that need to be tuned for the optimal model selection including two regularized parameters λ_1 and λ_2, number of clusters K, and number of latent factors Q. The classic BIC criterion is adopted for model selection to balance the model fitness and complexity, and the model with lowest BIC is preferred. The BIC is formally defined as $BIC = -2\log(L) + log(N) \times v$, where L is the maximized value of the likelihood function given the estimated parameters, N is the sample size, and v is the number of free parameters to be estimated by the model. This chapter adopts a 3-D grid search strategy for λ_1, λ_2, and the other two integer

hyperparameters. For efficient computing, parallel resources are used to perform the exhaustive search of the 3-D grid (Intel® Core™ i9-13900K Processor and 64 Gb memory).

2.5 Simulation

The application of the proposed method will be introduced in Chapter 2.6. Chapter 2.5 uses the simulation data to validate the performance of the proposed method. Chapter 2.5.1 introduces the simulation setup of five different experiments, i.e., Experiments 1-5, while Chapter 2.5.2 presents the clustering accuracy, and modality and feature selection accuracy of the proposed model in comparison with several benchmark advanced model-based clustering algorithms including Rmixmod (Lebret et al., 2015), clustMD (McParland and Gormley, 2016), and VarSelLCM (Marbac et al., 2020). The details of the benchmarks can be found from the literature review in Chapter 2.2.

2.5.1 Simulation Setup

The simulation study includes five different experiments to test the model's performance in different settings. Experiment 1 produces a dataset with a similar number of subjects and features to the real-world dataset to be discussed in Chapter 2.6. We first introduce the simulation setup of Experiment 1 in detail followed by a brief discussion on the simulation setup of Experiments 2-5.

Experiment 1 includes 400 subjects from three clusters with relative percentages of 30%, 40%, and 30% by assuming $w_1 = 0.3$, $w_2 = 0.4$, and $w_3 = 0.3$. Each modality includes 20 features with only 25% informative features, i.e., associated with non-zero loadings on latent factors. To demonstrate M2-SEM's performance of feature selection, sparsity is induced into the loading coefficients. Each modality has 20 features, among which only the first five features have non-zero loadings that are randomly generated from

uniform distributions by setting the loading matrix as $\mathbf{H}_m = \begin{bmatrix} \left(\mathbf{H}_m^{(1)}\right)_{5\times 2} \\ \mathbf{0}_{15\times 2} \end{bmatrix}$ for $m = 1, 2, \ldots, M_1 + M_2$. Each subject has a total of eight modalities with only four modalities being informative, i.e., contributing to cluster differentiation. Each of the four informative modalities depends on two latent factors, i.e., Q = 2, with cluster-specific parameter distributions as shown in Table 2. For ease of discussion, we assume that Modalities 1-4 and Modalities 5-8 are numerical and categorical modalities, respectively, among which Modalities 1 and 2 and Modalities 5 and 6 are informative modalities. Table 2 depicts the distributions of latent factors across three clusters for both informative modalities and non-informative modalities, which demonstrates how clusters can be differentiated by latent factors among informative modalities. As shown in Table 2, it is obvious that the three clusters are non-separable in the latent factor space for non-informative modalities, i.e., $m = 3,4,6,$ and 7, while the three clusters are noticeably separable for the informative modalities, i.e., $m =1$, 2, 3, and 4.

Table 2: Cluster-specific parameters in distributions of latent factors

Modality		**Cluster 1**		**Cluster 2**		**Cluster 3**	
		$\boldsymbol{u}_{m,1}$	$\boldsymbol{\Sigma}_{m,1}$	$\boldsymbol{u}_{m,2}$	$\boldsymbol{\Sigma}_{m,2}$	$\boldsymbol{u}_{m,3}$	$\boldsymbol{\Sigma}_{m,3}$
Informative numerical modalities	m=1	$\begin{pmatrix} 0.20 \\ 1.80 \end{pmatrix}$	$\begin{pmatrix} 0.25 & -0.21 \\ -0.21 & 0.37 \end{pmatrix}$	$\begin{pmatrix} -1.50 \\ -1.50 \end{pmatrix}$	$\begin{pmatrix} 0.25 & -0.21 \\ -0.21 & 0.37 \end{pmatrix}$	$\begin{pmatrix} 1.10 \\ -1.50 \end{pmatrix}$	$\begin{pmatrix} 0.25 & -0.21 \\ -0.21 & 0.37 \end{pmatrix}$
	m=2	$\begin{pmatrix} 1.55 \\ -1.75 \end{pmatrix}$	$\begin{pmatrix} 0.23 & 0.15 \\ 0.15 & 0.3 \end{pmatrix}$	$\begin{pmatrix} -1.15 \\ 1.75 \end{pmatrix}$	$\begin{pmatrix} 0.23 & 0.15 \\ 0.15 & 0.3 \end{pmatrix}$	$\begin{pmatrix} 1.75 \\ 1.75 \end{pmatrix}$	$\begin{pmatrix} 0.23 & 0.15 \\ 0.15 & 0.30 \end{pmatrix}$
Informative categorical modalities	m=5	$\begin{pmatrix} -1.59 \\ 0.77 \end{pmatrix}$	$\begin{pmatrix} 0.17 & 0.08 \\ 0.08 & 0.14 \end{pmatrix}$	$\begin{pmatrix} 1.50 \\ 0.16 \end{pmatrix}$	$\begin{pmatrix} 0.16 & -0.08 \\ -0.08 & 0.12 \end{pmatrix}$	$\begin{pmatrix} -0.01 \\ -1.15 \end{pmatrix}$	$\begin{pmatrix} 0.10 & -0.01 \\ -0.01 & 0.09 \end{pmatrix}$
	m=6	$\begin{pmatrix} 0.96 \\ -1.35 \end{pmatrix}$	$\begin{pmatrix} 0.12 & -0.07 \\ -0.07 & 0.17 \end{pmatrix}$	$\begin{pmatrix} -1.15 \\ 1.54 \end{pmatrix}$	$\begin{pmatrix} 0.09 & -0.06 \\ -0.06 & 0.13 \end{pmatrix}$	$\begin{pmatrix} 1.11 \\ 1.37 \end{pmatrix}$	$\begin{pmatrix} 0.15 & 0.04 \\ 0.04 & 0.11 \end{pmatrix}$
Non-informative modalities	m=3,4,7,8	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

Considering Experiment 1 as the baseline setting, two scenarios are further designed to test the model's performance as the number of features or number of modalities increases. The rest of the parameter settings are similar to those in Experiment 1. In the

first scenario, Experiment 1 includes 20 features for each modality, which is increased to 40 and 80 in Experiments 2 and 3, respectively. Each modality contains only 25% informative features and 75% non-informative features. We can observe how the model's performance, especially feature selection accuracy, varies as the number of features increases, by comparing Experiments 1, 2, and 3. In the second scenario, Experiments 4 and 5 are designed by increasing the number of modalities from 8 (as set in Experiment 1) to 16 and 24 modalities, respectively, by adding noise modalities. We can observe how the model's performance, especially modality selection accuracy, varies as the number of modalities increases by comparing Experiments 1, 4, and 5.

2.5.2 Simulation results

To test the robustness of the proposed model, 20 replicates are randomly generated based on the simulation settings in Chapter 2.5.1, and the proposed M2-SM is applied to each replicate. Table 3 summarizes the basic setup of the five experiments, as well as clustering accuracy and feature and modality selection accuracy averaged over all the replicates. Specifically, selection accuracy is evaluated by both sensitivity and specificity. Sensitivity is defined as the percentage of parameters estimated to be non-zero among parameters associated with all the informative features, while specificity is defined as the percentage of parameters estimated to be zero among parameters associated with the non-informative features. Sensitivity and specificity for modality selection are defined in a similar way. As discussed in Chapter 2.3.2, if Modality m is non-informative, the distributions of its latent factors are invariant across clusters. Given the identifiability constraints, it is equivalent to that $\boldsymbol{\mu}^{(m,k)} = \mathbf{0}_{Q\times 1}$ for $k = 1, \ldots, K$, i.e., $\mathbf{U}^{(m)} = \mathbf{0}_{Q\times K}$.

Therefore, if all the elements in $\mathbf{U}^{(m)}$ are estimated to be zero, Modality m is considered being excluded from the model.

Table 3 summarizes the simulation results of two different scenarios to test the model's performance with an increased number of features per modality and an increased number of modalities, respectively. Experiment 1 achieves a satisfactory accuracy in modality selection with the sensitivity and specificity being $(98.75 \pm 5.59)\%$ and $(97.50 \pm 7.69)\%$, respectively. As the number of modalities increases from 8 to 16 and 24, we can observe the expected decrease in modality selection accuracy reflected in specificities by comparing Experiments 1, 4, and 5. It is not unexpected to observe that the specificity decreases as more and more noise modalities are added to the dataset. Note that the relatively high standard deviations in selection accuracy are most likely attributed to a limited number of replicates. Additionally, Experiment 1 achieves $(98.75 \pm 6.11)\%$ sensitivity and $(77.00 \pm 24.30)\%$ specificity in feature selection. As the number of features in each modality increases from 20 to 40 and 80, the proposed method maintains at least 98% and 77% in sensitivity and specificity, respectively, in Experiments 2 and 3.

Table 3: Simulation results of 20 replicates in Experiments 1-5

Experiment	**Settings**		**Clustering Accuracy (%)**	**Modality selection**		**Feature selection**	
	# Modalities	**# Features**		**Sensitivity (%)**	**Specificity (%)**	**Sensitivity (%)**	**Specificity (%)**
Scenario 1: Test the model with different # features							
1	8	20	97.23 ± 6.59	98.75 ± 5.59	97.50 ± 7.69	98.63 ± 6.11	77.00 ± 24.30
2	8	40	100.00 ± 0.00	100.00 ± 0.00	90.00 ± 24.87	99.06 ± 1.96	79.20 ± 21.23
3	8	80	98.20 ± 5.55	100.00 ± 0.00	83.75 ± 21.88	98.89 ± 1.84	81.56 ± 18.77
Scenario 2: Test the model with different # modalities							
1	8	20	97.23 ± 6.59	98.75 ± 5.59	97.50 ± 7.69	98.63 ± 6.11	77.00 ± 24.30
4	16	20	97.30 ± 6.61	100.00 ± 0.00	90.95 ± 19.82	98.55 ± 4.06	78.13 ± 22.72
5	24	20	97.26 ± 6.60	97.50 ± 11.18	83.50 ± 30.18	98.19 ± 5.14	75.26 ± 25.78

The proposed model achieves at least 97% clustering accuracy, which is robust to the number of features and the number of modalities across Experiments 1-5. Moreover, the proposed model is compared with a few competing methods that are advanced MBC methods in existing R packages such as Rmixmod, clustMD, and VarSelLCM. Simulated data in Experiment 1 is used for comparison. Since existing methods do not consider the multi-modal data structure, three combined datasets are created by combining all four numeric modalities, all four categorical modalities, and all eight modalities, respectively. As shown in Figure 2, among competing benchmarks, Rmixmod achieves the highest clustering accuracy across all the three combined datasets. In particular, Rmixmod's clustering accuracy is $(87.40 \pm 3.94)\%$ for the combined dataset with eight modalities over the 20 replicates, but it is still significantly less than that of the proposed method, which is $(97.23 \pm 6.59)\%$.

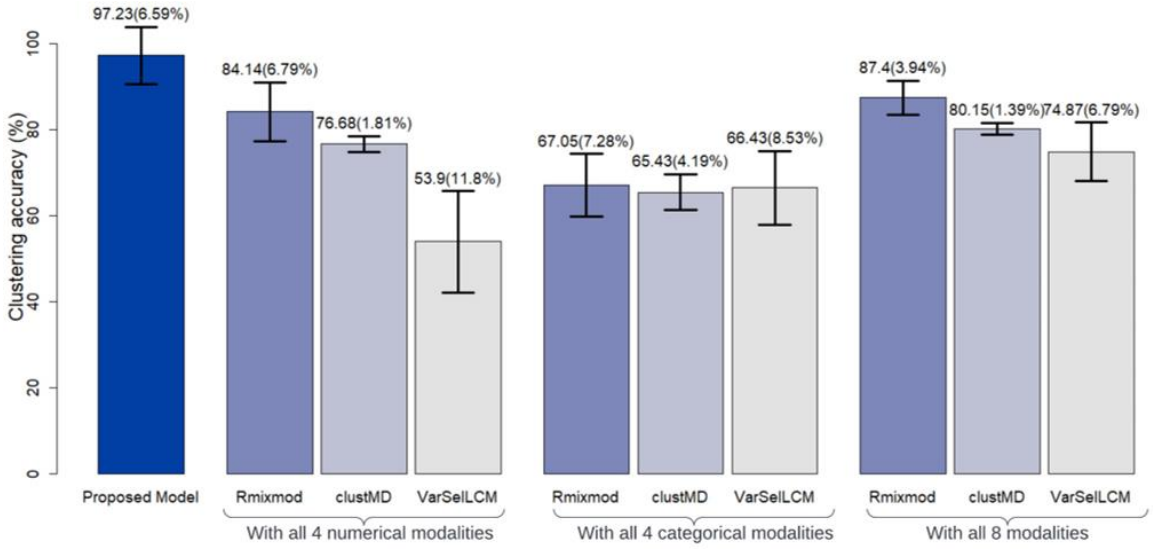

Figure 2: Comparison of clustering accuracy between the proposed method and benchmarks

2.6 Application

2.6.1 Data description and preprocessing

The data used for this application is from a large public dataset collected in the ongoing Hispanic Community Health Study (HCHS). This study has been exempt from the requirement for approval by an Institutional Review Board (IRB), because it is a secondary analysis of de-identified data with an approved data use agreement. A total of 1,052 subjects with 10 data modalities including four continuous and six categorical modalities were included in the analyses. To discover CM subgroups from HCHS data, this study includes subjects' demographics such as age and gender as well as their sleep health, dietary information, physical activities, and mental health, which are characterized by nine CM-related data modalities including medical history, lab results, sleep monitoring data, neurocognitive measures, Alternative Healthy Eating Indices (AHEIs), HCHS Acculturation questions, Center for Epidemiologic Studies Depression Scales (CES-D),

State-Trait Anxiety Inventories (STAIs), Epworth Sleepiness Scales (ESS), and Women's Health Initiative Insomnia Rating Scales (WHIIRSs).

To examine the model's performance in a real-world application, we preprocess the HCHS data. Although it is not feasible to test the clustering accuracy as the group truth of cluster membership is unknown, we plan to validate the model's performance on sparse selection. However, the challenge is that many of the CM data modalities have no sparsity within each modality by design. That is, if one modality is considered significant, all its features are significant and none of the features can be excluded from clustering. The reason is that the most commonly used health questionnaires such as the CES-D, STAI, and WHIIRS are well-developed tools to jointly measure a health condition of interest using multiple individual questions. Such questionnaire tools have been designed and tested in standard procedures (Björgvinsson et al., 2013; Spielberger, 1983) so that all questions are relevant to the measured health condition but no individual question can dominate the overall measurement. Therefore, additional noise features are added to each modality and the enhanced HCHS data include a total of 178 features within 10 modalities, consisting of 10, 32, 16, 10, 20, 20, 20, 20, 20, and 10 features from medical history, lab results, sleep monitoring data, neurocognitive measures, AHEI, HCHS Acculturation questions, CES-D, STAI, ESS, and WHIIRS, respectively. In each modality, the first 50% of the features are original features from HCHS and the last 50% of the features are noise features that we add.

2.6.2 Results and discussion of medical findings

The proposed M2-SEM was applied to the enhanced HCHS data for CM subgroup discovery, adjusted for age and gender, and identified 3 clusters within the 1,052 subjects.

Although it is not feasible to examine precisely the selection accuracy of all the features due to the lack of ground truth, we did inspect the estimates of all the noise features in each modality which indicated that 89% of noise features were identified as noise and excluded from the clustering by the proposed method. Among the 10 modalities, medical history, lab results, neurocognitive measures, AHEI, HCHS Acculturation questions, ESS, and WHIIRS are excluded from clustering, in the presence of the other more important CM-related modalities including sleep monitoring data, CES-D, and STAI. For ease of discussion, we refer to the three identified clusters as Cluster 1-3 with 456, 449, and 147 subjects in each cluster, respectively. To highlight the need to cluster multi-modal data, Figure 3 displays the cluster separation based on each individual modality and multi-modalities, in which the multi-modal data result in much better cluster separation. Clusters 1 and 2 differ significantly in depression and anxiety, not in sleep monitoring, whereas Cluster 3 differs significantly from the other two clusters in sleep monitoring.

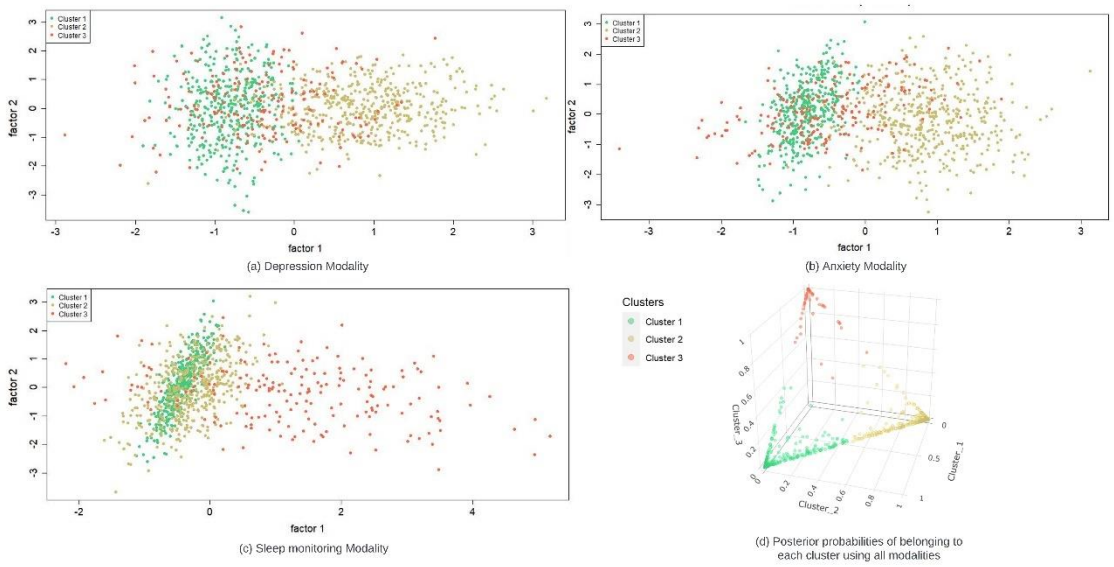

Figure 3: Comparison of cluster separation between (a-c) single-modal data and (d) multi-modal data

CES-D is a self-report measure of depression. The total score is calculated by finding the sum of 10 items and any score equal to or above 10 is considered depressed. Cluster 2 is the most depressed group with an average CES-D score of 12.4 and Cluster 1 is considered not depressed with an average score of 7.0. To eliminate the potential gender bias in depression and anxiety measures, i.e., females tend to be associated with higher levels of depression and anxiety, a two-sample proportion test is conducted that indicates there is no significant difference in gender distribution between Clusters 1 and 2 (P-value = 0.1), likely due to the adjustment for gender that is automatically considered in the proposed M2-SEM. Moreover, it is worth noting that there is no significant difference in age across the three clusters as well (P-values $\geq$ 0.6) suggesting that these findings that differentiate the clusters are not attributable to age differences. Similarly, the anxiety modality also indicates that Cluster 2 is more anxious than Cluster 1, consistent with the depression modality. That is, Cluster 2 is more depressed and anxious than Cluster 1, whereas Cluster 3 is in between for both modalities of mental health. Sleep monitoring is

the gold standard for the diagnosis of sleep disorders and assessment of sleep health. There is no significant difference in any feature between Clusters 1 and 2 (P-values ≥ 0.2). In contrast, Cluster 3 differs significantly from Clusters 1 and 2 across all features except for the minimum heart rate. Specifically, Cluster 3 is associated with a higher Apnea/Hypopnea Index (3% desat), a lower minimum and mean SpO2, a lower percent time SpO2 < 90, a higher maximum and mean heart rate, and a higher total time spent in loud snoring (P-values < 0.001), compared with both Clusters 1 and 2. It is likely that the increased heart rate in Cluster 3 may reflect the effects of apnea, in that at the end of apnea there is a marked increase in heart rate (Somers et al, 1995); The lower oxygen and loud snoring are very likely explanations for the worse sleep quality in Cluster 3 and for the faster heart rate as noted earlier. Therefore, it is not a surprise that the sparse selection of M2-SEM excludes only the dummy features and no real features are excluded from the clustering. In other words, almost all the features that we selected under the sleep monitoring modality play an important role in cluster identification. Overall, Cluster 3 has the worst clinical features as reflected in the sleep monitoring data compared with Clusters 1 and 2. In summary, Cluster 1 can be considered as the "healthiest" baseline group, i.e., with the most positive mental health status and no major concerns regarding sleep monitoring data, against which Clusters 2 and 3 can be compared; Cluster 2 has significantly worse mental status reflected in depression and anxiety, and Cluster 3 has significantly worse sleep quality.

Finally, we further validate the clinical relevance of the identified clusters by correlating the three clusters with several clinical characteristics not used for cluster identification. First, the Apnea/Hypopnea Index (4% desat), a clinical characteristic to

evaluate the severity of sleep disorders, is not considered in clustering and thus can be used to independently validate the clinical relevance of the identified clusters. Indeed, Cluster 3 has an average Apnea/Hypopnea Index (4%) as high as 42 events per hour, considered as "severe sleep disorders", but this average index is 12 events per hour for both Clusters 1 and 2, considered as "mild sleep disorders". There are abundant medical studies that suggest that sleep disorders and relevant features are independent risk factors for cardiovascular morbidity and mortality, constituting important aspects of cardiometabolic health. This further suggests that individuals in Cluster 3 have worse cardiometabolic health and might be at high risk of adverse health outcomes in later life, compared with individuals in Clusters 1-2. Accordingly, the Framingham Risk Score (FRS), a widely used clinical algorithm to estimate the 10-year cardiovascular risk of an individual, differs by Cluster. Cluster 1, the "healthiest" baseline group, is associated with the lowest FRS 33% among the three clusters; Cluster 2, the "worst" mental health group, and Cluster 3, the "worst" sleep disruption group, are associated with significantly higher FRSs of 45% and 48%, respectively.

2.7 Conclusion and Discussion

With the increased availability of health data from large biobanks, EHR systems, wearable sensors, etc., substantial data heterogeneity presents a common phenomenon in multi-modal health data. As the domain knowledge of underlying mechanisms in the medical field is often too scarce to explicitly articulate the patient-to-patient similarities and dissimilarities, data-driven clustering methods have been receiving increasing attention to subtype discovery from multi-modal heterogeneous health data to delineate heterogeneity and facilitate knowledge discovery. Among existing clustering approaches,

model-based clustering is a popular choice due to its statistical rigor in model inference and selection, but most models along this line do not consider the existence of latent factors within each data modality or take multi-modal data structure into account. Structural equation modeling encompasses a set of statistical models to analyze the interconnected relationship of multi-modal data and features, but this type of model relies heavily on the normality assumption that is required by model-based clustering models. The reason is when employed for clustering purposes, structural equation modeling assumes the cluster membership as a latent factor that follows a multinomial distribution, which therefore obviously violates the normality assumption. Although there are few studies that can accommodate categorical variables in structural equation modeling, most of them do not consider sparsity in model estimation and are computationally expensive in modeling high-dimensional data.

This chapter develops a novel Multi-modal Mixed-type Structural Equation Model (M2-SEM) with structured sparsity for precise subgroup discovery from multi-modal, mixed-type, high-dimensional data. The proposed M2-SEM results in a complex objective function with both observed data and latent variables that motivates the development of a novel GH-enabled Expectation Majorization Minimization algorithm (GH-EMM) for model estimation. The GH-EMM algorithm adopts the conventional EM framework but innovates as follows: instead of the Monte-Carlo (MC) EM algorithm, an efficient numerical optimal approach, i.e., Gauss-Hermite (GH) Quadrature, is leveraged to approximate the non-analytical terms in the expectation terms of the E-steps to provide a tractable, computationally efficient solution; the GH-enabled analytical terms are proven to have good computational properties such as Lipschitz-continuity and concavity, which

can be efficiently solved by the Majorization Minimization (MM) algorithm that is 5~10 times faster than conventional optimizers. The proposed M2-SEM and GH-EMM are examined in simulation studies under different settings and demonstrate robust accuracy in modality selection and model selection. Compared with several competing benchmarks, the proposed method achieves the highest clustering accuracy as well.

The proposed method is applied to a real-world dataset for CM subgroup discovery from multi-modal mixed-type CM data including 1,052 subjects with four continuous modalities including lab results, sleep monitoring data, and neurocognitive measures and six categorical modalities including medical history, AHEI, HCHS acculturation survey, CES-D, STAI, ESS, and WHIRS. The selected modalities that contribute to cluster differentiation the most are sleep monitoring data, CES-D, and STAI, from which three clusters are identified. Clusters 1 and 2 significantly differ in depression and anxiety, not in sleep characteristics, whereas Cluster 3 is significantly different from the other two clusters in sleep characteristics. Compared with single-modal data, multi-modal data clearly results in much better cluster separation highlighting the importance of clustering multi-modal data. Moreover, Cluster 1 can be considered as the "healthiest" baseline group, i.e., with the most positive mental health status and no adverse features in sleep monitoring data, against which Clusters 2 and 3 can be compared; Cluster 2 has significantly worse mental status reflected in depression and anxiety and Cluster 3 has significantly worse sleep quality. Last but not least, the clinical relevance of the clusters identified by the proposed M2-SEM is validated using two independent clinical characteristics. The findings of the proposed model are consistent with medical intuition and domain knowledge, e.g., the

commonly-used Framingham Risk Scores, that shed light on precise CM risk stratification and CM health promotion in the population.

Chapter 3 Federated Function-on-Function Regression with an Efficient Gradient Boosting Algorithm for Privacy-Preserving Telemedicine

3.1 Introduction

Federated Learning (FL) is an emerging computing paradigm to collaboratively train Machine Learning (ML) models by leveraging multi-source data without data exchange, thus removing many barriers to data sharing. FL is motivated by the growing need for data sharing and privacy-preserving of ML models. In the era of big data, the past decades have witnessed ML methods' success and rapid growth in modern society. Big data is required to empower ML methods at a large scale, but such data is often hard to obtain due to data privacy and ownership concerns. In the healthcare system, while one ML model that generalizes across heterogeneous, unharmonized Electronic Healthcare Record (EHR) data of different hospitals is desirable to facilitate clinical decision-making at the population level, it is challenging to combine the datasets across hospitals because health data is highly sensitive, and its usage is tightly regulated. The same phenomenon commonly exists in manufacturing systems, in which, for example, each company may collect and store its sensory data in the local server during production to monitor, inspect, and control the quality of its products. Such commercial data may be collected with a cost and thus poses challenges in data sharing with privacy and data ownership hurdles. FL is known to maintain the governance of data locally with only model parameters shared to enable collaboratively learning across multiple datasets, alleviating the privacy concern, and thus has already shown great promise in a variety of applications.

The function-on-function regression aims at predicting a functional response from other functional variables and receives more and more attention in functional data analysis. Functional data also referred to as time-series data, is a commonly encountered type of data in many research fields, such as healthcare, engineering, and economics. Take the health telemonitoring of Obstructive Sleep Apnea (OSA) as an example. OSA is a prevalent cardiac syndrome characterized by abnormal respiratory patterns during sleep, and its diagnosis involves an overnight recording of patients' multi-channel bio-signals, such as ECG and EEG, via wearable sensors (Alramadeen et al., 2023). The long-term recordings will later be manually scrutinized and scored by certified medical technicians to derive the frequency of adverse respiratory events, which is a labor-intensive procedure. Therefore, it is of clinical interest for a prediction model that automatically predicts the frequency of adverse respiratory events that occur in a certain time interval, i.e., epoch, from the bio-signals features extracted within the same epoch. Although the function-on-function regression model that can make a prediction of the functional response, e.g., frequency of adverse respiratory events in all epochs, is a natural choice for such a prediction model, there is no existing studies on FL of the function-on-function regression.

The major challenge of "meaningful" implementation of FL for any ML model is how to guarantee that the “federated model” can achieve a satisfactory performance comparable to the “global model” trained using the combined data as well as superior to each “local model” that can only see and use its own local data. Most existing FL methods focus on the empirical comparison of model performance between “federated model”, “global model”, and “local models” in comparative simulation studies, while the theoretical guarantee of the FL methods' performance is challenging and limited. This project

contributes the first-of-its-kind federated Gradient Boosting algorithm with the Least Squares Approximation (fed-GB-LSA) for efficient, privacy-preserving federated learning of the function-on-function regression. The proposed fed-GB-LSA will be tested in simulation studies and applied in a real-world dataset for OSA telemedicine.

The original contributions of the proposed fed-GB-LSA are summarized as follows. 1. The GB-based algorithm is flexible in the sense that it allows the inclusion and sparse selection of multivariate functional and non-functional features in the function-on-function regression prediction, which is not straightforward in functional regression. 2. The parameter estimation by the GB algorithm results in separate sub-optimization problems with explicitly analytical solutions for each of the features, providing a "computationally efficient" estimation algorithm for the function-on-function regression. 3. The LSA-enabled fed-GB provides a "one-shot" approach for FL that is "communicationally- and statistically- efficient". That is, the LSA-enabled aggregator is proven to enjoy the same asymptotic normality as the global estimator, offering theoretical guarantees to the performance of the federated model with a "one-shot" update strategy.

The rest of this chapter is organized as follows. Chapter 3.2 reviews the relevant work; Chapter 3.3 presents the model formulation of the function-on-function regression model; Chapter 3.4 introduces the federated model estimation based on fed-GB-LSA; Chapter 3.5 presents the simulation studies to evaluate the empirical performance of the proposed method; Chapter 3.6 applies the proposed method for health telemonitoring of OSA; Chapter 3.7 concludes this chapter.

3.2 Literature Review

Functional Data Analysis (FDA) encompasses a collection of statistical models increasingly being used to better analyze, model, and predict functional data, also commonly referred to as time-series data (Wang et al., 2016). Among a variety of FDA methods, functional regression is of great interest due to its flexibility and capacity in statistical modeling and predictive analytics inherited from traditional regression models. Depending on whether the responses or predictors are functional data or non-functional scalar data, there are scalar-on-function (Du and Wang, 2014; Cardot and Sarda, 2005; Müller and Yao, 2008; Wang et al., 2017), function-on-scalar (Zhang et al., 2022), or function-on-function regression (Chiou et al., 2016; Ivanescu et al., 2015; Luo and Qi, 2017; Imaizumi and Kato, 2018; Joseph et al., 2021). The function-on-function regression represents functional data with basis function systems and is often integrated with numerical optimization for approximation, thus resulting in a large number of parameters in basis functions to be estimated as well as a large number of knots for approximation to be considered in model training. A clear obstacle towards scalable computing of the function-on-function regression is how to efficiently process and estimate such a high number of knots and parameters in the high-dimensional data setting.

Boosting, which originated in the 1990s, is a data-driven model estimation approach that aggregates many weak learners, instead of a strong model, to get strong prediction results (Freund and Schapire, 1995). AdaBoost, the first well-known boosting algorithm, achieved outstanding binary classification results in its age (Freund and Schapire, 1995). The main idea of gradient boosting is sequentially adding new models by a gradient-descent-based algorithm to form an additive model (Friedman et al., 2000).

XGBoost, or eXtreme Gradient Boosting, (Chen et al., 2015) has demonstrated to be a reliable and efficient machine learning challenge solver and has been consistently placed among the top contenders in Kaggle competitions for a long time under various topics of data analysis tasks. Essentially, Boosting is a popular machine-learning technique, that focuses on improving the overall performance of model fitting by combining a collection of simple models (Friedman et al., 2000). By presenting boosting from a statistical point of view, Bühlmann and Hothorn (2007) extended the basic idea of boosting into a Gradient Boosting (GB) algorithm. GB trains model parameters by minimizing an empirical loss function without restrictions on the form of mathematical formulations or types of ML models, thus providing an efficient and flexible tool for model training (Hothorn et al., 2014).

Federated Learning (FL) has gained increasing attention due to the privacy concerns associated with the development of centralized learning in the era of big data. Unlike traditional centralized ML, FL, introduced by Google (McMahan et al., 2017; Konečný et al., 2016), enables knowledge sharing with a distributed training approach without data sharing, which allows individuals in different geographical locations to collaborate on developing ML models. Depending on data partitions across local servers, FL can be classified into horizontal Federated Learning (HFL), Vertical Federated Learning (VFL), and Federated Transfer Learning (FTL) (Yang et al., 2019). In HFL, the datasets of different local servers have the same feature space but with little to no intersection of sample space (McMahan et al., 2016); VFL comes in when the local servers are exposed to different feature spaces but with similar or the same sample space (Hardy et al., 2017); and FTL is a relatively rare architecture with a hybrid data partition in both

feature and sample spaces (Liu et al., 2020). HFL is most common among all FL architectures and is also the focus of this study. Moreover, FL can be categorized into cross-device and cross-silo FL based on the nature and scale of its participating local servers. Cross-device Federated Learning involves numerous small, distributed entities such as smartphones and wearable devices, with each entity holding a limited amount of data (Yang et al., 2019; Kairouz et al., 2021). This method hinges on the participation of a large number of these devices for successful training. On the other hand, cross-silo FL is characterized by its local servers, which are generally large organizations or corporations, such as hospitals and banks (Huang et al., 2022). This type of FL includes fewer local servers, but each server is heavily involved throughout the entire training process. The focus of this chapter is on cross-silo horizontal FL for privacy-preserving telemedicine, targeting various patient cohorts across different hospitals.

There are several studies that adopted GB for efficient estimation of the function-on-function regression (Brockhaus and Rügamer, 2016; Brockhaus et al., 2017). However, none of the studies has explored this problem in the FL setting. There is one study (Shen et al., 2022) that investigated the FL of the general GB algorithm, but it does not discuss the FL of the GB algorithm in the context of functional regression. Moreover, it used the Federated Averaging algorithm that consists of individual parameter update at each local server, followed by a model averaging update at the central server (Konečnỳ et al., 2016; McMahan et al., 2017). Although the Federated Averaging algorithm is the most commonly used method to obtain a global estimator by aggregating local estimators in FL, it requires multiple rounds of communication between the central server and local servers to refine the aggregated federated estimator thus being computationally and

communicationally expensive (Li et al., 2019; Yuan and Ma, 2020). Alternatively, the "one-shot" strategy requires one round of communication only. Despite the high efficiency, the "one-shot" approach might not achieve a satisfactory model performance in most FL studies. Wang and Leng (2007) proposed a Least Squares Approximation (LSA) method that can transfer many objective functions into their asymptotically equivalent least squares problems using the standard Taylor series expansion. The LSA method is naturally friendly for model estimation in FL in the sense that, when applied in the FL setting, the aggregated estimator by taking a weighted average of local estimators is guaranteed to be statistically as efficient as the global estimator, which requires only one-round of communication between the central and local servers, a.k.a., "one-shot" FL approach (Guha et al., 2019; Li et al., 2020; Zhu et al., 2021). However, no existing study has leveraged the LSA method for "one-shot" FL of the functional regression models.

3.3 Model Formulation

This subchapter introduces the model formulation of the function-on-function regression model with functional observations $\{\boldsymbol{y}, \mathbf{X}\}$, in which $\boldsymbol{y} = \{y_1(t), y_2(t), \cdots, y_N(t)\}^T$ and $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_N\}^T$ where $\boldsymbol{x}_n = \{x_{n1}(t), \cdots, x_{np}(t), \cdots, x_{nP}(t)\}^T$ for $n = 1, \ldots, N$. N denotes the number of observations and P denotes the number of predictors. The sampling period is T so that $t \in T$. Note that the function-on-function regression and its variations have been studied in recent years (Chiou et al., 2016; Ivanescu et al., 2015; Luo and Qi, 2017; Imaizumi and Kato, 2018; Joseph et al., 2021). This subchapter aims to provide a brief introduction of its classic mathematical formulation (Ramsay and Silverman, 2002) to facilitate the later discussion

on the development of a novel fed-GB-LSA algorithm for privacy-preserving model estimation in Chapter 3.4.

The function-on-function regression model extends the conventional linear regression by defining a bivariate coefficient function $\beta_p(s,t)$ for $p = 1, \dots, P$. For a given subject n, the function-on-function regression model can be written as

$$y_n(t) = \sum_{p=1}^{P} \int_{s \in T} x_{np}(s)\, \beta_p(s,t) ds + \varepsilon_n(t), \tag{3.1}$$

where $\beta_p(s,t)$ is the bivariate coefficient function for the p-th functional predictor, and $\varepsilon_n(t)$ is the random error function that follows a normal distribution. Similar to the conventional regression model, the intercept term in the function-on-function regression model (3.1) can be dropped after centering the functional response and predictors without loss of generality.

To approach the functional variables, the function-on-function regression introduces various basis systems such as Fourier, monomial, and B-Spline (Ramsay and Silverman, 2002). The basis representation results in a set of coefficients that are easy to estimate, and the appropriate choice of basis system can account for nearly any curve features in the functional variables. We assume the bivariate coefficient function $\beta_p(s,t)$ has a double expansion on one basis system $\boldsymbol{\theta}$ with K_1 functions and another basis system $\boldsymbol{\eta}$ with K_2 functions, i.e., $\beta_p(s,t) = \boldsymbol{\theta}(s)^T \mathbf{B}_p \boldsymbol{\eta}(t)$ in which $\boldsymbol{\theta}(t) = \left(\theta_1(t), \dots, \theta_{K_1}(t)\right)^T$, $\boldsymbol{\eta}(t) = \left(\eta_1(t), \dots, \eta_{K_2}(t)\right)^T$, and $\mathbf{B}_p \in \boldsymbol{R}^{K_1 \times K_2}$. After substituting the double expansion of $\beta_p(s,t)$ into (3.1), the function-on-function regression model can be rewritten as

$$y_n(t) = \sum_{p=1}^{P} \int_{s \in T} x_{np}(s)\, \boldsymbol{\theta}(s)^T ds\, \mathbf{B}_p \boldsymbol{\eta}(t) + \varepsilon(t), \tag{3.2}$$

in which we define a $1 \times K_1$ row vector $\boldsymbol{z}_{np} = \int_{s \in T} x_{np}(s)\, \boldsymbol{\theta}(s)^T ds$ that can be pre-calculated from the data of the p-th predictor and the selected basis functions. Specifically, let's assume the predictors $\mathbf{X}$ and response $\boldsymbol{y}$ are collected simultaneously throughout the sampling period T with a sampling interval of Δ. The sampling sequence throughout T is $\{t_1, t_2 \cdots, t_i, \cdots T\}$. The term $\boldsymbol{z}_{np}$ can be approximated by its Riemann sum (Hughes-Hallett et al., 2020), i.e., $\boldsymbol{z}_{np} = \int_{s=0}^{T} x_{np}(s)\, \boldsymbol{\theta}(s)^T ds \approx \Delta \sum_i x_{np}(s_i) \boldsymbol{\theta}(s_i)^T$. It is also noteworthy that the function-on-function regression model can accommodate both functional and non-functional predictors. For example, if the p-th predictor is non-functional and takes a fixed value for $n = 1, \ldots, N$, $x_{np}(s)$ becomes a constant with respect to s. Therefore, the formulation in (3.2) allows us to adopt a simple and unified formulation for both functional and non-functional predictors. After replacing the integral in (3.2) with the computed $\boldsymbol{z}_{np}$, the model can be rewritten as

$$y_n(t) = \sum_{p=1}^{P} h_p(t) + \varepsilon(t), \tag{3.3}$$

where $h_p(t) = \boldsymbol{z}_{np} \mathbf{B}_p \boldsymbol{\eta}(t)$ for $p \in \{1, \cdots, P\}$. $h_p(t)$ is defined as the "base learner" to facilitate the later discussion on using GB algorithm for model estimation. The formulation in (3.3) clearly reveals that the model estimation of the function-on-function regression can be achieved by selecting a sequence of appropriate "base learners" $h_p(t)$ to estimate the functional response $y_n(t)$ with an additive form of the selected base learners. More details on estimating the model in (3.3) are provided in Chapter 3.4.

3.4 A Novel fed-GB-LSA for Federated Model Estimation

The privacy-preserving model estimation is achieved by a novel federated Gradient Boosting (GB) algorithm integrated with the Least Square Approximation (fed-GB-LSA).

The proposed fed-GB-LSA innovatively employs the GB algorithm for efficient estimation of the function-on-function regression in Chapter 3.4.1 and integrates the LSA to enable "one-shot' federated learning of GB in Chapter 3.4.2.

3.4.1 Gradient Boosting (GB)

GB is a boosting-type ensemble method that uses a divide-and-conquer approach to optimize the loss function (Bühlmann and Hothorn, 2007; Hothorn et al., 2014). GB iteratively selects the best base learner from $\{h_1(\cdot), \cdots, h_P(\cdot)\}$ to update the model until reaching a pre-set maximal number of iterations q. Specifically, given prediction $f(t, \boldsymbol{z}_n)$ where $\boldsymbol{z}_n = \{\boldsymbol{z}_{n1}, \boldsymbol{z}_{n2}, \cdots, \boldsymbol{z}_{nP}\}$ and $n = 1, \ldots, N$, GB minimizes the loss function l over prediction function f. Compared with most boosting methods that are purely heuristic, the GB model results in an additive form of multiple simple models, e.g., $h_1(\cdot), \ldots, h_p(\cdot)$, that fit the negative gradient of the loss function to minimize the loss function along the steepest gradient descent in each iteration. Moreover, GB considers the variable selection during the model fitting process without relying on heuristic stepwise variable selection or lasso-type non-smooth penalties.

We first briefly review the conventional GB (Bühlmann and Hothorn, 2007; Hothorn et al., 2014) and then derive how GB framework can be extended to estimate the function-on-function regression. By choosing the least square function as the loss function, we define the loss function as $l(\boldsymbol{y}, f|\mathrm{X}) = \sum_{n=1}^{N} \int_{t \in T} (y_n(t) - f(t, \boldsymbol{z}_n))^2 dt$. Then, GB aims to solve the following optimization

$$f^* = argmin_f l(\boldsymbol{y}, f|\mathrm{X}). \tag{3.4}$$

The step 0 of GB initializes the estimation with offset values $f^{[0]}$, which is a vector of length N. Each element of $f^{[0]}(t)$ is a functional variable $\bar{y}(t) = \frac{1}{N}\sum_{n=1}^{N} y_n(t)$.

Next, we derive how to extend the conventional GB algorithm to efficiently estimate the function-on-function regression model in (3.3). In the ω-th iteration, $\omega \in \{1, \cdots, q\}$, GB first computes the negative gradient of risk function in (3.4) with respect to f, i.e., $\boldsymbol{u}^{(\omega)} \in R^{N\times 1} = -\frac{\partial l}{\partial f}\Big|_{f=f^{[\omega-1]}}$, and then fits each candidate base learner to the negative gradient $\boldsymbol{u}^{(\omega)}$ by solving the following optimization problems for $p = 1, \ldots, P$:

$$\widehat{\mathbf{B}}_p^{(\omega)} = \underset{\mathbf{B}_p}{\operatorname{argmin}} \sum_{n=1}^{N} \int_{t\in T} \left(u_n^{(\omega)}(t) - \boldsymbol{z}_{np}\mathbf{B}_p\boldsymbol{\eta}(t)\right)^2 dt, \tag{3.5}$$

where $u_n^{(\omega)}(t)$ is the n-th element in the negative gradient $\boldsymbol{u}^{(\omega)}$. The optimal solution of the problem in (3.5) can be used to derive the fitted base learner $\hat{\mathrm{h}}_p^{(\omega)}(t) = \boldsymbol{z}_{np}\widehat{\mathbf{B}}_p^{(\omega)}\boldsymbol{\eta}(t)$ for $p \in \{1, \cdots, P\}$. Among all the fitted base learners, GB selects the best base learner by minimizing the Residual Sum of Squares (RSSs) defined as

$$RSS_p = \sum_{n=1}^{N} \int_{t\in T} \left(u_n^{(\omega)}(t) - \boldsymbol{z}_{np}\widehat{\mathbf{B}}_p^{(\omega)}\boldsymbol{\eta}(t)\right)^2 dt \text{ for } p = 1, \ldots, P. \tag{3.6}$$

Using the best base learner $h_{p^*}^{(\omega)}$ with the minimal RSS, GB updates the model by $f^{(\omega)}(t) = f^{(\omega-1)}(t) + \nu h_{p^*}^{(\omega)}$, in which ν is the pre-set learning rate and $h_{p^*}^{(\omega)} = \boldsymbol{z}_{np}\widehat{\mathbf{B}}_{p^*}^{(\omega)}\boldsymbol{\eta}(t)$. The GB algorithm iterates until the stopping criterion, e.g., maximal number of iterations q, is reached.

Last but not least, it is noteworthy that although the optimization problem in (3.5) seems complicatedly structured, its optimal solution can be derived as an analytical form as shown in Proposition 3.1 below. Therefore, in each GB iteration, all the base learners

can be efficiently fitted by analytical forms without the need for any other iterative optimizers, which contributes to the improvement of the computational efficiency of the GB algorithm. Next, we present Proposition 3.1 and its proof in detail.

Proposition 3.1 (The optimization problem in (3.5) has an analytical form). Given $\mathbf{B} \in \boldsymbol{R}^{K_1 \times K_2}$, $\mathbf{Z} \in \boldsymbol{R}^{N \times K_1}$, and two functional vectors $\boldsymbol{u}(t)$ and $\boldsymbol{\eta}(t)$, the problem in (3.5) results in a general optimization problem:

$$\mathbf{B}^* = \underset{\mathbf{B}}{\operatorname{argmin}} \int_{t \in T} \|\mathbf{u}(t) - \mathbf{Z}\mathbf{B}\boldsymbol{\eta}(t)\|^2 dt, \tag{3.7}$$

Where $\boldsymbol{u}(t) = \left(u_{\mathbf{1}}(t), \cdots, u_N(t)\right)^T$ and $\boldsymbol{\eta}(t) = \left(\eta_{\mathbf{1}}(t), \cdots, \eta_N(t)\right)^T$. The optimal solution is

$$vec(\mathbf{B}^*) = \left(J_{\eta\eta} \otimes (\mathbf{Z}^T\mathbf{Z})\right)^{-1} vec\left(\mathbf{Z}^T \int_t \boldsymbol{u}(t)\boldsymbol{\eta}^T(t)dt\right), \tag{3.8}$$

where $J_{\eta\eta} = \int_{t \in T} \boldsymbol{\eta}(t)\boldsymbol{\eta}^T(t) dt$. Note that $vec(\cdot)$ writes a matrix as a vector column wise.

Proof: The loss function of the optimization problem can be written as

$$l = \int_{t \in T} \|\boldsymbol{u}(t) - \mathbf{Z}\mathbf{B}\boldsymbol{\eta}(t)\|^2 dt = \int_{t \in T} \left(\boldsymbol{u}(t) - \mathbf{Z}\mathbf{B}\boldsymbol{\eta}(t)\right)^T \left(\boldsymbol{u}(t) - \mathbf{Z}\mathbf{B}\boldsymbol{\eta}(t)\right) dt$$

$$= \int_{t \in T} \boldsymbol{u}(t)^T\boldsymbol{u}(t)dt + \int_{t \in T} \boldsymbol{\eta}(t)^T\mathbf{B}^T\mathbf{Z}^T\mathbf{Z}\mathbf{B}\boldsymbol{\eta}(t)dt - 2\int_{t \in T} \boldsymbol{u}(t)^T\mathbf{Z}\mathbf{B}\boldsymbol{\eta}(t)dt \ .$$

In the loss function, the operations of summation and integration are interchangeable. Based on the property of matrix's trace, for example, $trace(\boldsymbol{ABC}) = trace(\boldsymbol{CAB})$, we have

$$l = \int_{t\in T}^{\square} \boldsymbol{u}(t)^T \boldsymbol{u}(t)dt + \int_{t\in T}^{\square} trace(\boldsymbol{\eta}(t)^T \mathbf{B}^T \mathbf{Z}^T)dt - 2\int_{t\in T}^{\square} trace\big(\boldsymbol{u}(t)^T \mathbf{Z}\mathbf{B}\boldsymbol{\eta}(t)\big)dt$$

$$= \int_{t\in T}^{\square} \boldsymbol{u}(t)^T \boldsymbol{u}(t)dt + \int_{t\in T}^{\square} trace(\mathbf{Z}^T \mathbf{Z}\mathbf{B}\boldsymbol{\eta}(t)\boldsymbol{\eta}(t)^T \mathbf{B}^T)dt$$

$$- 2\int_{t\in T}^{\square} trace(\mathbf{B}\boldsymbol{\eta}(t)\boldsymbol{u}(t)^T \mathbf{Z})dt.$$

By defining $\mathbf{J}_{\eta\eta} \overset{\text{def}}{\triangleq} \int_{t\in T}^{\square} \boldsymbol{\eta}(t)\boldsymbol{\eta}^T(t)\mathrm{d}t$, we have

$$l = \int_{t\in T}^{\square} \boldsymbol{u}(t)^T \boldsymbol{u}(t)dt + trace\big(\mathbf{Z}^T \mathbf{Z}\mathbf{B}\mathbf{J}_{\eta\eta}\, \mathbf{B}^T\big) - 2trace\left(\mathbf{B}\int_{t\in T}^{\square} \boldsymbol{\eta}(t)\boldsymbol{u}(t)^T dt\, \mathbf{Z}\right).$$

Taking the derivative of the loss function with respect to **B**, we have

$$\frac{\partial l}{\partial \mathbf{B}} = \frac{trace\big(\mathbf{Z}^T \mathbf{Z}\mathbf{B}\mathbf{J}_{\eta\eta}\, \mathbf{B}^T\big)}{\partial \mathbf{B}} - 2\frac{\partial trace\left(\mathbf{B}\int_{t\in T}^{\square} \boldsymbol{\eta}(t)\boldsymbol{u}(t)^T dt\, \mathbf{Z}\right)}{\partial \mathbf{B}}$$

$$= 2\mathbf{Z}^T \mathbf{Z}\mathbf{B}\mathbf{J}_{\eta\eta} - 2\mathbf{Z}^T \int_{t\in T}^{\square} \boldsymbol{\eta}(t)\boldsymbol{u}(t)^T dt.$$

Making the derivative equal to zero results in

$$\mathbf{Z}^T \mathbf{Z}\mathbf{B}^* \mathbf{J}_{\eta\eta} = \mathbf{Z}^T \int_{t\in T}^{\square} \boldsymbol{\eta}(t)\boldsymbol{u}(t)^T dt.$$

Using the Kronecker product, we have

$$vec\big(\mathbf{Z}^T \mathbf{Z}\mathbf{B}^* \mathbf{J}_{\eta\eta}\big) = \big(\mathbf{J}_{\eta\eta} \otimes \mathbf{Z}^T \mathbf{Z}\big)vec(\mathbf{B}^*) = vec\left(\mathbf{Z}^T \int_{t\in T}^{\square} \boldsymbol{\eta}(t)\boldsymbol{u}(t)^T dt\right).$$

Thus, the optimization problem's solution can be written as

$$vec(\mathbf{B}^*) = \big(\mathbf{J}_{\eta\eta} \otimes \mathbf{Z}^T \mathbf{Z}\big)^{-1} vec\left(\mathbf{Z}^T \int_{t\in T}^{\square} \boldsymbol{\eta}(t)\boldsymbol{u}(t)^T dt\right). \quad \blacksquare$$

3.4.2 Federated Gradient Boosting by Least Square Approximation (fed-GB-LSA)

This subchapter introduces how to implement FL in the GB algorithm to preserve privacy in model training. In particular, the LSA is integrated within the GB algorithm to provide a "one-shot" aggregated estimator that is both communicationally and statistically efficient for FL. Assuming a total of N subjects distributed in K local servers, we define $S = \{1, \dots, N\}$ that is the union of K mutually exclusive subsets, i.e., $S = \mathrm{U}_{k=1}^{K} S_k$ where S_k contains subjects in the k-th local server. Then, we assume the number of subjects in server k is $|S_k| = N_k$, so that $N = \sum_{k=1}^{K} N_k$.

As described in Chapter 3.4.1, the non-federated GB needs to solve the optimization problem in (3.5) for all the base learners by leveraging the entire dataset, which is not feasible under the FL setting. Hereinafter, this subchapter focuses on how to solve the optimization problem in (3.5) with FL. For simplified notation, we can drop the subscript ω in (3.5) and rewrite this optimization problem as

$$\widetilde{\mathbf{B}}_p^* = \underset{\mathbf{B}_p}{\operatorname{argmin}}\, N^{-1} \sum_{k=1}^{K} \sum_{n \in S_k} l_{n,p}(\mathbf{B}_p), \tag{3.9}$$

where $\widetilde{\mathbf{B}}_p^*$ is the optimal solution of the "global model" that can use the entire dataset for model estimation, and $l_{n,p}(\mathbf{B}_p) = \int_{t \in T} \left(u_n^{(\omega)}(t) - \boldsymbol{z}_{np} \mathbf{B}_p \boldsymbol{\eta}(t) \right)^2 dt$ as in (3.5). Unfortunately, it is not feasible to obtain $\widetilde{\mathbf{B}}_p^*$ for the global model in FL because each local sever k can see its own portion of data, i.e., $\left\{ u_n^{(\omega)}(t), \boldsymbol{z}_{np} \right\}_{n \in S_k}$. Alternatively, the optimization problem can be rewritten for each local server as follows:

$$\widetilde{\mathbf{B}}_{p,k}^* = \underset{\mathbf{B}_p}{\operatorname{argmin}}\, {N_k}^{-1} \sum_{n \in S_k} l_{n,p}(\mathbf{B}_p). \tag{3.10}$$

However, since each server k only has access to its own subset of the data, the optimal solution $\widetilde{\mathbf{B}}_{p,k}^{*}$ at each local server is known to be statistically less efficient than the optimal solution $\widetilde{\mathbf{B}}_{p}^{*}$ of the global model, for $k = 1, \dots, K$.

FL is designed to approach this problem in (3.9) with a near-optimal solution under the constraint of the mutually exclusive distributed data. Let's denote $\widetilde{\mathbf{B}}_{p}$ as the federated solution of (3.9) for ease of discussion. Ideally, the federated estimator $\widetilde{\mathbf{B}}_{p}$ should outperform all the local estimators $\widetilde{\mathbf{B}}_{p,k}^{*}$ for $k = 1, \dots, K$ while providing comparable performance to the global estimator $\widetilde{\mathbf{B}}_{p}^{*}$. To obtain the federated estimator $\widetilde{\mathbf{B}}_{p}$, an intuitive solution is to aggregate all the local estimators $\widetilde{\mathbf{B}}_{p,k}^{*}$ for $k = 1, \dots, K$ and use their average as the federated estimator. However, this so-called "Federated Average algorithm" is heuristic, and its resulted aggregated estimator is obviously statistically inefficient thus requiring many rounds of communication between the local and global servers to iterate this averaging procedure. To overcome these challenges, this study proposes a "one-shot" approach by approximating the objective function in (3.9) with LSA. Although the local estimators are aggregated on the cloud only once, the proposed one-shot aggregated estimator is still considered statistically efficient and is shown to empirically outperform the estimator obtained by the federated average algorithm which is much more communicationally costly.

We first introduce the conventional LSA (Wang and Leng, 2007) and then derive how LSA can be leveraged for FL of the function-on-function regression. Specifically, let's denote $L\left(\mathbf{B}_{p}\right) = N^{-1} \sum_{k=1}^{K} \sum_{n \in S_k} l_{n,p}\left(\mathbf{B}_{p}\right)$ and apply Taylor's expansion to each $l_{n,p}\left(\mathbf{B}_{p}\right)$ at $\widetilde{\mathbf{B}}_{p,k}^{*}$ for $k = 1, \dots, K$. We have

$$L(\mathbf{B}_p) = N^{-1} \sum_{k=1}^{K} \sum_{n \in S_k} l_{n,p}(\mathbf{B}_p)$$

$$\approx N^{-1} \sum_{k=1}^{K} \sum_{n \in S_k} l_{n,p}(\widetilde{\mathbf{B}}_{p,k}^{*}) + N^{-1} \sum_{k=1}^{K} \sum_{n \in S_k} l'_{n,p}(\widetilde{\mathbf{B}}_{p,k}^{*})^{T} \left(vec(\mathbf{B}_p) - vec(\widetilde{\mathbf{B}}_{p,k}^{*})\right) + \frac{1}{2} N^{-1} \sum_{k=1}^{K} \sum_{n \in S_k} \left(vec(\mathbf{B}_p) - vec(\widetilde{\mathbf{B}}_{p,k}^{*})\right)^{T} l''_{n,p}(\widetilde{\mathbf{B}}_{p,k}^{*}) \left(vec(\mathbf{B}_p) - vec(\widehat{\mathbf{B}}_{p,k}^{*})\right), \quad (3.11)$$

where the first term $N^{-1} \sum_{k=1}^{K} \sum_{n \in S_k} l_{n,p}(\widetilde{\mathbf{B}}_{p,k}^{*})$ does not contain $\mathbf{B}_p$ and the second term is zero because $\widetilde{\mathbf{B}}_{p,k}^{*}$ is the optimal solution of the local model in (3.10) resulting in $l'_{n,p}(\widetilde{\mathbf{B}}_{p,k}^{*}) = \mathbf{0}$. Therefore, the optimization problem in (3.11) is reduced to the optimization of the third term $N^{-1} \sum_{k=1}^{K} \sum_{n \in S_k} \left(vec(\mathbf{B}_p) - vec(\widetilde{\mathbf{B}}_{p,k}^{*})\right)^{T} l''_{n,p}(\widetilde{\mathbf{B}}_{p,k}^{*}) \left(vec(\mathbf{B}_p) - vec(\widetilde{\mathbf{B}}_{p,k}^{*})\right)$ with respect to $\mathbf{B}_p$.

As $\widetilde{\mathbf{B}}_{p,k}^{*}$ is the minimizer of $\sum_{n \in S_k} l_{n,p}(\mathbf{B}_p)$, $\widetilde{\mathbf{B}}_{p,k}^{*}$ is $\sqrt{N_k}$-consistent and asymptotically normal, i.e., $\sqrt{N_k}\left(vec(\widetilde{\mathbf{B}}_{p,k}^{*}) - vec(\mathbf{B}_{p,0})\right) \rightarrow_d N(\mathbf{0}, \mathbf{\Sigma}_{p,k})$ for the true parameter $\mathbf{B}_{p,0}$. Moreover, given certain statistical assumptions, it is plausible that $\mathbb{E}(N_k^{-1} \sum_{n \in S_k} l''_{n,p}(\widetilde{\mathbf{B}}_{p,k}^{*})) \approx \mathbf{\Sigma}_{p,k}^{-1}$ where $\mathbf{\Sigma}_{p,k}^{-1}$ is the asymptotic covariance matrix of $vec(\widetilde{\mathbf{B}}_{p,k}^{*})$. Therefore, we can use $\widehat{\mathbf{\Sigma}}_{p,k}^{-1}$ as a natural estimator of $N_k^{-1} \sum_{n \in S_k} l''_{n,p}(\widetilde{\mathbf{B}}_{p,k}^{*})$ in equation (3.11) (Wang and Leng, 2007).

Next, we derive how to extend the conventional LSA algorithm to efficiently estimate the function-on-function regression model under FL. Based on Propositions 3.1, we have

$$vec(\widetilde{\mathbf{B}}_{p,k}^{*}) = \left(\mathbf{J}_{\eta\eta} \otimes \left(\mathbf{Z}_{p,k}{}^{T} \mathbf{Z}_{p,k}\right)\right)^{-1} vec\left(\mathbf{Z}_{p,k}{}^{T} \int_{t} \boldsymbol{u}_{p,k}(t) \boldsymbol{\eta}^{T}(t) dt\right), \quad (3.12)$$

and

$$\widehat{\boldsymbol{\Sigma}}_{p,k} = N_k \widehat{cov}\left(vec\left(\widetilde{\mathbf{B}}^*_{p,k}\right)\right)$$

$$= N_k \left(\mathbf{J}_{\eta\eta} \otimes \left(\mathbf{Z}_{p,k}{}^T \mathbf{Z}_{p,k}\right)\right)^{-1} vec\left(\mathbf{Z}_{p,k}{}^T \int_t^{\square} \boldsymbol{\varepsilon}_{p,k}(t)\boldsymbol{\eta}^T(t)dt\right)\left(vec\left(\mathbf{Z}_{p,k}{}^T \int_t^{\square} \boldsymbol{\varepsilon}_{p,k}(t)\boldsymbol{\eta}^T(t)dt\right)\right)^T \left(\left(\mathbf{J}_{\eta\eta} \otimes \left(\mathbf{Z}_{p,k}{}^T \mathbf{Z}_{p,k}\right)\right)^{-1}\right)^T \quad (3.13)$$

in which $\mathbf{Z}_{p,k} = \left\{\mathbf{Z}_{1,p,k}, \cdots \mathbf{Z}_{n_k,p,k}\right\}^T$, $\mathbf{Z}_{n_k,p,k} = \int_{s=0}^{T} x_{n_k,p}(s)\,\boldsymbol{\theta}(s)^T ds$, $\boldsymbol{\varepsilon}_{p,k}(t) = \left(\boldsymbol{\varepsilon}_{1,p,k}(t), \cdots, \boldsymbol{\varepsilon}_{n_k,p,k}(t)\right)^T$, and $\boldsymbol{\varepsilon}_{n_k,p,k}(t) = \boldsymbol{u}_{n_k,p,k}(t) - \mathbf{Z}_{n_k,p,k}\widetilde{\mathbf{B}}^*_{p,k}\boldsymbol{\eta}(t)$. Consequently, after replacing the $\sum_{n\in S_k}^{\square} l''_{n,p}\left(\widetilde{\mathbf{B}}^*_{p,k}\right)$ with its natural estimator $N_k\widehat{\boldsymbol{\Sigma}}^{-1}_{p,k}$ and dropping the constants, the general LSA term in (3.11) can be rewritten as

$$\tilde{L}\left(\mathbf{B}_p\right) = N^{-1}\sum_{k=1}^{K}\sum_{n\in S_k}^{\square}\left(vec(\mathbf{B}_p) - vec(\widetilde{\mathbf{B}}^*_{p,k})\right)^T l''_{n,p}\left(\widetilde{\mathbf{B}}^*_{p,k}\right)\left(vec(\mathbf{B}_p) - vec(\widetilde{\mathbf{B}}^*_{p,k})\right)$$

$$= \sum_{k=1}^{K}\frac{N_k}{N}\left(vec(\mathbf{B}_p) - vec(\widetilde{\mathbf{B}}^*_{p,k})\right)^T \widehat{\boldsymbol{\Sigma}}^{-1}_{p,k}\left(vec(\mathbf{B}_p) - vec(\widetilde{\mathbf{B}}^*_{p,k})\right), \quad (3.14)$$

which is referred to as the Least Squares Approximation (LSA) of the objective function $L\left(\mathbf{B}_p\right)$. Instead of minimizing $L\left(\mathbf{B}_p\right)$ in (3.11), this study proposes to minimize its LSA, i.e., $\tilde{L}\left(\mathbf{B}_p\right)$ in (3.14) resulting in an analytical minimal solution of $\mathbf{B}_p$ that takes the following form:

$$\widehat{\mathbf{B}}_p = \left(\sum_{k=1}^{K}\frac{N_k}{N}\widehat{\boldsymbol{\Sigma}}^{-1}_{p,k}\right)^{-1}\left(\sum_{k=1}^{K}\frac{N_k}{N}\widehat{\boldsymbol{\Sigma}}^{-1}_{p,k}\,\widetilde{\mathbf{B}}^*_{p,k}\right). \quad (3.15)$$

More importantly, we have proven the global asymptotic normality of the LSA aggregation estimator in (3.15) in Theorem 3.1. Essentially, this theorem confirms that the LSA aggregator $\widehat{\mathbf{B}}_p$ is proven to enjoy the same asymptotic normality as the global estimator $\widetilde{\mathbf{B}}^*_p$. In other words, it is statistically as efficient as the global estimator in terms of resulting bias and variance compared to the true global parameters, providing theoretical

guarantee for the proposed "one-shot" FL approach that necessitates only one-round communication between the central and local servers.

Theorem 3.1 (Global asymptotic normality). We denote the asymptotic covariance matrix of the global estimator $\widetilde{\mathbf{B}}_p^*$ as $\mathbf{\Sigma}_p$, i.e., $\mathbf{\Sigma}_p = Ncov(vec(\widetilde{\mathbf{B}}_p^*))$. Given certain statistical regularity conditions and $K \ll \sqrt{N}$, we have $\sqrt{N}\big(vec(\widehat{\mathbf{B}}_p) - vec(\mathbf{B}_{p,0})\big) \rightarrow_d N(0, \mathbf{\Sigma}_p)$, which indicates that the proposed LSA estimator $\widehat{\mathbf{B}}_p$ achieves the same asymptotic normality as the global estimator $\widetilde{\mathbf{B}}_p^*$.

(*Due to the limitation of space, the detailed proof is listed in Appendix G.*)

Consequently, the aggregated estimator in (3.15), a weighted average of local estimates $\widehat{\mathbf{B}}_{p,k}$ and $\widehat{\mathbf{\Sigma}}_{p,k}$ for $p = 1, \dots, P$ and $k = 1, \dots, K$, provides a natural "one-shot" FL solution in each GB iteration. Specifically, Table 4 presents the overview of the proposed fed-GB-LSA, in which each iteration of the fed-GB-LSA consists of two steps, i.e., Step 1 obtains the global aggregator $\left\{\widehat{\mathbf{B}}_p^{(\omega)}\right\}_{p=1}^{P}$ for all base learners without data sharing across local servers and Step 2 that calculates RSSs to select the best base learner. As shown in Table 4, it is obvious that the proposed fed-GB-LSA requires one round of communication between the local and global servers in Step 1. In contrast, the conventional fed-GB-Average, without the theoretical guarantee provided by Theorem 3.1, relies on the gradient descent algorithm (see appendix H for details on fed-GB-Average) and thus requires iterative communication to obtain a relatively robust aggregator, incurring higher communication costs.

Moreover, the "one-shot" FL strategy is less susceptible to privacy risks arising from the communications between local and global servers. Despite the fact that the

proposed LSA-based aggregator transmits both the local estimator $\widetilde{\mathbf{B}}_{p,k}^{*}$ and its covariance matrix $\widehat{\boldsymbol{\Sigma}}_{p,k}^{-1}$ to the central server for global aggregation incurring a higher amount of network data per iteration (Zhang et al., 2023), this difference becomes less pronounced when considering the overall communicational costs for Step 1. This is because, to complete Step 1, the proposed "one-shot" fed-GB-LSA requires one-round of data transmission, whereas the classic FL methods such as Federated Averaging require multiple rounds of communication and data transmissions between the global and local servers (Liu et al., 2022). Therefore, the proposed "one-shot" fed-GB-LSA introduces significantly less privacy loss when comparing to other FL methods.

Table 4: Pseudo code for the fed-GB-LSA on local and central servers

Import: $\{y_n(t), x_{n1}(t), \dots, x_{nP}(t)\}_{n=1}^{N}$; $\{S_k\}_{k=1}^{K}$; set the step length ν and stopping threshold m.

Initialization: $f^{(0)}(t) = 0; \omega = 0$.

Iterate until $\omega = m$:

$\omega = \omega + 1$;

Step 1: "One-shot" update to obtain the global aggregator $\left\{\widehat{\mathbf{B}}_p^{(\omega)}\right\}_{p=1}^{P}$ for all base learners

Local servers with parallel computing for $k = 1, \dots, K$

- Local data: $\mathbf{D}_k = \{y_n(t), x_{n1}(t), \dots, x_{nP}(t)\}_{n \in S_k}$
- Compute the negative gradient $\boldsymbol{u}^{(k,\omega)}$ using data $\mathbf{D}_k$
- Fit $\boldsymbol{u}^{(k,\omega)}$ with base learners by applying (3.8) to $\mathbf{D}_k$ to estimate $\widetilde{\mathbf{B}}^{*(\omega)}_{p,k}$ for $p = 1, \dots, P$
- Send local parameters $\left\{\widetilde{\mathbf{B}}^{*(\omega)}_{p,k}, \widehat{\sum}^{(\omega)}_{p,k}\right\}_{p=1}^{P}$ to the central server

Central server for

- Receive the local parameters $\left\{\left\{\widetilde{\mathbf{B}}^{*(\omega)}_{p,k}, \widehat{\sum}^{(\omega)}_{p,k}\right\}_{p=1}^{P}\right\}_{k=1}^{K}$ from all K local servers
- Calculate the global parameters $\widehat{\mathbf{B}}_p^{(\omega)}$ by (3.15) for $p = 1, \dots, P$.
- Send the aggregated parameters $\left\{\widehat{\mathbf{B}}_p^{(\omega)}\right\}_{p=1}^{P}$ to local servers

Step 2: Update of the additive model with the best base learner $p^{(\omega)}$

Local servers with parallel computing for $k = 1, \dots, K$

- Receive the global parameters $\left\{\widehat{\mathbf{B}}_p^{(\omega)}\right\}_{p=1}^{P}$
- Calculate $RSS_p^{(k,\omega)}$ by inserting $\widehat{\mathbf{B}}_p^{(\omega)}$ and $\mathbf{D}_k$ into (3.6) for $p = 1, \dots, P$
- Send local parameters $\left\{ RSS_p^{(k,\omega)}\right\}_{p=1}^{P}$ to the central server

Central server

- Receive the local parameters $\left\{RSS_p^{(k,\omega)}\right\}_{k=1}^{K}$ from all the K local servers
- Calculate the global residuals $RSS_p^{(agg,\omega)} = \sum_{k=1}^{K} \mathrm{RSS}_p^{(k,\omega)}$ for $p = 1, \dots, P$ and select the best base learner $p^{(\omega)}$
- Send the aggregated parameters $\left\{p^{(\omega)}\right\}$ to local servers

Local servers with parallel computing for $k = 1, \dots, K$

- Receive the global parameters $\left\{p^{(\omega)}\right\}$

- Update the additive model by $f^{(\omega)}(t) = f^{(\omega-1)}(t) + \nu h^{(\omega)}_{p^{(\omega)}}$

3.5 Simulation Studies

Chapter 3.5 validates the performance of the proposed fed-GB-LSA using simulated data before applying it for the real-world application in Chapter 3.6. Chapter 3.5.1 examines the prediction accuracy of the function-on-function regression model using the proposed fed-GB-LSA in comparison with the global and local models, and demonstrates that the proposed federated model's performance is comparable to the performance of the global model and much better than the performance of local models. Chapter 3.5.2 compares the prediction accuracy of the function-on-function regression model between two different FL methods, i.e., the proposed fed-GB-LSA and the conventional fed-GB-Average and shows that the proposed fed-GB-LSA outperforms the conventional fed-GB-Average in a challenging FL setting with heterogeneous data across local servers.

3.5.1 Performance of the fed-GB-LSA

There are $N = 1000$ observations generated for simulation study. Each observation contains 20 predictors, i.e., $P = 20$, with each predictor being generated following the equation as below:

$$x_{np}(t) = \sum_k c_{pk}\varphi_k(t) + \varepsilon(t) \text{ for } n = 1, \dots, N, p = 1, \dots, P, \text{ and } k = 1, \dots, K. \quad (3.16)$$

$\varphi_k(t)$ is the B-spline basis function and 20 basis functions are assumed, i.e., K = 20. $\varepsilon(t) = \sum_k e_k\varphi_k(t)$ is the noise term with e_k being randomly sampled from a normal distribution $N(0,1)$ for $k \in \{1, \cdots, 20\}$ and $t \in [0, 100]$. The coefficients of all 20 predictors form a $K \times P$ matrix, i.e., $\mathbf{C} = [\mathbf{C}_1, \cdots \mathbf{C}_P]$. Each element of $\mathbf{C}_P$ is randomly generated by $U(-1,1) + e^{N(0.1p,1)}$ for $p = 1, \dots, 20$, and $k = 1, \dots, 20$. Moreover, the

response variable $y_n(t)$ is generated from the predictors based on $y_n(t_i) = \Delta \sum_{p=1}^{P} \sum_{i'} x_p(s_{i'}) \beta_p(s_{i'}, t_i) + \varepsilon_n(t_i)$ with $\Delta = 1$. B-spline basis functions $\boldsymbol{\varphi}(\cdot) = \left(\varphi_1(\cdot), \ldots, \varphi_K(\cdot)\right)^T$ are also used as the basis system of $\beta_p(s,t)$, resulting in $\beta_p(s,t) = \boldsymbol{\varphi}(s)^T \mathbf{B}_p \boldsymbol{\varphi}(t)$. The coefficient matrix of all the predictors is denoted as $\mathbf{B} = [\mathbf{B}_1, \cdots, \mathbf{B}_{20}]^{\boldsymbol{T}}$. To test the sparse selection performance of the proposed method, the first five predictors are assumed to be effective with each element of the coefficient matrix following a normal distribution $N(1, 0.5)$, while the remaining 15 predictors are assumed to be dummy features with each element of the coefficient matrix being zero.

To examine the model's performance under different numbers of local servers, the generated dataset with 1000 subjects is randomly divided into k servers for $k = 1, \ldots, 48$. The Mean Absolute Percentage Error (MAPE) is chosen to be the performance measure of prediction error. The MAPE is defined as $MAPE = \frac{100\%}{NT} \sum_{n=1}^{N} \sum_{t=1}^{T} \left| \frac{Y_{nt} - F_{nt}}{Y_{nt}} \right|$, where Y_{nt} is the actual value of response n in the testing set evaluated at time t, F_{nt} is the corresponding prediction value evaluated at time t. The sampling period length T is 101, and the sample points are $\{0, 1, 2, \cdots 100\}$. Figure 4 summarizes the Cross-validated MAPEs obtained by the proposed federated model, the global model that can access the entire dataset, and the local models that can see its own data only. As shown in Figure 4, it is clear that the average of the local models indicated by the red color produces the highest MAPEs among the three methods, and this prediction error increases drastically when the number of local servers increases, and the sample size used for each local serve reduces. As expected, the global model indicated by the yellow color consistently achieves the lowest MAPEs. When the number of servers is limited, meaning that each local server has sufficient samples for estimation, the proposed federated model indicated by the green color is able to obtain

similar MAPEs to the global model. When the number of servers keeps increasing, the performance of the proposed model declines as expected but is still significantly better than local models.

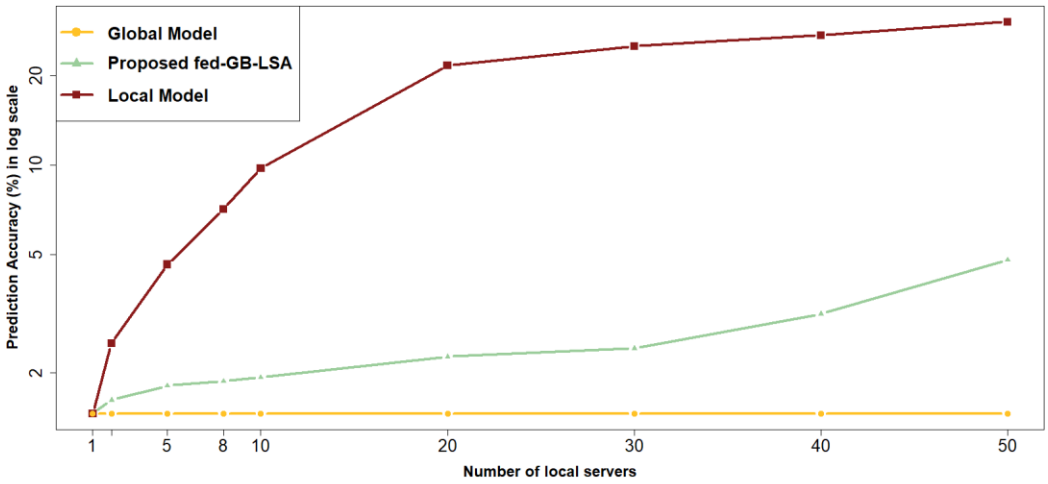

Figure 4: Compare the proposed federated model with both global and local models in cross-validated prediction errors for the simulation data with 20 replicates

3.5.2 Comparison with the fed-GB-Average

This subchapter aims to compare the proposed model with LSA with the conventional federated average model. Essentially, we have two different aggregated estimators, i.e., LSA-based estimator in (3.15) written as $\left(\sum_{k=1}^{K} \frac{N_k}{N} \widehat{\boldsymbol{\Sigma}}_{p,k}^{-1}\right)^{-1} \left(\sum_{k=1}^{K} \frac{N_k}{N} \widehat{\boldsymbol{\Sigma}}_{p,k}^{-1} \widetilde{\mathbf{B}}_{p,k}^{*}\right)$ and a simple averaged estimator written as $\sum_{k=1}^{K} \frac{N_k}{N} \widetilde{\mathbf{B}}_{p,k}^{*}$. In each GB iteration, since the proposed fed-GB-LSA is considered as a "one-shot" approach, its aggregated estimator is allowed to be computed only once, while the competing fed-GB-Average is allowed to iteratively refine the aggregated estimator until it converges. This is designed to confirm the efficacy of the proposed "one-shot" approach, compared with the federated averaging algorithm. A similar data generation

approach is adopted. Instead of fixing the total sample size N and changing the number of local servers K, Chapter 3.5.2 sets the number of samples per server as 100 but uses different number of local servers K with K = 2, 4, 6, 8, and 10.

Table 5 presents the comparison between fed-GB-LSA and fed-GB-Average with 4-fold Cross-validation across 20 replicates. The proposed fed-GB-LSA consistently outperforms the competing method with significantly lower MAPEs for different numbers of local servers. It is worth noting that the Standard Deviations of the predictions are slightly higher for the proposed method than the competing method, resulting in comparable worst-case MAPEs between the two methods. Overall, it achieves higher sensitivities and specificities in variable selection; The variable selection accuracies increase as the number of local servers increases and the proposed fed-GB-LSA reached a high sensitivity and specificity of 93% and 85%, respectively. While the computational runtime increases as the sample size increases, the proposed "one-shot" approach results in less computational runtime compared with Fed-GB-Average. In summary, compared with fed-GB-Average, the proposed fed-GB-LSA demonstrates an overall superior performance in terms of prediction accuracy and variable selection accuracy with much less computational and communicational costs.

Table 5: Comparison of fed-GB-LSA (LSA) and fed-GB-Average (Avg) with 20 replicates

	MAPE						Selection Accuracy				Computational Runtime (min)	
	Mean		SD		Worst Case		Sensitivity		Specificity			
	LSA	Avg	LSA	Avg	LSA	Avg	LSA	Avg	LSA	Avg	LSA	Avg
K=2	2.59	2.87	0.40	0.32	3.78	3.59	0.85	0.69	0.77	0.81	1.7	2.4
K=4	2.25	2.82	0.49	0.22	3.47	3.38	0.85	0.63	0.86	0.83	3.6	5.1
K=6	2.13	2.76	0.44	0.19	3.17	3.22	0.89	0.65	0.85	0.83	5.9	8.6
K=8	1.90	2.81	0.44	0.17	3.08	3.33	0.90	0.77	0.85	0.82	8.5	12.3
K=10	1.90	2.77	0.41	0.15	3.22	3.06	0.93	0.77	0.85	0.82	10.9	16.8

3.6 Application to Telemonitoring of OSA

This subchapter introduces the application of the proposed method on the data collected in the Sleep Heart Health Study (SHHS) (Zhang et al., 2018). SHHS is a popular study on OSA epidemiology and its health consequences in the United States. This study included 408 subjects and each subject contains 41 functional features with 13 and 28 features extracted from ECG and EEG signals, respectively, in addition to several non-functional patient-specific features including age, gender, BMI, and ethnicity. To prepare the functional features, each subject's ECG and EEG are aligned based on the time points and divided into consecutive 5-minute intervals from which ECG- and EEC-based features are extracted, resulting in multiple functional features to represent information within this interval. In total, the 310-minute ECG and EEG recording provides 63 longitudinal measures for each of the 408 subjects, i.e., $N = 408$ and $T = 63$. Specifically, ECG features are derived by the Heart Rate Variability (HRV) analysis (Vest et al., 2018; Goldberger et al., 2000) that reveals cardiac modulation in sleep by quantifying cardiovascular modulation under varying healthy and pathogenic conditions; EEG features

are based on the Power Spectral Density (PSD) analysis that decomposes the EEG signal into different frequency sub-bands and estimates the average spectral power of the sub-bands within each interval (Alramadeen et al., 2022). Note that PhysioNet Cardiovascular Signal Toolbox (https://physionet.org/content/pcst/1.0.0/) and National Sleep Research Resource Luna (https://zzz.bwh.harvard.edu/luna/) are used for feature extraction of ECG and EEG, respectively. Last but not least, the response variable of interest is defined as the frequency of adverse respiratory events occurring in each time interval. An overview of the features is depicted in Table 6.

Table 6: Description of variables included in the study

Variables		**Summary statistics**
Non-functional independent variables		
Age (Unit: year)		59.5 ± 10.7
Gender (Female: 0; Male: 1)		51.9% ± 50.0%
Ethnicity (Hispanic: 0; Non-Hispanic: 1)		0.9 ± 0.3
BMI (Unit: kg/m^2)		27.3 ± 3.9
Functional independent variables		
ECG features (13 variables)	AVNN: Average of all NN intervals (Unit: ms)	
	SDNN: Standard deviation of NN intervals (Unit: ms)	
	rMSSD: Square root of the mean of squares of the difference between adjacent NN intervals	
	pNNx: Percentage of differences between adjacent NN intervals that are greater than x ms. (x = 10, 20, 30, 40, and 50)	
	NN_RR: Ratio of consecutive normal sinus beats (NN) overall cardiac inter-beat (RR) intervals	
	VLF/LF/HF: Relative spectral power for very low frequency (0.003-0.04 Hz), low frequency (0.04-0.15 Hz), and High frequency (0.15-0.4 Hz)	
	LF_HF: Ratio of low to high-frequency power	
EEG features (28 variables)	Slow/Delta/Theta/Alpha/Sigma/Beta/Gamma Max/Min/Average/SD: Relative spectral power for Slow (0.5-1 Hz), Delta (1-4 Hz), Theta (4-8 Hz), Alpha (8-12 Hz), Sigma (12-15 Hz), Beta (15-30 Hz), and Gamma (30+Hz)	
Functional response variable		
DSI (Number of adverse events per interval)		

The proposed method is applied to the prepared SHHS dataset for DSI prediction from 4 non-functional and 41 functional features among the 408 subjects. The functional features are smoothed by the B-spline functions before being fed into the proposed method. MAPE is used to evaluate the prediction accuracy. The global model without FL results in 21.6% MAPE with 10-fold Cross-Validation. This prediction performance is considered satisfactory given the difficulty in predicting a functional response variable, which demonstrates the efficacy of GB in the estimation of the function-on-function regression. Moreover, the global model selects a few significant variables including Gender, AVNN, SDNN, pNN10, VLF, LF_HF, Alpha_Max, and Beta_Max among the 45 features.

Although the SHHS data is not collected in the FL setting, it is randomly divided into local servers to evaluate the proposed fed-GB-LSA with application to a real-world dataset. Specifically, the complete dataset with 408 subjects is randomly divided into a testing set with 41 subjects and a training set with 367 subjects. The training set is further randomly partitioned into several “local servers”. By dividing the training set into 1, 2, 5, 10, and 15 local servers, respectively, the sample size used for each local server gradually decreases. As shown in Figure 5, the proposed fed-GB-LSA’s prediction error increases as expected, but it is still drastically lower than the local models.

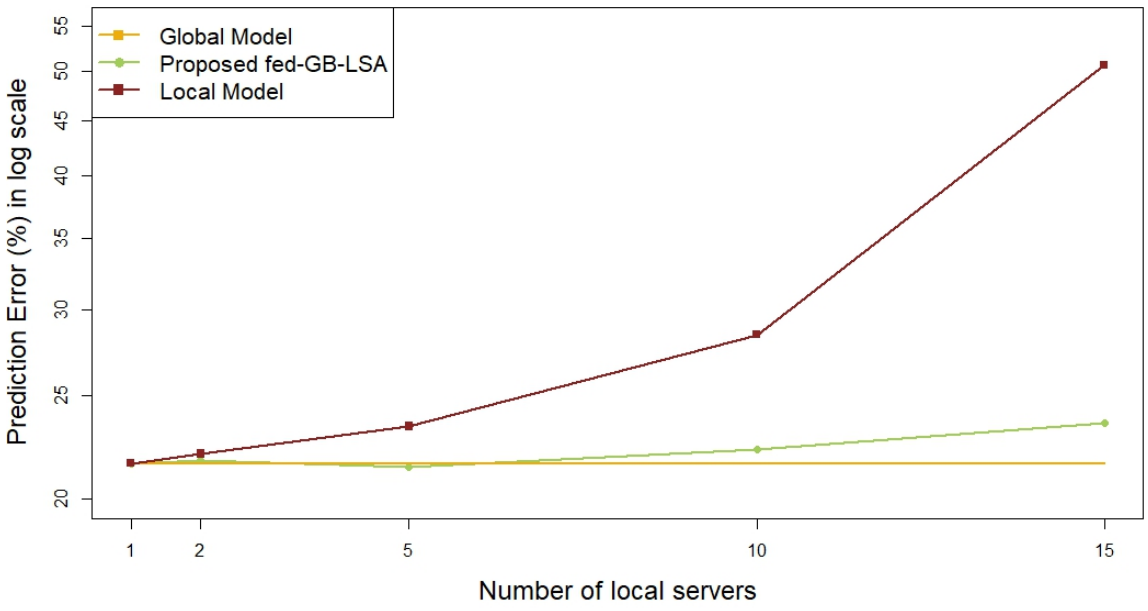

Figure 5: Compare the proposed federated model with both global and local models in prediction errors for the SHHS data

3.7 Conclusion and Discussion

Big data is often distributed in separate environments and thus is hard to combine due to data privacy and ownership concerns. For example, in the healthcare system, it is challenging to combine the datasets across hospitals because health data is highly sensitive, and its usage is tightly regulated. Consequently, Federated Learning (FL) receives more and more attention as it can leverage multi-source big data from local servers without data exchange for privacy-preserving training. Among other complicatedly structured data, functional data commonly exists in healthcare. The function-on-function regression aiming to predict a functional response from other non-function and functional variables is of great interest. Take the telemedicine of Obstructive Sleep Apnea (OSA) as an example. As a prevalent cardiac syndrome characterized by abnormal respiratory patterns during sleep, the diagnosis of OSA relies on an overnight recording of patients' multi-channel bio-signals, such as ECG and EEG, via wearable sensors, in addition to several patient-specific

characteristics. Such data is stored at separate sleep labs across different hospitals. Therefore, it is of clinical interest for a privacy-preserving model that automatically predicts the functional disease indicator from the multivariate function and non-functional variables. However, there is no existing research on FL of the function-on-function regression.

The major challenge in FL of the function-on-function regression is a lack of a "meaningful" approach to collaboratively train multiple local models and generate a federated estimator without data sharing. The conventional Federated Average model requires many rounds of communication which is communicationally costly. To address this, this study proposed the first-of-its-kind federated Gradient Boosting algorithm with the Least Squares Approximation (fed-GB-LSA) for efficient, privacy-preserving federated learning of the function-on-function regression. The methodological contribution of the proposed method is multi-fold. First, the GB-based algorithm allows a sparse selection of multivariate functional and non-functional features in the function-on-function regression model prediction, which tackles a long-standing challenge in functional regression. Second, the parameter estimation by the GB algorithm is efficient and results in separate sub-optimization problems with explicitly analytical solutions. Last but not least, the LSA provides a "one-shot" approach for FL that is proven to enjoy global asymptotic normality, which ensures communicational- and statistical efficiency.

The proposed "one-shot" fed-GB-LSA was tested in both simulation studies and a real-world dataset for OSA telemedicine, which demonstrates that the proposed federated model's performance is comparable to the performance of the global model and much better than the performance of local models. In a more challenging FL setting with

considerable heterogeneity across local servers, the proposed fed-GB-LSA significantly outperforms the conventional fed-GB-Average. The superior performance of the proposed method was also demonstrated in a real-world dataset for OSA telemedicine.

Chapter 4 Vertical Federated Functional Gradient Boosting with Differential Privacy

4.1 Introduction

Vertical Federated Learning (VFL) has been recognized as an instrumental mechanism that facilitates data collaboration among enterprises, which complies with strict privacy regulations such as the General Data Protection Regulation (GDPR) established by the European Union (Voigt & Von dem Bussche, 2017). Unlike Horizontal Federated Learning (HFL), where decentralized datasets largely share a common feature space with minimal overlap in sample space, VFL involves significant overlap in sample space coupled with diverse feature sets among different organizational participants.

This overlap necessitates a collaborative approach to model development in VFL, contrasting/which contrasts with HFL, where participants may independently develop models using their localized datasets. In VFL, the primary data protection strategy shifts from protecting gradients during their transmission to a central aggregator—secured through mechanisms such as differential privacy (Dwork, 2011; Dwork et al., 2006; Dwork and Roth, 2014; Sheffet, 2017; Lee et al., 2019) and secret sharing (Bonawitz et al., 2017)—to sharing intermediate results while maintaining control over local data.

In the healthcare sector, the application of VFL to the analysis of Electronic Health Records (EHR) from multiple hospitals exemplifies its potential. In particular, in the context of health telemonitoring for Obstructive Sleep Apnea (OSA), employment of a function-on-function regression model is advantageous. OSA, a condition marked by

abnormal respiratory patterns during sleep, is typically diagnosed by analyzing overnight multi-channel bio-signal recordings such as electrocardiograms (ECG) and electroencephalograms (EEG), which are collected using wearable sensors (Alramadeen et al., 2023). These recordings are manually examined by technicians to determine the frequency of adverse respiratory events, which is a labor-intensive and time-consuming process.

Employment of a function-on-function regression model in a VFL framework can autonomously predict the frequency of these events across all epochs from the bio-signal features extracted within the same epochs. This ability not only streamlines the diagnostic process but also improves the accuracy of clinical assessments at the population level. Despite its evident suitability, there is a notable absence of research on the application of VFL to function-on-function regression models. This gap in the research represents a significant opportunity to advance the efficient and effective diagnosis and management of conditions such as OSA. By leveraging VFL, hospitals can collaboratively develop models that generalize across diverse EHR data, which thereby enhances clinical decision-making while adhering to strict data privacy regulations. This method preserves data sovereignty for each participant and leverages collective insights from shared analyses, which underscores VFL's potential to transform data-sensitive sectors such as healthcare.

A major challenge in the meaningful implementation of federated learning for any machine learning model lies in ensuring that the federated model performs satisfactorily, particularly as the implementation of differential privacy can significantly degrade performance. This project contributes the first-of-its-kind Vertical Federated Learning

Functional Regression with Gradient Boosting, an approach designed for efficient, privacy-preserving federated learning of function-on-function regression models.

The rest of this chapter is organized as follows: Chapter 4.2 reviews the relevant work; Chapter 4.3 presents the model formulation of the Vertical Federated Learning function-on-function regression model with Gradient Boosting; Chapter 4.4 presents the simulation studies to evaluate the empirical performance of the proposed method with respect to the prediction accuracy and privacy-preserving; Chapter 4.5 conducts a case study in Obstructive Sleep Apnea (OSA) research. Chapter 4.6 concludes this chapter.

4.2 Literature Review

4.2.1 Functional Regression

In recent years, there has been an increase in the collection of large datasets, which are gathered either continuously or at predetermined intervals. These datasets are categorized as "functional data," and their analysis is becoming increasingly common. Functional Data Analysis (FDA) focuses on creating statistical methods specifically designed to analyze this type of data, which has a particular emphasis on functional regression. Studies by Ramsay and Silverman (2005), Ferraty (2006), Bosq (2012), Horvath and Kokoszka (2012), and Hsing and Eubank (2015) have reviewed various FDA techniques.

Several models have been developed to address different types of functional data relationships: scalar-on-function regression involves scalar responses and functional predictors; function-on-scalar regression deals with functional responses and scalar predictors; and function-on-function regression involves both functional responses and predictors. In scalar-on-function regression, Ramsay and Dalzell's (1991) integrated

generalized linear modeling and principal component analysis using L-spline theory applied to random function data. Brown et al. (2001) developed a model using wavelet coefficients and Bayesian variable selection techniques, while Ratcliffe et al. (2002) explored binary response modeling with functional and additional scalar covariates. Ramsay and Silverman (2005) further refined scalar-on-function regression through penalized least squares estimation and basis expansion. Reiss and Ogden (2007) introduced a functional version of principal component regression and partial least squares, and Goldsmith et al. (2012) extended the generalized linear mixed model to incorporate functional predictors. Additionally, Yao and Müller (2010), and McLean et al. (2014) introduced nonlinear approaches to scalar-on-function regression. In function-on-scalar regression, Guo et al. (2003) presented a smoothing spline ANOVA model, Lin et al. (2004) demonstrated that smoothing spline estimators are asymptotically equivalent to kernel estimators, and Reiss et al. (2010) developed a generalized ridge regression estimator through penalized generalized least squares. In function-on-function regression, Ramsay and Silverman (2005) formulated the bivariate coefficient function as a double expansion of basis functions. Yao and Müller (2005) and Wu and Müller (2011) explored a specific double expansion using eigenfunctions of the covariance functions of functional covariates and predictors. Ivanescu et al. (2015) treated multivariate function-on-function regression as a penalized additive model, while Luo and Xi (2017) approached the problem using eigenfunctions and solved it as a penalized generalized functional eigenvalue problem. Ding et al. (2019) proposed a semi-parametric model for degradation curves analysis.

4.2.2 Gradient Boosting

In many machine learning scenarios, the task often revolves around developing non-parametric models for regression or classification, particularly when specific expert-driven models are not feasible. Non-parametric methods, which include neural networks and support vector machines, become essential because they define the relationships between input variables without fixed assumptions. A commonly used strategy in data-driven modeling is the ensemble approach, which involves combining numerous simpler, weaker models to form a stronger, more accurate ensemble prediction. Boosting is a notable technique within this framework, which is characterized by its incremental assembly of models. With each iteration, a new, initially weak base learner model is introduced that is designed to minimize the previous errors of the ensemble, which thereby refines the prediction progressively.

The development of boosting methods that utilize gradient descent was formalized in key works by Freund and Schapire (1997), Friedman et al. (2000), and Friedman (2001), which led to the creation of gradient boosting machines. This methodology not only provided a mathematical framework but also rationalized the choice of model hyperparameters, which thus establishes a solid foundation for the ongoing development of gradient boosting models. In recent years, gradient-boosting-based algorithms have demonstrated significant efficacy across various domains. Notably, XGBoost or eXtreme Gradient Boosting (Chen and Guestrin, 2016) has performed well in numerous Kaggle competitions. LightGBM (Ke et al., 2017) is recognized for its efficiency in training speed, and CatBoost (Prokhorenkova et al., 2018) has been effective in enhancing generalization accuracy. Brockhaus et al. 2017 proposed a gradient boosting algorithm that takes several functional regression models as the base learners.

4.2.3 Vertical Federated Learning

Federated Learning (FL) has advanced as a prominent model in the field of distributed learning, notably for its contribution to data privacy preservation. This model is differentiated into three primary forms based on the data distribution method among participating entities: Horizontal, Vertical, and Hybrid. Vertical Federated Learning (VFL) is particularly applicable where data across entities is segmented vertically. In typical scenarios, one entity holds the outcomes, or labels, alongside specific attributes of data points, while other entities contribute additional attributes for the same data points, which thereby maintains data confidentiality and prevents exposure to other participants. VFL finds its utility in various sectors, such as collaborations between telecommunications firms and entertainment service providers, or between airlines and car rental companies, which facilitates the enhancement of service quality and customer experience through protected data insights.

The architecture of Vertical Federated Learning can be structured with or without a coordinator. In the coordinated model (Hardy et al., 2017), a central coordinator oversees the training process via encrypted communications with the participants and does not access the raw data. Typically, this coordinator role is assumed by an active participant or a trusted third party (Li et al., 2023). In contrast, the architectures proposed by He et al. (2021) and Sun et al. (2022) dispense with the coordinator, which thereby reduces the system's complexity. This framework has been extended to include multiple collaborating parties (Cheng et al., 2021; Zhao et al., 2022), where participants exchange public keys and intermediate results while keeping the computations of gradients, loss functions, and model updates localized.

According to Li et al. (2023), VFL can be classified into non-split VFL, split VFL, and customized VFL, depending on the extent of model sharing. In non-split VFL, each participant possesses a complete model and calculates gradients based on their local data and intermediate information shared by others. This configuration can further be divided into systems with a coordinator (Jin et al., 2021; Benmalek et al., 2022; Sun et al., 2021) and those without (Zhang 2021; Zhang & Jiang, 2022; Liu et al., 2020). In the split VFL configuration (Zhang et al., 2021; Hashemi et al., 2021; Kang et al., 2020), the model is partitioned into a top model and several bottom models, in which each party, whether active or passive, maintains their respective bottom models. The division typically occurs at a fully connected layer. For instance, in one configuration, parties A and B may manage three and four features, respectively, in which each operates a part of the network. In customized VFL, the algorithm involves a super participant who holds the label and other passive participants who hold different features to jointly train the best classification tree model (Wu et al. 2020, Luo et al. 2021, Tan et al. 2020). Li et al. (2023) and Wang et al. (2023) also provide the applications of cybersecurity on IoT systems.

4.2.4 Differential Privacy

In the context of Vertical Federated Learning (VFL), addressing privacy concerns is crucial, which necessitates the design of schemes that effectively safeguard participant privacy. Differentially Private (DP) data release is a promising technique to disseminate data without compromising the privacy of data subjects. Differential privacy (Dework, 2011, Dework et al. 2006, Dework and Roth, 2014, Sheffet. 2017, Lee et al. 2019) is shown to be a promising direction to release datasets while protecting individual privacy. It is now widely accepted as a strong and rigorous notion of data privacy. It has received acclaim in

theory, which won the 2017 Gödel Prize and the 2016 TCC Test-of-Time Award. At the same time, it has now been seen in practice in many organizations, including Apple, Google, Microsoft, the US Census Bureau, and many more.

Differential Privacy (DP) is a protocol that mitigates the risk of privacy breaches during learning tasks by adding noise to the original data or results, though it must be carefully calibrated to avoid significant performance degradation or insufficient privacy protection. Wang et al. (2020) developed a DP-based algorithm that maintains data confidentiality for VFL participants, which achieves outcomes comparable to those obtained in non-private VFL settings that use generalized linear models. Additionally, Xu et al. (2021) proposed a multi-party learning framework for vertically partitioned datasets that achieves differential privacy by adding noise directly to the objective function, which requires only a single round of noise addition and secure aggregation.

Beyond model training, DP can also be applied during the model evaluation phase to protect against the leakage of private label information. Sun et al. (2022) proposed two algorithms that allow for accurate computation of the AUC (area under curve) metric using the label DP (Ghazi et al. 2021) in VFL. Moreover, Tian et al. (2020) and Li et al. (2022) have applied differentially private noise to federated gradient-based decision trees in customized ways to achieve an optimal privacy-utility trade-off. Chen et al. (2022) integrated Graph Neural Networks (GNN) into the split VFL framework, which utilizes DP-enhanced additive secret sharing to enhance data privacy.

4.3 Proposed Method

The functional observation dataset is $\mathcal{D} = \{\boldsymbol{y}, \mathbf{X}\}$, in which $\boldsymbol{y} = \{y_1(t), y_2(t), \cdots, y_N(t)\}^T$ and $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_P\}$ where $\boldsymbol{x}_p =$

$\{x_{1p}(t), \cdots, x_{np}(t), \cdots, x_{Np}(t)\}^T$ for $n = 1, \ldots, N$. N denotes the number of observations and P denotes the number of predictors. The sampling period is T so that $t \in T$. Using the framework of Vertical Federated Learning without coordinator, we assume there are $K + 1$ parties with $K \geq 1$. The party that holds the functional response $\{\boldsymbol{y}\}$ is called **active party**. All the other parties who hold the functional predictor are called **passive parties**. Each functional covariate vector $\boldsymbol{x}_p = \{x_{1p}(t), \cdots, x_{Np}(t)\}^T$ in $\mathcal{D}$ is distributed among K passive parties $\left\{\{\boldsymbol{x}_{p_k} \in \mathbb{R}^{N \times T}\}_{p_k \in s_k}\right\}_{k=1}^{K}$, where s_k is the set of predictors that stored in party k. The size of s_k is $|s_k| = P_k$, for $k \in \{1, 2 \cdots, K\}$, such that $\sum_{k=1}^{K} p_k = P$. Then we have $\bigcup_{i=1}^{K} s_k = \{1, 2 \cdots, P\}$ and $s_i \cap s_j = \emptyset$ for any $i, j \in \{1, 2 \cdots, K\}$ and $i \neq j$.

In this chapter, we consider the following function-on-function linear regression model:

$$y_n(t) = \sum_{p=1}^{P} \int_{s \in T} x_{np}(s)\, \beta_p(s,t) ds + \varepsilon_n(t), \tag{4.1}$$

where $\beta_p(s,t)$ is the bivariate coefficient function for the p-th functional predictor, and $\varepsilon_n(t)$ is the random error function that follows a normal distribution. Model (4.1) assumes that both the functional response and the predictors are centered. Centering refers to adjusting each function so that its mean over the observed domain T is zero. This is a standard practice in functional data analysis because it normalizes the data, which ensures that all functional observations have a common baseline.

4.3.1 Functional Regression with Gradient Boosting

In this subchapter, we introduce the methodology for solving model (4.1) via gradient boosting (GB), which specifically excludes considerations of vertical federated

learning. Gradient boosting constructs the predictive model through sequential aggregation of base leaners with the form as

$$\mathbb{E}(y_n(t)|\boldsymbol{x}_N) = \sum_{m=1}^{M} h_m(t), \tag{4.2}$$

in which $h_m(t)$ is the additive effect that depends on one functional predictor $\boldsymbol{x}_p$, $p \in \{1, 2 \cdots, P\}$. M is assumed to be the number of additive effects.

Equation (4.2) can be solved iteratively. For each iteration, we have the candidate base learners as $\{h^p(t)\}_{p=1}^{P}$, in which $h^p(t) = \int_{s\in T} x_{np}(s)\,\beta_p(s,t)ds$. Each candidate base learner models the historical effect from the corresponding functional predictor, i.e., $x_{np}(s)$.

We assume the bivariate coefficient function $\beta_p(s,t)$ has a double expansion on one basis system $\boldsymbol{\theta}$ with K_1 functions and another basis system $\boldsymbol{\eta}$ with K_2 functions, i.e., $\beta_p(s,t) = \boldsymbol{\theta}(s)^T\mathbf{B}_p\boldsymbol{\eta}(t)$ in which $\boldsymbol{\theta}(t) = \left(\theta_1(t), \dots, \theta_{K_1}(t)\right)^T$, $\boldsymbol{\eta}(t) = \left(\eta_1(t), \dots, \eta_{K_2}(t)\right)^T$, and $\mathbf{B}_p \in \boldsymbol{R}^{K_1\times K_2}$. Using a $1 \times K_1$ row vector $\boldsymbol{z}_{np} = \int_{s\in T} x_{np}(s)\,\boldsymbol{\theta}(s)^T ds$ to precalculated n-th observation of the p-th predictor and the selected basis functions, candidate base learners can be rewritten as $h^p(t) = \boldsymbol{z}_{np}\mathbf{B}_p\boldsymbol{\eta}(t)$.

Given any functional regression model $f(t, \boldsymbol{z}_n)$, the empirical loss function for the functional response and regression model is formulated as $l(\boldsymbol{y}, f|\mathrm{X}) = \sum_{n=1}^{N}\int_{t\in T}(y_n(t) - f(t,\boldsymbol{z}_n))^2 dt$, which leads GB to solve the following optimization problem:

$$f^* = \underset{f}{\operatorname{argmin}}\, l(\boldsymbol{y}, f|\mathbf{X}) \tag{4.3}$$

In the ω-th iteration of GB, $\omega \in \{1, \cdots, M\}$, the algorithm computes the negative gradient of the risk function with respect to the current model estimation f, i.e., $\boldsymbol{u}^{(\omega)} \in$

$R^{N\times 1} = -\frac{\partial l}{\partial f}\Big|_{f=\hat{y}_n^{[\omega-1]}}$. Subsequently, GB regresses the negative gradient $\boldsymbol{u}^{(\omega)}$ to each of the P candidate base learners by addressing the following optimization problems:

$$\widehat{\mathbf{B}}_p^{(\omega)} = \underset{\mathbf{B}_p}{\operatorname{argmin}} \sum_{n=1}^{N} \int_{t\in T} \left(u_n^{(\omega)}(t) - h^p(t)\right)^2 dt. \tag{4.4}$$

Here, $u_n^{(\omega)}(t)$ is the n-th component of $\boldsymbol{u}^{(\omega)}$. The optimal solution from problem (4.4) facilitates the derivation of the fitted base learners $\left\{\hat{h}^p(t) = \boldsymbol{z}_{np}\widehat{\mathbf{B}}_p^{(\omega)}\boldsymbol{\eta}(t)\right\}_{p=1}^{P}$. Among these fitted base learners, GB selects the one that minimizes the Residual Sum of Squares (RSS):

$$RSS_p = \sum_{n=1}^{N} \int_{t\in T} \left(u_n^{(\omega)}(t) - \hat{h}^p(t)\right)^2 dt \text{ for } p = 1, \dots, P. \tag{4.5}$$

Thus, $h_m(t)$ in equation (4.2) is the fitted base learner with minimal RSS of (4.5) in the ω-th iteration. GB updates the model by $f^{(\omega)}(t) = f^{(\omega-1)}(t) + \nu h_m(t)$, in which ν is the preset learning rate. The GB algorithm continues to iterate until a predefined stopping criterion is met, i.e., reaching the maximum number of iterations M. Table 7 below summarizes the implementation steps of gradient boosting to solve the problem (4.3), which do not account for Vertical Federated Learning conditions.

Table 7: Pseudo code for the Gradient Boosting Functional Regression

Import: $\mathcal{D} = \{y_n(t), x_{n1}(t), \dots, x_{nP}(t)\}_{n=1}^{N}$;
Initialization: $f^{(0)}(t) = \frac{1}{N}\sum_{n=1}^{N} y_n(t)$; $\omega = 0$. Set stopping threshold m and learning rate ν
Iterate until $\omega = M$:
$\omega = \omega + 1$;
• Compute the negative gradient $\boldsymbol{u}^{(\omega)}$ given $f^{(\omega-1)}(t)$ and $\mathcal{D}$ • Fit $\boldsymbol{u}^{(\omega)}$ with base learners $h^p(t)$ by solving problem (4.4) to estimate $\widehat{\mathbf{B}}_p^{(\omega)}$ for $p = 1, \dots, P$ • Compute the Residual Sum of Squares (RSS) by (4.5) for $p = 1, \dots, P$ • Select best base learner $h_m(t)$ with minimal RSS among $\{\hat{h}^p(t)\}_{p=1}^{P}$ • Update the additive model by $f^{(\omega)}(t) = f^{(\omega-1)}(t) + \nu h_m(t)$

4.3.2 Vertical Federated Functional Regression with Gradient Boosting

In this chapter, we explore the application of gradient boosting to solve the function-on-function regression problem (4.1) within the framework of Vertical Federated Learning (VFL), which particularly emphasizes privacy protection. VFL allows various parties, which are referred to as active or passive, to collaboratively learn a predictive model without necessitating a central coordinator. This decentralized setup, while beneficial in many respects, inherently risks the exposure of sensitive data. This includes both raw and intermediate data exchanged between parties, which can lead to significant privacy breaches.

To address this substantial concern, we incorporate Differential Privacy (DP) into our VFL framework. Initially conceptualized by Dwork et al. in 2006 and subsequently elaborated upon in further studies (Dwork, 2011; Dwork and Roth, 2014), DP introduces a robust method to safeguard individual data points within a dataset. By employing DP, we establish mathematical guarantees that limit the information about any single data entry

that can be inferred by participation in the federated model. This approach quantifies and confines the extent of privacy leakage during the collaborative learning process, which thus preserves the confidentiality of sensitive information across distributed data sources in the absence of centralized control.

Definition 4.1 ((ε, δ)-differential privacy). A randomized mechanism $\mathcal{M}: \mathcal{D} \to \mathcal{R}$ with domain $\mathcal{D}$ and range $\mathcal{R}$ satisfies (ε, δ)-differential privacy if for any two adjacent datasets $D, D' \in \mathcal{D}$, which differ at exactly one data point; for any subset of output $S \subseteq \mathcal{R}$, it holds that

$$\mathbb{P}[\mathcal{M}(D) \in S] \leq e^{\varepsilon}\mathbb{P}[\mathcal{M}(D') \in S] + \delta.$$

Definition 4.1 articulates a quantitative measure to evaluate the risk associated with privacy leakage, which utilizes parameters ε and δ. Lower values of these parameters signify a higher degree of privacy protection, whereas increasing values indicate a weakening of these protections. In the domain of differential privacy, the Gaussian Mechanism (Dework and Roth, 2014) is a post-hoc mechanism that converts a deterministic real-valued function $f: \mathcal{D} \to \mathbb{R}^m$ to a randomized algorithm with differential privacy guarantee. It relies on sensitivity of f, as given by Definition 4.2.

Definition 4.2. Let $f: \mathcal{D} \to \mathbb{R}^m$. The ℓ^2-sensitivity of f is

$$S_f = \max_{\mathcal{D},\mathcal{D}'} \|f(\mathcal{D}) - f(\mathcal{D}')\|_2,$$

where $\mathcal{D}, \mathcal{D}'$ are any two adjacent datasets.

This sensitivity quantifies the greatest potential change in the output of f resulting from a single alteration in its input dataset $\mathcal{D}$. In response to this measure, the Gaussian Mechanism applies noise, sourced from a Gaussian distribution, at a magnitude proportional to the sensitivity. The distribution's parameters are strategically chosen based

on ε and δ, which ensure compliance with differential privacy standards. This strategic application of noise acts as a safeguard, which substantially mitigates the risk of privacy breaches through controlled data obfuscation. Given Definition 4.2, we have Gaussian Mechanism in Lemma 4.1.

Lemma 4.1 (Gaussian Mechanism). For any deterministic real-valued function $f: \mathcal{D} \rightarrow \mathbb{R}^m$ with sensitivity S_f, we can define a randomized function by adding Gaussian noise to f:

$$f^{dp} := f(\mathcal{D}) + \boldsymbol{r},$$

where $\boldsymbol{r}$ is sampled from a multivariate normal distribution $\mathcal{N}(0, S_f^2\sigma^2 \cdot \mathbf{I})$. When $\sigma \geq \frac{\sqrt{2\log(1.25/\delta)}}{\varepsilon}$, f^{dp} is (ε, δ)-differential privacy for $0 < \varepsilon \leq 1$ and $\delta > 0$.

Proof of Lemma 4.1 can be found in (Dwork and Roth, 2014). For simplicity, denote $\sigma_{\varepsilon,\delta} = \frac{\sqrt{2\log(1.25/\delta)}}{\varepsilon}$ in the following paragraphs.

In the iterations of Algorithm in Table 7, active party computes the negative gradient is $\boldsymbol{u}^{(\omega)} = (u_1^{(\omega)}, \cdots, u_N^{(\omega)})^T$, based on the model ensemble formed in the previous iteration. Subsequently, each passive party engages in regression on this negative gradient that utilizes its designated base learners. These procedural steps, however, pose a risk of contravening the stringent privacy stipulations inherent in Vertical Federated Learning (VFL) because the negative gradient may include sensitive private information.

To mitigate this privacy concern, Differential Privacy (DP) is integrated into each iteration of the gradient boosting process. This integration ensures that the disclosure of the negative gradient adheres to privacy-preserving protocols. Specifically, DP mechanisms are employed to introduce a carefully calibrated amount of noise to the

negative gradient before it is shared among the parties. This approach allows the continuation of collaborative learning under the VFL framework while effectively safeguarding sensitive information against unauthorized disclosure, which thus aligns the process with the required privacy standards.

In iteration ω of gradient boosting in Table 7, we use basis system $\boldsymbol{\eta}$ to express $u_n^{(\omega)}$ of $\boldsymbol{u}^{(\omega)}$ as $u_n^{(\omega)}(t) = \boldsymbol{c}_n^T\boldsymbol{\eta}$, where $\boldsymbol{c}_n \in \mathbb{R}^{Q_2}$ as Q_2 is the number of basis functions in $\boldsymbol{\eta}$. Thus, the sensitive information of $u_n^{(\omega)}$ is captured by $\boldsymbol{c}_n$ given basis system $\boldsymbol{\eta}$. The coefficient matrix of $\boldsymbol{u}^{(\omega)}$ is $\mathbf{C} = (\boldsymbol{c}_1, \cdots, \boldsymbol{c}_N)^T \in \mathbb{R}^{N\times Q_2}$. To bound the elements of $\mathbf{C}$, we clip it as

$$\tilde{\mathbf{C}} = \frac{\mathbf{C}}{\|\mathbf{C}\|_{\max}}. \tag{4.6}$$

The sensitivity of $\tilde{\mathbf{C}}$ is $S_{\tilde{\mathbf{C}}} = \max_{\tilde{\mathbf{C}},\tilde{\mathbf{C}}'}\left\|\tilde{\mathbf{C}} - \tilde{\mathbf{C}}'\right\|_2 = 2Q_2^{1/2}$. Given $\tilde{\mathbf{C}}$ as the coefficient matrix of negative gradient, the solution of problem (4.4) for predictor p is

$$vec(\mathbf{B}_1) = \left(\mathbf{J}_{\eta\eta} \otimes \mathbf{Z}_p^T\mathbf{Z}_p\right)^{-1} vec\left(\mathbf{Z}_p^T\tilde{\mathbf{C}}\mathbf{J}_{\eta\eta}\right), \tag{4.7}$$

in which $\mathbf{J}_{\eta\eta} = \int_{t\in T}\boldsymbol{\eta}(t)\boldsymbol{\eta}^T(t)\mathrm{d}t$, $\mathbf{Z}_p = (\boldsymbol{z}_{1p}, \cdots, \boldsymbol{z}_{Np})^T \in \mathbb{R}^{N\times Q_1}$, and $\otimes$ represents Kronecker product.

To ensure compliance with the privacy-preserving properties specified by (ε, δ)-differential privacy, the release of negative gradient from active party in each iteration should meet the differential privacy requirement as Definition 4.1. The implementation of differential privacy involves the strategic injection of noise into the negative gradient before its dissemination. This noise addition is meticulously calibrated based on the sensitivity of the gradient function and the parameters. Applying the Gaussian Mechanism

as outlined in Lemma 4.1 to the matrix $\tilde{\mathbf{C}}$, we have the differential privacy-enabled release matrix:

$$\mathbf{C}^{\mathrm{dp}} = \tilde{\mathbf{C}} + \mathbf{R}, \tag{4.8}$$

in which $\mathbf{R} \in \mathbb{R}^{N \times Q_2}$. The entries within $\mathbf{R}$ are independently and identically distributed (i.i.d.), each sampled from a Gaussian distribution $\mathcal{N}(0, 4Q_2\sigma_{\varepsilon,\delta}^2)$. Thus, the corresponding solution of $\mathbf{C}^{dp}$ is

$$vec(\mathbf{B}_2) = \left(\mathbf{J}_{\eta\eta} \otimes \mathbf{Z}_p^T\mathbf{Z}_p\right)^{-1} vec\left(\mathbf{Z}_p^T\mathbf{C}^{\mathrm{dp}}\mathbf{J}_{\eta\eta}\right). \tag{4.9}$$

For simplicity, we drop the subscript p in $\mathbf{Z}_p$ in the further analysis.

The asymptotic optimal of $vec(\mathbf{B}_2)$ in equation (4.9) is implied by Theorem 4.1 below.

Theorem 4.1. Given $\mathbf{Z}$, and $\mathbf{J}_{\eta\eta}$, for any given β, we have

$$\mathbb{P}[\|vec(\mathbf{B}_2) - vec(\mathbf{B}_1)\| > \beta] \leq O\left(\exp\left(-\frac{N^2\beta^2}{8Q_2^2\sigma_{\varepsilon,\delta}^2\|\mathbf{Z}\|_F^2}\right)\right).$$

The norm $\|\cdot\|$ denote ℓ^2 norm.

(*Due to the limitation of space, the detailed proof is listed in Appendix J.*)

Given Theorem 4.1, we can further derive Corollary 4.1, which indicates $vec(\mathbf{B}_2)$ converges in probability to $vec(\mathbf{B}_1)$.

Corollary 4.1. Given $\mathbf{Z}$, and $\mathbf{J}_{\eta\eta}$, we have

$$\operatorname*{plim}_{n\to\infty} vec(\mathbf{B}_2) = \operatorname*{plim}_{n\to\infty} vec(\mathbf{B}_1).$$

(*Due to the limitation of space, the detailed proof is listed in Appendix K.*)

Theorem 4.1 and Corollary 4.1 provide the properties of the proposed differential privacy estimator presented in equation (4.9), in contrast to the estimator in equation (4.7). These results suggest that the influence of the Gaussian Mechanism, as described in

equation (4.8), diminishes as the sample size increases. In addition to equation (4.8), we limit sensitivity by clipping the coefficients of the negative gradient as detailed in equation (4.6). To examine the effect of this clipping, we consider the estimator without clipping, expressed as:

$$vec(\mathbf{B}_3) = \left(\mathbf{J}_{\eta\eta} \otimes \mathbf{Z}_p^T \mathbf{Z}_p\right)^{-1} vec\left(\mathbf{Z}_p^T \mathbf{C} \mathbf{J}_{\eta\eta}\right). \tag{4.10}$$

The fitted base learner based on the estimator in equation (4.10) is given by: $h_3(t) = \mathbf{Z}_p \mathbf{B}_3 \boldsymbol{\eta}(t)$. In comparison to the fitted base learner based on the estimator in equation (4.7), which is $h_2(t) = \mathbf{Z}_p \mathbf{B}_2 \boldsymbol{\eta}(t)$, we have $h_3(t) = \|\mathbf{C}\|_{\max} h_2(t)$. To maintain the performance consistency of gradient boosting, we dynamically adjust the step length in each iteration $\omega \in \{1, \cdots, M\}$ to counteract the impact of clipping as:

$$\eta^{(\omega)} = \nu \|\mathbf{C}\|_{\max}. \tag{4.11}$$

Furthermore, since the negative gradient is clipped before being broadcast to each passive party, the active party retains the information about $\|\mathbf{C}\|_{\max}$, Consequently, equation (4.11) adheres to the privacy-preserving requirement.

After each passive party has conducted regression on the released negative gradient using their designated base learners, they compute the Residual Sum of Squares (RSS) from this regression as equation (4.5). These RSS values are then transmitted to the active party. Subsequently, the passive party whose base learner has demonstrated the most effective performance, as indicated by the lowest RSS, sends the estimated coefficients of their model to the active party. These steps ensure the collaborative updating of the model while minimizing data exchange. The privacy concerns associated with these two data transmissions are highlighted by Lemma 4.2.

Table 8 below outlines the implementation details of Vertical Federated Learning Functional Regression with Gradient Boosting.

Table 8: Pseudo code for Vertical Federated Learning Functional Regression with Gradient Boosting on active and passive parties

Import: $\mathcal{D} = \left\{\boldsymbol{y}, \left\{\boldsymbol{x}_p(t) \in \mathbb{R}^{N\times p_k\times T}\right\}_{k=1}^{K}\right\}$;
Initialization: Active Party $f^{(0)}(t) = \frac{1}{N}\sum_{n=1}^{N} y_n(t)$; $\omega = 0$; Set stopping threshold m
Iterate until $\omega = M$:
$\omega = \omega + 1$;
Active Party • Compute the negative gradient $\boldsymbol{u}^{(\omega)}$ given $f^{(\omega-1)}(t)$ and $\boldsymbol{y}$ • Clip the negative gradient coefficient as (4.6) • Release the negative gradient with DP $\mathbf{C}^{(\omega,dp)}\boldsymbol{\eta}$ as (4.8) to each passive party Passive Parties with parallel computing for $k = 1, \cdots, K$ • Regress negative gradient $\mathbf{C}^{(\omega,dp)}\boldsymbol{\eta}$ on each base learner $h_p(t)$ in $\{h_p(t)\}_{p\in s_k}$ as (4.9) • Compute $\left\{RSS_p^{(k,\omega)}\right\}_{p\in s_k}$ as (4.5) and send to the active party Active Party • Select the best base learner p^* with minimal RSS Passive Party • Send $h_{p^*}^{(\omega)}(t)$ to active party Active Party • Compute $\eta^{(\omega)}$ as (4.11) • Update the additive model by $f^{(\omega)}(t) = f^{(\omega-1)}(t) + \eta^{(\omega)}h_{p^*}^{(\omega)}(t)$

Lemma 4.2 (Post-Processing). If $M: \mathcal{X} \to \mathcal{Y}$ is (ε, δ)-differential privacy and $W: \mathcal{Y} \to \mathcal{Z}$ is any randomized function, then the algorithm $W \circ M$ is (ε, δ)-differential privacy.

Proof of Lemma 4.2 can be found in Dwork et al. 2006. By Lemma 4.2, we have the releases of $\left\{RSS_p^{(k,\omega)}\right\}_{p\in s_k}$ and $h_{p^*}^{(\omega)}$ in Algorithm in Table 8 are (ε, δ)-differential privacy. Thus, all data transmissions in Algorithm in Table 8 are (ε, δ)-differential privacy.

4.4 Simulation Studies

There are $N = 6{,}250$ observations generated for the simulation study. Each observation contains 20 predictors, i.e., $P = 20$, with each predictor being generated following the equation as below:

$x_{np}(t) = \sum_q a_{pq}\varphi_q(t) + \varepsilon(t)$ for $n = 1, \dots, N, p = 1, \dots, P$, and $q = 1, \dots, Q$.

$\varphi_k(t)$ is the B-spline basis function and 20 basis functions are assumed, denoted as Q = 20. $\varepsilon(t) = \sum_q e_q\varphi_q(t)$ is the noise term with e_k being randomly sampled from a normal distribution $N(0, 1)$ for $q \in \{1, \cdots, 20\}$ and $t \in [0, 100]$. The coefficients of all 20 predictors form a $Q \times P$ matrix, i.e., $\mathbf{A} = [\mathbf{A}_1, \cdots \mathbf{A}_{\mathrm{P}}]$.

Assume there are two passive parties, and each passive party will have 10 predictors. To distinguish between predictors and passive parties, the coefficient matrix, $\mathbf{A}_{p_k}$, of each predictor is randomly generated by $N(k, 0.5) + e^{N(0.1p_k, 0.1k)}$ for $p_k = 1, \dots, 10$, $k = 1, 2$, in which k indicates the passive party to which this predictor belongs.

The response variable $y_n(t)$ is generated from the predictors based on $y_n(t_i) = \Delta \sum_{p=1}^{P} \sum_{i'} x_p(s_{i'})\beta_p(s_{i'}, t_i) + \varepsilon_n(t_i)$ with $\Delta = 1$. B-spline basis functions $\boldsymbol{\varphi}(\cdot) = \left(\varphi_1(\cdot), \dots, \varphi_Q(\cdot)\right)^T$ are also used as the basis system of $\beta_p(s, t)$, which results in $\beta_p(s, t) = \boldsymbol{\varphi}(s)^T \mathrm{B}_p \boldsymbol{\varphi}(t)$. The coefficient matrix of all the predictors is denoted as $\mathbf{B} = [\mathbf{B}_{1_1}, \cdots, \mathbf{B}_{P_1}, \mathbf{B}_{1_2}, \cdots, \mathbf{B}_{P_2}]^T$. To test the sparse selection performance of the proposed method, the first two predictors $p_k = 1, 2$ of each passive party, i.e., $k = 1, 2$, are assumed to be effective with each element of the coefficient matrix following a normal distribution $N(10, 1)$, while the remaining eight predictors, $p_k = 3, \cdots, 10$, of each passive party are assumed to be dummy features with each element of the coefficient matrix being zero.

Prior to the evaluation of the predictive performance, the convergence of the proposed method is assessed. As depicted in Figure 6, the regression residuals of the proposed method diminish as the iteration count increases. Notably, the algorithm terminates sooner than the gradient boosting algorithm, which does not incorporate Vertical Federated Learning (VFL) and privacy-preserving mechanisms, as outlined in Algorithm 4.1. This premature termination may result in the potential underfitting of the proposed method, which subsequently leads to a decline in prediction performance, as evidenced by the subsequent simulation study.

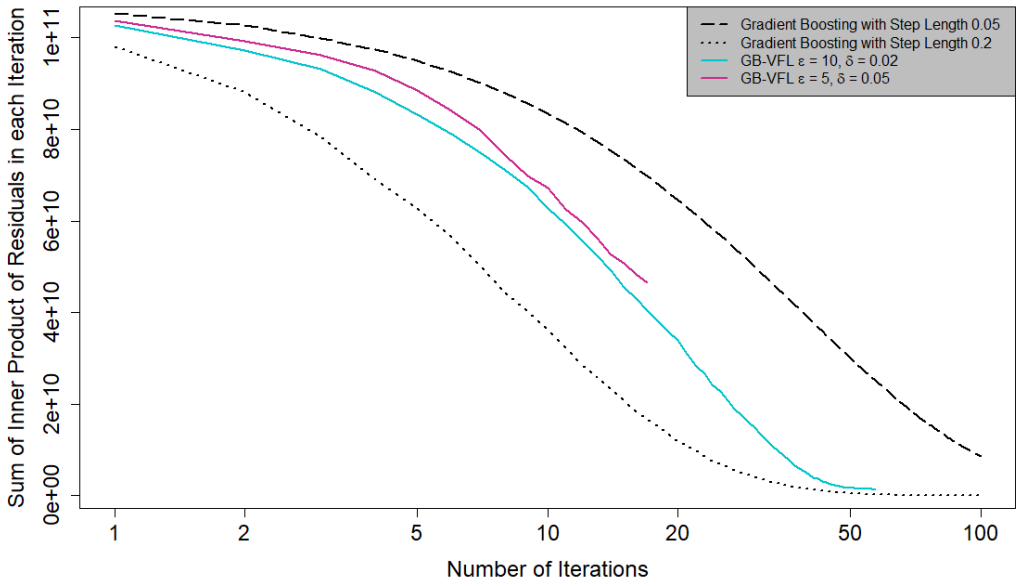

Figure 6: Convergence of the Proposed Method in Table 8 in comparison with Gradient Boosting Functional Regression in Table 7

Among 6,250 observations, 80%, i.e., 5,000 samples, are selected as the training set, and 20%, i.e., 1,250 samples, are selected as the test set. Five-fold cross-validation is performed for the prediction performance comparisons of the proposed method. To ensure privacy-preserving, we apply (ε, δ)-differential privacy in the release of negative gradient.

Different settings of (ε, δ) are chosen to test the prediction performance with different privacy budgets.

We deploy a self-adaptive step length for the gradient boosting algorithm. This is the step length that is decided by the MAX norm of coefficient matrix of negative gradient before clipping, which is $\eta^{(\omega)} = \min\left(0.1\|\mathbf{C}\|_{\text{MAX}}, \eta^{(\omega-1)}\right)$. An upper bound for step length is used to prevent $\|\mathbf{C}\|_{\text{MAX}}$ goes to extremely large.

Four different settings are given to check the influence of differential privacy regarding the test performance. Mean Absolute Percentage Error (MAPE) is chosen to be the performance measure of prediction error. The MAPE is defined as $MAPE = \frac{100\%}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\left|\frac{Y_{nt}-F_{nt}}{Y_{nt}}\right|$, where Y_{nt} is the actual value of response n in the testing set evaluated at time t, and F_{nt} is the corresponding prediction value evaluated at time t. The sampling period length T is 101, and the sample points are $\{0, 1, 2, \cdots 100\}$. For each test, 20 duplications are performed and the average MAPE is reported as the prediction accuracy with 5-fold cross-validation. The results are summarized in Table 9

Table 9: Prediction Performance by MAPE of the Proposed Method

Setting	ε	δ	MAPE in %
Exp 1	10	0.02	0.677
Exp 2	10	0.05	0.554
Exp 3	5	0.02	4.424
Exp 4	5	0.05	3.363

We employ a black-box Membership Inference Attack (MIA) to assess the privacy-preserving capabilities of the proposed method. For the sake of simplicity, we have altered

the data generation parameters. A dataset of 5,000 samples is generated. Each element of the coefficient matrix used to generate the p_k-th predictors of passive party k, $\mathbf{A}_{p_k}$, is drawn from a normal distribution $N(k-10, 0.05)$ for the first 2,500 samples for $p_k = 1, \dots, 10,\ k = 1, 2$. The remaining 2,500 samples' $\mathbf{A}_{p_k}$ are drawn from a normal distribution $N(10-k, 0.05)$. The coefficient matrix that captures the relationship between the p_k-th predictors of passive party k and the response, $\mathbf{B}_{p_k}$, is generated from $N(1, 0.05)$ for $p_k = 1, 2$, $k = 1, 2$, and 0 for $p_k = 3, \dots, 10$, $k = 1, 2$.

MIA aims to identify the presence or absence of individual records within the training data of a data owner. Such attacks are particularly pertinent when the nature of the training dataset exposes sensitive information, such as a medical dataset that contains patients with various types of cancer, or a dataset utilized to predict the stage of pregnancy based on shopping cart data. In the context of Vertical Federated Learning, where all passive parties are assumed to be malicious, it is crucial to evaluate the performance of MIA with respect to each release of negative gradient from the active party, which specifically focuses on the DP component as defined in equation (4.8).

The MIA assumes an honest-but-curious adversary with access to the trained model (referred to as the target model), the distribution of input data, knowledge about hyperparameters, and DP mechanisms employed during training. The attacker trains an attack model, typically a binary classifier, which is capable of accurately classifying data points as members or non-members of the target model's training dataset. MIA can be categorized into black-box and white-box approaches, depending on whether the attacker has access to the target model's learned parameters and architecture. In this simulation, we

utilize a black-box MIA, which supposes that the attacker lacks access to the learned parameters and architecture of the target model.

Training a binary classifier effectively poses a significant challenge in this context. However, Shokri et al. (2017) introduced a notable technique called shadow training, which presents a viable solution. This approach involves creating multiple shadow models that emulate the behavior of the target model. It operates under the premise that the attacker possesses detailed knowledge of the target model's structure and learning algorithm. For these shadow models, the attacker has their shadow training datasets $\mathcal{D}_{\text{train}}^{\text{sallow}}$ and shadow test datasets $\mathcal{D}_{\text{test}}^{\text{sallow}}$. The sallow models make predictions given $\mathcal{D}_{\text{train}}^{\text{sallow}}$ and $\mathcal{D}_{\text{test}}^{\text{sallow}}$ as $\mathcal{P}_{\text{train}}^{\text{sallow}}$ and $\mathcal{P}_{\text{test}}^{\text{sallow}}$. Together with the ground true of membership of these dataset $\mathcal{L}_{\text{train}}^{\text{sallow}}$ and $\mathcal{L}_{\text{test}}^{\text{sallow}}$, the attacker can train the binary classifier-based attack model by training set $\mathcal{D}_{\text{train}} = \mathcal{D}^1 \cup \mathcal{D}^2$, where $\mathcal{D}^1 \subseteq \{\mathcal{D}_{\text{train}}^{\text{sallow}}, \mathcal{P}_{\text{train}}^{\text{sallow}}, \mathcal{L}_{\text{train}}^{\text{sallow}}, \mathbf{1}\}$ and $\mathcal{D}^2 \subseteq \{\mathcal{D}_{\text{test}}^{\text{sallow}}, \mathcal{P}_{\text{test}}^{\text{sallow}}, \mathcal{L}_{\text{test}}^{\text{sallow}}, \mathbf{0}\}$. In $\mathcal{D}_{\text{train}}$, $\{\mathbf{1}, \mathbf{0}\}$ is the response for the training of attack model, where $\mathbf{1}$ represents the corresponding samples that are "in" the training set of target model, and $\mathbf{0}$ otherwise. Figure 7 provides an overview of the training of the attack model.

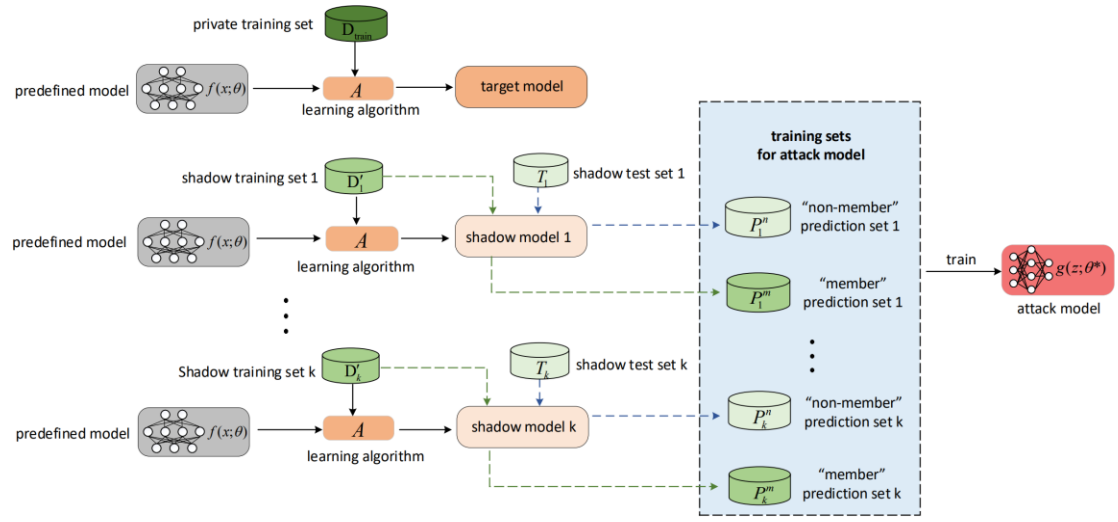

Figure 7: Overview of the shadow training technique

To demonstrate the impact of the privacy parameters ε and δ on the performance of the MIA, we focus on two critical metrics that assess the identification of training data: precision, which measures the proportion of records predicted as "in" that actually belong to the training dataset; and recall, which measures the proportion of truly contained records that are correctly predicted as "in". As shown in Figure 8, a decrease in ε and δ results in the attacker model's performance that approaches that of random guessing, where both precision and recall stabilize around 50%. These results suggest that the proposed DP method can effectively prevent privacy leakage with a small (ε, δ) pair.

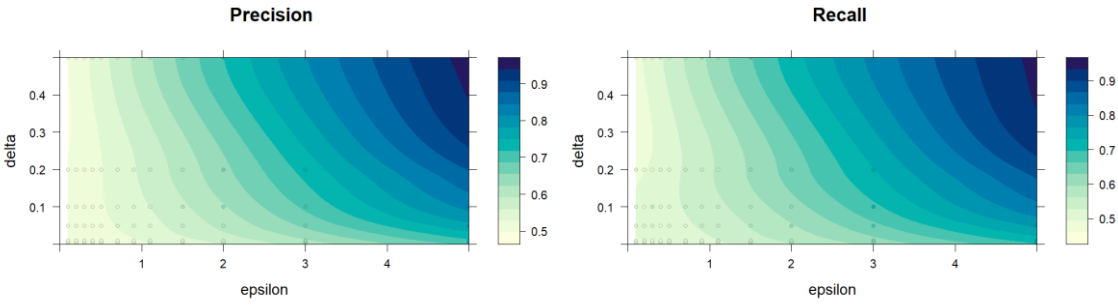

Figure 8: Precision and recall of MIA under different DP settings

4.5 Case Study

This subchapter introduces the application of the proposed method on the data collected in the Sleep Heart Health Study (SHHS). The SHHS, a seminal epidemiological study of Obstructive Sleep Apnea (OSA) in the United States, examines its broader health implications. The study encompasses a cohort of 2,338 participants, each of whom is characterized by 41 functional features. These features comprise 13 and 28 features extracted from electrocardiogram (ECG) and electroencephalogram (EEG) signals, respectively. In addition, the dataset includes 20 non-functional, patient-specific features such as age, gender, body mass index (BMI), ethnicity, and others. A comprehensive summary of all features included in this case study is presented in Table 10.

Table 10: Description of variables included in the study

Variables		Summary statistics
Non-functional independent variables		
Age (Unit: year)		59.5 ± 10.7
Gender (Female: 0; Male: 1)		51.9% ± 50.0%
Ethnicity (Hispanic: 0; Non-Hispanic: 1)		0.9 ± 0.3
BMI (Unit: kg/m^2)		27.3 ± 3.9
chol (Cholesterol milligrams per deciliter mg/dl)		203.5 ± 35.1
hdl (High-density lipoprotein cholesterol milligrams per deciliter mg/dl)		49.8 ± 15.1
trig (Triglycerides milligrams per deciliter mg/dl)		152.4 ± 143.1
fev1 (Forced Expiratory Volume in One Second liters)		2.9 ± 0.8
fvc (Forced Vital Capacity liters liters)		3.8 ± 1.0
attabl02 (Fall asleep while at the dinner table 1-4)		1.0 ± 0.2
drive02 (Fall asleep while driving 1-4)		1.2 ± 0.4
ess_s1 (Epworth Sleepiness Scale score 0-24)		7.3 ± 4.2
incar02 (Fall asleep while in a car 1-4)		1.1 ± 0.4
lydwn02 (Fall asleep while lying down in the afternoon 1-4)		2.8 ± 1.0
pgrcar02 (Fall asleep while a passenger in a car 1-4)		1.8 ± 0.9
sitlch02 (Fall asleep while sitting quietly after lunch 1-4)		1.8 ± 0.9
sitpub02 (Fall asleep while sitting inactive in a public place 1-4)		1.6 ± 0.8
sitrd02 (Fall asleep while sitting and reading 1-4)		2.4 ± 1.0
sittlk02 (Fall asleep while sitting and talking 1-4)		1.1 ± 0.4
watv02 (Fall asleep while watching TV 1-4)		2.6 ± 1.0
Functional independent variables		
ECG features (13 variables)	AVNN: Average of all NN intervals (Unit: ms)	
	SDNN: Standard deviation of NN intervals (Unit: ms)	
	rMSSD: Square root of the mean of squares of difference between adjacent NN intervals	
	pNNx: Percentage of differences between adjacent NN intervals that are greater than x ms. (x = 10, 20, 30, 40, and 50)	

	NN_RR: Ratio of consecutive normal sinus beats (NN) over all cardiac inter-beat (RR) intervals
	VLF/LF/HF: Relative spectral power for very low frequency (0.003-0.04 Hz), low frequency (0.04-0.15 Hz), and High frequency (0.15-0.4 Hz)
	LF_HF: Ratio of low to high frequency power
EEG features (28 variables)	Slow/Delta/Theta/Alpha/Sigma/Beta/Gamma Max/Min/Average/SD: Relative spectral power for Slow (0.5-1 Hz), Delta (1-4 Hz), Theta (4-8 Hz), Alpha (8-12 Hz), Sigma (12-15 Hz), Beta (15-30 Hz), and Gamma (30+Hz)
Functional response variable	
DSI (Number of adverse events per interval)	

The 61 functional and non-functional features identified are distributed among five passive parties, as illustrated in Table 11. Initially, to establish a comparative framework, the gradient boosting functional regression is executed without the integration of Vertical Federated Learning. The outcomes of this test are detailed in Table 12. The prediction error is quantified using MAPE and validated through a five-fold cross-validation process.

Table 11: Variables' distribution in VFL

Party	Features
Active Party	DSI
Passive Party 1	Age, Gender, Ethnicity, BMI
Passive Party 2	ECG features(13 variables), EEG features(28 variables)
Passive Party 3	chol, hdl, trig
Passive Party 4	fev1, fvc
Passive Party 5	attabl02, drive02, ess_s1, incar02, lydwn02, pgrcar02, sitlch02, sitpub02, sitrd02, sittlk02, watv02

In Table 12, the term 'global model' denotes the gradient boosting algorithm capable of selecting base learners from the entire array of both functional and non-functional features. The designations PP1 through PP5 correspond to models that selectively incorporate features from only their respective passive parties.

Table 12: MAPE of different models without VFL

	Global Model	PP1	PP2	PP3	PP4	PP5
MAPE in %	17.79	21.35	18.86	26.09	32.58	25.35

Table 13 provides the prediction error associated with our proposed Vertical Federated Gradient Boosting Functional Regression Model. We have implemented differential privacy parameters with ε set at 5 and δ at 0.05. Given the integration of random noise within the model, we conducted 10 replicates of the test. The results, which encompass both the mean and the standard deviation of the MAPE*s*, are derived from a five-fold cross-validation.

Table 13: MAPE of different models with VFL

	Global Model	PP1	PP2	PP3	PP4	PP5
MAPE in %	20.67 ± 0.84	41.32 ± 13.17	23.37 ± 1.50	61.72 ± 35.39	47.65 ± 16.80	45.22 ± 22.53

4.6 Conclusion and Discussion

In this study, we explored a novel approach to Vertical Federated Learning (VFL) by integrating Gradient Boosting for Functional Regression with Differential Privacy (DP). This innovative methodology facilitates secure and efficient collaborative learning across multiple organizations while ensuring data privacy, a critical requirement in the healthcare sector. Our focus was on addressing the complexities of function-on-function regression within a federated learning framework, particularly emphasizing the healthcare domain with a case study on Obstructive Sleep Apnea (OSA).

Our approach extends the capabilities of traditional regression models by accommodating function-on-function relationships. This is particularly relevant in medical contexts where continuous monitoring data, such as ECG and EEG signals, are used for predictive modeling. By incorporating DP into the federated learning framework, we

addressed one of the major challenges in VFL—privacy preservation. The differentially private gradient sharing mechanism ensures that individual data points remain secure, mitigating the risks associated with data breaches and privacy violations.

The application of our proposed model on the Sleep Heart Health Study (SHHS) dataset provided practical insights into its effectiveness. The SHHS dataset, rich with functional and non-functional features, served as an ideal test bed. Our model's performance in predicting OSA outcomes demonstrated the feasibility and advantages of our approach. The federated model achieved performance metrics comparable to a global model trained on centralized data, while significantly outperforming local models trained in isolation. This indicates that VFL, when combined with DP, does not compromise on predictive accuracy despite the stringent privacy constraints.

One of the key benefits of incorporating differential privacy in our model is its ability to defend against membership inference attacks. These attacks aim to determine whether a particular data point was part of the training dataset, posing significant privacy risks. The DP mechanism effectively obfuscates individual data contributions, making it exceedingly difficult for an attacker to ascertain the presence of specific records in the training set. Our experiments confirmed that even with advanced membership inference techniques, the privacy-preserving capability of our DP-enhanced model remained robust, thus providing strong guarantees against such attacks.

The incorporation of differential privacy successfully protected individual data points during the training process. Our results confirmed that the DP mechanism effectively obfuscated sensitive information, reducing the risk of privacy breaches without degrading model performance. The proposed approach is scalable and can be generalized to various

other domains beyond healthcare. The principles of combining functional regression with DP in a federated setting are applicable wherever sensitive functional data are involved.

The implications of this study are profound for the healthcare industry. By enabling secure, privacy-preserving predictive modeling, healthcare providers can collaborate more effectively, leading to improved diagnostic tools and patient outcomes. This is especially crucial in areas like OSA, where continuous monitoring data are critical for accurate predictions. Our approach aligns with global privacy regulations such as GDPR, highlighting its potential for widespread adoption in regulated industries. As privacy concerns continue to grow, the adoption of VFL frameworks with integrated DP will likely become a standard practice.

Future research can explore further enhancements in the differential privacy mechanism to improve its efficiency and robustness. Additionally, expanding the functional regression techniques to handle more complex data structures and relationships will enhance the applicability of our model. While our case study focused on OSA, the methodology is applicable to a wide range of scenarios involving functional data. Applications in finance, environmental monitoring, and other fields can benefit from this approach.

In conclusion, this study advances the field of Vertical Federated Learning by demonstrating that it is possible to achieve high-performance predictive models while preserving data privacy. The integration of Gradient Boosting for Functional Regression with Differential Privacy offers a robust solution to the challenges of collaborative learning in privacy-sensitive domains. By specifically addressing the threat of membership inference attacks, our approach ensures that individual data contributions remain

confidential, fostering greater trust and collaboration among data-sharing entities. This work lays the groundwork for future innovations, fostering a more secure and collaborative data science landscape.

Chapter 5 Discussion and Future Work

This dissertation presents three significant contributions to the field of machine learning and its application in health data analysis, each addressing unique challenges and proposing novel solutions.

In Topic I, we developed the Multi-modal Mixed-type Structural Equation Model (M2-SEM) with structured sparsity for subgroup discovery from heterogeneous health data. The integration of the Gauss-Hermite-enabled Expectation-Majorization-Minimization (GH-EMM) algorithm within the Expectation Maximization framework demonstrated the model's capability to handle high-dimensional, multi-modal mixed-type data effectively. Our simulation studies and real-world applications, specifically in identifying subgroups at risk for adverse cardiometabolic outcomes, illustrate the model's potential to enhance targeted health interventions and promote population health.

Topic II introduced the Federated Function-on-Function Regression with an efficient Gradient Boosting algorithm (fed-GB-LSA) for privacy-preserving telemedicine. The federated learning paradigm ensures data privacy while achieving performance comparable to global models. The innovative GB-based algorithm facilitates sparse selection of multivariate functional and non-functional features, providing an efficient estimation method. The application to telemonitoring of Obstructive Sleep Apnea (OSA) underscores the practical relevance and effectiveness of our approach.

In Topic III, we extended our research to Vertical Federated Learning (VFL) with Differential Privacy for function-on-function regression models. By integrating differential privacy into the federated gradient boosting process, we addressed the trade-off between model performance and privacy protection. The empirical results from simulation studies

and the case study on OSA validate the robustness of our method, highlighting its applicability in privacy-sensitive healthcare environments.

Overall, this dissertation advances the field of machine learning by developing innovative models and algorithms that address the complexities of multi-modal, mixed-type, and functional data in health research, while ensuring data privacy and efficient computation.

The research presented in this dissertation opens several avenues for future exploration and development. Enhanced model interpretability is a critical area for future work. While our proposed models and algorithms have demonstrated robust performance, there remains a need for improved interpretability. Future work can focus on developing methods to provide more transparent insights into the decision-making processes of these complex models, particularly for clinical applications where understanding the rationale behind predictions is crucial.

Another important direction is scalability to larger datasets. As health data continues to grow in volume and variety, ensuring the scalability of our models is essential. Future research should aim to optimize computational efficiency and memory usage to handle even larger datasets, potentially through distributed computing frameworks.

Integration with real-time data presents an exciting opportunity to enhance the applicability of our models. The incorporation of real-time data from wearable sensors and mobile health applications will allow the development of algorithms that can adapt to streaming data and provide real-time predictions and interventions, which would be a valuable extension of this work.

Broadening the applicability of our methods to diverse health conditions is another promising avenue. While this dissertation focused on cardiometabolic risk factors and Obstructive Sleep Apnea, the methodologies developed can be extended to other health conditions. Future studies should explore the application of these models to a wider range of diseases and health outcomes, further validating their utility and robustness.

Ethical and regulatory considerations will continue to be of paramount importance. With the increasing emphasis on data privacy and ethical AI, future research should address these considerations, particularly in federated learning frameworks. Developing and validating models that comply with evolving privacy regulations and ethical standards will be critical for their acceptance and deployment in real-world settings.

Finally, creating collaborative platforms for health data sharing is an important area for future work. Encouraging collaboration across institutions while preserving data privacy is a significant challenge. Future research can focus on developing secure, federated platforms that facilitate the sharing and analysis of health data across different organizations, enabling more comprehensive and inclusive health research.

In conclusion, the advancements made in this dissertation lay a strong foundation for future research and development in the intersection of machine learning, health data analysis, and privacy preservation. By continuing to innovate and address the challenges outlined, we can significantly contribute to improving health outcomes and advancing personalized medicine.

Appendix

A. Derivation of the expectation terms φ_1, φ_3, and φ_4 in function (2.10)

A1. Deriving the expectation φ_1 in function (2.10)

From the function (2.10), we have $\varphi_1\left(\{\boldsymbol{\alpha}^{(m)}, \mathbf{L}^{(m)}, \mathbf{B}^{(m)}, \boldsymbol{\Psi}^{(m)}\}_{m=1}^{M_1}\right) = \sum_{m=1}^{M_1}\left\{-\sum_{i=1}^{N} E_{\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\,\widehat{\boldsymbol{\Theta}}^{(j-1)}}\left[\log\left(f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)}, \boldsymbol{z}_i^{(m)}; \boldsymbol{\Theta}_{1m}\right)\right)\right] + \lambda_1 \sum_{p=1}^{P}\left\|\boldsymbol{l}_p^{(m)}\right\|_2\right\}$. Below we present how to derive $E_{\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\,\widehat{\boldsymbol{\Theta}}^{(j-1)}}\left[\log\left(f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)}, \boldsymbol{z}_i^{(m)}; \boldsymbol{\Theta}_{1m}\right)\right)\right]$.

From the equation (2.3), given $\boldsymbol{\Theta}_{1m}$, the conditional distribution of $\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)}, \boldsymbol{z}_i^{(m)}$ is $N(L^{(m)}\boldsymbol{\eta}_i^{(m)} + B^{(m)}\boldsymbol{z}_i^{(m)}, \boldsymbol{\Psi}^{(m)})$, i.e., a multivariate Gaussian distribution. After dropping the constants, we have

$$E_{\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\,\widehat{\boldsymbol{\Theta}}^{(j-1)}}\left[\log\left(f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)}, \boldsymbol{z}_i^{(m)}; \boldsymbol{\Theta}_{1m}\right)\right)\right]$$

$$= \frac{1}{2}\log|\boldsymbol{\Psi}^{(m)}| + \frac{1}{2}\left(\boldsymbol{x}_i^{(m)} - L^{(m)}E\left(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right) - B^{(m)}\boldsymbol{z}_i^{(m)}\right)^T \boldsymbol{\Psi}^{(m)^{-1}}\left(\boldsymbol{x}_i^{(m)} - L^{(m)}E\left(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right) - B^{(m)}\boldsymbol{z}_i^{(m)}\right)$$

$$+\frac{1}{2}tr\left(L^{(m)^T}\boldsymbol{\Psi}^{(m)^{-1}}L^{(m)}\left(E\left(\boldsymbol{\eta}_i^{(m)}\boldsymbol{\eta}_i^{(m)^T}|\boldsymbol{x}_i^{(m)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right) - E\left(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)E\left(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)^T\right)\right),$$

where $E\left(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right) = \sum_{k=1}^{K} E\left(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)}, s_{i,k}=1;\widehat{\boldsymbol{\Theta}}^{(j-1)}\right) f\left(s_{i,k}=1|\boldsymbol{x}_i^{(1)},\cdots\boldsymbol{x}_i^{(M)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)$ and $E\left(\boldsymbol{\eta}_i^{(m)}\boldsymbol{\eta}_i^{(m)^T}|\boldsymbol{x}_i^{(m)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right) = \sum_{k=1}^{K} E\left(\boldsymbol{\eta}_i^{(m)}\boldsymbol{\eta}_i^{(m)^T}|\boldsymbol{x}_i^{(m)}, s_{i,k}=1;\widehat{\boldsymbol{\Theta}}^{(j-1)}\right) f\left(s_{i,k}=1|\boldsymbol{x}_i^{(1)},\cdots\boldsymbol{x}_i^{(M)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)$.

Moreover, because $\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)}, s_{i,k}=1;\widehat{\boldsymbol{\Theta}}^{(j-1)}$ follows a Gaussian distribution, i.e., $N\left(\frac{\hat{L}^{(m)^T}\widehat{\boldsymbol{\Psi}}^{(m)^{-1}}\boldsymbol{x}_i^{(m)} + \widehat{\boldsymbol{\Sigma}}^{(m,k)^{-1}}\widehat{\boldsymbol{\mu}}^{(m,k)}}{\hat{L}^{(m)^T}\widehat{\boldsymbol{\Psi}}^{(m)^{-1}}\hat{L}^{(m)} + \widehat{\boldsymbol{\Sigma}}^{(m,k)^{-1}}}, \left(\hat{L}^{(m)^T}\widehat{\boldsymbol{\Psi}}^{(m)^{-1}}\hat{L}^{(m)} + \widehat{\boldsymbol{\Sigma}}^{(m,k)^{-1}}\right)^{-1}\right)$, we have

$E\left(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)}, s_{i,k}=1;\widehat{\boldsymbol{\Theta}}^{(j-1)}\right) = \frac{\hat{L}^{(m)^T}\widehat{\boldsymbol{\Psi}}^{(m)^{-1}}\boldsymbol{x}_i^{(m)} + \widehat{\boldsymbol{\Sigma}}^{(m,k)^{-1}}\widehat{\boldsymbol{\mu}}^{(m,k)}}{\hat{L}^{(m)^T}\widehat{\boldsymbol{\Psi}}^{(m)^{-1}}\hat{L}^{(m)} + \widehat{\boldsymbol{\Sigma}}^{(m,k)^{-1}}}$, and $E\left(\boldsymbol{\eta}_i^{(m)}\boldsymbol{\eta}_i^{(m)^T}|\boldsymbol{x}_i^{(m)}, s_{i,k}=1;\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)$

$= \left(\hat{L}^{(m)^T}\widehat{\boldsymbol{\Psi}}^{(m)^{-1}}\hat{L}^{(m)} + \widehat{\boldsymbol{\Sigma}}^{(m,k)^{-1}}\right)^{-1} - \frac{\hat{L}^{(m)^T}\widehat{\boldsymbol{\Psi}}^{(m)^{-1}}\boldsymbol{x}_i^{(m)} + \widehat{\boldsymbol{\Sigma}}^{(m,k)^{-1}}\widehat{\boldsymbol{\mu}}^{(m,k)^T}}{\hat{L}^{(m)^T}\widehat{\boldsymbol{\Psi}}^{(m)^{-1}}\hat{L}^{(m)} + \widehat{\boldsymbol{\Sigma}}^{(m,k)^{-1}}} \frac{\hat{L}^{(m)^T}\widehat{\boldsymbol{\Psi}}^{(m)^{-1}}\boldsymbol{x}_i^{(m)} + \widehat{\boldsymbol{\Sigma}}^{(m,k)^{-1}}\widehat{\boldsymbol{\mu}}^{(m,k)}}{\hat{L}^{(m)^T}\widehat{\boldsymbol{\Psi}}^{(m)^{-1}}\hat{L}^{(m)} + \widehat{\boldsymbol{\Sigma}}^{(m,k)^{-1}}}$.

Lastly, we have $f\left(s_{i,k}=1\middle|\boldsymbol{x}_i^{(1)},\cdots\boldsymbol{x}_i^{(M)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)=\frac{\widehat{w}_k\prod_{m=1}^{M_1}f(\boldsymbol{x}_i^{(m)}|s_{i,k}=1;\widehat{\boldsymbol{\Theta}}^{(j-1)})}{\sum_{k=1}^{K}\widehat{w}_k\prod_{m=1}^{M_1}f(\boldsymbol{x}_i^{(m)}|s_{i,k}=1;\widehat{\boldsymbol{\Theta}}^{(j-1)})}$, in which $\boldsymbol{x}_i^{(m)}|s_{i,k}=1;\widehat{\boldsymbol{\Theta}}^{(j-1)}\sim N(\hat{L}^{(m)}\widehat{\boldsymbol{\mu}}^{(m,k)}+\hat{B}^{(m)}\boldsymbol{z}_i^{(m)},\hat{L}^{(m)}\widehat{\boldsymbol{\Sigma}}^{(m,k)}\hat{L}^{(m)^T}+\widehat{\boldsymbol{\Psi}}^{(m)})$ for $m=1,\ldots,M_1$.

■

A2. Deriving the expectation φ_3 in function (2.10)

Based on the function in (2.9), we have

$\varphi_3\left(\left\{\left\{\boldsymbol{\mu}^{(m,k)},\boldsymbol{\Sigma}^{(m,k)}\right\}_{k=1}^{K}\right\}_{m=1}^{M}\right)=\sum_{m=1}^{M}\left\{-\sum_{i=1}^{N}E_{\boldsymbol{\eta}_i^{(m)},\boldsymbol{s}_i|\boldsymbol{x}_i^{(1)},\cdots\boldsymbol{x}_i^{(M)};\widehat{\boldsymbol{\Theta}}^{(j-1)}}\log\left(f\left(\boldsymbol{\eta}_i^{(m)}|\ \boldsymbol{s}_i;\boldsymbol{\Theta}_{3m}\right)\right)+\lambda_2\left\|\mathbf{u}^{(m)}\right\|_2\right\}$ .To obtain an explicit form of φ_3 , we need to derive $E_{\boldsymbol{\eta}_i^{(m)},\boldsymbol{s}_i|\boldsymbol{x}_i^{(1)},\cdots\boldsymbol{x}_i^{(M)};\widehat{\boldsymbol{\Theta}}^{(j-1)}}\log\left(f\left(\boldsymbol{\eta}_i^{(m)}|\ \boldsymbol{s}_i;\boldsymbol{\Theta}_{3m}\right)\right)$ as follows. By function (2.2), the conditional distribution of $\boldsymbol{\eta}_i^{(m)}$ given $\boldsymbol{s}_{i,k}=1$ is $N(\boldsymbol{\mu}^{(m,k)},\boldsymbol{\Sigma}^{(m,k)})$ that is a multivariate Gaussian distribution. After dropping the constants, this expectation can be rewritten as

$$E_{\boldsymbol{\eta}_i^{(m)},\boldsymbol{s}_i|\boldsymbol{x}_i^{(1)},\cdots\boldsymbol{x}_i^{(M)};\widehat{\boldsymbol{\Theta}}^{(j-1)}}\log\left(f\left(\boldsymbol{\eta}_i^{(m)}|\ \boldsymbol{s}_i;\boldsymbol{\Theta}_{3m}\right)\right)=$$

$$\sum_{k=1}^{K}\left\{\begin{array}{c}\frac{1}{2}\log\left|\boldsymbol{\Sigma}^{(m,k)}\right|\\+\frac{1}{2}\left(E\left(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)},s_{i,k}=1;\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)-\boldsymbol{\mu}^{(m,k)}\right)^T\boldsymbol{\Sigma}^{(m,k)^{-1}}\left(E\left(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)},s_{i,k}=1;\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)-\boldsymbol{\mu}^{(m,k)}\right)\\+\frac{1}{2}tr\left(E\left(\boldsymbol{\eta}_i^{(m)}\boldsymbol{\eta}_i^{(m)^T}|\boldsymbol{x}_i^{(m)},s_{i,k}=1;\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)\right)\end{array}\right\}f\left(s_{i,k}=1|\boldsymbol{x}_i^{(1)},\cdots\boldsymbol{x}_i^{(M)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right).$$

For numerical modalities with $m=1,\ldots,M_1$, we can derive $E\left(\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)},s_{i,k}=1;\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)$ and $E\left(\boldsymbol{\eta}_i^{(m)}\boldsymbol{\eta}_i^{(m)^T}|\boldsymbol{x}_i^{(m)},s_{i,k}=1;\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)$ as shown in A.1. For categorical modalities with $m=M_1+1,\ \ldots,\ M_1+M_2$, these expectations can be approximated by GH Quadrature following a similar approach to the equation (2.16). Last, based on Bayes' Theorem, the derivation of $f\left(s_{i,k}=1\middle|\boldsymbol{x}_i^{(1)},\cdots\boldsymbol{x}_i^{(M)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)$ relies on the term

$f\left(\boldsymbol{x}_i^{(m)}\middle|s_{i,k}=1;\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)$ for $m=1,\dots,M$. For numerical modalities with $m=1,\dots,M_1$, this term is the probability density function of a Gaussian distribution; For categorical modalities with $m=M_1+1,\ \dots,\ M_1+M_2$, this term can be approximated based on Proposition 2.2.

■

A3. Deriving the expectation φ_4 in function (2.10)

From the function (2.10), we have $\varphi_4(\boldsymbol{w}) = -\sum_{i=1}^{N} E_{\boldsymbol{s}_i|\boldsymbol{x}_i^{(1)},\cdots\boldsymbol{x}_i^{(M)};\,\widehat{\boldsymbol{\Theta}}^{(j-1)}}\left[\log\left(f(\boldsymbol{s}_i;\boldsymbol{\Theta}_4)\right)\right] = -\sum_{i=1}^{N}\sum_{k=1}^{K}\log(w_k) f(s_{i,k}=1|\boldsymbol{x}_i^{(1)},\cdots\boldsymbol{x}_i^{(M)};\,\widehat{\boldsymbol{\Theta}}^{(j-1)})$, in which the GH approximation of $f\left(s_{i,k}=1\middle|\boldsymbol{x}_i^{(1)},\cdots\boldsymbol{x}_i^{(M)};\widehat{\boldsymbol{\Theta}}^{(j-1)}\right)$ is given by Proposition 2.2.

■

B. Proof of Proposition 2.1

Proof: We denote the GH approximation in Definition 2.1 as $GH(g) \triangleq \sum_{t_1=1}^{T}\cdots\sum_{t_Q=1}^{T}\omega_{t_1}\cdots\omega_{t_Q} g(\boldsymbol{z}_t)$ and the approximation error can be written as $error(g) = \int_{\boldsymbol{S}} exp\{-\boldsymbol{z}^T\boldsymbol{z}\}g(\boldsymbol{z})d\boldsymbol{z} - GH(g)$. Authors (Hilderbrand, 1987) proved that $error(g) = \frac{T!\sqrt{\pi}}{2^T(2T)!}g^{(2T)}(\xi)$ for $\xi \in \mathbb{R}^Q$.

By the Weierstrass Approximation Theorem (Stone, 1948), there exists a polynomial function $p(\boldsymbol{z})$ with the order of T'such that $|g(\boldsymbol{z}) - p(\boldsymbol{z})| < \varepsilon$ for every $\varepsilon > 0$ and $\boldsymbol{z} \in \boldsymbol{S}$, where $\boldsymbol{S}$ is an arbitrary closed subset of $\mathbb{R}^Q$. Accordingly, the error term can be rewritten as the sum of three bounded terms as follows:

$$error(g) = \left(\int_S exp\{-\boldsymbol{z}^T\boldsymbol{z}\}g(\boldsymbol{z})d\boldsymbol{z} - \int_S exp\{-\boldsymbol{z}^T\boldsymbol{z}\}p(\boldsymbol{z})d\boldsymbol{z}\right) + \left(\int_S exp\{-\boldsymbol{z}^T\boldsymbol{z}\}p(\boldsymbol{z})d\boldsymbol{z} - GH(p)\right) + \left(GH(p) - GH(g)\right). \quad \text{(B.1)}$$

The first term in (B.1) is bounded, because we have

$$\left|\int_S exp\{-\boldsymbol{z}^T\boldsymbol{z}\}g(\boldsymbol{z})d\boldsymbol{z} - \int_S exp\{-\boldsymbol{z}^T\boldsymbol{z}\}p(\boldsymbol{z})d\boldsymbol{z}\right| \le \int_S exp\{-\boldsymbol{z}^T\boldsymbol{z}\}|g(\boldsymbol{z}) - p(\boldsymbol{z})|d\boldsymbol{z} < \varepsilon \int_S exp\{-\boldsymbol{z}^T\boldsymbol{z}\}d\boldsymbol{z}. \quad \text{(B.2)}$$

The second term is zero for a sufficiently large T. Specifically, the second term is equivalent to the GH approximation error of the polynomial function $p(\boldsymbol{z})$ with the order of T', i.e., $error(p) = \int_S exp\{-\boldsymbol{z}^T\boldsymbol{z}\}p(\boldsymbol{z})d\boldsymbol{z} - GH(p)$. Therefore, if the order of the Hermite polynomial, i.e., T, is sufficiently larger than T'such that $2T \ge T'$, we have

$$error(p) = \frac{T!\sqrt{\pi}}{2^T(2T)!}p^{(2T)}(\xi) = 0. \quad \text{(B.3)}$$

Last, it is straightforward to show the third term is bounded by a small quantity. That is,

$$|GH(p) - GH(g)| = \sum_{t_1=1}^{T}\cdots\sum_{t_Q=1}^{T}\omega_{t_1}\cdots\omega_{t_Q}|p(\boldsymbol{z}_t) - g(\boldsymbol{z}_t)| < \quad \sum_{t_1=1}^{T}\cdots\sum_{t_Q=1}^{T}\omega_{t_1}\cdots\omega_{t_Q}\varepsilon. \quad \text{(B.4)}$$

We substitute (B.2) and (B.3) into (B.1) and have $|error(g)| < \varepsilon\left(\int_S exp\{-\boldsymbol{z}^T\boldsymbol{z}\}d\boldsymbol{z} + \sum_{t_1=1}^{T}\cdots\sum_{t_Q=1}^{T}\omega_{t_1}\cdots\omega_{t_Q}\right)$. Since $\varepsilon > 0$ can be arbitrarily small, we can conclude that the GH error is zero for a sufficiently large T.

■

C. Proof of Proposition 2.2

Proof: Following Bayes' Theorem, we have

$$f\left(s_{ik} = 1|\boldsymbol{x}_i^{(m)};\ \widehat{\boldsymbol{\Theta}}^{(j)}\right) = \frac{f\left(\boldsymbol{x}_i^{(m)}|s_{ik} = 1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)f\left(s_{ik} = 1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)}{\sum_{k=1}^{K}f\left(\boldsymbol{x}_i^{(m)}|s_{ik} = 1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)f\left(s_{ik} = 1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)}$$

including a non-analytical term $f\left(\boldsymbol{x}_i^{(m)}|s_{ik} = 1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)$ that can be rewritten as

$$f\left(\boldsymbol{x}_i^{(m)}|s_{ik} = 1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right) = \int f\left(\boldsymbol{\eta}_i^{(m)}|s_{ik} = 1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)};\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)d\boldsymbol{\eta}_i^{(m)}.$$

Because $\boldsymbol{\eta}_i^{(m)}|(s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)})$ follows a multivariate Gaussian distribution, i.e., $N(\boldsymbol{\mu}^{(m,K)},\boldsymbol{\Sigma}^{(m)})$, we have

$$\int f\left(\boldsymbol{\eta}_i^{(m)}|s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right) f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)};\ \widehat{\boldsymbol{\Theta}}^{(j)}\right) d\boldsymbol{\eta}_i^{(m)} =$$

$$\int_{\mathbb{R}^Q} (2\pi)^{-Q/2}|\boldsymbol{\Sigma}^{(m,k)}|^{-1/2}\exp\{-\frac{(\boldsymbol{\eta}_i^{(m)}-\boldsymbol{\mu}^{(m,K)})^T\boldsymbol{\Sigma}^{(m,k)^{-1}}(\boldsymbol{\eta}_i^{(m)}-\boldsymbol{\mu}^{(m,K)})}{2}\} f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)};\ \widehat{\boldsymbol{\Theta}}^{(j)}\right) d\boldsymbol{\eta}_i^{(m)}\ .$$

Then denote $\widetilde{\boldsymbol{\eta}}_i^{(m)^T}\widetilde{\boldsymbol{\eta}}_i^{(m)} = \frac{(\boldsymbol{\eta}_i^{(m)}-\boldsymbol{\mu}^{(m,K)})^T\boldsymbol{\Sigma}^{(m,k)^{-1}}(\boldsymbol{\eta}_i^{(m)}-\boldsymbol{\mu}^{(m,K)})}{2}$ such that $\boldsymbol{\eta}_i^{(m)} = \sqrt{2}\boldsymbol{\Sigma}^{(m,k)^{\frac{1}{2}}}\widetilde{\boldsymbol{\eta}}_i^{(m)} + \boldsymbol{\mu}^{(m,K)}$, we have

$$\int_{\mathbb{R}^Q} (2\pi)^{-\frac{Q}{2}}|\boldsymbol{\Sigma}^{(m,k)}|^{-\frac{1}{2}}\exp\left\{-\frac{(\boldsymbol{\eta}_i^{(m)}-\boldsymbol{\mu}^{(m,K)})^T\boldsymbol{\Sigma}^{(m,k)^{-1}}(\boldsymbol{\eta}_i^{(m)}-\boldsymbol{\mu}^{(m,K)})}{2}\right\} f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)};\ \widehat{\boldsymbol{\Theta}}^{(j)}\right) d\boldsymbol{\eta}_i^{(m)}$$

$$= (\pi)^{-\frac{Q}{2}}\int_{\mathbb{R}^Q}\exp\left\{-\widetilde{\boldsymbol{\eta}}_i^{(m)^T}\widetilde{\boldsymbol{\eta}}_i^{(m)}\right\} f\left(\boldsymbol{x}_i^{(m)}|\sqrt{2}\boldsymbol{\Sigma}^{(m,k)^{\frac{1}{2}}}\widetilde{\boldsymbol{\eta}}_i^{(m)} + \boldsymbol{\mu}^{(m,K)};\ \widehat{\boldsymbol{\Theta}}^{(j)}\right) d\widetilde{\boldsymbol{\eta}}_i^{(m)}. \qquad \text{(C.1)}$$

Applying the GH approximation to (C.1), the non-analytical term $f\left(\boldsymbol{x}_i^{(m)}|s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)$ can be explicitly approximated as follows:

$$\tilde{f}\left(\boldsymbol{x}_i^{(m)}|s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right) = (\pi)^{-Q/2}\sum_{t_1=1}^{T}\dots\sum_{t_Q=1}^{T} w_{t_1}\dots w_{t_Q} f\left(\boldsymbol{x}_i^{(m)}|\sqrt{2}\boldsymbol{\Sigma}^{(m,k)^{\frac{1}{2}}}\widetilde{\boldsymbol{\eta}}_{i,t}^{(m)} + \boldsymbol{\mu}^{(m,K)};\ \widehat{\boldsymbol{\Theta}}^{(j)}\right),$$

where $\widetilde{\boldsymbol{\eta}}_{i,t}^{(m)} = \left(\widetilde{\eta}_{i,t_1}^{(m)},\cdots,\widetilde{\eta}_{i,t_Q}^{(m)}\right)$ are the roots of the Hermite polynomial of order T, T is the number of quadrature points of $\boldsymbol{\eta}_{i,t_q}^{(m)}$, and the weights are given by $\omega_{t_q} = \frac{2^{T+1}\mathrm{T}!\sqrt{\pi}}{[H_{T+1}(\widetilde{\eta}_{i,t_q}^{(m)})]^2}$.

■

D. Proof of Proposition 2.3

Proof: Following the similar proof for Proposition 2.2, because $\boldsymbol{\eta}_i^{(m)}|(s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)})$ follows a multivariate Gaussian distribution, i.e., $N(\boldsymbol{\mu}^{(m,K)},\boldsymbol{\Sigma}^{(m,k)})$, the left side of the equation (2.25) can be rewritten as

$$\int_{\boldsymbol{\eta}}^{\square}\left(log\left(f\left(\boldsymbol{x}_{ip}^{(m)}|\boldsymbol{\eta}_i^{(m)};\ \boldsymbol{\Theta}_{2m}\right)\right)f\left(\boldsymbol{\eta}_i^{(m)}|s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)\right)d\boldsymbol{\eta}_i^{(m)}$$

$$=\int_{\mathbb{R}^Q}(2\pi)^{-Q/2}|\boldsymbol{\Sigma}^{(m,k)}|^{-1/2}\exp\{-\frac{(\boldsymbol{\eta}_i^{(m)}-\boldsymbol{\mu}^{(m,K)})^T\boldsymbol{\Sigma}^{(m,k)^{-1}}(\boldsymbol{\eta}_i^{(m)}-\boldsymbol{\mu}^{(m,K)})}{2}\}log\left(f\left(\boldsymbol{x}_{ip}^{(m)}|\boldsymbol{\eta}_i^{(m)};\ \boldsymbol{\Theta}_{2m}\right)\right)d\boldsymbol{\eta}_i^{(m)}.$$

We denote $\widetilde{\boldsymbol{\eta}}_i^{(m)^T}\widetilde{\boldsymbol{\eta}}_i^{(m)}=\frac{(\boldsymbol{\eta}_i^{(m)}-\boldsymbol{\mu}^{(m,K)})^T\boldsymbol{\Sigma}^{(m,k)^{-1}}(\boldsymbol{\eta}_i^{(m)}-\boldsymbol{\mu}^{(m,K)})}{2}$ such that $\boldsymbol{\eta}_i^{(m)}=\sqrt{2}\boldsymbol{\Sigma}^{(m,k)\frac{1}{2}}\widetilde{\boldsymbol{\eta}}_i^{(m)}+\boldsymbol{\mu}^{(m,K)}$, and then have

$$\int_{\boldsymbol{\eta}}^{\square}\left(log\left(f\left(\boldsymbol{x}_{ip}^{(m)}|\boldsymbol{\eta}_i^{(m)};\ \boldsymbol{\Theta}_{2m}\right)\right)f\left(\boldsymbol{\eta}_i^{(m)}|s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)\right)d\boldsymbol{\eta}_i^{(m)}$$

$$=(\pi)^{-\frac{Q}{2}}\int_{\mathbb{R}^Q}\exp\left\{-\widetilde{\boldsymbol{\eta}}_i^{(m)^T}\widetilde{\boldsymbol{\eta}}_i^{(m)}\right\}\log\left(f\left(\boldsymbol{x}_i^{(m)}|\sqrt{2}\boldsymbol{\Sigma}^{(m,k)\frac{1}{2}}\widetilde{\boldsymbol{\eta}}_i^{(m)}+\boldsymbol{\mu}^{(m,K)};\ \boldsymbol{\Theta}_{2m}\right)\right)d\widetilde{\boldsymbol{\eta}}_i^{(m)}.$$

Applying the GH approximation in Definitaion 2.1, we have the approximation as

$$\int_{\boldsymbol{\eta}}^{\square}\left(log\left(f\left(\boldsymbol{x}_{ip}^{(m)}|\boldsymbol{\eta}_i^{(m)};\ \boldsymbol{\Theta}_{2m}\right)\right)f\left(\boldsymbol{\eta}_i^{(m)}|s_{ik}=1;\ \widehat{\boldsymbol{\Theta}}^{(j)}\right)\right)d\boldsymbol{\eta}_i^{(m)}$$

$$\approx(\pi)^{-Q/2}\sum_{t_1=1}^{T}\dots\sum_{t_Q=1}^{T}w_{t_1}\dots w_{t_Q}\,log\left(f\left(\boldsymbol{x}_i^{(m)}|\sqrt{2}\boldsymbol{\Sigma}^{(m,k)\frac{1}{2}}\widetilde{\boldsymbol{\eta}}_{i,t}^{(m)}+\boldsymbol{\mu}^{(m,K)};\ \boldsymbol{\Theta}_{2m}\right)\right),$$

where $\boldsymbol{\eta}_{i,t}^{(m)}=\left(\widetilde{\eta}_{i,t_1}^{(m)},\cdots,\widetilde{\eta}_{i,t_Q}^{(m)}\right)$ are the roots of the Hermite polynomial of order T, T is the number of quadrature points of $\widetilde{\eta}_{i,t_q}^{(m)}$, and the weights are given by $\omega_{t_q}=\frac{2^{T+1}\mathrm{T}!\sqrt{\pi}}{[H_{T+1}(\widetilde{\eta}_{i,t_q}^{(m)})]^2}$.

■

E. Proof of Theorem 2.1

Proof: Since$\nabla\varphi(\boldsymbol{l}|\mathbf{D})$ is Lipschitz continuous, there exists $\theta\in\mathbb{R}$ such that $\|\nabla\varphi(\boldsymbol{l}|\mathbf{D})-\nabla\varphi(\boldsymbol{l}'|\mathbf{D})\|\leq\theta\|\boldsymbol{l}-\boldsymbol{l}'\|,\forall\boldsymbol{l},\boldsymbol{l}'$. Therefore,$g(\boldsymbol{l})=\frac{\theta}{2}\boldsymbol{l}^T\boldsymbol{l}-\varphi(\boldsymbol{l}|\mathbf{D})$ is a convex function, because $\nabla^2\varphi(\boldsymbol{l}|\mathbf{D})\preccurlyeq\theta\mathbf{I}$. By the first order condition of convex function, we have$g(\boldsymbol{l})\geq g(\boldsymbol{l}')+\nabla g(\boldsymbol{l}')^T(\boldsymbol{l}-\boldsymbol{l}')$, which is $\frac{\theta}{2}\boldsymbol{l}^T\boldsymbol{l}-\varphi(\boldsymbol{l}|\mathbf{D})\geq\frac{\theta}{2}\boldsymbol{l}'^T\boldsymbol{l}'-\varphi(\boldsymbol{l}'|\mathbf{D})+$

$\left(\theta \boldsymbol{l}' - \nabla\varphi(\boldsymbol{l}'|\mathbf{D})\right)^T(\boldsymbol{l} - \boldsymbol{l}')$. By changing the order and regrouping items, we have the following inequality

$$\varphi(\boldsymbol{l}|\mathbf{D}) \leq \varphi(\boldsymbol{l}'|\mathbf{D}) + \nabla\varphi(\boldsymbol{l}'|\mathbf{D})^T(\boldsymbol{l} - \boldsymbol{l}') + \frac{\theta}{2}\|\boldsymbol{l} - \boldsymbol{l}'\|^2. \tag{E.1}$$

Inequality (E.1) is called the "quadratic bound property" of a function with Lipschitz continuous gradient.

Define the surrogate function of MM algorithm as $Q(\boldsymbol{l}|\mathbf{D}) = \varphi(\boldsymbol{l}'|\mathbf{D}) + \nabla\varphi(\boldsymbol{l}'|\mathbf{D})^T(\boldsymbol{l} - \boldsymbol{l}') + \frac{\theta}{2}\|\boldsymbol{l} - \boldsymbol{l}'\|^2 + \lambda\|\boldsymbol{l}\|_2$. Due to the subgradient optimality condition, i.e., $\mathbf{0} \in \nabla Q(\boldsymbol{l}|\mathbf{D})$, we have the following solution

$$\tilde{\boldsymbol{l}} = \underset{\boldsymbol{l}}{\operatorname{argmin}} Q(\boldsymbol{l}|\mathbf{D}) =$$

$$\begin{cases} \frac{1}{\theta}(-\nabla\varphi(\boldsymbol{l}'|\mathbf{D}) + \theta\boldsymbol{l}')\left(1 - \frac{\lambda}{\|-\nabla\varphi(\boldsymbol{l}'|\mathbf{D}) + \theta\boldsymbol{l}'\|_2}\right) & \text{if } \|-\nabla\varphi(\boldsymbol{l}'|\mathbf{D}) + \theta\boldsymbol{l}'\|_2 > \lambda \\ \mathbf{0} & \text{if } \|-\nabla\varphi(\boldsymbol{l}'|\mathbf{D}) + \theta\boldsymbol{l}'\|_2 \leq \lambda \end{cases}.$$

To prove convergence of MM algorithm, we will first show the strict decent property in each iteration. In a certain iteration that updates $\boldsymbol{l}$ by $\tilde{\boldsymbol{l}}$, we have

$$\varphi\left(\tilde{\boldsymbol{l}}|\mathbf{D}\right) + \lambda\left\|\tilde{\boldsymbol{l}}\right\|_2 \leq Q\left(\tilde{\boldsymbol{l}}|\mathbf{D}\right) \leq Q(\boldsymbol{l}'|\mathbf{D}) = \varphi(\boldsymbol{l}'|\mathbf{D}) + \lambda\|\boldsymbol{l}'\|_2.$$

The first inequality is due to the quadratic bound property (E.1) of $\varphi(\boldsymbol{l}|\mathbf{D})$. Note that the inequality strictly holds unless $\tilde{\boldsymbol{l}} = \boldsymbol{l}'$. The second inequality is due to the optimality of $\tilde{\boldsymbol{l}}$ with respect to $Q(\boldsymbol{l}|\mathbf{D})$, and the third equality is based on the definition of $Q(\boldsymbol{l}|\mathbf{D})$. Thus, within each iteration beginning with $\boldsymbol{l}'$, the objective function will decrease if we update $\boldsymbol{l}'$ with $\tilde{\boldsymbol{l}}$.

Therefore, if we have $\boldsymbol{l}' = \tilde{\boldsymbol{l}}$ in any iteration, we have $\begin{cases} \nabla\varphi(\boldsymbol{l}'|\mathbf{D}) + \lambda\frac{\boldsymbol{l}'}{\|\boldsymbol{l}'\|_2} = \mathbf{0} & \text{if } \boldsymbol{l}' \neq \mathbf{0} \\ \|-\nabla\varphi(\boldsymbol{l}'|\mathbf{D})\|_2 \leq \lambda & \text{if } \boldsymbol{l}' = \mathbf{0} \end{cases}$, which is the stationarity of the KKT condition. In other words, if the objective function remains unchanged, i.e., $\boldsymbol{l}' = \tilde{\boldsymbol{l}}$, it indicates that the algorithm has converged to the optimal solution.

■

F. Proof of Proposition 2.4

Proof: We will use φ_2 in (2.29) as an example to show that the objective function in (2.29) is joint convex with a Lipschitz continuous gradient with respect to $\{\mathbf{L}^{(m)}, \mathbf{B}^{(m)}\}$ for $m = M_1 + 1, \ldots, M_1 + M_2$, respectively. Therefore, the optimization problem in (2.29) can be efficiently optimized by the MM algorithm in Theorem 2.1. Similar discussions follow for (2.28) and (2.30).

Since the optimization problem in (2.29) is separable for $m = M_1 + 1, \ldots, M_1 + M_2$, we need to provide proofs for only the sub-optimization problem with the objective function $\varphi_{2m} = \sum_{i=1}^{N} E_{\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\, \widehat{\boldsymbol{\Theta}}^{(j-1)}}\left[\log\left(f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)}, \boldsymbol{z}_i^{(m)}\right)\right)\right] + \lambda_1 \sum_{p=1}^{P}\left\|\boldsymbol{l}_p^{(m)}\right\|_2$ with respect to $\{\boldsymbol{\alpha}^{(m)}, \mathbf{L}^{(m)}, \mathbf{B}^{(m)}\}$. Based on the equation (2.4), the feasible region of φ_{2m} is obviously convex with respect to $\{\boldsymbol{\alpha}^{(m)}, \mathbf{L}^{(m)}, \mathbf{B}^{(m)}\}$. Next, we can prove the objective function φ_{2m} is convex. The l-21 norm $\sum_{p=1}^{P}\left\|\boldsymbol{l}_p^{(m)}\right\|_2$ is clearly convex with respect to $\mathbf{L}^{(m)}$ (Boyd et al., 2004). Moreover, because expectation preserves convexity, we only need to prove the term $\log\left(f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)}, \boldsymbol{z}_i^{(m)}\right)\right)$ within the expectation is convex with respect

to $\{\boldsymbol{\alpha}^{(m)}, \mathbf{L}^{(m)}, \mathbf{B}^{(m)}\}$. Based on the model in (2.4), this term is the log-likelihood function of the ordinal logistic regression that is proven to be convex (Kim, 2004).

Next, we prove that $\sum_{i=1}^{N} E_{\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\, \hat{\boldsymbol{\Theta}}^{(j-1)}} \left[\log\left(f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)}, \boldsymbol{z}_i^{(m)}\right)\right)\right]$ has the Lipschitz continuous gradient with respect to $\{\mathbf{L}^{(m)}, \mathbf{B}^{(m)}\}$. To facilitate later discussion, we first prove that, given a function $h(\boldsymbol{x}, \boldsymbol{y})$ which is Lipschitz continuous with respect to $\boldsymbol{x}$, we have $H(\boldsymbol{x}) = \int_{\boldsymbol{y}\in\boldsymbol{Y}} h(\boldsymbol{x}, \boldsymbol{y}) g(\boldsymbol{y}) d\boldsymbol{y}$ is also Lipschitz continuous, in which $g(\boldsymbol{y})$ is the probability density function of $\boldsymbol{y}$. Based on the definition of the Lipschitz continuity, $\forall\, \boldsymbol{x}_1, \boldsymbol{x}_2 \in \boldsymbol{X}$, there exists a positive constant K such that $|h(\boldsymbol{x}_1, \boldsymbol{y}) - h(\boldsymbol{x}_2, \boldsymbol{y})| \leq K|\boldsymbol{x}_1 - \boldsymbol{x}_2|$. Therefore, considering the sum rule of integration, we have

$$
\begin{aligned}
|H(\boldsymbol{x}_1) - H(\boldsymbol{x}_2)| &= \left|\int_{\boldsymbol{y}\in\boldsymbol{Y}} [h(\boldsymbol{x}_1, \boldsymbol{y}) - h(\boldsymbol{x}_2, \boldsymbol{y})] g(\boldsymbol{y}) d\boldsymbol{y}\right| \\
&\leq \int_{\boldsymbol{y}\in\boldsymbol{Y}} |h(\boldsymbol{x}_1, \boldsymbol{y}) - h(\boldsymbol{x}_2, \boldsymbol{y})| g(\boldsymbol{y}) d\boldsymbol{y} \\
&\leq \int_{\boldsymbol{y}\in\boldsymbol{Y}} K|\boldsymbol{x}_1 - \boldsymbol{x}_2| g(\boldsymbol{y}) d\boldsymbol{y} \\
&= K|\boldsymbol{x}_1 - \boldsymbol{x}_2| \int_{\boldsymbol{y}\in\boldsymbol{Y}} g(\boldsymbol{y}) d\boldsymbol{y} = K|\boldsymbol{x}_1 - \boldsymbol{x}_2|.
\end{aligned}
$$

In other words, if $h(\boldsymbol{x}, \boldsymbol{y})$ is Lipschitz continuous with respect to $\boldsymbol{x}$, we have proved that $H(\boldsymbol{x})$ is also Lipschitz continuous. Consequently, if we can prove that $\nabla_{\{\mathbf{L}^{(m)}, \mathbf{B}^{(m)}\}} \log\left(f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)}, \boldsymbol{z}_i^{(m)}\right)\right)$ is Lipschitz continuous with respect to $\{\mathbf{L}^{(m)}, \mathbf{B}^{(m)}\}$, it indicates that the term $\nabla_{\{\mathbf{L}^{(m)}, \mathbf{B}^{(m)}\}} \sum_{i=1}^{N} E_{\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\, \hat{\boldsymbol{\Theta}}^{(j-1)}} \left[\log\left(f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)}, \boldsymbol{z}_i^{(m)}\right)\right)\right]$ is Lipschitz continuous.

Last, According to the statement in page 19 of Wheeden et al. (2015), if $\nabla_{\{\mathbf{L}^{(m)},\mathbf{B}^{(m)}\}} \log\left(f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)},\boldsymbol{z}_i^{(m)};\boldsymbol{\Theta}_{2m}\right)\right)$ has a continuous derivative with respect to $\{\mathbf{L}^{(m)},\mathbf{B}^{(m)}\}$, then $\nabla_{\{\mathbf{L}^{(m)},\mathbf{B}^{(m)}\}} \log\left(f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)},\boldsymbol{z}_i^{(m)};\boldsymbol{\Theta}_{2m}\right)\right)$ is Lipschitz continuous. By checking the property of second derivates in Kim (2004), we have that $\nabla_{\{\mathbf{L}^{(m)},\mathbf{B}^{(m)}\}} \log\left(f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)},\boldsymbol{z}_i^{(m)};\boldsymbol{\Theta}_{2m}\right)\right)$ is continuous differentiable that implies $\nabla_{\{\mathbf{L}^{(m)},\mathbf{B}^{(m)}\}} \log\left(f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)},\boldsymbol{z}_i^{(m)};\boldsymbol{\Theta}_{2m}\right)\right)$ is Lipschitz continuous. Therefore, $\sum_{i=1}^{N} E_{\boldsymbol{\eta}_i^{(m)}|\boldsymbol{x}_i^{(m)};\,\widehat{\boldsymbol{\Theta}}^{(j-1)}}\left[\log\left(f\left(\boldsymbol{x}_i^{(m)}|\boldsymbol{\eta}_i^{(m)},\boldsymbol{z}_i^{(m)}\right)\right)\right]$ has the Lipschitz continuous gradient with respect to $\{\mathbf{L}^{(m)},\mathbf{B}^{(m)}\}$.

■

G. Proof of Theorem 3.1

Proof: For simplicity, we define $\hat{\mathbf{b}}_p \triangleq vec(\widehat{\mathbf{B}}_p)$, then we have $vec\left(\widetilde{\mathbf{B}}_{p,k}^{*}\right) = \hat{\mathbf{b}}_{p,k}^{*}$. Similarly, we can rewrite the true parameter as $vec(\mathbf{B}_{p,0}) = \mathbf{b}_{p,0}$. First, we prove that $\sqrt{N}\left(\hat{\mathbf{b}}_p - \mathbf{b}_{p,0}\right) = V(\mathbf{b}_{p,0})$ given $K \ll \sqrt{N}$. Specifically, $\mathbb{E}\left(V(\mathbf{b}_{p,0})\right) = \mathbf{0}$, $cov\left(V(\mathbf{b}_{p,0})\right) = \boldsymbol{\Sigma}_p$, where $\boldsymbol{\Sigma}_p = \left(\sum_{k=1}^{K} \frac{N_k}{N} \boldsymbol{\Sigma}_{p,k}^{-1}\right)^{-1}$ and $\boldsymbol{\Sigma}_{p,k} = N_k cov(\hat{\mathbf{b}}_{p,k}^{*})$.

Before analyzing the asymptotic property of $\hat{\mathbf{b}}_p$, we shall prove $\boldsymbol{\Sigma}_p$ above is the asymptotic covariance matrix of $vec(\widetilde{\mathbf{B}}_p^{*})$.

It is clear that $\boldsymbol{\Sigma}_{p,k}$ is the asymptotic covariance matrix of $\hat{\mathbf{b}}_{p,k}^{*}$, which means $\mathbb{E}\left(N_k^{-1} \sum_{n \in S_k} l''_{n,p}\left(\hat{\mathbf{b}}_{p,k}^{*}\right)\right) = \mathbb{E}\left(N_k^{-1} \sum_{n \in S_k} l''_{n,p}\left(\mathbf{b}_{p,0}\right)\right) \approx \boldsymbol{\Sigma}_{p,k}^{-1}$. The first equality comes from $\hat{\mathbf{b}}_{p,k}^{*}$ is an unbiased estimator of $\mathbf{b}_{p,0}$. Therefore,

$$\boldsymbol{\Sigma}_p^{-1} = \sum_{k=1}^{K} \frac{N_k}{N} \boldsymbol{\Sigma}_{p,k}^{-1} \approx \sum_{k=1}^{K} \frac{N_k}{N} \mathbb{E}\big(N_k^{-1} \sum\nolimits_{n \in S_k} l''_{n,p}\big(\mathbf{b}_{p,0}\big)\big) = \mathbb{E}\big(N^{-1} \sum_{k=1}^{K} \sum\nolimits_{n \in S_k} l''_{n,p}\big(\mathbf{b}_{p,0}\big)\big).$$

The last equality above comes from the samples are i.i.d across different local servers. It implies $\boldsymbol{\Sigma}_p^{-1} \approx \mathbb{E}\big(N^{-1} \sum_{k=1}^{K} \sum\nolimits_{n \in S_k} l''_{n,p}\big(\mathbf{b}_{p,0}\big)\big) = \mathbb{E}\big(N^{-1} \sum_{k=1}^{K} \sum\nolimits_{n \in S_k} l''_{n,p}\big(vec(\widetilde{\mathbf{B}}_p^*)\big)\big)$, which means $\boldsymbol{\Sigma}_p$ is the asymptotic covariance matrix of $vec(\widetilde{\mathbf{B}}_p^*)$.

Note that $\hat{\mathbf{b}}_p - \mathbf{b}_{p,0} = \left(\sum_{k=1}^{K} \frac{N_k}{N} \widehat{\boldsymbol{\Sigma}}_{p,k}^{-1}\right)^{-1} \left(\sum_{k=1}^{K} \frac{N_k}{N} \widehat{\boldsymbol{\Sigma}}_{p,k}^{-1} \left(\hat{\mathbf{b}}_{p,k}^* - \mathbf{b}_{p,0}\right)\right)$. By Slutsky's Theorem, it suffices to prove the following two statements.

$$\sum_{k=1}^{K} \frac{N_k}{N} \widehat{\boldsymbol{\Sigma}}_{p,k}^{-1} \rightarrow_p \boldsymbol{\Sigma}_p^{-1}, \tag{G.1}$$

$$\sqrt{N}\left(\sum_{k=1}^{K} \frac{N_k}{N} \widehat{\boldsymbol{\Sigma}}_{p,k}^{-1} \big(\hat{\mathbf{b}}_{p,k}^* - \mathbf{b}_{p,0}\big)\right) = V'\big(\mathbf{b}_{p,0}\big), \tag{G.2}$$

in which $cov\left(V'\big(\mathbf{b}_{p,0}\big)\right) = \boldsymbol{\Sigma}_p^{-1}$, and $\mathbb{E}\left(V'\big(\mathbf{b}_{p,0}\big)\right) = \mathbf{0}$.

As defined in equation (3.13), $\widehat{\boldsymbol{\Sigma}}_{p,k}$ is the estimator of $\boldsymbol{\Sigma}_{p,k}$. In addition, $\widehat{\boldsymbol{\Sigma}}_{p,k}^{-1} - \boldsymbol{\Sigma}_{p,k}^{-1} = O_p\left(N_k^{-\frac{1}{2}}\right)$. Then, we have

$$\sum_{k=1}^{K} \frac{N_k}{N} \widehat{\boldsymbol{\Sigma}}_{p,k}^{-1} - \boldsymbol{\Sigma}_p^{-1} = \sum_{k=1}^{K} \frac{N_k}{N} \big(\widehat{\boldsymbol{\Sigma}}_{p,k}^{-1} - \boldsymbol{\Sigma}_{p,k}^{-1}\big) = O_p\left(N^{-\frac{1}{2}}\right) = o_p(1).$$

In other words, this proves that $\sum_{k=1}^{K} \frac{N_k}{N} \widehat{\boldsymbol{\Sigma}}_{p,k}^{-1} \rightarrow_p \boldsymbol{\Sigma}_p^{-1}$ as shown in (G.1). Note that $O_p(\cdot)$ is a shorthand means of characterizing the convergence in probability of a set of random variables, as well as $o_p(\cdot)$ refers to convergence in probability towards zero.

Recall $\hat{\mathbf{b}}_{p,k}$ is the minimizer of $\sum\nolimits_{n \in S_k} l_{n,p}\big(\mathbf{b}_p\big)$. Since $\frac{\partial^2 \sum\nolimits_{n \in S_k} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p d\mathbf{b}_p{}^T} = 2J_{\eta\eta}{}^T \otimes \mathbf{Z}_{p,k}{}^T \mathbf{Z}_{p,k}$, then $\nabla^i \sum\nolimits_{n \in S_k} l_{n,p}\big(\mathbf{b}_p\big) = 0$ for $i \in \{3, 4, \cdots, \infty\}$. By applying Taylor's expansion on $\frac{\partial \sum\nolimits_{n \in S_k} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p}$ at $\mathbf{b}_{p,0}$, we have

$$\mathbf{0} = \left.\frac{\partial \sum_{n\in S_k}^{\square} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p}\right|_{\mathbf{b}_p=\hat{\mathbf{b}}_{p,k}^{*}} = \left.\frac{\partial \sum_{n\in S_k}^{\square} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p}\right|_{\mathbf{b}_p=\mathbf{b}_{p,0}} + \left.\frac{\partial^2 \sum_{n\in S_k}^{\square} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p d\mathbf{b}_p{}^T}\right|_{\mathbf{b}_p=\mathbf{b}_{p,0}} (\hat{\mathbf{b}}_{p,k}^{*} - \mathbf{b}_{p,0}). \quad \text{(G.3)}$$

By standard arguments,

$$\left.\frac{\partial^2 \sum_{n\in S_k}^{\square} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p d\mathbf{b}_p{}^T}\right|_{\mathbf{b}_p=\mathbf{b}_{p,0}} = \mathbb{E}\left(\left.\frac{\partial^2 \sum_{n\in S_k}^{\square} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p d\mathbf{b}_p{}^T}\right|_{\mathbf{b}_p=\mathbf{b}_{p,0}}\right) + O_p\left(N_k^{\frac{1}{2}}\right)$$

$$= N_k \boldsymbol{\Sigma}_{p,k}^{-1} + O_p\left(N_k^{\frac{1}{2}}\right).$$

Equation (G.3) can be rewritten as

$$\mathbf{0} = \left.\frac{\partial \sum_{n\in S_k}^{\square} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p}\right|_{\mathbf{b}_p=\mathbf{b}_{p,0}} + \left(N_k \boldsymbol{\Sigma}_{p,k}^{-1} + O_p\left(N_k^{\frac{1}{2}}\right)\right)(\hat{\mathbf{b}}_{p,k}^{*} - \mathbf{b}_{p,0}).$$

Furthermore, given that $\hat{\mathbf{b}}_{p,k}^{*} - \mathbf{b}_{p,0} = O_p\left(N_k^{-\frac{1}{2}}\right)$, equation (G.3) implies

$$\hat{\mathbf{b}}_{p,k}^{*} - \mathbf{b}_{p,0} = -{N_k}^{-1} \boldsymbol{\Sigma}_{p,k} \left.\frac{\partial \sum_{n\in S_k}^{\square} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p}\right|_{\mathbf{b}_p=\mathbf{b}_{p,0}} - O_p(N_k^{-1}).$$

Therefore, given $\widehat{\boldsymbol{\Sigma}}_{p,k}^{-1} - \boldsymbol{\Sigma}_{p,k}^{-1} = O_p\left(N_k^{-\frac{1}{2}}\right)$, we have

$$\sqrt{N}\left(\sum_{k=1}^{K} \frac{N_k}{N} \widehat{\boldsymbol{\Sigma}}_{p,k}^{-1}(\hat{\mathbf{b}}_{p,k} - \mathbf{b}_{p,0})\right)$$

$$= \sqrt{N}\left(\sum_{k=1}^{K} \frac{N_k}{N} \boldsymbol{\Sigma}_{p,k}^{-1}(\hat{\mathbf{b}}_{p,k}^{*} - \mathbf{b}_{p,0})\right) + \sqrt{N}\left(\sum_{k=1}^{K} \frac{N_k}{N} (\widehat{\boldsymbol{\Sigma}}_{p,k}^{-1} - \boldsymbol{\Sigma}_{p,k}^{-1})(\hat{\mathbf{b}}_{p,k}^{*} - \mathbf{b}_{p,0})\right)$$

$$= \sqrt{N}\left(\sum_{k=1}^{K} \frac{N_k}{N} \boldsymbol{\Sigma}_{p,k}^{-1}(\hat{\mathbf{b}}_{p,k}^{*} - \mathbf{b}_{p,0})\right) + O_p\left(\frac{K}{\sqrt{N}}\right)$$

$$= \sqrt{N}\left(\sum_{k=1}^{K} \frac{N_k}{N} \boldsymbol{\Sigma}_{p,k}^{-1}\left(-N_k^{-1} \boldsymbol{\Sigma}_{p,k} \left.\frac{\partial \sum_{n\in S_k}^{\square} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p}\right|_{\mathbf{b}_p=\mathbf{b}_{p,0}} - O_p(N_k^{-1})\right)\right)$$

$$+ O_p\left(\frac{K}{\sqrt{N}}\right)$$

$$= -\frac{1}{\sqrt{N}}\sum_{k=1}^{K}\frac{\partial\sum_{n\in S_k}^{\square} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p}\Big|_{\mathbf{b}_p=\mathbf{b}_{p,0}} + O_p\left(\frac{K}{\sqrt{N}}\right). \quad \text{(G.4)}$$

As we assume $K \ll \sqrt{N}$, $O_p\left(\frac{K}{\sqrt{N}}\right) = o_p(1)$. Define

$$V'(\mathbf{b}_{p,0}) = -\frac{1}{\sqrt{N}}\sum_{k=1}^{K}\frac{\partial\sum_{n\in S_k}^{\square} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p}\Big|_{\mathbf{b}_p=\mathbf{b}_{p,0}}.$$

We have

$$\mathbb{E}\left(V'(\mathbf{b}_{p,0})\right) = \mathbf{0}, \quad \text{(G.5)}$$

and

$$cov\left(V'(\mathbf{b}_{p,0})\right) = \frac{1}{N}\sum_{k=1}^{K} cov\left(\frac{\partial\sum_{n\in S_k}^{\square} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p}\Big|_{\mathbf{b}_p=\mathbf{b}_{p,0}}\right)$$

$$= \frac{1}{N}\sum_{k=1}^{K}\mathbb{E}\left(\frac{\partial\sum_{n\in S_k}^{\square} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p}\Big|_{\mathbf{b}_p=\mathbf{b}_{p,0}}\left(\frac{\partial\sum_{n\in S_k}^{\square} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p}\Big|_{\mathbf{b}_p=\mathbf{b}_{p,0}}\right)^T\right).$$

Since equation (3.10) is a minimization problem, the corresponding information equality is

$$\mathbb{E}\left[\left(\frac{\partial\left(-l_{n,p}(\mathbf{b}_p)\right)}{\partial \mathbf{b}_p}\right)\left(\frac{\partial\left(-l_{n,p}(\mathbf{b}_p)\right)}{\partial \mathbf{b}_p}\right)^T\right] = -\mathbb{E}\left[\frac{\partial^2\left(-l_{n,p}(\mathbf{b}_p)\right)}{\partial \mathbf{b}_p \partial \mathbf{b}_p{}^T}\right],$$

which indicates that $\mathbb{E}\left[\left(\frac{\partial l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p}\right)\left(\frac{\partial l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p}\right)^T\right] = \mathbb{E}\left[\frac{\partial^2 l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p \partial \mathbf{b}_p{}^T}\right]$. Thus,

$$cov\left(V'(\mathbf{b}_{p,0})\right) = \frac{1}{N}\sum_{k=1}^{K}\mathbb{E}\left(\frac{\partial^2\sum_{n\in S_k}^{\square} l_{n,p}(\mathbf{b}_p)}{\partial \mathbf{b}_p d\mathbf{b}_p{}^T}\Big|_{\mathbf{b}_p=\mathbf{b}_{p,0}}\right) = \sum_{k=1}^{K}\frac{N_k}{N}\,\boldsymbol{\Sigma}_{p,k}^{-1} = \boldsymbol{\Sigma}_p^{-1}. \quad \text{(G.6)}$$

For a short summary, condition on $K \ll \sqrt{N}$, equation (G.4) gives

$$\sqrt{N}\left(\sum_{k=1}^{K}\frac{N_k}{N}\widehat{\boldsymbol{\Sigma}}_{p,k}^{-1}\left(\hat{\mathbf{b}}_{p,k}-\mathbf{b}_{p,0}\right)\right)=-\frac{1}{\sqrt{N}}\sum_{k=1}^{K}\left.\frac{\partial\sum_{n\in S_k}l_{n,p}(\mathbf{b}_p)}{\partial\mathbf{b}_p}\right|_{\mathbf{b}_p=\mathbf{b}_{p,0}}=V'\left(\mathbf{b}_{p,0}\right).$$

Equation (G.5) gives $\mathbb{E}\left(V'\left(\mathbf{b}_{p,0}\right)\right)=\mathbf{0}$, and equation (G.6) gives $cov\left(V'\left(\mathbf{b}_{p,0}\right)\right)=\boldsymbol{\Sigma}_p^{-1}$. Together with (G.4, G.5, G.6), we can prove (G.2).

By the Central Limit Theorem, we have $V'\left(\mathbf{b}_{p,0}\right)\rightarrow_d N(\mathbf{0},\boldsymbol{\Sigma}_p^{-1})$. By Slutsky's Theorem, we have $\sqrt{N}\left(\hat{\mathbf{b}}_p-\mathbf{b}_{p,0}\right)=\frac{V'(\mathbf{b}_{p,0})}{\boldsymbol{\Sigma}_p^{-1}}$, where $\boldsymbol{\Sigma}_p=\left(\sum_{k=1}^{K}\frac{N_k}{N}\boldsymbol{\Sigma}_{p,k}^{-1}\right)^{-1}$.

Thus, we have $\sqrt{N}\left(\hat{\mathbf{b}}_p-\mathbf{b}_{p,0}\right)\rightarrow_d N(\mathbf{0},\boldsymbol{\Sigma}_p)$. This indicates that the proposed LSA estimator $\widehat{\mathbf{B}}_p$ achieves the same asymptotic normality as the global estimator $\widetilde{\mathbf{B}}_p^*$.

■

H. Pseudo code for the fed-GB-Average in Step 1

Table 14: Pseudo code for the fed-GB-Average on local and central servers in Step 1

Step 1: Iterative update to obtain the global aggregator $\left\{\widehat{\mathbf{B}}_p^{(\omega)}\right\}_{p=1}^{P}$ for all base learners
Initialization: $n_1 = 0$; $\widehat{\mathbf{B}}_p^{(0)}$ for $p = 1, \dots, P$ **Iterate until** $n_1 = N_1$ Local servers with parallel computing for $k = 1, \dots, K$ **Initialization**: $n_2 = 0$; $\widehat{\mathbf{B}}_{p,k}^{(0)} = \widehat{\mathbf{B}}_p^{(n_1)}$ for $p = 1, \dots, P$ **Iterate until** $n_2 = N_2$ • Compute local step length η_k • $\widehat{\mathbf{B}}_{p,k}^{(n_2+1)} = \widehat{\mathbf{B}}_{p,k}^{(n_2)} - \eta_k l'_{n,p}\left(\widehat{\mathbf{B}}_{p,k}^{(n_2)}\right)$ for $p = 1, \dots, P$ • $n_2 = n_2 + 1$ Send local parameters $\widehat{\mathbf{B}}_{p,k}^{(N_2)}$ to the central server Central server • Receive the local parameters $\left\{\widehat{\mathbf{B}}_{p,k}^{(N_2)}\right\}_{k=1}^{K}$ for $p = 1, \dots, P$ • Aggregate local parameters by $\widehat{\mathbf{B}}_p^{(n_1)} = \frac{1}{N}\sum_{k=1}^{K} N_k \widehat{\mathbf{B}}_{p,k}^{(N_2)}$ for $p = 1, \dots, P$ • Send parameters $\widehat{\mathbf{B}}_p^{(n_1)}$ to the local servers • $n_1 = n_1 + 1$ Send the aggregated parameters $\left\{\widehat{\mathbf{B}}_p^{(\omega)}\right\}_{p=1}^{P} = \left\{\widehat{\mathbf{B}}_p^{(N_1)}\right\}_{p=1}^{P}$ to the local servers

I. Lemma 4.3 and Proof

Lemma 4.3 Given a constant $c > 0$ and vector $\boldsymbol{x} \in \mathbb{R}^d \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Then the following inequality holds:

$$\mathbb{P}[\|\boldsymbol{x}\| > c] \leq 2\exp\left(-\frac{c^2}{2\mathrm{tr}(\boldsymbol{\Sigma})}\right).$$

Proof of Lemma 4.3:

Let $\mathbf{\Gamma\Lambda\Gamma}^T$ be the eigen decomposition of $\mathbf{\Sigma}$. Then $\|\boldsymbol{x}\|_2 =_d \left\|\mathbf{\Lambda}^{1/2}\boldsymbol{y}\right\|_2$, where $\boldsymbol{y}\sim\mathcal{N}(\mathbf{0}, \mathbf{I}_{d\times d})$. For all $c > 0$ and $s > 0$, we have

$$\begin{aligned}
\mathbb{P}[\|\boldsymbol{x}\|_2 > c] &= \mathbb{P}\left[\left\|\mathbf{\Lambda}^{1/2}\boldsymbol{y}\right\|_2 > c\right] \\
&\leq \mathbb{P}\left[\left\|\mathbf{\Lambda}^{1/2}\boldsymbol{y}\right\|_1 > c\right] \\
&= \mathbb{P}\left[\exp\left(s\textstyle\sum_{i=1}^d \sqrt{\lambda_i}|y_i|\right) > \exp(sc)\right] \\
&\leq \exp(-sc)\,\mathbb{E}\left[\exp\left(s\textstyle\sum_{i=1}^d \sqrt{\lambda_i}|y_i|\right)\right] \\
&= \exp(-sc)\textstyle\prod_{i=1}^d \mathbb{E}\left[\exp\left(s\sqrt{\lambda_i}|y_i|\right)\right] \\
&= 2\exp\left(-sc + \frac{s^2}{2}\textstyle\sum_{i=1}^d \lambda_i\right),
\end{aligned}$$

where $\lambda_1, \cdots, \lambda_d$ are the eigenvalues of Λ. The first inequality comes from the comparison of ℓ^1 norm and ℓ^2 norm. The second inequality comes from Markov's inequality. The second equality comes from the monotonicity of exponential function. The third equality comes from the independence of $y_1, \cdots, y_d$. The last equality comes from the moment generating function of folded normal distribution, which is $\mathbb{E}\left[\exp\left(s\sqrt{\lambda_i}|y_i|\right)\right] = 2\exp\left(\frac{s^2\lambda_i}{2}\right)$.

Taking $s = \frac{c}{\sum_{i=1}^d \lambda_i}$, we have $\mathbb{P}[\|\mathbf{x}\|_2 > c] \leq 2\exp\left(-\frac{c^2}{2\mathrm{tr}(\Sigma)}\right)$, which comes from the sum of the eigenvalues of Σ equals to the trace of Σ.

■

J. Proof of Theorem 4.1:

Denote $\|\cdot\|$ as ℓ^2 norm and $\|\cdot\|_F$ as Frobenius norm. By Cauchy-Schwarz Inequality, we have

$$\begin{aligned}\|vec(\mathbf{B}_2) - vec(\mathbf{B}_1)\| &= \left\|\left(\frac{1}{N}\mathbf{J}_{\eta\eta} \otimes \mathbf{Z}^T\mathbf{Z}\right)^{-1} vec\left(\frac{1}{N}\mathbf{Z}^T\mathbf{R}\mathbf{J}_{\eta\eta}\right)\right\| \\ &\leq \left\|\left(\frac{1}{N}\mathbf{J}_{\eta\eta} \otimes \mathbf{Z}^T\mathbf{Z}\right)^{-1}\right\| \left\|vec\left(\frac{1}{N}\mathbf{Z}^T\mathbf{R}\mathbf{J}_{\eta\eta}\right)\right\| \\ &= \left\|\left(\mathbf{J}_{\eta\eta} \otimes \frac{\mathbf{Z}^T\mathbf{Z}}{N}\right)^{-1}\right\| \left\|\frac{1}{N}\mathbf{Z}^T\mathbf{R}\mathbf{J}_{\eta\eta}\right\|_F \\ &\leq \left\|\left(\mathbf{J}_{\eta\eta} \otimes \frac{\mathbf{Z}^T\mathbf{Z}}{N}\right)^{-1}\right\| \left\|\frac{1}{N}vec(\mathbf{Z}^T\mathbf{R})\right\| \|\mathbf{J}_{\eta\eta}\|_F,\end{aligned}$$

in which, for an entry r_{ij} of $\mathbf{R}$, we have $r_{ij} \sim \mathcal{N}(0, 4Q_2\sigma_{\varepsilon,\delta}^2)$.

Denote $\mathbf{K} = \mathbf{Z}^T\mathbf{R} \in \mathbb{R}^{Q_1 \times Q_2}$ with $k_{ij} = \boldsymbol{z}_i^T\boldsymbol{r}_j$, where $\boldsymbol{z}_i$ is the ith column of $\mathbf{Z}$, and $\boldsymbol{r}_j$ is the j th column of $\mathbf{R}$. Given $\mathbf{Z}$, we have $k_{ij} \sim \mathcal{N}(0, \sum_{l=1}^{N} z_{il}^2 4Q_2\sigma_{\varepsilon,\delta}^2)$. Then, $vec(\mathbf{K}) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{K}})$, where $\boldsymbol{\Sigma}_{\mathbf{K}} \in \mathbb{R}^{Q_1Q_2 \times Q_1Q_2}$ is a diagonal matrix with $\mathrm{tr}(\boldsymbol{\Sigma}_{\mathbf{K}}) = 4Q_2^2\sigma_{\varepsilon,\delta}^2 \sum_{i=1}^{Q_1}\sum_{l=1}^{N} z_{il}^2 = 4Q_2^2\sigma_{\varepsilon,\delta}^2\|\mathbf{Z}\|_F^2$. Thus, we have

$$\mathbb{P}\left[\left\|\frac{1}{N}vec(\mathbf{Z}^T\mathbf{R})\right\| > \beta\right] = \mathbb{P}[\|vec(\mathbf{Z}^T\mathbf{R})\|_2 > N\beta] \leq 2\exp\left(-\frac{N^2\beta^2}{8Q_2^2\sigma_{\varepsilon,\delta}^2\|\mathbf{Z}\|_F^2}\right).$$

The above inequality is implied by Lemma 4.3. Moreover, we have $\frac{\mathbf{Z}^T\mathbf{Z}}{N} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{z}_i \cdot \boldsymbol{z}_i^T$, where $\boldsymbol{z}_i$ is the ith row of $\mathbf{Z}$. This means $\left\|\left(\mathbf{J}_{\eta\eta} \otimes \frac{\mathbf{Z}^T\mathbf{Z}}{N}\right)^{-1}\right\| = O(1)$.

Therefore, we can conclude that $\mathbb{P}[\|vec(\mathbf{B}_2) - vec(\mathbf{B}_1)\| > \beta] \leq O\left(\exp\left(-\frac{N^2\beta^2}{8Q_2^2\sigma_{\varepsilon,\delta}^2\|\mathbf{Z}\|_F^2}\right)\right)$.

■

K. Proof of Corollary 4.1:

To prove $\operatorname{plim}_{n\to\infty} vec(\mathbf{B}_2) = \operatorname{plim}_{n\to\infty} vec(\mathbf{B}_1)$, it equals to prove

$$\lim_{n\to\infty} \mathbb{P}[\|vec(\mathbf{B}_2) - vec(\mathbf{B}_1)\| > \beta] = 0$$

for any arbitrary small β. Given theorem 4.1, it equals to prove

$$\lim_{n\to\infty} O\left(\exp\left(-\frac{N^2\beta^2}{8Q_2^2\sigma_{\varepsilon,\delta}^2\|\mathbf{Z}\|_F^2}\right)\right) = 0.$$

It is equivalent to prove $\lim_{n\to\infty} \frac{\|\mathbf{Z}\|_F^2}{N^2} = 0$. Since $\|\mathbf{Z}\|_F^2 = \sum_{i=1}^{N}\sum_{j=1}^{Q_1} z_{ij}^2 \leq NQ_1 z_{\text{MAX}}^2$, where $z_{\text{MAX}} = \|\mathbf{Z}\|_{\text{MAX}}$. Because $\mathbf{Z} \in \mathcal{R}^{N\times Q_1}$, we have $\lim_{n\to\infty} \frac{\|\mathbf{Z}\|_F^2}{N^2} = 0$, which proves Corollary 4.1.

■

References

Agresti, A. (2010). *Analysis of ordinal categorical data*, Volume 656. John Wiley & Sons. Allen, T. T., Z.

Alramadeen, W., Ding, Y., Costa, C., & Si, B. (2023). A novel sparse linear mixed model for multi-source mixed-frequency data fusion in telemedicine. *IISE transactions on healthcare systems engineering*, *13*(3), 215-225.

Alramadeen, W., Rababa, S., Costa, C., & Si, B. (2022, August). Multi-level multi-channel bio-signal analysis for health telemonitoring. In *2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)* (pp. 1539-1544). IEEE.

American College of Cardiology (2021). *Cardiometabolic initiatives*. Washington, DC. Available at https://www.acc.org/tools-and-practice-support/quality-programs/cardiometabolic-health-alliance.

American Diabetes Association (2018). Economic costs of diabetes in the us in 2017. *Diabetes care* 41 (5), 917–928.

Barndorff-Nielsen, O. (2014). *Information and exponential families: in statistical theory*. John Wiley & Sons.

Benmalek, M., Benrekia, M. A., & Challal, Y. (2022). Security of federated learning: Attacks, defensive mechanisms, and challenges. Revue des Sciences et Technologies de l'Information-Série RIA: Revue d'Intelligence Artificielle, 36(1), 49-59.

Björgvinsson T, Kertz SJ, Bigda-Peyton JS, McCoy KL, Aderka IM (2013). Psychometric properties of the CES-D-10 in a psychiatric sample. *Assessment* 20(4), 429-36.

Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017, October). Practical secure aggregation for privacy-preserving machine learning. In *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1175-1191).

Bosq, D. (2000). Linear processes in function spaces: theory and applications (Vol. 149). Springer Science & Business Media.

Boyd, S., S. P. Boyd, and L. Vandenberghe (2004). *Convex optimization*. Cambridge university press.

Brockhaus, S., & Rügamer, D. (2016). FDboost: Boosting functional regression models. *R package version 0.2-0, URL https://CRAN. R-project. org/package= FDboost*.

Brockhaus, S., Rügamer, D., & Greven, S. (2017). Boosting functional regression models with FDboost. arXiv preprint arXiv:1705.10662.

Brown, P. J., Fearn, T., & Vannucci, M. (2001). Bayesian wavelet regression on curves with application to a spectroscopic calibration problem. Journal of the American Statistical Association, 96(454), 398-408.

Bühlmann, P., & Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting.

Buxton, O. M., Lee, S., Marino, M., Beverly, C., Almeida, D. M., & Berkman, L. (2018). Sleep health and predicted cardiometabolic risk scores in employed adults from two industries. *Journal of Clinical Sleep Medicine*, *14*(3), 371-383.

Cai, X., J. A. Bazerque, and G. B. Giannakis (2013). Inference of gene regulatory networks with sparse structural equation models exploiting genetic perturbations. *PLoS computational biology 9* (5), e1003068.

Cardot, H., & Sarda, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, *92*(1), 24-41.

Chen, C., Zhou, J., Zheng, L., Wu, H., Lyu, L., Wu, J., ... & Zheng, X. (2020). Vertically federated graph neural network for privacy-preserving node classification. arXiv preprint arXiv:2005.11903.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Zhou, T. (2015). Xgboost: extreme gradient boosting. *R package version 0.4-2*, *1*(4), 1-4.

Cheng, K., Fan, T., Jin, Y., Liu, Y., Chen, T., Papadopoulos, D., & Yang, Q. (2021). Secureboost: A lossless federated learning framework. IEEE Intelligent Systems, 36(6), 87-98.

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

Davis, P. J. and P. Rabinowitz (2007). *Methods of numerical integration*. Courier Corporation.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological) 39* (1), 1–22.

Dennis, J. E. and Schnabel, R. B. (1983). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations. Prentice-Hall, Englewood Cliffs, NJ.*

Ding, Y., Yang, Q., King, C. B., & Hong, Y. (2019). A general accelerated destructive degradation testing model for reliability analysis. *IEEE Transactions on Reliability*, *68*(4), 1272-1282.

Dobson, A. J. and A. G. Barnett (2018). *An introduction to generalized linear models.* CRC press.

Du, P., & Wang, X. (2014). Penalized likelihood functional regression. *Statistica Sinica*, 1017-1041.

Dwork, C. (2011). A firm foundation for private data analysis. *Communications of the ACM*, *54*(1), 86-95.

Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, *9*(3–4), 211-407.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). 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* (pp. 265-284). Springer Berlin Heidelberg

Ehrich S (2002). On stratified extensions of Gauss–Laguerre and Gauss–Hermite quadrature formulas. *Journal of computational and applied mathematics* 140(1-2):291-9.

Ferraty, F. (2006). Nonparametric functional data analysis. Springer.

Freund, Y., & Schapire, R. E. (1995, March). A desicion-theoretic generalization of on-line learning and an application to boosting. In *European conference on computational learning theory* (pp. 23-37). Berlin, Heidelberg: Springer Berlin Heidelberg.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

Friedman, J. (2001). Greedy boosting approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, *28*(2), 337-407.

Fryar, C. D., T.-C. Chen, and X. Li (2012). *Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999-2010*. Number 103. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.

Geyer, R. C., Klein, T., & Nabi, M. (2017). Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557.

Ghazi, B., Golowich, N., Kumar, R., Manurangsi, P., & Zhang, C. (2021). Deep learning with label differential privacy. Advances in neural information processing systems, 34, 27131-27145.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, *101*(23), e215-e220.

Goldsmith, J., Crainiceanu, C. M., Caffo, B., & Reich, D. (2012). Longitudinal penalized functional regression for cognitive outcomes on neuronal tract measurements. Journal of the Royal Statistical Society Series C: Applied Statistics, 61(3), 453-469.

Guha, N., Talwalkar, A., & Smith, V. (2019). One-shot federated learning. *arXiv preprint arXiv:1902.11175*.

Guo, W., Dai, M., Ombao, H. C., and von Sachs, R. (2003), "Smoothing Spline ANOVA for Time-Dependent Spectral Analysis," Journal of the American Statistical Association, 98, 643–652.

Hardy, S., Henecka, W., Ivey-Law, H., Nock, R., Patrini, G., Smith, G., & Thorne, B. (2017). Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. arXiv preprint arXiv:1711.10677.

Hashemi, N., Safari, P., Shariati, B., & Fischer, J. K. (2021, September). Vertical federated learning for privacy-preserving ML model development in partially disaggregated networks. In 2021 European Conference on Optical Communication (ECOC) (pp. 1-4). IEEE.

He, D., Du, R., Zhu, S., Zhang, M., Liang, K., & Chan, S. (2021). Secure logistic regression for vertical federated learning. IEEE Internet Computing, 26(2), 61-68.

Heiser, W. J. (1995). Convergent computation by iterative majorization. *Recent advances in descriptive multivariate analysis*, 157–189.

Hildebrand, F. B. (1987). *Introduction to numerical analysis*. Courier Corporation.

Hildreth, C. (1957). A quadratic programming procedure. *Naval research logistics quarterly 4* (1), 79–85.

Horváth, L., & Kokoszka, P. (2012). Inference for functional data with applications (Vol. 200). Springer Science & Business Media.

Hothorn, T., Kneib, T., & Bühlmann, P. (2014). Conditional transformation models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *76*(1), 3-27.

Hsing, T., & Eubank, R. (2015). Theoretical foundations of functional data analysis, with an introduction to linear operators (Vol. 997). John Wiley & Sons.

Huang, P.-H. (2018). A penalized likelihood method for multi-group structural equation modelling. *British Journal of Mathematical and Statistical Psychology 71* (3), 499–522.

Huang, P.-H. (2020). lslx: Semi-confirmatory structural equation modeling via penalized likelihood. *Journal of Statistical Software 93*, 1–37.

Huang, P.-H., H. Chen, and L.-J. Weng (2017). A penalized likelihood method for structural equation modeling. *psychometrika 82* (2), 329–354.

Hughes-Hallett, D., Gleason, A. M., & McCallum, W. G. (2020). *Calculus: Single and multivariable*. John Wiley & Sons.

Hunter DR, Lange K (2004). A tutorial on MM algorithms. *The American Statistician* 58(1):30-7.

Imaizumi, M., & Kato, K. (2018). PCA-based estimation for functional linear regression with functional responses. *Journal of multivariate analysis*, *163*, 15-36.

Ivanescu, A. E., Staicu, A.-M., Scheipl, F., and Greven, S. (2015), "Penalized Function-on-Function Regression," Computational Statistics, 30, 539– 568.

Jackel, P. (2005). *A note on multivariate gauss-hermite quadrature*. London: ABN-Amro. Re.

Jacobucci, R., K. J. Grimm, and J. J. McArdle (2016). Regularized structural equation modeling. *Structural equation modeling: a multidisciplinary journal 23* (4), 555–566.

Jiang, L., Ding, Y., Sutherland, M. A., Hutchinson, M. K., Zhang, C., & Si, B. (2022). A novel sparse model-based algorithm to cluster categorical data for improved health screening and public health promotion. *IISE Transactions on Healthcare Systems Engineering*, *12*(2), 137-149.

Jin, X., Chen, P. Y., Hsu, C. Y., Yu, C. M., & Chen, T. (2021). Cafe: Catastrophic data leakage in vertical federated learning. Advances in Neural Information Processing Systems, 34, 994-1006.

Joseph, A., Wu, J., Yu, K., Jiang, L., Cady, N., & Si, B. (2021, August). Function-on-function regression for trajectory prediction of small-scale particles towards next-generation neuromorphic computing. In *2021 IEEE 17th International Conference on Automation Science and Engineering (CASE)* (pp. 1997-2002). IEEE.

Kang, Y., Liu, Y., & Chen, T. (2020). Fedmvt: semi-supervised vertical federated learning with multiview training (2020). arXiv preprint arXiv:2008.10838.

Kang, Y., Luo, J., He, Y., Zhang, X., Fan, L., & Yang, Q. (2022). A framework for evaluating privacy-utility trade-off in vertical federated learning. arXiv preprint arXiv:2209.03885.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

Kim, H. S. (2004). *Topics in ordinal logistic regression and its applications*. Texas A&M University.

Kirkland, E. B., M. Heincelman, K. G. Bishu, S. O. Schumann, A. Schreiner, R. N. Axon, P. D. Mauldin, and W. P. Moran (2018). Trends in healthcare expenditures among us adults with hypertension: national estimates, 2003–2014. *Journal of the American Heart Association 7* (11), e008731.

Konečný, J., McMahan, H. B., Ramage, D., & Richtárik, P. (2016). Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527*.

Lebret, R., S. Iovleff, F. Langrognet, C. Biernacki, G. Celeux, and G. Govaert (2015). Rmixmod: The r package of the model-based unsupervised, supervised, and semi-supervised classification mixmod library. *Journal of Statistical Software 67*, 1–29.

Lee, K., Kim, H., Lee, K., Suh, C., & Ramchandran, K. (2019, July). Synthesizing differentially private datasets using random mixing. In *2019 IEEE International Symposium on Information Theory (ISIT)* (pp. 542-546). IEEE.

Li, K., Wang, Z., Wang, Y., Luo, B., & Li, F. (2023, November). Poster: ethics of computer security and privacy research-trends and standards from a data perspective. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security* (pp. 3558-3560).

Li, Q., He, B., & Song, D. (2020). Practical one-shot federated learning for cross-silo setting. *arXiv preprint arXiv:2010.01017*.

Li, Q., Thapa, C., Ong, L., Zheng, Y., Ma, H., Camtepe, S. A., ... & Gao, Y. (2023). Vertical federated learning: Taxonomies, threats, and prospects. arXiv preprint arXiv:2302.01550.

Li, X., Hu, Y., Liu, W., Feng, H., Peng, L., Hong, Y., ... & Qin, Z. (2022). OpBoost: a vertical federated tree boosting framework based on order-preserving desensitization. arXiv preprint arXiv:2210.01318.

Li, X., Huang, K., Yang, W., Wang, S., & Zhang, Z. (2019). On the convergence of fedavg on non-iid data. *arXiv preprint arXiv:1907.02189*.

Lin, X., Wang, N., Welsh, A. H., and Carroll, R. J. (2004), "Equivalent Kernels of Smoothing Splines in Nonparametric Regression for Clustered/Longitudinal Data," Biometrika, 91, 177–193.

Liu, C., S. Wu, and X. Pan (2021). Clustering of cardio-metabolic risk factors and prediabetes among us adolescents. *Scientific Reports 11* (1), 1–7.

Liu, J., S. Ji, J. Ye, et al. (2009). Slep: Sparse learning with efficient projections. *Arizona State University 6* (491), 7.

Liu, T., Di, B., Wang, B., & Song, L. (2022). Loss-privacy tradeoff in federated edge learning. *IEEE Journal of Selected Topics in Signal Processing*, *16*(3), 546-558.

Liu, Y., Zhang, X., & Wang, L. (2020). Asymmetrical vertical federated learning. arXiv preprint arXiv:2004.07427.

Luo, R., & Qi, X. (2017). Function-on-function linear regression by signal compression. Journal of the American Statistical Association, 112(518), 690-705.

Luo, X., Wu, Y., Xiao, X., & Ooi, B. C. (2021, April). Feature inference attack on model predictions in vertical federated learning. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)* (pp. 181-192). IEEE.

Mairal J (2015). Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*.;25(2):829-55.

Marbac, M., M. Sedki, and T. Patin (2020). Variable selection for mixed data clustering: application in human population genomics. *Journal of Classification 37*, 124–142.

McLean, M. W., Hooker, G., Staicu, A. M., Scheipl, F., & Ruppert, D. (2014). Functional generalized additive models. Journal of Computational and Graphical Statistics, 23(1), 249-269.

McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017, April). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273-1282). PMLR.

McParland, D. and I. C. Gormley (2016). Model based clustering for mixed data: clustmd. *Advances in Data Analysis and Classification 10* (2), 155–169.

Mueller, S. D., Sutherland, M. A., Hutchinson, M. K., Si, B., Ding, Y., & Connolly, S. L. (2024). Student Health Services at Historically Black Colleges and Universities and Predominantly Black Institutions in the United States. Health Equity, 8(1), 226-234.

Mugunthan, V., Peraire-Bueno, A., & Kagal, L. (2020, October). Privacyfl: A simulator for privacy-preserving and secure federated learning. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management (pp. 3085-3092).

Müller, H. G., & Yao, F. (2008). Functional additive models. *Journal of the American Statistical Association, 103*(484), 1534-1544.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. (2018). CatBoost: unbiased boosting with categorical features. Advances in neural information processing systems, 31.

Ramsay, J. O., & Dalzell, C. (1991). Some tools for functional data analysis. Journal of the Royal Statistical Society Series B: Statistical Methodology, 53(3), 539-561.

Ramsay, J. O., and Silverman, B. W. (2005), Functional Data Analysis (2nd ed.), New York: Springer

Ratcliffe, S. J., Heller, G. Z., & Leader, L. R. (2002). Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression. Statistics in medicine, 21(8), 1115-1127.

Redline, S., Sotres-Alvarez, D., Loredo, J., Hall, M., Patel, S. R., Ramos, A., ... & Daviglus, M. L. (2014). Sleep-disordered breathing in Hispanic/Latino individuals of diverse backgrounds. The Hispanic community health study/study of Latinos. *American journal of respiratory and critical care medicine, 189*(3), 335-344.

Reiss, P. T., & Ogden, R. T. (2007). Functional principal component regression and functional partial least squares. Journal of the American Statistical Association, 102(479), 984-996.

Reiss, P. T., Huang, L., & Mennes, M. (2010). Fast function-on-scalar regression with penalized basis expansions. The international journal of biostatistics, 6(1)

Reynolds, D. A. (2009). Gaussian mixture models. *Encyclopedia of biometrics, 741*(659-663).

Ryu, S., H. Choi, H. Lee, and H. Kim (2019). Convolutional autoencoder based feature extraction and clustering for customer load analysis. *IEEE Transactions on Power Systems 35* (2), 1048–1060.

Sauer, T. and Y. Xu (1995). On multivariate hermite interpolation. *Advances in Computational Mathematics 4* (1), 207.

Schnabel, R. B., Koontz, J. E. and Weiss, B. E. (1985). A modular system of algorithms for unconstrained minimization. *ACM Transactions on Mathematical Software, 11, 419--440. 10.1145/6187.6192.*

Shah, N. S., Lloyd-Jones, D. M., O'Flaherty, M., Capewell, S., Kershaw, K., Carnethon, M., & Khan, S. S. (2019). Trends in cardiometabolic mortality in the United States, 1999-2017. *Jama, 322*(8), 780-782.

Sheffet, O. (2017, July). Differentially private ordinary least squares. In *International Conference on Machine Learning* (pp. 3105-3114). PMLR.

Shen, Z., Hassani, H., Kale, S., & Karbasi, A. (2022, May). Federated functional gradient boosting. In *International Conference on Artificial Intelligence and Statistics* (pp. 7814-7840). PMLR.

Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017, May). Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)* (pp. 3-18). IEEE.

Si, C., S. Xu, C. Wan, D. Chen, W. Cui, and J. Zhao (2021). Electric load clustering in smart grid: Methodologies, applications, and future trends. *Journal of Modern Power Systems and Clean Energy 9* (2), 237–252.

Silverman, B. W., & Ramsay, J. O. (2002). Applied functional data analysis: methods and case studies.

Somers, V. K., Dyken, M. E., Clary, M. P., & Abboud, F. M. (1995). Sympathetic neural mechanisms in obstructive sleep apnea. *The Journal of clinical investigation*, *96*(4), 1897-1904.

Spielberger, C. D., Gorsuch, R. L., Lushene, R. E., Vagg, P. R., & Jacobs, G. A. (2015). *State-trait anxiety inventory for adults: Manual, instrument and scoring guide*. Mind Garden.

Stone, M. H. (1948). The generalized weierstrass approximation theorem. *Mathematics Magazine* 21 (5), 237–254.

Sui, and N. L. Parker (2017). Timely decision analysis enabled by efficient social media modeling. *Decision Analysis* 14 (4), 250–260.

Sun, H., Wang, Z., Huang, Y., & Ye, J. (2022, January). Privacy-preserving vertical federated logistic regression without trusted third-party coordinator. In Proceedings of the 2022 6th International Conference on Machine Learning and Soft Computing (pp. 132-138).

Sun, J., Yang, X., Yao, Y., Xie, J., Wu, D., & Wang, C. (2022). Differentially private AUC computation in vertical federated learning. arXiv preprint arXiv:2205.12412.

Sun, J., Yang, X., Yao, Y., Zhang, A., Gao, W., Xie, J., & Wang, C. (2021). Vertical federated learning without revealing intersection membership. arXiv preprint arXiv:2106.05508.

Sun, Y., P. Babu, and D. P. Palomar (2016). Majorization-minimization algorithms in signal processing, communications, and machine learning. *IEEE Transactions on Signal Processing 65* (3), 794–816.

Sutherland, M. A., Hutchinson, M. K., Si, B., Ding, Y., Liebermann, E., Connolly, S. L., ... & Mueller, S. D. (2024). Health screenings in college health centers: Variations in practice. *Journal of American College Health*, 1-8.

Tan, P.-N., M. Steinbach, and V. Kumar (2016). *Introduction to data mining*. Pearson Education India.

Tian, Z., Zhang, R., Hou, X., Liu, J., & Ren, K. (2020). Federboost: Private federated learning for gbdt. arXiv preprint arXiv:2011.02796.

Vest, A. N., Da Poian, G., Li, Q., Liu, C., Nemati, S., Shah, A. J., & Clifford, G. D. (2018). An open source benchmarked toolbox for cardiovascular waveform and interval analysis. *Physiological measurement*, *39*(10), 105004.

Voigt, P., & Von dem Bussche, A. (2017). The eu general data protection regulation (gdpr). *A Practical Guide, 1st Ed., Cham: Springer International Publishing*, *10*(3152676), 10-5555.

Wang, C., Liang, J., Huang, M., Bai, B., Bai, K., & Li, H. (2020). Hybrid differentially private federated learning on vertically partitioned data. arXiv preprint arXiv:2009.02763.

Wang, C., Liang, J., Huang, M., Bai, B., Bai, K., & Li, H. (2020). Hybrid differentially private federated learning on vertically partitioned data. arXiv preprint arXiv:2009.02763.

Wang, H., & Leng, C. (2007). Unified LASSO estimation by least squares approximation. *Journal of the American Statistical Association*, *102*(479), 1039-1048.

Wang, J. and X. Wang (2019). *Structural equation modeling: Applications using Mplus*. John Wiley & Sons.

Wang, J. L., Chiou, J. M., & Müller, H. G. (2016). Functional data analysis. *Annual Review of Statistics and its application*, *3*, 257-295.

Wang, X., Zhu, H., & Alzheimer's Disease Neuroimaging Initiative. (2017). Generalized scalar-on-image regression models via total variation. *Journal of the American Statistical Association*, *112*(519), 1156-1168.

Wang, Z., Luo, B., & Li, F. (2023, May). SmartAppZoo: a Repository of SmartThings Apps for IoT Benchmarking. In *Proceedings of the 8th ACM/IEEE Conference on Internet of Things Design and Implementation* (pp. 448-449).

Warga, J. (1963). Minimizing certain convex functions. *Journal of the Society for Industrial and Applied Mathematics 11* (3), 588–593.

Wheeden, R. L. (2015). *Measure and integral: an introduction to real analysis* (Vol. 308). CRC press.

Wu, S., and Müller, H.-G. (2011), "Response-Adaptive Regression for Longitudinal Data," Biometrics, 67, 852–860.

Wu, Y., Cai, S., Xiao, X., Chen, G., & Ooi, B. C. (2020). Privacy preserving vertical federated learning for tree-based models. *arXiv preprint arXiv:2008.06170*.

Xu, D., Yuan, S., & Wu, X. (2021, December). Achieving differential privacy in vertically partitioned multiparty learning. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 5474-5483). IEEE.

Yang, S., Ren, B., Zhou, X., & Liu, L. (2019). Parallel distributed logistic regression for vertical federated learning without third-party coordinator. arXiv preprint arXiv:1911.09824

Yao, F., & Müller, H. G. (2010). Functional quadratic regression. Biometrika, 97(1), 49-64.

Yao, F., Müller, H.-G., and Wang, J.-L. (2005), "Functional Linear Regression Analysis for Longitudinal Data," The Annals of Statistics, 33, 2873– 2903.

Yuan, H., & Ma, T. (2020). Federated accelerated stochastic gradient descent. *Advances in Neural Information Processing Systems*, *33*, 5332-5344.

Zhang, G. Q., Cui, L., Mueller, R., Tao, S., Kim, M., Rueschman, M., ... & Redline, S. (2018). The National Sleep Research Resource: towards a sleep data commons. *Journal of the American Medical Informatics Association*, *25*(10), 1351-1358.

Zhang, J., & Jiang, Y. (2022). A data augmentation method for vertical federated learning. Wireless Communications and Mobile Computing, 2022, 1-16.

Zhang, Q., Gu, B., Dang, Z., Deng, C., & Huang, H. (2021, October). Desirable companion for vertical federated learning: New zeroth-order gradient based algorithm. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management (pp. 2598-2607).

Zhang, Q., Gu, B., Deng, C., & Huang, H. (2021, May). Secure bilevel asynchronous vertical federated learning with backward updating. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 12, pp. 10896-10904).

Zhang, X., Kang, Y., Chen, K., Fan, L., & Yang, Q. (2023). Trading Off Privacy, Utility, and Efficiency in Federated Learning. *ACM Transactions on Intelligent Systems and Technology*, *14*(6), 1-32.

Zhang, Z., Wang, X., Kong, L., & Zhu, H. (2022). High-dimensional spatial quantile function-on-scalar regression. *Journal of the American Statistical Association*, *117*(539), 1563-1578.

Zhao, D., Yao, M., Wang, W., He, H., & Jin, X. (2022, February). Ntp-vfl-a new scheme for non-3rd party vertical federated learning. In Proceedings of the 2022 14th International Conference on Machine Learning and Computing (pp. 134-139).

Zhou, X. and X. Cai (2022). Joint eqtl mapping and inference of gene regulatory network improves power of detecting both cis-and trans-eqtls. *Bioinformatics 38* (1), 149–156.

Zhu, X., Li, F., & Wang, H. (2021). Least-square approximation for a distributed system. *Journal of Computational and Graphical Statistics*, *30*(4), 1004-1018.