DIMENSION REDUCTION AND EFFICIENT RECOMMENDER SYSTEM FOR LARGE-SCALE COMPLEX DATA

BY

XUAN BI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2016

Urbana, Illinois

Doctoral Committee:

Professor Annie Qu, Chair
Assistant Professor Xiaohui Chen
Professor Jeffrey Douglas
Professor Douglas Simpson

Abstract

Large-scale complex data have drawn great attention in recent years, which play an important role in information technology and biomedical research. In this thesis, we address three challenging issues: sufficient dimension reduction for longitudinal data, nonignorable missing data with refreshment samples, and large-scale recommender systems.

In the first part of this thesis, we incorporate correlation structure in sufficient dimension reduction for longitudinal data. Existing sufficient dimension reduction approaches assuming independence may lead to substantial loss of efficiency. We apply the quadratic inference function to incorporate the correlation information and apply the transformation method to recover the central subspace. The proposed estimators are shown to be consistent and more efficient than the ones assuming independence. In addition, the estimated central subspace is also efficient when the correlation information is taken into account. We compare the proposed method with other dimension reduction approaches through simulation studies, and apply this new approach to an environmental health study.

In the second part of this thesis, we address nonignorable missing data which occur frequently in longitudinal studies and can cause biased estimations. Refreshment samples which recruit new subjects in subsequent waves from the original population could mitigate the bias. In this thesis, we introduce a mixed-effects estimating equation approach which enables one to incorporate refreshment samples and recover missing information. We show that the proposed method achieves consistency and asymptotic normality for fixed-effect estimation under shared-parameter models, and we extend it to a more general nonignorable-missing framework. Our finite sample simulation studies show the effectiveness and robustness of the proposed method under different missing mechanisms. In addition, we apply our method to election poll longitudinal survey data with refreshment samples from the 2007-2008 Associated Press–Yahoo! News.

In the third part of this thesis, we develop a novel recommender system which track users' preferences and recommend items of interest effectively. In this thesis, we propose a group-specific method to utilize dependency information from users and items which share similar characteristics under the singular value decomposition framework. The new approach is effective for the "cold-start" problem, where new users and new items' information is not available from the existing data collection. One advantage of the proposed model is that we are able to incorporate information from the missing mechanism and group-specific features through clustering based on variables associated with missing patterns. In addition, we propose a new algorithm that embeds a back-fitting algorithm into alternating least squares, which avoids large matrices operation and big memory storage, and therefore makes it feasible to achieve scalable computing. Our simulation studies and MovieLens data analysis both indicate that the proposed group-specific method improves prediction accuracy significantly compared to existing competitive recommender system approaches.

Dedicated to my family,

for their unconditional love, knowledge and patience.

Acknowledgements

First of all, I would like to express my sincerest gratitude to my advisor Professor Annie Qu for her continuous support, encouragement and inspiration. Her patience, passion and great knowledge stimulate, facilitate and significantly improve the efficacy and efficiency of my Ph.D. study and research. I cannot be more appreciative of her heuristic guidance which builds the foundation of my research interests, broadens my horizon in methodology learning and computing, and drastically improves my scientific writing skill. Beyond the scope of research, her constant encouragement and countless inspirations make the five years of my doctoral life strong and confident, and carry me through all the difficulties. I cannot imagine my doctoral research and life without her mentorship and support.

Special thanks to my committee members Professor Xiaohui Chen, Professor Jeffrey Douglas and Professor Douglas Simpson for their constant support in my career development and valuable suggestions to my thesis research. My doctoral life cannot be strong and smooth without their generous help. Furthermore, I owe my thanks to Professor Xiaotong Shen and Professor Junhui Wang for being wonderful collaborators. Their deep insight, great knowledge and critical comments substantially improve my understanding of machine learning and scientific computing. Besides, I would like to thank Professor Peng Wang and Professor Yunzhang Zhu for helpful discussions.

My thanks are given to the faculty and staff members in the Statistics Department for providing a friendly environment. I am also grateful to the fellow students who spent years with me, especially Xiwei Tang, Xiaolu Zhu, Peibei Shi, Xianyang Zhang, Xueying Zheng, Jianjun Hu, Jin Wang, Yunbo Ouyang, Srijan Sengupta, Yifeng He and Matthew Ulm. Thank you for being fantastic company and bringing me countless memorable experiences.

Finally, I want to thank my mother, my fiancée and my best friends for their constant love, encouragement, and most importantly, being in my life.

Table of Contents

1 Introduction **1**
1.1 Dimension Reduction 1
1.2 Missing Data 2
1.3 Recommender System 3

2 Sufficient Dimension Reduction for Longitudinal Data **5**
2.1 Introduction 5
2.2 Quadratic Inference Function 8
2.3 Sufficient Dimension Reduction for Longitudinal Data 10
2.3.1 Theoretical Properties 11
2.3.2 Efficiency 13
2.4 Implementation 15
2.4.1 Estimation of Structural Dimension 15
2.4.2 Algorithm 16
2.4.3 Implementation with Unbalanced Data 17
2.5 Simulation 17
2.5.1 Study 1: Binary Responses with One Set of Parameters 18
2.5.2 Study 2: Continuous Responses with Multiple Sets of Parameters . . 19
2.6 Asthma Data 21
2.7 Discussion 23
2.8 Proofs of Theoretical Results 24
2.8.1 Proof of Theorem 1 24
2.8.2 Proof of Corollary 1(transformation) 26
2.8.3 Proof of Lemma 1 26
2.8.4 Proof of Theorem 2 27
2.9 Tables and Figures 30

3 A Mixed-Effects Estimating Equation Approach to Nonignorable Missing Longitudinal Data with Refreshment Samples **34**
3.1 Introduction 34
3.2 Notation and Basic Assumptions 36
3.3 The General Method 38
3.3.1 Construction of Unbiased Estimating Equations 38
3.3.2 Estimation of Mixed Effects 42
3.3.3 Asymptotic Properties 44
3.3.4 Tuning Parameter Selection 47
3.4 Simulation Studies 48
3.4.1 Study 1: Count Responses under the SPM Assumption 49
3.4.2 Study 2: Binary Responses under the CMAR Assumption 50

3.5 Application . . . 51
3.6 Discussion . . . 53
3.7 Proofs of Theoretical Results . . . 54
3.7.1 Notation and Regularity Conditions . . . 54
3.7.2 Proofs of Lemma 2 and Theorem 3 . . . 56
3.8 Tables and Figures . . . 60

4 A Group-Specific Recommender System . . . 64
4.1 Introduction . . . 64
4.2 Background and Model Framework . . . 67
4.2.1 Background . . . 67
4.2.2 Model Framework . . . 68
4.3 The General Method . . . 70
4.3.1 Parameter Estimation . . . 70
4.3.2 Algorithm . . . 72
4.3.3 Implementation . . . 74
4.4 Theory . . . 75
4.5 Simulation Studies . . . 80
4.5.1 Comparison under the "Cold-Start" Problem . . . 80
4.5.2 Robustness against Cluster Misspecification . . . 82
4.6 MovieLens Data . . . 83
4.7 Discussion . . . 85
4.8 Proofs of Theoretical Results . . . 86
4.8.1 Proof of Lemma 3 . . . 86
4.8.2 Proof of Lemma 4 . . . 87
4.8.3 Proof of Theorem 4 . . . 88
4.8.4 Proof of Theorem 5 . . . 90
4.8.5 Proof of Corollary 3 . . . 92
4.8.6 Proof of Corollary 4 . . . 92
4.9 Tables and Figures . . . 94

References . . . 96

Chapter 1

Introduction

There has been a growing demand to develop effective and efficient methods to address large-scale complex data, which are prevalent in both information technology and biomedical sciences. Large-scale complex data contain high heterogeneous variation, multi-modality distribution and complicated data structure with high dimensional variables. In this thesis, we propose methods for complex data. Our contributions are mainly from three perspectives: dimension reduction, missing data and recommender systems.

1.1 Dimension Reduction

Dimension reduction, also known as dimensionality reduction or feature extraction in computer science, is a major technique to reduce the dimension of variables while maintain the crucial information. For example, in image recognition, one may be interested in developing the artificial intelligence to recognize a certain feature in a photo or targets in a video. This brings an issue that the resolution of the photos/videos might be high, and could cost large memory and computational resources. Dimension reduction is one of the possible solutions for this type of problems. The ideal dimension reduction methods are able to capture the pattern of the original data while reducing the number of variables. In addition to pattern recognition, dimension reduction is also an important data-preprocessing technique especially in machine learning. That is, dimension reduction can be conducted as an initial step of data analysis, then we perform desired methods on the reduced data for efficient computation, which can provide similar or even more accurate predictions. For instance, in genomic studies, researchers might have to search from millions of genes for their association with a

certain disease. Dimension reduction can accelerate the process through pre-selecting a few combinations of genes while maintain sound predicability.

Sufficient dimension reduction is one of the most important dimension reduction strategy, which utilizes information from the response variable. However, sufficient dimension reduction is mainly developed for independent data, and remains challenging if data are correlated. In Chapter 2, we propose a new sufficient dimension reduction method which incorporates correlation. Specifically, the proposed method applies the quadratic inference function (Qu et al., 2000) and utilizes a transformation method to approximate the central subspace. In contrast to existing methods where dependence information among repeated measures are ignored, the proposed method take intra-cluster correlation into account and hence improve estimation efficiency. In theory, we show that basis vectors found by the proposed method are in the central subspace, and the efficiency of parameter estimation leads to the efficiency of the central subspace estimation.

1.2 Missing Data

In addition to high-dimensional covariates that are frequently associated with large-scale complex data, another challenging issue is the existence of missing data. When the missing rate is not high, an easy way is to remove subjects with missing values and use complete cases only, which is called a complete-case analysis. However, when the missing rate is high, it is well-known that using complete cases may result in estimation bias because the missing data might not be missing randomly. For example, in a college class, it is usually students with low midterm scores that tend to drop; and in socioeconomic surveys, low-income participants tend not to report their salaries.

Rubin (1976) defines three missing mechanisms: missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR), where MNAR is also called nonignorable missingness in terms of estimation. The MCAR defines that the missingness

does not rely on any other variables, MAR states that missingness relies on observed values, and MNAR defines that missingness depends on additional unobserved values. There are extensive existing literature regarding statistical analysis under the MCAR and MAR. In contrast, the MNAR case is still challenging to handle, especially in the context of longitudinal studies where subjects are repeatedly measured.

In longitudinal studies, it is quite common that the drop-out rate gets higher as the study lasts longer. To offset the great data attrition, experiment designers usually consider refreshment samples. Here refreshment samples are subjects that are recruited at subsequent waves and are believed to be from the same population as the original samples. However, this strategy introduces a new type of missing pattern, that is, the refreshment samples miss the first few waves and hence their baseline information is not collected. This brings a new challenge as for most existing methods, such as inverse-weighting strategies or multiple imputation, baseline values are usually needed. In Chapter 3, we propose a mixed-effects estimating equation approach which incorporates unobserved latent variables through random effects to address nonignorable missing data. One advantage of the proposed method is that it does not require baseline values. We show that under an existing shared-parameter framework or a less restrictive new framework, the proposed method provides consistent and asymptotically normal estimation.

1.3 Recommender System

Recommender systems are another important technique in handling large-scale complex data. Nowadays, personalized products become increasingly popular, and recommender systems as one of its most important tools have drawn great attention. The recommender systems have a broad range of applications from shopping, dining, movie watching to personalized medicine. For example, one who purchased a product online may be automatically recommended relative accessories. In addition to item recommendations, many recommender-system methods

could have broader applications in other fields. For instance, the matrix completion method can be used for gene-expression prediction and image/video inpainting.

However, recommender system data also bring new and challenging issues in dealing with large-scale complex data. First, the data are of dynamic nature: new users and new items are recruited all the time and even after data collection. This leads to a problem that historical data might not be representative of future activities, which is called the "cold-start problem." Similar phenomenon has also been noticed in natural language processing where new words and internet slang cannot be found in traditional dictionaries. For example, in MovieLens data, 96% of the latest movie ratings are given by newly registered viewers or on recently released movies, whose information is not collected in the training data set. This leads to a critical issue that new users could only get "average" recommendations that are not personalized. Another critical issue is that, similar to the one described in Chapter 3, data may missing nonignorably. For example, popular items may attract more users, while users tend not to rate items they dislike. However, the major concern of missing data in recommender systems is that it affects prediction accuracy. For instance, a user gives five stars to the movies he/she watched does not necessarily indicate that he/she will give five stars to all other movies.

In Chapter 4, we propose a group-specific recommender system which targets at the two difficult issues mentioned above. The proposed method incorporates dependency among users and among items through group effects, which solves the "cold-start" problem effectively. In addition, we utilize missingness information through clustering and enhance prediction accuracy. In theory, we show that the proposed method has a fast convergence rate and has a smaller loss function than the methods without specifying group effects. In practice, we propose a scalable two-step algorithm which avoids large-matrix operations. We apply the proposed method to the MovieLens data which contain up to 10 million ratings. The proposed method improves prediction accuracy significantly comparing with existing competitive methods.

Chapter 2

Sufficient Dimension Reduction for Longitudinal Data

2.1 Introduction

Sufficient dimension reduction plays an important role in reducing the dimension of predictors and providing better modeling for response variables. The essential idea is to construct low-dimensional variables which can predict the response without loss of information. In contrast to the variable selection strategy, sufficient dimension reduction does not select or eliminate variables in a certain way. Instead, it extracts important information through optimally combining all predictors. Another advantage of sufficient dimension reduction is that it can be an effective way to visualize data (Li, 1991) through plotting the responses against the first several optimal combinations of covariates, which is especially important for handling high-dimensional data. Moreover, sufficient dimension reduction provides essential tools in analysis and curation for high-dimensional data, as it is able to reduce the original high-dimension of data to a moderate size without losing important information.

Existing methods of sufficient dimension reduction include, but are not limited to, ordinary least square (OLS; Li and Duan, 1989), slice inverse regression (SIR; Li, 1991), sliced average variance estimation (SAVE; Cook and Weisberg, 1991), principal Hessian direction (PHD; Li, 1992), discriminant analysis (Cook and Yin, 2001; Pardoe et al., 2007), minimum average variance estimation (MAVE; Xia et al., 2002), coutour regression (CR; Li et al., 2005), inverse regression estimation (IRE; Cook and Ni, 2005), directional regression (DR;

Li and Wang, 2007), sliced regression (SR; Wang and Xia, 2008), contour projection (CP; Luo et al., 2009), dimension reduction for non-elliptically distributed predictors (Li and Dong, 2009; Dong and Li, 2010), and dimension reduction based on canonical correlation (Fung et al., 2002; Zhou and He, 2008; Zhou, 2009). The study of sufficient dimension reduction for longitudinal data is still quite limited. With the prevalence of longitudinal study in biomedical, social, political, psychological, and environmental sciences, and with the increasing demand for handling high-dimensional data, it is of great importance to address sufficient dimension reduction problems under the longitudinal data framework.

For the longitudinal data setting, following Li et al. (2003)'s partial OLS, Li and Yin (2009) propose an analog partial OLS by conducting OLS at each time point and extracting a small subset of eigenvectors to achieve longitudinal data dimension reduction. However, their method does not incorporate intracluster correlation structure, and therefore leads to a significant loss of correlation information. In addition, their method is not able to exhaust the central subspace (Cook and Weisberg, 1994; Cook, 1996, 1998) if the cluster size is less than the structural dimension. Pfeiffer et al. (2012) propose a longitudinal first-moment-based sufficient dimension reduction method to solve these problems. They utilize a Kronecker-product space of clusters and predictors, and successfully accommodate the correlation structure of longitudinal covariates. However, their method is mainly applicable for handling longitudinal covariates, and not for longitudinal responses.

In this chapter, we apply the quadratic inference function (QIF; Qu et al., 2000) to longitudinal data sufficient dimension reduction, which can accommodate both longitudinal responses and correlation information. The QIF improves the generalized estimating equation (GEE; Liang and Zeger, 1986) without estimating nuisance parameters, and is shown to be efficient in regression parameter estimation for longitudinal data. In our approach, we first identify a group of transformation functions for the responses, then minimize the quadratic inference function which incorporates correlation information for transformed responses to obtain regression parameter estimators, and then apply eigen-decomposition to

extract information from a set of regression parameter estimators for the transformed responses.

The proposed method allows one to gain extra efficiency in parameter estimation for both continuous and discrete responses through incorporating correlation structure, while not requiring that the true correlation structure be known. Most importantly, we obtain parameter estimation from the entire cluster instead of performing regression separately at each time point, as in Li and Yin (2009). This leads to several advantages, such that the proposed method can still be efficient even for a small sample size, since we utilize information from repeated measurements within the same subject, and therefore the sample points used in our estimation are larger than the ones in Li and Yin (2009). In addition, the proposed method is computationally more efficient than existing methods, as the operation cost is lower for the same reason. In our approach the recovery of the central subspace does not depend on the cluster size, is in contrast to existing approaches which require the cluster size to be greater than the structural dimension.

In theory, we show that estimation through minimizing the QIF for the transformed data is still in the central subspace, and asymptotic efficiency can be improved by incorporating correlation structures. Another finding is that the efficiency of parameter estimation leads to the efficiency of the central subspace estimation. This is confirmed by our simulation studies, which show that the proposed method can improve accuracy and efficiency for sufficient dimension reduction in finite samples.

The remainder of this chapter is organized as follows. Section 2.2 provides background for the quadratic inference function. Section 2.3 introduces the proposed method for longitudinal dimension reduction using the QIF, and provides its theoretical foundation and properties. Section 2.4 illustrates how to recover the structural dimension and provides the implementation of the proposed method. Section 2.5 compares the proposed approach with existing work through simulation studies for normal and binary responses. Section 2.6 applies the proposed method to a longitudinal asthma study. Section 2.7 concludes our findings

and provides a brief discussion. Technical derivations are provided in Section 2.8.

2.2 Quadratic Inference Function

For longitudinal data, suppose y_{it} is the response of subject i at time t, and $\mathbf{x}_{it}$ is a p-dimensional covariate, where $i = 1, \ldots, n$ and $t = 1, \ldots, T_i$. To simplify notation, we set $T_i = T$ for all i; the unbalanced data case will be discussed in more details in Section 2.4. Let $\mu(\cdot)$ be an inverse link function satisfying $\mathrm{E}(y_{it}|\mathbf{x}_{it}) = \mu(\boldsymbol{\beta}'\mathbf{x}_{it})$, where $\boldsymbol{\beta}$ is a p-dimensional parameter. Define $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$, $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, and $\boldsymbol{\mu}_i = \mathrm{E}(\mathbf{y}_i|\mathbf{x}_i)$ for each i. If independence structure is assumed among subjects, the quasi-likelihood equation (Wedderburn, 1974; McCullagh, 1983) for solving $\boldsymbol{\beta}$ is

$$\sum_{i=1}^{n} \dot{\boldsymbol{\mu}}_i' \mathbf{V}_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

where $\dot{\boldsymbol{\mu}}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$ is a $T \times p$ matrix, and $\mathbf{V}_i$ is the covariance matrix of $\mathbf{y}_i$. In practice $\mathbf{V}_i$ is usually unknown. One common approach is to substitute the empirical estimator $\hat{\mathbf{V}}_i$ for $\mathbf{V}_i$. However, this involves many nuisance parameter estimations and thus $\hat{\mathbf{V}}_i$ can be unstable when T is large. Liang and Zeger (1986) introduced the working correlation matrix which reduces the number of correlation parameters significantly. They assume $\widetilde{\mathbf{V}}_i = \mathbf{A}_i^{\frac{1}{2}} \mathbf{R}(\boldsymbol{\alpha}) \mathbf{A}_i^{\frac{1}{2}}$, where $\mathbf{A}_i$ is a diagnal matrix of marginal variance of $\mathbf{y}_i$, $\mathbf{R}(\boldsymbol{\alpha})$ is the working correlation matrix, and $\boldsymbol{\alpha}$ contains a small number of correlation parameters.

The QIF approach (Qu et al., 2000) further avoids the estimation of $\boldsymbol{\alpha}$ by formulating $\mathbf{R}^{-1}$ as a linear combination of $\mathbf{M}_0, \mathbf{M}_1, \ldots, \mathbf{M}_{m-1}$, where $\mathbf{M}_0$ is a T-dimensional identity matrix. For example, if $\mathbf{R}(\boldsymbol{\alpha})$ is exchangeable, then $m = 2$ and $\mathbf{M}_1$ has 0 on the diagonal and 1 elsewhere. The idea of the QIF is to ensure the additional moment conditions $\sum_{i=1}^{n} \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-\frac{1}{2}} \mathbf{M}_r \mathbf{A}_i^{-\frac{1}{2}} (\mathbf{y}_i - \boldsymbol{\mu}_i)$ are as close to 0 as possible for $r = 1, \ldots, m-1$. Since the number of equations is greater than the number of parameters, the QIF utilizes the generalized method of moments (GMM; Hansen, 1982), where the specified moment conditions of

$\mathbf{b} \in \mathbb{R}^p$ for estimating $\boldsymbol{\beta}$ are

$$\mathbf{g}_i(\mathbf{b}) = \begin{pmatrix} (\dot{\boldsymbol{\mu}}_i)' \mathbf{A}_i^{-\frac{1}{2}} \mathbf{M}_0 \mathbf{A}_i^{-\frac{1}{2}} (\mathbf{y}_i - \boldsymbol{\mu}_i) \\ \vdots \\ (\dot{\boldsymbol{\mu}}_i)' \mathbf{A}_i^{-\frac{1}{2}} \mathbf{M}_{m-1} \mathbf{A}_i^{-\frac{1}{2}} (\mathbf{y}_i - \boldsymbol{\mu}_i) \end{pmatrix}, i = 1 \ldots, n. \tag{2.1}$$

The quadratic inference function is defined as

$$\hat{Q}(\mathbf{b}) = n \bar{\mathbf{g}}'(\mathbf{b}) \hat{\mathbf{W}}^{-1}(\mathbf{b}) \bar{\mathbf{g}}(\mathbf{b}), \tag{2.2}$$

where $\bar{\mathbf{g}}(\mathbf{b}) = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i(\mathbf{b})$, and $\hat{\mathbf{W}}(\mathbf{b}) = \frac{1}{n} \sum_{i=1}^n \mathbf{g}_i(\mathbf{b}) \mathbf{g}_i'(\mathbf{b})$. The corresponding QIF estimator is obtained as $\hat{\mathbf{b}} = \text{argmin}_b \hat{Q}(\mathbf{b})$. Qu et al. (2000) showed that $\hat{\mathbf{b}}$ is a $\sqrt{n}$-consistent estimator and is efficient if a linear combination of basis matrices $\mathbf{M}_0, \mathbf{M}_1, \ldots, \mathbf{M}_{m-1}$ contains the true correlation structure.

A critical issue regarding the QIF is the selection of the number m of basis matrices, which has been addressed by model selection for correlation structure in Zhou and Qu (2012). The basic idea is to approximate the inverse of the empirical correlation matrix by a group of basis matrices, which contain only 0 and 1 as entries. Then a Euclidean-norm measuring the difference between two estimating functions, one based on the empirical correlation information and the other on the model-based approximation, is minimized. Through a groupwise penalty on the basis matrices, an appropriate number m of basis matrices can be selected such that sufficient correlation information is captured. In theory, the selected correlation structure is consistent if the candidate basis matrices are from a sufficiently rich class to represent the true structure.

In general, the moment condition $\mathbf{g}_i(\mathbf{b}) = \mathbf{g}(\mathbf{b}'\mathbf{x}_i, \mathbf{y}_i)$ is required to satisfy $\text{E}(\mathbf{g}_i) = \mathbf{0}, i = 1, \ldots, n$, to identify the true parameter $\boldsymbol{\beta}$. The population version of the QIF is $Q(\mathbf{b}) = (\text{E}\mathbf{g})' \mathbf{W}^{-1} (\text{E}\mathbf{g})$, where $\mathbf{W} = \text{Var}(\mathbf{g})$. Therefore, $Q(\mathbf{b}) \geq 0$, and the equality holds if and only if $\mathbf{b} = \boldsymbol{\beta}$.

2.3 Sufficient Dimension Reduction for Longitudinal Data

In this section, we propose the QIF approach for sufficient dimension reduction in the longitudinal data setting.

Let $\mathbf{X}$ be a $p \times T$-dimensional covariate matrix and $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_T)'$ be a T-dimensional response. Both $\mathbf{X}$ and $\mathbf{Y}$ can be random. The main purpose of sufficient dimension reduction (SDR; Li, 1991; Cook, 1998) is to seek a minimal dimension-reduction subspace with a $p \times d$ basis matrix $\mathbf{B}$, where $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$, $d \leq p$, such that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{B}'\mathbf{X}$. Here $\perp\!\!\!\perp$ indicates independence. Under some regularity conditions (Cook, 1998), the minimal subspace exists and is unique, that is, the central subspace of the regression of $\mathbf{Y}$ on $\mathbf{X}$, denoted by $\mathcal{S}_{Y|\mathbf{X}}$. Suppose rank($\mathbf{B}$)= d, then $d = \dim(\mathcal{S}_{Y|\mathbf{X}})$ is also called the structural dimension of regression. The central subspace is the smallest subspace of $\mathbb{R}^p$ that captures all of the regression information of $\mathbf{Y}$ given $\mathbf{X}$, and therefore reduces the dimension of the predictors from $\mathbf{X}$ to $\mathbf{B}'\mathbf{X}$.

We propose to identify the central subspace by recovering its basis through minimizing the QIF. If the dimension of the central subspace is $d = 1$, then the problem of identifying the central subspace is equivalent to a parameter estimation problem, and thus the QIF estimator alone can capture the central subspace completely, since the fact that $\mathcal{S}_{Y|\mathbf{X}} = \text{Span}(\boldsymbol{\beta}_1)$. When $d \geq 2$, $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ may not be identifiable (Li, 1991).

Alternatively, to recover the central subspace, we propose to minimize the QIF for transformed responses. This approach does not have the identifiability constraint for $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$. Suppose we have a group of transformations h_j's for responses, $h_j : \mathbb{R} \to \mathbb{R}$, $j = 1, \ldots, s$. Let $\mathbf{h}_j = (h_j, \ldots, h_j)'$ be a T-dimensional transformation function vector on the response vector $\mathbf{Y}$. Take

$$Q_j(\mathbf{b}) = \{\mathrm{E}\mathbf{g}(\mathbf{b}'\mathbf{X}, \mathbf{h}_j(\mathbf{Y}))\}' \, \mathbf{W}^{-1} \, \{\mathrm{E}\mathbf{g}(\mathbf{b}'\mathbf{X}, \mathbf{h}_j(\mathbf{Y}))\}, \tag{2.3}$$

with minimizer $\boldsymbol{\gamma}_j = \text{argmin}_b\, Q_j(\mathbf{b})$, $j = 1, \ldots, s$. In Section 2.3.1, we show that $\boldsymbol{\gamma}_j$ is in the central subspace under certain conditions, and $\text{Span}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_s)$ approximates $\mathcal{S}_{Y|\mathbf{X}}$. Since d is typically unknown, we need a sufficiently large s to ensure that $s \geq d$. The selection of s is discussed in Section 2.4.2 in detail.

There are several strategies to choose the transformation function h_j. One common practice is to use the power transformation (Cook and Li, 2002; Yin and Cook, 2002; Zhu and Zhu, 2009; Yin and Li, 2011), $h_j(Y_t) = Y_t^j, j = 1, \ldots, s$. Other transformation methods include the slice indicator function proposed by Li (1991), which defines $h_j(Y_t) = 1$ if Y_t is in the jth slice and 0 otherwise, the covariance inverse regression method (Cook and Ni, 2006) defining $h_j(Y_t) = Y_t$ if Y_t is in the jth slice and 0 otherwise, and the normalized B-spline basis functions for Y_t (Fung et al., 2002). (Cook, 1998, p.114) shows that $\mathcal{S}_{h(Y)|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$ holds for any transformation function h, and $\mathcal{S}_{h(Y)|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}}$ holds if h is a one-to-one function.

The purpose of applying the transformation method is that, although minimizing the QIF from the original responses can only recover one basis vector for the central subspace, the transformation method can provide a group of transformed responses, and therefore recover a group of basis vectors that allow one to explore the central subspace to its largest extent.

2.3.1 Theoretical Properties

We assume the well-known *linearity condition* (Li and Duan, 1989) that states that $\text{E}(\mathbf{X}|\mathbf{B}'\mathbf{X})$ is linear in $\boldsymbol{\beta}_1'\mathbf{X}, \ldots, \boldsymbol{\beta}_d'\mathbf{X}$. This entails that the distribution of $\mathbf{X}$ be elliptically symmetric. Li and Dong (2009); Dong and Li (2010); Ma and Zhu (2012, 2013a,c) provide alternative strategies on how to relax this condition. On the other hand, the *constant conditional variance* assumption (Cook and Weisberg, 1991), where $\text{Var}(\mathbf{X}|\mathbf{B}'\mathbf{X})$ is a constant matrix, is not required.

Suppose $\mathbf{Y} \perp\!\!\!\perp \mathbf{X}|\mathbf{B}'\mathbf{X}$, and let $L(\mathbf{b}'\mathbf{X}, \mathbf{Y}) = \mathbf{g}'(\mathbf{b}'\mathbf{X}, \mathbf{Y})\mathbf{W}^{-1}\mathbf{g}(\mathbf{b}'\mathbf{X}, \mathbf{Y})$ be a loss function.

Then the following theorem shows that the QIF minimizer,

$$\boldsymbol{\gamma} = \underset{b}{\operatorname{argmin}}\, Q(\mathbf{b}), \boldsymbol{\gamma} \in \mathbb{R}^p, \tag{2.4}$$

is in the central subspace. In addition, the sample estimator

$$\hat{\boldsymbol{\gamma}} = \underset{b}{\operatorname{argmin}}\, \hat{Q}(\mathbf{b}) \tag{2.5}$$

is a strongly consistent estimator of $\boldsymbol{\gamma}$, where $\hat{Q}(\mathbf{b})$ is defined in (2.2).

Theorem 1. *Assume $L(\cdot,\cdot)$ is convex in its first argument, the linearity condition holds, and Var$(\mathbf{X})$ is positive definite. If $\boldsymbol{\gamma}$ in (2.4) exists and is unique, then $\boldsymbol{\gamma} \in \mathcal{S}_{Y|\mathbf{X}}$, and $\hat{\boldsymbol{\gamma}}$ in (2.5) converges to $\boldsymbol{\gamma}$ almost surely.*

The convexity condition of $L(\cdot,\cdot)$ in its first argument is easily satisfied in our approach since $\ddot{L} = \dot{\mathbf{g}}'\mathbf{W}^{-1}\dot{\mathbf{g}} + o_p(1)$ is a non-negative definite matrix, asymptotically. The strict convexity of L is a sufficient condition to ensure the uniqueness of $\boldsymbol{\gamma}$ in Theorem 1 (Li and Duan, 1989).

When $d = 1$ and $\mathbf{g}_i$ is defined in equation (2.1), Theorem 1 implies that the minimizer $\boldsymbol{\gamma}$ in (2.4) is the true parameter, and the sample minimizer $\hat{\boldsymbol{\gamma}}$ in (2.5) converges to the true parameter $\boldsymbol{\gamma}$ almost surely.

Theorem 1 does not require $\mathbf{g}$ to satisfy $\mathrm{E}(\mathbf{g}) = 0$, the strong consistency property is robust to the misspecification of the link functions. This is even more desirable when the conditional distribution of $\mathbf{Y}|\mathbf{X}$ is difficult to find. As for the efficiency argument in Section 2.3.2, however, a correctly specified link function is required to achieve an efficiency gain through incorporating correlation information.

Theorem 1 lays the foundation for formulating basis vectors for the central subspace. Suppose $\hat{Q}_j(\mathbf{b})$ is the sample version of $Q_j(\mathbf{b})$, and $\hat{\boldsymbol{\gamma}}_j = \operatorname{argmin}_b \hat{Q}_j(\mathbf{b})$ is the sample estimator of $\boldsymbol{\gamma}_j$.

Corollary 1. *Assume $L(\cdot,\cdot)$ is convex in its first argument, the linearity condition holds, and Var$(\mathbf{X})$ is positive definite. If $\boldsymbol{\gamma}_j$ exists and is unique, then $\boldsymbol{\gamma}_j \in \mathcal{S}_{Y|\mathbf{X}}$, and $\hat{\boldsymbol{\gamma}}_j$ converges to $\boldsymbol{\gamma}_j$ almost surely, $j = 1, \ldots, s$.*

Corollary 1 implies that each $\boldsymbol{\gamma}_j$ is a linear combination of $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$, and that Span$(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_s) \subseteq \mathcal{S}_{Y|\mathbf{X}}$. This provides an effective way to build a central subspace basis. If $\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_d$ are the eigenvectors corresponding to the largest d eigenvalues of $(\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_s)\ (\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_s)'$, then the basis for the central subspace can be taken as $\hat{\mathbf{B}} = (\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_d)$.

Recently, Yin and Li (2011) formulated the conditions to achieve exhaustiveness of the central subspace, which can accommodate power transformations as a special case. In their Theorem 2.1 and Example 2.1, they proved that given a sufficiently large s, the subspace spanned by $\mathcal{S}_{E(Y^j|\mathbf{X})}$ $(j = 1, \ldots, s)$ approaches the central subspace under mild conditions, where $\mathcal{S}_{E(Y|\mathbf{X})}$ denotes the central mean subspace of $\mathbf{Y}$ on $\mathbf{X}$ (Cook and Li, 2002). For each transformation $\mathbf{Y}^j$, the quadratic inference function (QIF) can recover one basis vector from $\mathcal{S}_{E(Y^j|\mathbf{X})}$. Therefore, a sufficient condition to achieve exhaustiveness, as mentioned in Yin and Cook (2002), is to assume that there exists a group of powers $k_1, \ldots, k_d$, such that $\dim(\mathcal{S}_{E(Y^{k_j}|\mathbf{X})}) = 1$ for $j = 1, \ldots, d$. Under such an assumption, the QIF approach with the transformed response $\mathbf{Y}^j, j = 1, 2, \ldots, k_d$, can exhaust the central subspace. When other types of tranformations are applied, a similar assumption should be satisfied accordingly. Exhaustiveness can then be achieved if the new tranformations follow the conditions of Theorem 2.1 in Yin and Li (2011). Alternatively, Ma and Zhu (2012, 2013b) propose a semiparametric estimating equation approach that avoids the aforementioned condition, but still achieves exhaustiveness by identifying and estimating the central subspace basis $(\boldsymbol{\beta}_1, \cdots, \boldsymbol{\beta}_d)$ using one estimating equation.

2.3.2 Efficiency

As long as the responses from the same subject are not independent, incorporating correlation information always leads to efficiency gain. In addition, the efficiency gain of parameter

estimation from the data with each transformation of the response variable provides an overall efficiency gain of the central subspace estimation.

For illustration, suppose there are two sets of moment conditions: $\mathbf{G}_l = \sum_{i=1}^{n} (\dot{\boldsymbol{\mu}}_i)' \mathbf{A}_i^{-\frac{1}{2}} \mathbf{M}_l \mathbf{A}_i^{-\frac{1}{2}} (\mathbf{y}_i - \boldsymbol{\mu}_i)$, $l = 1, 2$, where $\mathbf{M}_1$ and $\mathbf{M}_2$ are symmetric matrices, $\dot{\boldsymbol{\mu}}_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}$, and $\boldsymbol{\beta}$ is the true parameter. Let $\mathbf{G} = (\mathbf{G}_1', \mathbf{G}_2')'$, $\dot{\mathbf{G}} = \partial \mathbf{G} / \partial \boldsymbol{\beta}$, $\dot{\mathbf{G}}_1 = \partial \mathbf{G}_1 / \partial \boldsymbol{\beta}$, $\mathbf{C} = \mathrm{Var}(\mathbf{G})$, and $\mathbf{C}_{11} = \mathrm{Var}(\mathbf{G}_1)$. The empirical information matrices corresponding to $\mathbf{G}$ and $\mathbf{G}_1$ are $\dot{\mathbf{G}}' \mathbf{C}^{-1} \dot{\mathbf{G}}$ and $\dot{\mathbf{G}}_1' \mathbf{C}_{11}^{-1} \dot{\mathbf{G}}_1$, respectively. We show that incorporating a correlation structure leads to an increase of the empirical information matrix in the sense of the Loewner ordering (Beckenback and Bellman, 1965), which is equivalent to an improvement in parameter estimation efficiency.

Lemma 1. *If* $\mathbf{R}^{-1} = a_1 \mathbf{M}_1 + a_2 \mathbf{M_2}$ *is the true correlation matrix and* $E(\mathbf{G}) = \mathbf{0}$*, then* $\dot{\mathbf{G}}' \mathbf{C}^{-1} \dot{\mathbf{G}} \geq \dot{\mathbf{G}}_1' \mathbf{C}_{11}^{-1} \dot{\mathbf{G}}_1$*, in terms of the Loewner ordering for matrices. Equality holds if* $a_2 = 0$*.*

Lemma 1 indicates that we gain efficiency by incorporating additional correlation information; if $\mathbf{M}_1$ is an identity matrix, then the proposed dimension reduction method incorporating correlation structure is more efficient than those assuming independence. In simulation studies provided in Section 2.5, we illustrate that the performance of sufficient dimension reduction based on the QIF assuming independence is similar to other approaches such as the OLS or SIR, while the QIF incorporating correlation information can significantly improve the efficiency for sufficient dimension reduction.

The condition $\mathrm{E}(\mathbf{G}) = \mathbf{0}$ assumes that each moment condition has zero expectation. That is, use the conditional mean $\mathrm{E}(\mathbf{h}_j(\mathbf{Y})|\mathbf{X})$ as a link function for the transformed response $\mathbf{h}_j(\mathbf{Y})$ (Yin and Cook, 2002; Ma and Zhu, 2012, 2013a). In practice, however, the conditional mean $\mathrm{E}(\mathbf{h}_j(\mathbf{Y})|\mathbf{X})$ is usually unknown. As pointed out by Ma and Zhu (2013b), unless one uses a nonparametric approach, it might be difficult to find the correct link function. For the proposed method, this is even more challenging than in Ma and Zhu (2013b) case, since for each transformation the QIF can only generate one basis vector for the central subspace.

Unless we assume $\mathrm{E}(\mathbf{h}_j(\mathbf{Y})|\mathbf{X})$ is known and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d$ are identifiable, the link function of the QIF is typically misspecified.

A possible way to have a correctly specified link function might be to apply a nonparametric procedure, but this could complicate our method significantly. To avoid this, Ma and Zhu (2013b) also suggested imposing additional assumptions; for example, the linearity condition on $\mathbf{Y}$, or applying a common link function. In one simulation study, we apply a common (identity) link function. There are two practical justifications for this application. First, the response $\mathbf{Y}$ is continuous and could range from negative infinity to infinity. Second, using the identity link is a linear approximation of the true link function. Thus, even though the link function may not be exact, it will still achieve good efficiency in practice. In fact, we find that the proposed method with the identity link function indeed has an efficiency gain through incorporating correlation information. Other common link functions can be applied when the response is not continuous. Refer to Ma and Zhu (2013b, 2014) for more detail.

The consistency of the estimator for a central subspace vector is guaranteed by Corollary 1, and the efficiency gained by incorporating correlation information can be followed by Lemma 1.

Theorem 2. *Suppose $\hat{\boldsymbol{\gamma}}_j$ is an efficient estimator of $\boldsymbol{\gamma}_j$ corresponding to the j-th transformation function, where $\boldsymbol{\gamma}_j \in \mathcal{S}_{Y|\mathbf{X}}$, $j = 1, \ldots, s$. Then $(\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_s)$ is an efficient estimator of $(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_s)$, provided the information matrix corresponding to the true parameter $(\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_d')'$ is bounded.*

2.4 Implementation

2.4.1 Estimation of Structural Dimension

For selection of structural dimension d, several approaches have been proposed. Li (1991) provided an asymptotic chi-squared test, assuming that the covariates are normally distributed, and Cook and Yin (2001) built the foundation of the permutation test for the

structural dimension. In addition, Li and Wang (2007) introduced a sequential test, and Ye and Weiss (2003) proposed a bootstrap procedure. Luo et al. (2009) further suggested a quick and effective selection procedure called the *maximal eigenvalue ratio criterion*, which chooses

$$\hat{d} = \underset{1 \le q \le d_{max}}{\operatorname{argmax}} \hat{\lambda}_q / \hat{\lambda}_{q+1}. \tag{2.6}$$

In practice, $d_{max} = 5$ usually suffices. The intuition behind (2.6) can be explained. Suppose $\hat{\mathbf{B}}$ is a consistent estimator of $\mathbf{B}$, and therefore that each $\hat{\lambda}_q$ converges to λ_q consistently. Since $\dim(\mathbf{B}) = d$, λ_q's are nonzero if $q \le d$. As $\lim_{n\to\infty} \hat{\lambda}_d / \hat{\lambda}_{d+1} = +\infty$, choosing $\hat{d}$ to satisfy (2.6) is a sensible approach.

2.4.2 Algorithm

We provide an algorithm for sufficient dimension reduction for longitudinal data.

(i) Choose a transformation function $\mathbf{h}_j$, and transform the response $\mathbf{y}_i$ into $\mathbf{h}_j(\mathbf{y}_i)$, for $j = 1, \ldots, s$ and $i = 1, \ldots, n$.

(ii) For the transformed responses $\mathbf{h}_j(\mathbf{y}_1), \ldots, \mathbf{h}_j(\mathbf{y}_n)$, obtain $\hat{\boldsymbol{\gamma}}_j$ by minimizing $\hat{Q}_j(\mathbf{b})$.

(iii) Conduct a spectral decomposition for $(\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_s)(\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_s)'$, and obtain the structural dimension d based on (2.6).

(iv) Select eigenvectors $\hat{\boldsymbol{\eta}}_1, \ldots, \hat{\boldsymbol{\eta}}_d$ corresponding to the first to d-th largest eigenvalues, and formulate the basis of the central subspace $\mathcal{S}_{Y|\mathbf{X}}$.

The selection of s in (i) is similar to, but less critical than, the selection of the number of slices in SIR, still an open question (Wang and Xia, 2008). If $\lim_{s\to\infty} \operatorname{Span}(\gamma_1, \ldots, \gamma_s) = \mathcal{S}_{Y|X}$, the transformed QIF with a sufficiently large s could approximate the central subspace (compared to SIR where the number of slices may be restricted if the support of $\mathbf{Y}$ is finite).

On the other hand, a finite and fixed s may not be enough to exhaust the central subspace even if we are given $\lim_{s\to\infty} \text{Span}(\gamma_1, \ldots, \gamma_s) = \mathcal{S}_{Y|X}$ (Yin and Li, 2011), as difficult as the SIR in choosing the total number of slices.

In practice, the selection of s may not be very critical, similar to the selection of the total number of slices for many inverse regression methods, e.g., SIR, SAVE and SR. Our numerical studies indicate that the proposed method is rather robust against s: the simulation results did not change much once $s \geq d$. Currently if d can be detected by other methods, as in our data analysis for the asthma study in Section 2.6, then s can be selected accordingly.

2.4.3 Implementation with Unbalanced Data

In practice, unbalanced data are quite common. If the measurements from unbalanced data are regarded as cluster data without considering the order of lag time, then each $\boldsymbol{\mu}_i$ is a T_i-dimensional vector, and $\mathbf{M}_r$ is a $T_i \times T_i$ matrix for $i = 1, \ldots, n$ and $r = 0, 1, \ldots, m-1$.

If the lag time between measurements is considered important, we can define $T = \max(T_1, \ldots, T_n)$, and impose a $T \times T_i$ dimensional transformation matrix $\mathbf{U}_i$ for the i-th subject. Let $\mathbf{y}_i^* = \mathbf{U}_i \mathbf{y}_i$, $\boldsymbol{\mu}_i^* = \mathbf{U}_i \boldsymbol{\mu}_i$, $\dot{\boldsymbol{\mu}}_i^* = \mathbf{U}_i \dot{\boldsymbol{\mu}}_i$ and $\mathbf{A}_i^* = \mathbf{U}_i \mathbf{A}_i \mathbf{U}_i'$. Thus, we transform the unbalanced data to artificial balanced data where each component of $\mathbf{U}_i$ is an indicator of whether the data is observed or missing. Then we formulate moment conditions as in (2.1) for the newly created balanced data. The QIF estimator from minimizing (2.2) still has the right properties if the data are missing completely at random. See Zhou and Qu (2012) for more details.

2.5 Simulation

We report on simulation studies to illustrate the performance of the proposed method and existing approaches for longitudinal data sufficient dimension reduction. They show that incorporating a suitable correlation structure can improve the accuracy and efficiency of

estimation for both the parameters and the central subspace.

2.5.1 Study 1: Binary Responses with One Set of Parameters

We generated the covariate $\mathbf{x}_i$ as standard normal for subject $i = 1, \ldots, n$. For each $\mathbf{x}_i$, we assumed T repeated measurements $\mathbf{x}_i = (\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT})$, and that each $\mathbf{x}_{it}$ is a p-dimensional vector, $t = 1, \ldots, T$. We assumed independence among different subjects and different covariates, but an exchangeable correlation structure among T time points for each covariate, with $\rho_x = 0.2$.

In Study 1, we let $p = 50, T = 20$, with sample size n as 51, 100, or 200. The true parameter $\boldsymbol{\beta}$ was a p-dimensional vector with 1 in its first 10 components and 0 otherwise. We generated $\mathbf{v}_i$ based on the linear model $\mathbf{v}_i = 0.4\boldsymbol{\beta}'\mathbf{x}_i + \boldsymbol{\varepsilon}_i$ and $\boldsymbol{\varepsilon}_i \overset{\text{iid}}{\sim} N(0, \boldsymbol{\Sigma}_\varepsilon)$, $i = 1, \ldots, n$, where $\boldsymbol{\Sigma}_\varepsilon$ is a T-dimensional exchangeable correlation matrix with $\rho_\varepsilon = 0.2, 0.5$, or 0.8. We then generated $\mathbf{y}_i$ by utilizing an indicator function $y_{it} = \mathbf{1}_{A_{it}}$, where event $A_{it} = \{e^{v_{it}}/(1+e^{v_{it}}) > 0.5\}$ and v_{it} is the t-th component of $\mathbf{v}_i$, $t = 1, \ldots, T$. Since $\mathbf{y}_i|\mathbf{x}_i = \mathbf{y}_i|\boldsymbol{\beta}'\mathbf{x}_i$, the structural dimension is $d = 1$, and the central subspace is $\mathcal{S}_{Y|\mathbf{X}} = \text{Span}(\boldsymbol{\beta})$. It is straightforward that $\text{E}(y_{it}) = 0.5$ and $\text{E}(y_{it}|\mathbf{x}_{it}) = 1 - \Phi(0.4\boldsymbol{\beta}'\mathbf{x}_{it})$, where $\Phi(\cdot)$ is the standard normal distribution function. The correlation structure of $\mathbf{y}_i$ is close to, but not exactly, that of the exchangeable structure, as the correlation is mainly contributed by the error term $\boldsymbol{\varepsilon}_i$.

We measured the distance between central subspace basis matrix $\mathbf{B}$ and the estimated central subspace $\hat{\mathbf{B}}$ by $||\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}' - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'||_F$, where $||\cdot||_F$ is the Frobenius norm. We compared our method with the partial ordinary least square (partial OLS) by Li and Yin (2009), where the linear regression is conducted at each time point to recover parameter vectors for the central subspace, and d eigenvectors corresponding to the largest d eigenvalues are extracted through an eigen decomposition. We also compared with the "partial SIR," similar to Li and Yin's partial OLS except that at each time point linear regression is replaced by sliced inverse regression. Our simulation study shows that the partial SIR provides results similar to those of the partial OLS approach.

We generated simulation samples $N = 1000$. Table 2.1 provides the average distance, and the standard deviation (inside the parenthesis). The proposed dimension reduction method based on the QIF is significantly better than those from the partial OLS and partial SIR in the sense of accuracy and efficiency. For one, when $n = 51$ and $p = 50$, the partial OLS and the partial SIR provide estimators that are nearly orthogonal to the true parameter vector, while the proposed QIF is still robust, with much smaller distances between the true and estimated vectors. The linear regression at each time point has a sample size of 51, with a 50-dimensional parameter, so estimation is unstable. The proposed method utilizes data from all time points simultaneously, so the number of sample points is $51 \times 20 = 1020$, and this leads to a more precise estimation.

When the sample size is $n = 100$ or 200, the QIF assuming exchangeable correlation is still the best, though all methods converge to the true parameter space as the sample size increases. In an unreported simulation study, we found that the QIF converges faster than the other methods as the cluster size increases. The existing methods regress at each time point and have computing time dependent on the cluster size, while the QIF incorporates data from all time points simultaneously. As the cluster sizes increase, computational times of the proposed method and the existing approaches grow further apart.

Information on correlation has a strong influence on the estimations, and incorporating a correct correlation structure achieves higher accuracy and efficiency. The partial OLS and SIR approaches do not take correlation into account, and their results are relatively close to, but still worse than those estimated by the QIF dimension reduction approach assuming independence of data.

2.5.2 Study 2: Continuous Responses with Multiple Sets of Parameters

We investigated the performance of the new method when the dimension of central subspace d is greater than 1. Here we had two p-dimensional coefficient vectors $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, such

that $\mathbf{Y} \perp\!\!\!\perp \mathbf{X} \mid (\boldsymbol{\beta}_1'\mathbf{X}, \boldsymbol{\beta}_2'\mathbf{X})$, so $d = 2$. We set $p = 8$ or 15. When $p = 8$, we let $\boldsymbol{\beta}_1 = (1,1,1,1,1,1,1,1)'/\sqrt{8}$, and $\boldsymbol{\beta}_2 = (1,-1,1,-1,1,-1,1,-1)'/\sqrt{8}$; when $p = 15$, we set the rest of the 7 components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ to be 0. The continuous response variable y_{it} was generated using

$$\begin{aligned}
\text{Model I: } y_{it} &= \exp(\boldsymbol{\beta}_1'\mathbf{x}_{it}) + 2(1+\boldsymbol{\beta}_2'\mathbf{x}_{it})^2 + 0.5(\boldsymbol{\beta}_1'\mathbf{x}_{it})\tau_{it};\\
\text{Model II: } y_{it} &= (0.45\boldsymbol{\beta}_1'\mathbf{x}_{it})/\{0.5 + (1.5+\boldsymbol{\beta}_2'\mathbf{x}_{it})^2\} + 0.5\varepsilon_{it};\\
\text{Model III: } y_{it} &= \sin(\boldsymbol{\beta}_1'\mathbf{x}_{it}/4) + \exp(2\boldsymbol{\beta}_2'\mathbf{x}_{it}/3) + 0.5\varepsilon_{it}.
\end{aligned}$$

In Model I, we took ρ_x, the correlation of $\mathbf{x}_i$, as 0.2, 0.5, or 0.8, and took the error $\boldsymbol{\tau}_i = (\tau_{i1},\ldots,\tau_{iT})' \overset{iid}{\sim} N(\mathbf{0}, \mathbf{I}_T)$, $i = 1,\ldots,n$. Because of heteroscedasticity, the responses in Model I are highly correlated, even though τ_{it}'s are independent. In Models II and III, we generated each $\mathbf{x}_i$ the same way as in Study 1, except that the correlation parameter was replaced by $\rho_x = 0.5$. The error $\boldsymbol{\varepsilon}_i$ was generated as in Study 1, with exchangeable correlation $\rho_\varepsilon = 0.2, 0.5$, or 0.8.

For the partial OLS and the partial SIR approaches, we applied the same procedure as in Study 1. For the proposed method, we used a power transformation to recover basis vectors for the central subspace: let $h_j(y_{it}) = y_{it}^j, j = 1,\ldots,s$. Here we set $s = 2$. In an unreported simulation study, we found that increasing s does not make much difference for central subspace estimation. Alternative transformation methods provided in Section 2.3.1 can also be applied here.

Tables 2.2, 2.3, and 2.4 list the distance under the configurations $(n,p) = (100,8)$ and $(n,p) = (300,15)$ for Models I, II, and III. Evidently, the proposed QIF methods are better than the partial OLS and the partial SIR, and the QIF assuming exchangeable correlation is the best. When the correlation of responses increases, either through the correlations of covariate $\mathbf{x}_i$ in Model I or through the error $\boldsymbol{\varepsilon}_i$ in Models II and III, the proposed method with exchangeable correlation structure is most accurate, while methods assuming independence

structure perform poorly. Meanwhile, the QIF assuming AR-1 structure provides very similar estimation as the one assuming exchangeable correlation, because, although we generate both $\mathbf{x}_i$ and $\boldsymbol{\varepsilon}_i$ using the exchangeable correlation structure the combined correlation structure of $\mathbf{y}_i$ is neither exchangeable nor AR-1, due to the nonlinear relationship of the response and covariates.

In general, the proposed QIF dimension reduction method is still applicable if $T < d$, but the partial OLS is not feasible.

2.6 Asthma Data

We applied the proposed method to an asthma study conducted in Windsor, Ontario, Canada in 1992. This study intends to measure the impact of air pollution on asthmatic patients. The data were originally provided by Professor Paul Corey of the University of Toronto and the Ontario Ministry of Health, and were investigated for model selection in the GEE (Fu, 2003) and partial OLS dimension reduction by Li and Yin (2009). This data set consists of 39 asthmatic patients who were observed on 21 consecutive days. Patients' asthmatic status on difficulty of breathing is recorded as 1 (presence) or 0 (absence) daily, where difficulty of breathing is determined by patients' daily forced expiratory volume. The predictors are daily mean humidity (HUMD), daily mean temperature (TEMP), and seven air pollutants: nitrogen oxide (NO), nitrogen dioxide (NO2), mixture of NO and NO2 (NOX), carbon monoxide (CO), ozone level (OZ), total reduced sulphur (TRS) and coefficient of haze (COH). The data thus contains $n = 39$ patients with cluster size $T = 21$, and dimension of covariates $p = 9$.

We applied the partial OLS by Li and Yin (2009). The scree plot of the eigenvalues of $(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T)(\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_T)'$ is shown in Figure 2.1, where each $\boldsymbol{\eta}_t$ is the OLS estimator at each time point $t = 1, \ldots, 21$. To select the structural dimension d, we applied the maximal eigenvalue ratio criterion (Luo et al., 2009) discussed in Section 2.4.1, and $\hat{d} = 1$ was selected.

The choice of d can also be observed directly by examing the scree plot in Figure 2.1, where a sharp drop occurs right after the largest eigenvalue. The corresponding eigenvector associated with the maximum eigenvalue is $\hat{\boldsymbol{\beta}}_{OLS} = (-0.0012, 0.4303, -0.8608, 0.2503, -0.0596, -0.0015, -0.0804, 0.0264, -0.0220)'$, and therefore Span$(\hat{\boldsymbol{\beta}}_{OLS})$ is an estimated central subspace. We also observe that the mean sample correlation of the intracluster correlation matrix for the responses is 0.6992, and pair correlations among the 21 measurements are quite similar, suggesting a non-negligible exchangeable correlation structure.

For the proposed method, we took $\hat{\boldsymbol{\beta}}_{OLS}$ as an initial value and $\hat{d} = 1$, then calculated the basis for the central subspace using the proposed QIF dimension reduction approach. We employed the QIF assuming the exchangeable, AR-1, and independence correlation structures. The estimated results were:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{Indep} &= (-0.0665, -0.0058, -0.0046, -0.0331, -0.0254, -0.0243, 0.0725, -0.0071, 0.0513)';\\
\hat{\boldsymbol{\beta}}_{Ar1} &= (0.0247, -0.0020, -0.0478, -0.0160, 0.0187, -0.0149, 0.0552, -0.0635, 0.0019)';\\
\hat{\boldsymbol{\beta}}_{Exch} &= (-0.0954, -0.0047, 0.0341, -0.0199, -0.0111, 0.0031, 0.0765, -0.0041, -0.0273)'.
\end{aligned}$$

These differ from the partial OLS estimate. For instance, the angle between $\hat{\boldsymbol{\beta}}_{OLS}$ and $\hat{\boldsymbol{\beta}}_{Exch}$ is 71.82°, indicating a weak correlation between these two estimators.

We conducted logistic regressions of y_{it} given $\hat{\boldsymbol{\beta}}'\mathbf{x}_{it}$ to investigate which method provides the best prediction, where $\hat{\boldsymbol{\beta}}$ is the estimator based on the partial OLS, or the QIF assuming exchangeable, AR-1, or independent correlation, respectively. Table 2.5 provides the estimators, standard errors, and p-values for the slope of each regression. The QIF dimension reduction with independent and exchangeable correlation structures fits the data better than the other approaches.

The QIF assuming the exchangeable structure is the most accurate, with the smallest MSE compared to other three methods. At each level of the continuous explanatory variable $\mathbf{x}_{it}$, there is only one observation of the response, so the log-odds of receiving $y_{it} = 1$ at each

level of $\mathbf{x}_{it}$ is usually infinity or negative infinity. To pool information of adjacent $\hat{\boldsymbol{\beta}}'\mathbf{x}_{it}$, we divided the range of $\hat{\boldsymbol{\beta}}'\mathbf{x}_{it}$ into K intervals of equal length based on the distribution of $\hat{\boldsymbol{\beta}}'\mathbf{x}_{it}$, where $K = 25, 26, 30$, or 25 was applied to each method, respectively. We then calculated $\hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_k$, the average of $\hat{\boldsymbol{\beta}}'\mathbf{x}_{it}$ for the k-th interval, and $\text{logit}(\bar{y}_k)$, the log-odds of $\bar{y}_k$, where $\bar{y}_k$ is the average of y_{it} corresponding to $\hat{\boldsymbol{\beta}}'\mathbf{x}_{it}$ in the k-th interval, $k = 1, \ldots, K$. Table 2.6 lists the correlation between $(\text{logit}(\bar{y}_1), \ldots, \text{logit}(\bar{y}_K))$ and $(\hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_1, \ldots, \hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_K)$ and the MSE of $(\alpha_0 + \alpha_1\hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_1, \ldots, \alpha_0 + \alpha_1\hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_K)$, where α_0 and α_1 are the logistic regression coefficients. The QIF method assuming exchangeable correlation structure achieves the highest magnitude of regression correlation with a smaller MSE, compared with the QIF assuming independence structure.

Scatterplots of $(\text{logit}(\bar{y}_1), \ldots, \text{logit}(\bar{y}_K))$ against $(\hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_1, \ldots, \hat{\boldsymbol{\beta}}'\bar{\mathbf{x}}_K)$ for each method are provided in Figure 2.2. The slope of the partial OLS method (upper-left panel) are very sensitive to the two influential points on the top left, leading to a potentially unstable estimator; while the QIF assuming the exchangeable correlation structure provides a better fitted regression line overall.

2.7 Discussion

We have addressed the sufficient dimension reduction problem for longitudinal data, with the goal of showing that incorporating intracluster correlation information can achieve more efficiency than assuming independence in both parameter and central subspace estimations. We used the quadratic inference function to incorporate correlation structures and a transformation method to formulate basis vectors for the central subspace. These basis vectors were shown to be consistent and more efficient than estimators assuming independence. The proposed method achieves an overall efficiency for central subspace estimation through combining each efficient estimator of an individual basis vector. Our simulation studies show that the proposed method is quite effective for both binary and continuous data for small

and large sample sizes, compared with existing approaches that do not take intracluster correlation into consideration.

Simulation show that even if the correlation structure is misspecified, the efficiency of the proposed estimator is higher than the one assuming independence; our method is quite robust under a small sample size, due to utilizing the entire cluster information for dimension reduction. The proposed method is able to recover the central subspace even when the cluster size is small, it can handle unbalanced data, and is computationally efficient when the cluster size is large.

Further investigation is needed regarding a tuning procedure to select the number of transformations s by minimizing the distance between γ_s and $\text{Span}(\gamma_1, \ldots, \gamma_{s-1})$, along with a penalty function. Another possible research direction is sufficient dimension reduction for binary data (or data with finite support) when the structural dimension is greater than 1. Binary sufficient dimension reduction is quite challenging, since the binary response is invariant for most types of transformation methods. Pooling similar covariate information together so that the log-odds of $Y = 1$ have sufficient variability for estimation is a possible approach.

2.8 Proofs of Theoretical Results

2.8.1 Proof of Theorem 1

Suppose $\mathcal{S}_\zeta$ is an arbitrary dimension-reduction subspace. Define $\mathbf{B}_\zeta$ as the basis matrix of $\mathcal{S}_\zeta$, and $\mathbf{P}_{B_\zeta} = \mathbf{B}_\zeta(\mathbf{B}_\zeta'\mathbf{B}_\zeta)^{-1}\mathbf{B}_\zeta'$ is the projection matrix. Note that the population version of QIF is $Q(\mathbf{b}) = (\text{E}\mathbf{g})'\mathbf{W}^{-1}(\text{E}\mathbf{g})$, where $\mathbf{g}$ is a (mp)-dimensional estimating function.

We first show that $Q(\mathbf{b}) \geq Q(\mathbf{P}_{B_\zeta}\mathbf{b})$ for any p-dimensional parameter $\mathbf{b} \in \mathbb{R}^p$. This implies that the minimizer of $Q(\mathbf{b})$, denoted as $\boldsymbol{\gamma}$, must lie in $\mathcal{S}_\zeta$, and thus lie in the central subspace $\mathcal{S}_{Y|\mathbf{X}} = \cap_\zeta \mathcal{S}_\zeta$. This is similar to the argument of Theorem 2.1 in Li and Duan (1989), and Proposition 8.1 in (Cook, 1998, p.144).

Since $Q(\mathbf{b}) = \{\mathrm{E}(\mathbf{W}^{-\frac{1}{2}}\mathbf{g})\}'\{\mathrm{E}(\mathbf{W}^{-\frac{1}{2}}\mathbf{g})\}$, we define $\mathbf{g}^* = \mathbf{W}^{-\frac{1}{2}}\mathbf{g}$. Then,

$$\mathrm{Var}(\mathbf{g}^*) = \mathbf{I}_{mp} = \mathrm{E}(\mathbf{g}^*\mathbf{g}^{*\prime}) - (\mathrm{E}\mathbf{g}^*)(\mathrm{E}\mathbf{g}^*)'.$$

Therefore,

$$\begin{aligned} Q(\mathbf{b}) &= (\mathrm{E}\mathbf{g}^*)'(\mathrm{E}\mathbf{g}^*) = \mathrm{tr}\{(\mathrm{E}\mathbf{g}^*)'(\mathrm{E}\mathbf{g}^*)\} \\ &= \mathrm{tr}\{(\mathrm{E}\mathbf{g}^*)(\mathrm{E}\mathbf{g}^*)'\} = \mathrm{tr}\{\mathrm{E}(\mathbf{g}^*\mathbf{g}^{*\prime}) - \mathbf{I}_{mp}\} \\ &= \mathrm{E}\{\mathrm{tr}(\mathbf{g}^*\mathbf{g}^{*\prime}) - mp\} = \mathrm{E}(\mathbf{g}^{*\prime}\mathbf{g}^*) - mp \\ &= \mathrm{E}[\mathrm{E}\{\mathbf{g}^{*\prime}(\mathbf{b}'\mathbf{X}, \mathbf{Y})\mathbf{g}^*(\mathbf{b}'\mathbf{X}, \mathbf{Y})\}|\mathbf{Y}, \mathbf{B}_\zeta'\mathbf{X}] - mp. \end{aligned}$$

Note that $L(\mathbf{b}'\mathbf{X}, \mathbf{Y}) = \mathbf{g}^{*\prime}(\mathbf{b}'\mathbf{X}, \mathbf{Y})\mathbf{g}^*(\mathbf{b}'\mathbf{X}, \mathbf{Y})$ is convex with respect to its first argument. Therefore,

$$\begin{aligned} Q(\mathbf{b}) &= \mathrm{E}[\mathrm{E}\{L(\mathbf{b}'\mathbf{X}, \mathbf{Y})\}|\mathbf{Y}, \mathbf{B}_\zeta'\mathbf{X}] - mp \\ &\geq \mathrm{E}\{L(\mathrm{E}(\mathbf{b}'\mathbf{X}|\mathbf{Y}, \mathbf{B}_\zeta'\mathbf{X}), \mathbf{Y})\} - mp \\ &= \mathrm{E}\{\mathbf{g}^{*\prime}(\mathrm{E}(\mathbf{b}'\mathbf{X}|\mathbf{Y}, \mathbf{B}_\zeta'\mathbf{X}), \mathbf{Y})\mathbf{g}^*(\mathrm{E}(\mathbf{b}'\mathbf{X}|\mathbf{Y}, \mathbf{B}_\zeta'\mathbf{X}), \mathbf{Y})\} - mp. \end{aligned}$$

Because $\mathbf{B}_\zeta$ is the basis matrix of $\mathcal{S}_\zeta$, we have $\mathbf{X}|(\mathbf{Y}, \mathbf{B}_\zeta'\mathbf{X}) \stackrel{\mathrm{d}}{=} \mathbf{X}|\mathbf{B}_\zeta'\mathbf{X}$; and the linearity condition implies that $\mathrm{E}(\mathbf{X}|\mathbf{B}_\zeta'\mathbf{X}) = \mathbf{P}_{B_\zeta}\mathbf{X}$. Hence,

$$\begin{aligned} Q(\mathbf{b}) &\geq \mathrm{E}\{\mathbf{g}^{*\prime}(\mathrm{E}(\mathbf{b}'\mathbf{X}|\mathbf{B}_\zeta'\mathbf{X}), \mathbf{Y})\mathbf{g}^*(\mathrm{E}(\mathbf{b}'\mathbf{X}|\mathbf{B}_\zeta'\mathbf{X}), \mathbf{Y})\} - mp \\ &= \mathrm{E}\{\mathbf{g}^{*\prime}((\mathbf{P}_{B_\zeta}\mathbf{b})'\mathbf{X}, \mathbf{Y})\mathbf{g}^*((\mathbf{P}_{B_\zeta}\mathbf{b})'\mathbf{X}, \mathbf{Y})\} - mp \\ &= Q(\mathbf{P}_{B_\zeta}\mathbf{b}). \end{aligned}$$

Next, we show that $\hat{\boldsymbol{\gamma}}$ is a strongly consistent estimator of $\boldsymbol{\gamma}$. This follows Theorem 5.1 of Li and Duan (1989), which states that the minimizer of the sample loss function converges to the minimizer of the risk function almost surely, if the objective loss function is convex

with respect to its first arument. ■

2.8.2 Proof of Corollary 1(transformation)

Following (Cook, 1998, p.115), if $\mathbf{h}$ is a function of $\mathbf{Y}$, then $\mathcal{S}_{\mathbf{h}(Y)|\mathbf{X}} \subseteq \mathcal{S}_{Y|\mathbf{X}}$; and if $\mathbf{h}$ is one-to-one, then $\mathcal{S}_{\mathbf{h}(Y)|\mathbf{X}} = \mathcal{S}_{Y|\mathbf{X}}$. Then Corollary 1 follows immediately from Theorem 1. ■

2.8.3 Proof of Lemma 1

The first part of this proof shows that we gain more information and achieve higher efficiency by incorporating additional correlation information formulated by the moment condition $\mathbf{G}_2$.

We first orthogonalize $\mathbf{G}_2$ from $\mathbf{G}_1$ as

$$\mathbf{G}_2^* = \mathbf{G}_2 - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{G}_1,$$

where $\mathbf{C}_{21} = \mathrm{Cov}(\mathbf{G}_2, \mathbf{G}_1)$ and $\mathbf{C}_{11} = \mathrm{Var}(\mathbf{G}_1)$. After orthogonalization, $\mathrm{Cov}(\mathbf{G}_2^*, \mathbf{G}_1) = \mathbf{0}$. Let $\mathbf{G}^* = (\mathbf{G}_1', \mathbf{G}_2^{*\prime})'$, $\mathbf{C}^* = \mathrm{Var}(\mathbf{G}^*)$, and $\mathbf{C}_{22}^* = \mathrm{Var}(\mathbf{G}_2^*)$, then $\mathbf{C}_{22}^* = \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}$, where $\mathbf{C}_{22} = \mathrm{Var}(\mathbf{G}_2)$ and $\mathbf{C}_{12} = \mathrm{Cov}(\mathbf{G}_1, \mathbf{G}_2)$. Since $\dot{\mathbf{G}}_2^* = \dot{\mathbf{G}}_2 - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\dot{\mathbf{G}}_1$, the information matrix of the estimator by minimizing $\mathbf{G}'\mathbf{C}^{-1}\mathbf{G}$ is proportional to

$$\begin{aligned}\dot{\mathbf{G}}'\mathbf{C}^{-1}\dot{\mathbf{G}} &= (\dot{\mathbf{G}}^*)'(\mathbf{C}^*)^{-1}(\dot{\mathbf{G}}^*) \\ &= \dot{\mathbf{G}}_1'\mathbf{C}_{11}^{-1}\dot{\mathbf{G}}_1 + (\dot{\mathbf{G}}_2^*)'(\mathbf{C}_{22}^*)^{-1}(\dot{\mathbf{G}}_2^*).\end{aligned}$$

Note that $\mathbf{C}_{22}^*$ is non-negative definite, so in the sense of Loewner ordering for matrices,

$$\dot{\mathbf{G}}'\mathbf{C}^{-1}\dot{\mathbf{G}} \geq \dot{\mathbf{G}}_1'\mathbf{C}_{11}^{-1}\dot{\mathbf{G}}_1.$$

The following argument shows that if $\mathbf{G}_1$ contains all information about the parameter, adding additional moment conditions will not improve efficiency. That is, if $\mathbf{M}_1$ is proportional to $\mathbf{R}^{-1}$, then $\dot{\mathbf{G}}'\mathbf{C}^{-1}\dot{\mathbf{G}} = \dot{\mathbf{G}}_1'\mathbf{C}_{11}^{-1}\dot{\mathbf{G}}_1$.

The detailed proof is provided as follows. Recall that $\mathbf{G}_l = \sum_{i=1}^{n} \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-\frac{1}{2}} \mathbf{M}_l \mathbf{A}_i^{-\frac{1}{2}} (\mathbf{y}_i - \boldsymbol{\mu}_i)$, $l = 1, 2$. Assume $\mathbf{R}^{-1} = a_1 \mathbf{M}_1$, then

$$\dot{\mathbf{G}}_1 = -\frac{1}{a_1} \sum_{i=1}^{n} \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-\frac{1}{2}} \mathbf{R}^{-1} \mathbf{A}_i^{-\frac{1}{2}} \dot{\boldsymbol{\mu}}_i + o_p(1), \text{and } \dot{\mathbf{G}}_2 = -\sum_{i=1}^{n} \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-\frac{1}{2}} \mathbf{M}_2 \mathbf{A}_i^{-\frac{1}{2}} \dot{\boldsymbol{\mu}}_i + o_p(1).$$

In addition,

$$\mathbf{C}_{11} = \frac{1}{a_1^2} \sum_{i=1}^{n} \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-\frac{1}{2}} \mathbf{R}^{-1} \mathbf{A}_i^{-\frac{1}{2}} \dot{\boldsymbol{\mu}}_i, \text{and } \mathbf{C}_{21} = \frac{1}{a_1} \sum_{i=1}^{n} \dot{\boldsymbol{\mu}}_i' \mathbf{A}_i^{-\frac{1}{2}} \mathbf{M}_2 \mathbf{A}_i^{-\frac{1}{2}} \dot{\boldsymbol{\mu}}_i.$$

Thus, $\mathbf{C}_{11} = -\frac{1}{a_1} \dot{\mathbf{G}}_1 + o_p(1)$ and $\mathbf{C}_{21} = -\frac{1}{a_1} \dot{\mathbf{G}}_2 + o_p(1)$, and this results in $\dot{\mathbf{G}}_2^* = o_p(1)$. Therefore,

$$\dot{\mathbf{G}}' \mathbf{C}^{-1} \dot{\mathbf{G}} = \dot{\mathbf{G}}_1' \mathbf{C}_{11}^{-1} \dot{\mathbf{G}}_1 + o_p(1).$$

■

2.8.4 Proof of Theorem 2

Theorem 18.11 of (Kosorok, 2008, p.341) shows that the marginal efficiency of two estimators leads to their joint efficiency on product spaces, given the condition that the two estimated parameters are differentiable with respect to their tangent space. The main goal of this proof is to verify this condition under the sufficient dimension reduction framework for longitudinal data, and thus the estimators by the proposed method have joint efficiency, leading to the efficiency of the central subspace. The definition of tangent space and differentiability with respect to the tangent space are provided in the following two paragraphs respectively.

Without loss of generosity, we assume $\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d$ are linearly independent. Then $\boldsymbol{\gamma}_j \in \mathcal{S}_{Y|\mathbf{X}}$ implies $\text{Span}(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d) = \mathcal{S}_{Y|\mathbf{X}}$, $j = 1, \ldots, d$. Set $s = d$ and $\mathbf{B}^* = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d)$. Suppose $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_d)$ is a $p \times d$ constant matrix in $\mathbb{R}^{p \times d}$, and let $\text{vec}(\mathbf{u}) = (\mathbf{u}_1', \ldots, \mathbf{u}_d')'$ denote the vectorization of $\mathbf{u}$. Suppose the transformed response $h_j(y_{it})$ is imposed for score

function $\mathbf{S}_j$ such that the solution $\hat{\boldsymbol{\gamma}}_j$ of $\mathbf{S}_j = \mathbf{0}$ is an efficient estimator of $\boldsymbol{\gamma}_j, j = 1, \ldots, d$. Let $\mathbf{S} = (\mathbf{S}_1', \ldots, \mathbf{S}_d')'$. Define the tangent function to be $H = \mathbf{S}'\text{vec}(\mathbf{u})$. Then a tangent set is $\mathcal{T} = \{H = \mathbf{S}'\text{vec}(\mathbf{u}) : \mathbf{u} \in \mathbb{R}^{p\times d}\}$. Since this tangent set is closed under linear combination, it is also a tangent space.

For an arbitrarily small $\delta \geq 0$ and fixed $\text{vec}(\mathbf{B}) = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_d')'$, suppose the model has a true parameter $\text{vec}(\mathbf{B}) + \delta\text{vec}(\mathbf{u})$. A parameter $\boldsymbol{\gamma}$ is differentiable with respect to the tangent space $\mathcal{T}$, if $d\boldsymbol{\gamma}/d\delta|_{\delta=0} = \dot{\boldsymbol{\psi}}(H)$, where $\dot{\boldsymbol{\psi}}(\cdot)$ is a bounded linear operator.

Since $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)$ and $\text{span}(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{X}}$, there exists a $d \times d$ matrix $\mathbf{D}$, such that $\mathbf{B}^* = \mathbf{BD}$. Since the $pd \times pd$ information matrix of $\text{vec}(\mathbf{B}) = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_d')'$ is bounded, the information matrix $\dot{\mathbf{S}}'\widetilde{\mathbf{C}}^{-1}\dot{\mathbf{S}}$ of $(\boldsymbol{\gamma}_1', \ldots, \boldsymbol{\gamma}_d')'$ is also bounded, where $\widetilde{\mathbf{C}} = \text{Var}(\mathbf{S})$.

For any direction $\mathbf{u} \in \mathbb{R}^{p\times d}$ and an arbitrarily small $\delta \geq 0$,

$$\begin{aligned}
\frac{d\boldsymbol{\gamma}_j}{d\delta} &= \frac{d\boldsymbol{\gamma}_j}{d\text{vec}(\mathbf{B}+\delta\mathbf{u})}\text{vec}(\mathbf{u}) \\
&= \frac{d\boldsymbol{\gamma}_j}{d\text{vec}(\mathbf{BD}+\delta\mathbf{uD})}\frac{d\text{vec}(\mathbf{BD}+\delta\mathbf{uD})}{d\text{vec}(\mathbf{B}+\delta\mathbf{u})}\text{vec}(\mathbf{u}) \\
&= \frac{d\boldsymbol{\gamma}_j}{d\text{vec}(\mathbf{BD}+\delta\mathbf{uD})}(\mathbf{D}'\otimes\mathbf{I}_p)\text{vec}(\mathbf{u}) \\
&= \frac{d\boldsymbol{\gamma}_j}{d\text{vec}(\mathbf{B}^*+\delta\mathbf{u}^*)}(\mathbf{D}'\otimes\mathbf{I}_p)\text{vec}(\mathbf{u}),
\end{aligned}$$

where $\mathbf{u}^* = \mathbf{uD}$, $j = 1, \ldots, d$ and $\otimes$ denotes the Kronecker product.

Similar to Lemma 1, we can show that $\widetilde{\mathbf{C}} = -\dot{\mathbf{S}} + o_p(1)$. And $\text{E}(\mathbf{S}) = \mathbf{0}$ implies $-\dot{\mathbf{S}}'\widetilde{\mathbf{C}}^{-1}\dot{\mathbf{S}} = \text{E}(\mathbf{SS}') + o_p(1)$. Therefore,

$$\begin{aligned}
\frac{d\boldsymbol{\gamma}_j}{d\delta} &= \frac{d\boldsymbol{\gamma}_j}{d\text{vec}(\mathbf{B}^*+\delta\mathbf{u}^*)}(\mathbf{D}'\otimes\mathbf{I}_p)(-\dot{\mathbf{S}}'\widetilde{\mathbf{C}}^{-1}\dot{\mathbf{S}})^{-1}\{\text{E}(\mathbf{SS}')\}\text{vec}(\mathbf{u}) + o_p(1) \\
&= \frac{d\boldsymbol{\gamma}_j}{d\text{vec}(\mathbf{B}^*+\delta\mathbf{u}^*)}(\mathbf{D}'\otimes\mathbf{I}_p)(-\dot{\mathbf{S}}'\widetilde{\mathbf{C}}^{-1}\dot{\mathbf{S}})^{-1}\{\text{E}(\mathbf{S}H)\} + o_p(1).
\end{aligned}$$

Define $\dot{\boldsymbol{\psi}}_j(H) = d\boldsymbol{\gamma}_j/d\delta|_{\delta=0}$ for any tangent function $H \in \mathcal{T}$. Since $\boldsymbol{\gamma}_j$ is the j-th column of $\mathbf{B}^*$, $d\boldsymbol{\gamma}_j/d\text{vec}(\mathbf{B}^*+\delta\mathbf{u}^*)|_{\delta=0}$ is bounded. Because $\mathbf{D}$ is a bounded linear transformation and

$(-\dot{\mathbf{S}}'\widetilde{\mathbf{C}}^{-1}\dot{\mathbf{S}})$ is also bounded, it follows that $\dot{\boldsymbol{\psi}}_j(\cdot)$ is a bounded linear operator. Therefore, $\boldsymbol{\gamma}_j$ is differentiable with respect to the tangent space $\mathcal{T}$, $j = 1, \ldots, d$.

Following Theorem 18.11 of (Kosorok, 2008, p.341), we conclude that $(\hat{\boldsymbol{\gamma}}_1, \ldots, \hat{\boldsymbol{\gamma}}_d)$ is an asymptotic efficient estimator of $(\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_d)$. ■

2.9 Tables and Figures

Table 2.1: Mean and standard deviation of $||\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}' - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'||_F$ for longitudinal binary data with $p = 50$ from 1000 simulations.

		$\rho_\varepsilon = 0.2$	$\rho_\varepsilon = 0.5$	$\rho_\varepsilon = 0.8$
		$n = 51$		
partial OLS	Independent	1.5507(0.1798)	1.5412(0.1632)	1.5603(0.1775)
partial SIR	Independent	1.5235(0.1955)	1.5209(0.1969)	1.5299(0.2007)
QIF	Independent	0.4142(0.0422)	0.4627(0.0497)	0.5070(0.0552)
	AR-1	0.4092(0.0416)	0.4260(0.0446)	0.4311(0.0436)
	Exchangeable	0.3981(0.0406)	0.3950(0.0416)	0.3871(0.0408)
		$n = 100$		
partial OLS	Independent	0.4208(0.0427)	0.4592(0.0468)	0.4963(0.0522)
partial SIR	Independent	0.4138(0.0423)	0.4521(0.0463)	0.4901(0.0517)
QIF	Independent	0.3008(0.0313)	0.3440(0.0371)	0.3833(0.0415)
	AR-1	0.2929(0.0300)	0.3073(0.0315)	0.3101(0.0305)
	Exchangeable	0.2830(0.0289)	0.2821(0.0287)	0.2752(0.0278)
		$n = 200$		
partial OLS	Independent	0.2427(0.0243)	0.2741(0.0270)	0.3029(0.0294)
partial SIR	Independent	0.2416(0.0243)	0.2731(0.0269)	0.3020(0.0292)
QIF	Independent	0.2153(0.0220)	0.2490(0.0251)	0.2792(0.0275)
	AR-1	0.2086(0.0212)	0.2193(0.0222)	0.2205(0.0216)
	Exchangeable	0.2005(0.0204)	0.1993(0.0203)	0.1931(0.0197)

Table 2.2: Mean and standard deviation of $||\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}' - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'||_F$ for longitudinal continuous data with $d = 2$ for model I from 1000 simulations.

	Model I	$\rho_x = 0.2$	$\rho_x = 0.5$	$\rho_x = 0.8$
		$n = 100$, $2p = 16$		
partial OLS	Independent	1.1180(0.0730)	0.9337(0.1075)	0.8814(0.2544)
partial SIR	Independent	1.0652(0.1216)	1.2080(0.1394)	1.1708(0.1884)
QIF	Independent	0.4836(0.0265)	0.9060(0.0184)	1.1154(0.0105)
	AR-1	0.8142(0.0439)	0.6762(0.0454)	0.5903(0.0674)
	Exchangeable	0.8231(0.0322)	0.6409(0.0309)	0.5501(0.0365)
		$n = 300$, $2p = 30$		
partial OLS	Independent	1.0846(0.0378)	1.0163(0.0405)	1.1000(0.0830)
partial SIR	Independent	1.2093(0.0696)	1.2090(0.0996)	1.3272(0.0938)
QIF	Independent	0.9548(0.0304)	0.9450(0.0204)	1.0431(0.0133)
	AR-1	0.6930(0.0428)	0.7508(0.0447)	0.8741(0.0527)
	Exchangeable	0.5825(0.0358)	0.5883(0.0311)	0.6203(0.0368)

Table 2.3: Mean and standard deviation of $||\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}' - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'||_F$ for longitudinal continuous data with $d = 2$ for model II from 1000 simulations.

	Model II	$\rho_\varepsilon = 0.2$	$\rho_\varepsilon = 0.5$	$\rho_\varepsilon = 0.8$
		$n = 100$, $2p = 16$		
partial OLS	Independent	1.3222(0.1456)	1.3406(0.1400)	1.3517(0.1393)
partial SIR	Independent	1.2555(0.2078)	1.2829(0.2045)	1.3060(0.1949)
QIF	Independent	0.8430(0.1255)	0.8975(0.1477)	0.9602(0.1687)
	AR-1	0.8015(0.1548)	0.7789(0.1600)	0.7233(0.1602)
	Exchangeable	0.7917(0.1555)	0.7787(0.1649)	0.7367(0.1665)
		$n = 300$, $2p = 30$		
partial OLS	Independent	1.3657(0.0906)	1.3772(0.0907)	1.3874(0.0926)
partial SIR	Independent	1.1809(0.2060)	1.1967(0.2022)	1.2212(0.1990)
QIF	Independent	0.5789(0.0720)	0.6430(0.0815)	0.7152(0.0919)
	AR-1	0.5757(0.0832)	0.5706(0.0830)	0.5357(0.0778)
	Exchangeable	0.5446(0.0755)	0.5414(0.0750)	0.5096(0.0716)

Table 2.4: Mean and standard deviation of $||\hat{\mathbf{B}}(\hat{\mathbf{B}}'\hat{\mathbf{B}})^{-1}\hat{\mathbf{B}}' - \mathbf{B}(\mathbf{B}'\mathbf{B})^{-1}\mathbf{B}'||_F$ for longitudinal continuous data with $d = 2$ for model III from 1000 simulations.

	Model III	$\rho_\varepsilon = 0.2$	$\rho_\varepsilon = 0.5$	$\rho_\varepsilon = 0.8$
		$n = 100$, $2p = 16$		
partial OLS	Independent	1.1462(0.2257)	1.1358(0.2271)	1.1130(0.2359)
partial SIR	Independent	1.2368(0.1883)	1.2293(0.1921)	1.2363(0.1911)
QIF	Independent	0.9571(0.0821)	0.9638(0.0959)	0.9713(0.1089)
	AR-1	0.8141(0.1584)	0.8038(0.1503)	0.7941(0.1385)
	Exchangeable	0.7713(0.1303)	0.7735(0.1247)	0.7755(0.1167)
		$n = 300$, $2p = 30$		
partial OLS	Independent	1.3290(0.0929)	1.3258(0.0947)	1.3205(0.0994)
partial SIR	Independent	1.3363(0.0922)	1.3307(0.0972)	1.3319(0.0978)
QIF	Independent	0.8916(0.0750)	0.9093(0.0893)	0.9272(0.0998)
	AR-1	0.9644(0.1244)	0.9466(0.1222)	0.9269(0.1162)
	Exchangeable	0.6716(0.1004)	0.6592(0.0918)	0.6437(0.0837)

Table 2.5: Logistic regression slope estimates, standard errors and p-values for each $\hat{\boldsymbol{\beta}}$ for the asthma data.

	partial OLS	QIF independent	QIF AR-1	QIF exchangeable
Estimate	-1.6884	0.6879	-0.2307	0.6111
Std. Error	0.9654	0.2002	0.6774	0.1991
p-value	0.0839	0.0006	0.7330	0.0021

Table 2.6: MSEs and correlations of four models between the log-odds and the predicted log-odds for the asthma data.

	partial OLS	QIF independent	QIF AR-1	QIF Exchangeable
absolute value of correlation	0.4474	0.3051	0.1039	0.5918
MSE	0.2013	0.5603	0.2361	0.2105

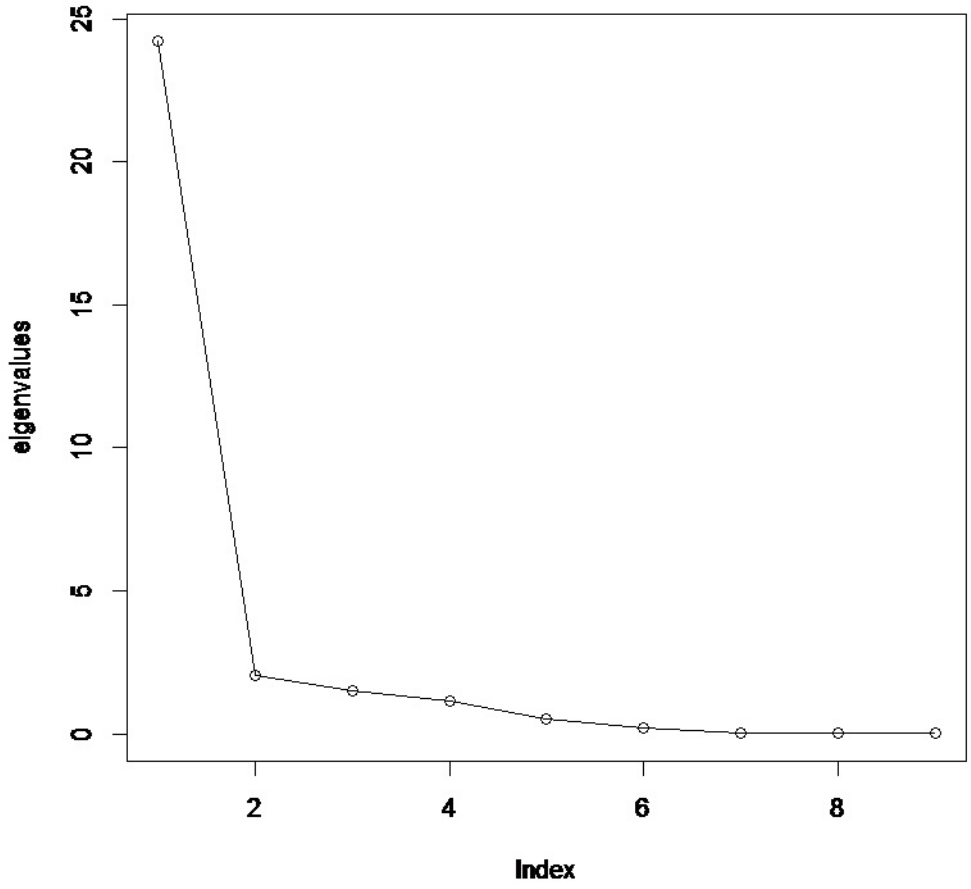

Figure 2.1: Scree plot of eigenvalues from the partial OLS method for the asthma data by Li and Yin (2009).

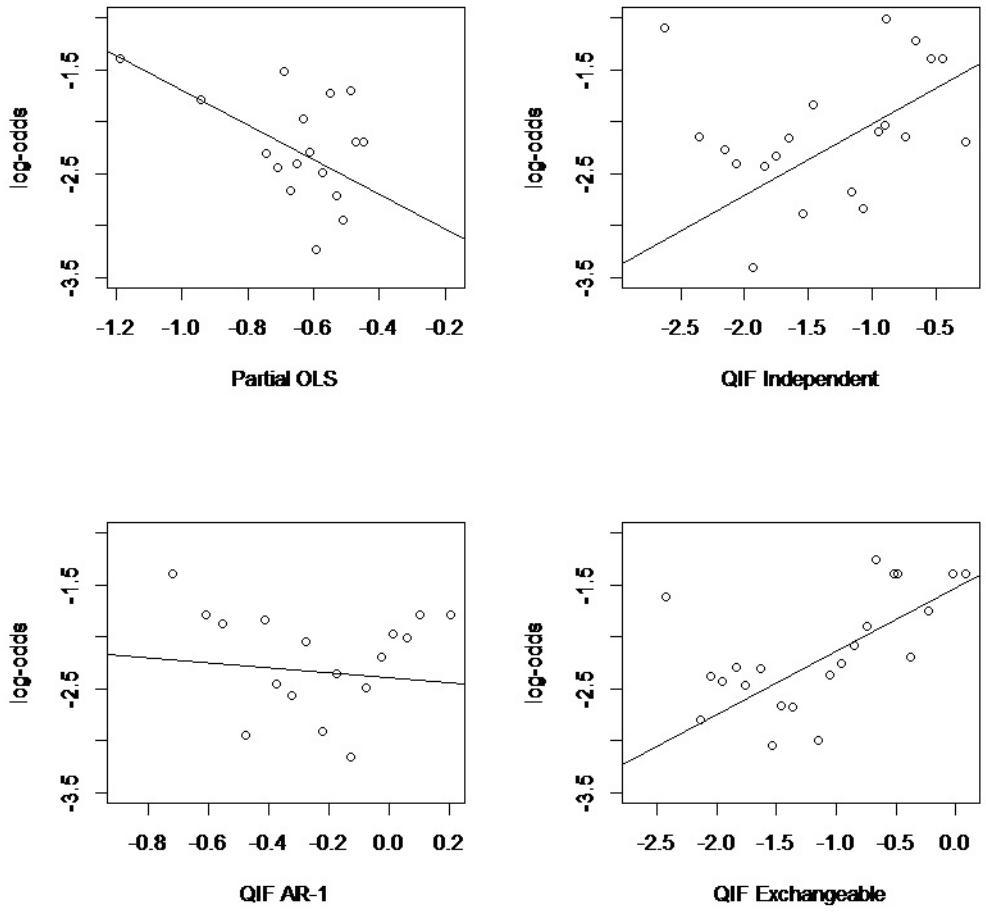

Figure 2.2: Scatterplots and regression lines after grouping, given by four different methods for the asthma data.

Chapter 3

A Mixed-Effects Estimating Equation Approach to Nonignorable Missing Longitudinal Data with Refreshment Samples

3.1 Introduction

Missing data are often encountered in longitudinal studies. Among all the missing mechanisms, missing not at random (MNAR; Rubin, 1976) is the most challenging one to handle. For example, in a public survey, participants with lower socioeconomic status may have a lower probability to release their annual income (Kim and Yu, 2012); and in AIDS clinical trials, subjects with a lower CD4 level may drop out prematurely due to death or pessimism about treatment. Estimation and inference procedures ignoring non-random missing mechanisms may lead to misleading and biased conclusions.

Existing literature on analyzing the MNAR mechanism for longitudinal data includes, but is not limited to, Diggle and Kenward (1994), Little (1994), Little (1995), Hogan and Laird (1997), Molenberghs et al. (1997), Ibrahim et al. (2001), Roy (2003), Stubbendick and Ibrahim (2003), Stubbendick and Ibrahim (2006), Lin et al. (2004), Vansteelandt et al. (2007), Zhou et al. (2010), Spagnoli et al. (2011), and Shao and Zhang (2015). Most of these methods are built under certain MNAR assumptions. However, the MNAR assumption is

typically difficult to verify in practice, since the information required for such a test is also missing (Van Buuren, 2012). Consequently it is challenging to assess model effectiveness and robustness under a general MNAR setting, and a sensitivity analysis (Rotnitzky et al., 1998; Robins et al., 2000) might be required.

An alternative strategy for handling nonignorable missing data is to introduce refreshment samples as part of the experimental design (Ridder, 1992), which recruits new subjects randomly from the same population in subsequent waves over time. Hirano et al. (2001) demonstrate that implementing refreshment samples can mitigate the effect of data attrition, and Deng et al. (2013) further show that refreshment samples are useful to adjust for bias. However, studies of statistical properties are still limited in application to refreshment samples, partially because baseline values from refreshment samples are typically missing. In addition, the existing methods are restricted to few waves with a small longitudinal cluster size, as it could be computationally intensive to handle refreshment samples with a large number of repeated measurements. Furthermore, the MNAR could still exist even after recruitment of refreshment samples.

In this chapter, we propose a mixed-effects estimating equation approach (MEEE) which preserves the advantages of estimating equations in addressing refreshment samples. The key idea is to reduce the estimation bias through utilizing unspecified random effects for MNAR data. In addition, our theoretical properties also confirm that the fixed-effects estimators solved through the MEEE are consistent and asymptotically normal under two different MNAR mechanisms. The proposed method has practical advantages as it is able to utilize a large number of repeated measurements from the same subject to achieve higher estimation accuracy for the random effects. This is in contrast to traditional methods which could be problematic if the cluster size of longitudinal data is large (Lipsitz et al., 2009).

The idea of unspecified random effects has also been considered in the existing literature under the likelihood framework (for example, Tsonaka et al., 2009, 2010; Li et al., 2012; Maruotti, 2015). However, the proposed method has several advantages compared to the

existing approaches. First, the proposed method does not require random effects to follow a discrete distribution with finite support points. In fact, our random effects are solved by estimating equations and their values are not restricted to a certain set. For existing approaches, the selection of the number of support points for the unspecified random-effects distribution remains an open question (Tsonaka et al., 2009). Second, the proposed method is not restricted to shared-parameter models (SPM) (e.g., Wu and Carroll, 1988; Wu and Bailey, 1989; Follmann and Wu, 1995) where the response variable and the missing process are linked through random effects. We demonstrate that the proposed method can be applied under both SPMs and an extended SPM where the missing process is related to observed responses in addition to random effects. Third, the proposed method does not require baseline observations. This is especially useful for longitudinal survey studies with refreshment samples. In such a case, existing methods requiring baseline observations are difficult to implement, while the proposed method is still applicable.

The rest of this chapter is organized as follows. Section 3.2 introduces notation and the shared-parameter model assumption. Section 3.3 illustrates how to construct unbiased estimating equations under MNAR mechanisms and provides theoretical properties. Section 3.4 demonstrates the performance of the proposed method through simulation studies. Section 3.5 applies the proposed method to election poll survey data provided by the 2007-2008 Associated Press–Yahoo! News. Section 3.6 presents concluding remarks and a brief discussion. All technical proofs are shown in Section 3.7.

3.2 Notation and Basic Assumptions

In this section, we introduce notation and basic assumptions for longitudinal missing data.

Let y_{it} denote the tth observation from the ith subject, $\mathbf{x}_{it}$ and $\mathbf{z}_{it}$ are p-dimensional fixed-effects and q-dimensional random-effects covariates, respectively, $i = 1, \ldots, n$, $t = 1, \ldots, T$. We assume that the responses and covariates are linked through a known inverse link function

μ:

$$\mathrm{E}(y_{it}|\mathbf{x}_{it}, \mathbf{z}_{it}, \mathbf{b}_i) = \mu(\mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{z}_{it}\mathbf{b}_i),$$

where $\boldsymbol{\beta}$ is a p-dimensional fixed-effects vector and $\mathbf{b}_i$ is a q-dimensional random-effects vector, $i = 1, \ldots, n$. Suppose $\mathbf{b}_1, \ldots, \mathbf{b}_n$ are true realizations of an unknown stochastic process, and denote $\mathbf{b} = (\mathbf{b}_1', \ldots, \mathbf{b}_n')'$. Notice that we do not impose any distribution assumption on $\mathbf{b}$.

Let δ_{it} be y_{it}'s missing indicator with $\delta_{it} = 1$ if y_{it} is observed and $\delta_{it} = 0$ otherwise. Denote $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$, $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{iT})'$, and $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{iT})'$, where $\mu_{it} = \mu(\mathbf{x}_{it}\boldsymbol{\beta} + \mathbf{z}_{it}\mathbf{b}_i)$. Let $n_i = \sum_{t=1}^{T} \delta_{it}$ be the number of measurements observed from the ith subject. We define a missing indicator matrix Δ_i as an $n_i \times T$-dimensional matrix, corresponding to the rows of identity matrix I_T for which $\mathbf{y}_i$ is observed.

The idea of creating Δ_i is to transform the hypothetical complete response $\mathbf{y}_i$ into an observed vector $\Delta_i\mathbf{y}_i$ (e.g., Paik, 1997). For example, if $\boldsymbol{\delta}_i = (1, 0, 1, 0, 1)'$ and $\mathbf{y}_i = (y_{i1}^o, y_{i2}^m, y_{i3}^o, y_{i4}^m, y_{i5}^o)'$ where the superscript "o" and "m" indicate "observed" and "missing" respectively, then $\Delta_i\mathbf{y}_i = (y_{i1}^o, y_{i3}^o, y_{i5}^o)'$. Each $\boldsymbol{\delta}_i$ determines a unique Δ_i, and thus Δ_i is a function of $\boldsymbol{\delta}_i$. We represent $\mathbf{y}_i^o = \Delta_i\mathbf{y}_i$ and $\boldsymbol{\mu}_i^o = \Delta_i\boldsymbol{\mu}_i$.

The shared-parameter model assumption (e.g., Follmann and Wu, 1995) is:

$$\mathbf{y}_i \perp\!\!\!\perp \boldsymbol{\delta}_i | \mathbf{b}_i, \tag{3.1}$$

which assumes that the missing process $\boldsymbol{\delta}_i$ and the measurement process $\mathbf{y}_i$ share the same random effect $\mathbf{b}_i$. Note that the missing mechanism satisfying (3.1) must be MNAR. We further discuss this assumption and provide an extension in Section 3.3.1.

3.3 The General Method

In this section, we propose an unbiased estimating equation approach to estimate the fixed effect $\boldsymbol{\beta}$ and the random effect $\mathbf{b}$. Throughout this chapter, we use unbiasedness to denote the conditional unbiasedness of an estimating equation given latent random effects.

A standard GEE can be formulated as follows:

$$\sum_{i=1}^{n} \dot{\boldsymbol{\mu}}_i' A_i^{-\frac{1}{2}} R^{-1} A_i^{-\frac{1}{2}} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{3.2}$$

where $\dot{\boldsymbol{\mu}}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}}$, A_i is a diagonal matrix of marginal variance of $\mathbf{y}_i$, and R is a working correlation matrix that contains fewer nuisance parameters than an unspecified correlation matrix.

Notice that the unbiasedness of estimating equation (3.2) leads to the consistency and asymptotic normality of the fixed-effect estimator $\hat{\boldsymbol{\beta}}$ (Liang and Zeger, 1986; Robins et al., 1994; Rotnitzky et al., 1998). Therefore, our goal is to build unbiased estimating equations in the presence of nonignorable missing data.

3.3.1 Construction of Unbiased Estimating Equations

In this subsection, we demonstrate the unbiasedness of the proposed MEEE by incorporating unspecified random effects. Specifically, if either the SPM assumption or a relaxed version of the SPM assumption is satisfied, we have conditionally unbiased estimating equations, which do not rely on a specification of the missing process.

Let $A_i^o = \Delta_i A_i \Delta_i'$, $R_i^o = \Delta_i R \Delta_i'$ and

$$\begin{aligned}
\bar{\mathbf{G}}_n = \frac{1}{n}\sum_{i=1}^{n} \mathbf{G}_i &= \frac{1}{n}\sum_{i=1}^{n} (\dot{\boldsymbol{\mu}}_i^o)' (A_i^o)^{-\frac{1}{2}} (R^o)^{-1} (A_i^o)^{-\frac{1}{2}} (\mathbf{y}_i^o - \boldsymbol{\mu}_i^o) \\
&= \frac{1}{n}\sum_{i=1}^{n} (\dot{\boldsymbol{\mu}}_i)' \Delta_i' (\Delta_i A_i \Delta_i')^{-1/2} (\Delta_i R \Delta_i')^{-1} (\Delta_i A_i \Delta_i')^{-1/2} \Delta_i \left(\mathbf{y}_i - \boldsymbol{\mu}_i\right).
\end{aligned}$$

We define a $T \times T$ weighting matrix conditional on the random effect $\mathbf{b}_i$:

$$K_i = K_i(\boldsymbol{\delta}_i|\mathbf{b}_i) = \Delta_i'(\Delta_i A_i \Delta_i')^{-1/2}(\Delta_i R \Delta_i')^{-1}(\Delta_i A_i \Delta_i')^{-1/2}\Delta_i, i = 1, \ldots, n.$$

Then $\mathbf{G}_i = (\dot{\boldsymbol{\mu}}_i)' K_i(\mathbf{y}_i - \boldsymbol{\mu}_i)$.

Suppose assumption (3.1) holds, then $\mathrm{E}\{(\mathbf{y}_i - \boldsymbol{\mu}_i)|\mathbf{b}_i\} = \mathbf{0}$ implies that

$$\mathrm{E}(\mathbf{G}_i|\mathbf{b}_i) = \mathrm{E}\{\dot{\boldsymbol{\mu}}_i' K_i(\mathbf{y}_i - \boldsymbol{\mu}_i)|\mathbf{b}_i\} = \dot{\boldsymbol{\mu}}_i' \mathrm{E}\{K_i|\mathbf{b}_i\}\mathrm{E}\{(\mathbf{y}_i - \boldsymbol{\mu}_i)|\mathbf{b}_i\} = \mathbf{0}. \tag{3.3}$$

Note that the above decomposition has no restriction on the weighting matrix $K_i(\boldsymbol{\delta}_i|\mathbf{b}_i)$, which contains the information of the MNAR mechanism. This brings up one advantage of the proposed method compared to existing SPM approaches, in that the distribution formulation of the missing process $\boldsymbol{\delta}_i$ is not needed in our approach.

The formulation of (3.3) still requires the SPM assumption (3.1). This assumption does not hold if the missingness is a function of past or current responses directly. Several methods have been developed to weaken the SPM assumption. Among them, Henderson et al. (2000); Rizopoulos et al. (2008) introduce different but correlated random effects for the measurement process and the missing process, where the two processes no longer "share" the same random effects. Little (2008); Yuan and Little (2009) propose a mixed-effects hybrid model which models the dropout process directly. Nevertheless, these works still require parametric assumptions on the random effects, and some are only applicable for the drop-out missing mechanism.

In this chapter, we propose to relax assumption (3.1) through strengthening the association between $\mathbf{y}_i$ and $\boldsymbol{\delta}_i$ if the random effect itself can not completely capture the missing information. Our relaxation does not require any parametric distribution on the random effects nor have restrictions on missing patterns. In particular, the relaxed assumption is applicable to the estimating equation framework where only the first two moments are known. We define a new missing mechanism called *conditionally missing at random (CMAR)* as

follows.

Definition 1. *A missing mechanism is conditionally missing at random if missingness does not depend on unobserved data, given the observed data and the random effects.*

Mathematically, assumption (3.1) states that $\boldsymbol{\delta}_i|\mathbf{b}_i, \mathbf{y}_i \stackrel{d}{=} \boldsymbol{\delta}_i|\mathbf{b}_i$, where $\stackrel{d}{=}$ denotes "equivalent in distribution." Then Definition 1 generalizes assumption (3.1) to

$$\boldsymbol{\delta}_i|\mathbf{b}_i, \mathbf{y}_i \stackrel{d}{=} \boldsymbol{\delta}_i|\mathbf{b}_i, \mathbf{y}_i^o. \tag{3.4}$$

It can be shown that the CMAR mechanism is still a MNAR mechanism. This generalization is analogous to the generalization from MCAR to MAR. That is, we allow the observed response $\mathbf{y}^o$ to carry out information of the missing mechanism as well. The new assumption offers more flexibility compared to assumption (3.1), since it no longer requires the random effect $\mathbf{b}$ to capture all information associating the missing process with the measurement process. This weakens the conditional independence assumption between $\mathbf{y}_i$ and $\boldsymbol{\delta}_i$ in (3.1). In the following we show that the estimating equation (3.2) is still unbiased under the new assumption.

Let $W_i = A_i^{\frac{1}{2}} R A_i^{\frac{1}{2}}$, then for each subject i:

$$\begin{aligned} \mathbf{0} &= \mathrm{E}\{\dot{\boldsymbol{\mu}}_i' W_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)|\mathbf{b}_i\} \\ &= \mathrm{E}\{\dot{\boldsymbol{\mu}}_i' W_i^{-1} \mathrm{E}(\mathbf{y}_i - \boldsymbol{\mu}_i|\mathbf{b}_i, \mathbf{y}_i^o, \boldsymbol{\delta}_i)|\mathbf{b}_i\}. \end{aligned}$$

Suppose $\mathbf{y}_i = (\mathbf{y}_i^o, \mathbf{y}_i^m)'$ and $\boldsymbol{\mu}_i = (\boldsymbol{\mu}_i^o, \boldsymbol{\mu}_i^m)'$, then based on the CMAR assumption in (3.4), $\mathrm{E}(\mathbf{y}_i^m - \boldsymbol{\mu}_i^m|\mathbf{b}_i, \mathbf{y}_i^o, \boldsymbol{\delta}_i) = \mathrm{E}(\mathbf{y}_i^m - \boldsymbol{\mu}_i^m|\mathbf{b}_i, \mathbf{y}_i^o)$ is no longer a function of $\boldsymbol{\delta}_i$, and thus can be modeled by available information through random effects. This follows similarly as the MAR definition.

There are several methods to impute missing values based on observed values. For example, Paik (1997) apply mean imputation for longitudinal data. Seaman and Copas

(2009) combine mean imputation with a weighting strategy to construct a doubly robust estimator. Qu et al. (2010) propose to impute missing values through utilizing the linear conditional mean method (LCM). Here we adopt the LCM under the context of the mixed-effects model for simplicity:

$$\mathrm{E}(\mathbf{y}_i^m - \boldsymbol{\mu}_i^m | \mathbf{b}_i, \mathbf{y}_i^o) = W_i^{21}(W_i^{11})^{-1}(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o), \tag{3.5}$$

where $W_i^{21} = \mathrm{Cov}(\mathbf{y}_i^m, \mathbf{y}_i^o | \mathbf{b}_i)$ and $W_i^{11} = \mathrm{Var}(\mathbf{y}_i^o | \mathbf{b}_i)$.

Given (3.5) and the fact that

$$W_i^{-1}\begin{pmatrix} I \\ W_i^{21}(W_i^{11})^{-1} \end{pmatrix} = \begin{pmatrix} (W_i^{11})^{-1} \\ 0 \end{pmatrix},$$

we have:

$$\begin{aligned} \mathbf{0} = \mathrm{E}\{\dot{\boldsymbol{\mu}}_i' W_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) | \mathbf{b}_i\} &= \mathrm{E}\{(\dot{\boldsymbol{\mu}}_i^o)'(W_i^{11})^{-1}(\mathbf{y}_i^o - \boldsymbol{\mu}_i^o) | \mathbf{b}_i\} \\ &= \mathrm{E}\{\dot{\boldsymbol{\mu}}_i' K_i(\mathbf{y}_i - \boldsymbol{\mu}_i) | \mathbf{b}_i\}. \end{aligned}$$

The above equation indicates that once the LCM method is valid, the estimating equation in (3.2) is unbiased without requiring the shared-parameter model assumption (3.1). Under either the SPM or the CMAR assumption, one can check the conditional unbiasedness of the estimating equation $\mathrm{E}\{\dot{\boldsymbol{\mu}}_i' K_i(\mathbf{y}_i - \boldsymbol{\mu}_i) | \mathbf{b}_i\} = \mathbf{0}$ through a chi-square test (Hansen, 1982; Qu et al., 2000) to test the null hypothesis for the mean zero assumption of the estimating functions.

The LCM imputation method (3.5) is based on the idea of first-order linear approximation. The imputed values are valid if the conditional distribution $\mathbf{y}_i | \mathbf{b}_i$ is multivariate normal, or bivariate binary (Qu et al., 2010). In addition, for a multivariate binary distribution, the LCM is also valid if it belongs to the conditional linear family (Qaqish, 2003) which assumes zero for the second and higher-order terms in Bahadur's representation (Bahadur, 1961).

For other circumstances such as multivariate count data, the LCM provides an approximate estimation with accuracy similar to linear regression. Nevertheless, more complicated imputation methods might be considered if one believes high-order approximations are necessary for the observed data.

3.3.2 Estimation of Mixed Effects

In this subsection we discuss how to solve the proposed MEEE and estimate both fixed effects and unspecified random effects. When the sample size is small or the missing rate is high, the empirical correlation matrix might be unstable or could be non-positive definite. In this case, we apply the following technique to avoid the estimation of such a matrix. Specifically, we formulate estimating functions based on the observed data as

$$\bar{\mathbf{g}}_n^f = \frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i^f(\boldsymbol{\beta}|\mathbf{b}_i) = \frac{1}{n}\begin{pmatrix} \sum_{i=1}^n(\dot{\boldsymbol{\mu}}_i^o)'(A_i^o)^{-1/2}M_{i1}(A_i^o)^{-1/2}\left(\mathbf{y}_i^o-\boldsymbol{\mu}_i^o\right) \\ \vdots \\ \sum_{i=1}^n(\dot{\boldsymbol{\mu}}_i^o)'(A_i^o)^{-1/2}M_{im}(A_i^o)^{-1/2}\left(\mathbf{y}_i^o-\boldsymbol{\mu}_i^o\right) \end{pmatrix},$$

where $M_{ij} = \Delta_i M_j \Delta_i'$ and $\{M_j\}_{j=1}^m$ is a matrix representation of R^{-1}, satisfying $R^{-1} = \sum_{j=1}^{m} a_j M_j$. Here M_j is a basis matrix containing only 0 and 1. See more details on selection of M_j's in Qu et al. (2000), and the number of basis matrices m in Zhou and Qu (2012).

The equality $M_{ij} = \Delta_i M_j \Delta_i'$ entails the assumption $(\Delta_i R \Delta_i')^{-1} = \Delta_i R^{-1}\Delta_i'$, which simplifies the matrix representation for R^{-1} of each subject. This representation does not affect the consistency of estimation when misspecified, and provides better efficiency compared to the one using the independence structure.

We further define

$$K_{ij} = K_{ij}(\boldsymbol{\delta}_i|\mathbf{b}_i) = \Delta_i'(\Delta_i A_i \Delta_i')^{-1/2}M_{ij}(\Delta_i A_i \Delta_i')^{-1/2}\Delta_i, i = 1,\ldots,n; j = 1\ldots,m.$$

Then solving $\bar{\mathbf{g}}_n^f = \mathbf{0}$ is equivalent to solving:

$$\bar{\mathbf{g}}_n^f(\boldsymbol{\beta}|\mathbf{b}) = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n (\dot{\boldsymbol{\mu}}_i)' K_{i1}\,(\mathbf{y}_i - \boldsymbol{\mu}_i) \\ \vdots \\ \frac{1}{n}\sum_{i=1}^n (\dot{\boldsymbol{\mu}}_i)' K_{im}\,(\mathbf{y}_i - \boldsymbol{\mu}_i) \end{pmatrix} = \mathbf{0}.$$

Notice that the relation between K_i and K_{ij} is $K_i = \sum_{j=1}^m a_j K_{ij}$.

For the fixed-effects estimation, since there are more estimating functions than number of parameters, we estimate $\boldsymbol{\beta}$ by applying the generalized method of moments (Hansen, 1982) conditional on $\mathbf{b}$:

$$\hat{\boldsymbol{\beta}} = \arg\min(\bar{\mathbf{g}}_n^f)'(\bar{C}_n^f)^{-1}(\bar{\mathbf{g}}_n^f), \tag{3.6}$$

where $\bar{C}_n^f = \dfrac{1}{n}\sum_{i=1}^n (\mathbf{g}_i^f)(\mathbf{g}_i^f)'$.

For the random-effects estimation, we solve the following equations:

$$\bar{\mathbf{g}}_n^r = \begin{pmatrix} (\frac{\partial \boldsymbol{\mu}_1}{\partial \mathbf{b}_1})' K_1(\mathbf{y}_1 - \boldsymbol{\mu}_1) \\ \vdots \\ (\frac{\partial \boldsymbol{\mu}_n}{\partial \mathbf{b}_n})' K_n(\mathbf{y}_n - \boldsymbol{\mu}_n) \\ \lambda P_A \mathbf{b} \end{pmatrix} = \mathbf{0},$$

where P_A is the projection matrix on the null space of $(I - P_X)Z$ and P_X is the projection matrix on X, respectively, and λ is a tuning parameter. The term $\lambda P_A \mathbf{b}$ is to ensure the identifiability of $\hat{\mathbf{b}}$. The random-effect estimator $\hat{\mathbf{b}}$ is obtained by:

$$\hat{\mathbf{b}} = \arg\min\{(\bar{\mathbf{g}}_n^r)'(\bar{\mathbf{g}}_n^r) + \lambda_1^2 \mathbf{b}'\mathbf{b}\}, \tag{3.7}$$

where $\lambda_1^2 \mathbf{b}'\mathbf{b}$ is an L_2-penalty term to control the magnitude of $\mathrm{Var}(\mathbf{b})$ in order to ensure the convergence in optimization. We estimate $\boldsymbol{\beta}$ and $\mathbf{b}$ by solving (3.6) and (3.7) iteratively. In

Section 3.3.4 we discuss in detail how the tuning parameters λ and λ_1 are selected.

We propose the following chi-square test to test the validity of the LCM imputation method. We construct two sets of estimating equations: one contains subjects with no missing response, and the other has missing responses imputed using the LCM method. Since the first set of estimating equations is always unbiased, a chi-square test can be conducted to test whether the second set of estimating equations is unbiased or not. Let

$$\mathcal{C} = \left\{ i \in \{1, \ldots, n\} : \sum_{t=1}^{T} \delta_{it} = T \right\}$$

be the set of complete subjects. Denote $\Phi_1 = \dfrac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} (\dot{\boldsymbol{\mu}}_i)' V^{-1} (\mathbf{y}_i - \boldsymbol{\mu}_i)$, where $|\cdot|$ denotes the cardinality of a set, and V is the covariance matrix calculated based on subjects with completed responses; and

$$\Phi_2 = \begin{pmatrix} \frac{1}{n-|\mathcal{C}|} \sum_{i \notin \mathcal{C}} (\dot{\boldsymbol{\mu}}_i)' K_{i1} \left(\mathbf{y}_i - \boldsymbol{\mu}_i \right) \\ \vdots \\ \frac{1}{n-|\mathcal{C}|} \sum_{i \notin \mathcal{C}} (\dot{\boldsymbol{\mu}}_i)' K_{im} \left(\mathbf{y}_i - \boldsymbol{\mu}_i \right) \end{pmatrix}$$

be the estimating functions using subjects with missing values. Let $\Phi = (\Phi_1', \Phi_2')'$. We apply Theorem 1 of Qu et al. (2011), then under the null hypothesis $H_0 : \mathrm{E}(\Phi) = \mathbf{0}$,

$$\Phi' \mathrm{Var}^{-1}(\Phi) \Phi \sim \chi^2_{mp}.$$

3.3.3 Asymptotic Properties

In this subsection, we investigate fixed-effects estimation consistency and asymptotic normality. Lemma 2 provides the asymptotic property when $\mathbf{b}$ is known or is consistently estimated, and Theorem 3 shows that the desirable properties still hold under certain conditions even

if $\mathbf{b}$ is unspecified. In the rest of this chaper, we use $\boldsymbol{\beta}_0$ and $\mathbf{b}_0 = (\mathbf{b}'_{01}, \ldots, \mathbf{b}'_{0n})'$ to denote the true fixed effect and the true random effect, respectively.

Lemma 2. *Given that assumption (3.1) is satisfied, conditional on* $\mathbf{b}_0$*,* $\hat{\boldsymbol{\beta}}$ *solved by (3.6) has the following asymptotic properties:*

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O_p(\frac{1}{\sqrt{n}}),$$

and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(0, \Sigma_0),$$

where Σ_0 *is derived in the proof of Lemma 2 provided in the Appendix.*

The conclusions in Lemma 2 still hold if $\mathbf{b}_0$ is replaced by a consistent estimator $\hat{\mathbf{b}}$. However, the consistency of $\hat{\mathbf{b}}$ requires that the cluster size T goes to infinity, which might be too restrictive in practice. Therefore, we provide the following weaker condition, and show that the properties stated in Lemma 2 are still valid. That is, we assume that $\hat{\mathbf{b}}$ satisfies

$$\frac{1}{n}\sum_{i=1}^{n} \mathbf{g}_i^f(\boldsymbol{\beta}_0|\hat{\mathbf{b}}) \to \mathbf{0} \text{ as } n \to \infty. \tag{3.8}$$

This condition implies that conditional on $\hat{\mathbf{b}}$, the sample mean of estimating equations for the fixed effect converges to 0 when the sample size n goes to infinity while the cluster size T is fixed.

In fact, condition (3.8) is weaker than the consistency of $\hat{\mathbf{b}}$. This is because, if $\hat{\mathbf{b}}$ is consistent, then condition (3.8) holds true for a large T. However, we can show a counterexample where condition (3.8) does not imply the consistency of $\hat{\mathbf{b}}$. Suppose

$$y_{it} = \beta_0 + b_i + \varepsilon_{it},$$

where $\beta_0 = 0$, $\mathrm{E}(b_i) = 0$, $\mathrm{E}(\varepsilon_{it}) = 0$, and $\mathrm{Corr}(\varepsilon_{it}, \varepsilon_{it'}) = 1$, for $i = 1, \ldots, n$ and $t, t' =$

$1, \ldots, T$. Then $y_{it} = y_{it'}$ with a probability of 1 and the corresponding quasi-likelihood equation is:

$$g_i^f(\beta_0|\hat{b}) = \dot{\boldsymbol{\mu}}_i'(\mathbf{y}_i - \boldsymbol{\mu}_i) = \sum_{t=1}^{T}(y_{it} - \hat{b}_i).$$

If condition (3.8) is satisfied, then $\hat{b}_i = \dfrac{1}{T}\sum_{t=1}^{T} y_{it} = y_{i1}$. However, $\hat{b}_i = y_{i1}$ is not a consistent estimator of b_i as $T \to \infty$.

Theorem 3. *Suppose that (3.1) and (3.8) hold, then conditional on* $\hat{\mathbf{b}}$, $\hat{\boldsymbol{\beta}}$ *solved by (3.6) has the following properties:*

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O_p(\frac{1}{\sqrt{n}});$$

and further,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(0, \Sigma),$$

where Σ *is derived in the proof of Theorem 3 provided in the Appendix.*

If $\hat{\mathbf{b}}$ is a consistent estimator of $\mathbf{b}_0$, then we have $\Sigma = \Sigma_0$ (Wang et al., 2012). Next, we relax the regular shared-parameter model assumption, and show that the fixed-effects estimator is still consistent and asymptotically normal under the CMAR mechanism described in Definition 1.

Corollary 2. *Suppose that (3.4) and (3.8) hold and the LCM imputation method (3.5) is valid, then conditional on* $\hat{\mathbf{b}}$, $\hat{\boldsymbol{\beta}}$ *solved by (3.6) satisfies*

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 = O_p(\frac{1}{\sqrt{n}}),$$

and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \to N(0, \Sigma).$$

Corollary 2 can be shown similarly as the derivation of Theorem 3 and is therefore omitted.

Remark: Given $\boldsymbol{\beta}_0$ or its consistent estimator $\hat{\boldsymbol{\beta}}$, the penalized random-effects estimator $\hat{\mathbf{b}}$ is consistent as the cluster size T goes to infinity, as discussed in Cho et al. (2016), given that regularity conditions are satisfied. Under the nonignorable missing data framework, the consistency property holds as long as either the SPM or the CMAR assumption is satisfied. The proof is quite similar to Cho et al. (2016), and is therefore omitted here. One notable condition is the L_2-mixingale condition (McLeish et al., 1975) which controls the serial correlation $\text{Cor}(\mathbf{y}_i|\mathbf{b}_i)$ to achieve the consistency property. That is, $\text{Cor}(y_{it}, y_{i,t+s})$ should be sufficiently small with an increase of s.

3.3.4 Tuning Parameter Selection

In this subsection, we discuss the selection of tuning parameters λ and λ_1 in (3.7). A large value of λ guarantees that the random-effects estimation is identifiable. However, a very large value of λ does not enhance identifiability significantly, and might result in slower convergence or non-convergence of the algorithm. In our numerical studies, we notice that $\lambda = \log(n)$ is sufficiently large to balance the needs of estimation identifiability and algorithm convergence.

On the other hand, the estimation is more sensitive towards the choice of λ_1, since a larger value of λ_1 leads to a smaller variance of $\hat{\mathbf{b}}$ that could affect the estimation of $\boldsymbol{\beta}$. The term $\lambda_1 \mathbf{b}'\mathbf{b}$ is essentially an L_2-penalty, which controls the bias-variance trade-off of $\mathbf{b}$. As a special case, when $\boldsymbol{\mu}$ is an identity mean function, $\hat{\mathbf{b}}$ is equivalent to a ridge regression estimator. We use a cross-validation method to select λ_1. Note that each subject has a unique random effect, and hence a classical K-fold cross-validation is not applicable. We propose a longitudinal K-fold cross-validation with $K = T$. That is, each time we remove measurements observed at time t from all subjects ($t = 1, \ldots, T$), and minimize the following objective function:

$$L(\lambda_1) = \frac{1}{\sum_{i=1}^{n} n_i} \sum_{i=1}^{n} (\mathbf{y}_i^o - \hat{\mathbf{y}}_i^o)' (W_i^o)^{-1} (\mathbf{y}_i^o - \hat{\mathbf{y}}_i^o),$$

where $\hat{\mathbf{y}}_i^o = \Delta_i(\hat{y}_{i1}^{(-1)}, \ldots, \hat{y}_{iT}^{(-T)})'$ and $\hat{y}_{it}^{(-t)}$ is the predicted value of y_{it} without using information from time t.

To calculate $\hat{y}_{it}^{(-t)}$, (Hastie et al., 2009, Chapter 7.2) suggest $\hat{y}_{it} = \arg\max_y f_{it,\lambda_1}(y)$, where f_{it,λ_1} is the probability density function or the probability mass function of y_{it} with parameter λ_1. For example, if y_{it} is normally distributed, then $\hat{y}_{it}^{(-t)} = \mathbf{x}_{it}\hat{\boldsymbol{\beta}}^{(-t)}$; and if y_{it} follows a Poisson distribution, then $\hat{y}_{it}^{(-t)} = [\hat{\mu}_{it}^{(-t)}]$, where $[\hat{\mu}_{it}^{(-t)}]$ denotes the largest integer not greater than $\hat{\mu}_{it}^{(-t)}$.

3.4 Simulation Studies

We conduct simulation studies to examine the performance of the proposed method. To make a fair comparison to existing approaches using the SPM, we compare the proposed MEEE with generalized linear mixed-effects models (GLMMs), where the estimating equations are also unbiased under the SPM assumption (3.1). The difference is that the GLMM assumes normality of the random effects and independence of repeated measurements, given that random effects are taken into account. The GLMM can be implemented through the penalized quasi-likelihood (PQL; Breslow and Clayton, 1993) or the adaptive Gaussian-Hermite quadrature approach (GHQ; Anderson and Aitkin, 1985). In addition, we also compare the proposed method with two marginal estimating equation approaches, namely, the weighted generalized estimating equations (WGEE; Robins et al., 1995), and the multiple imputation method for longitudinal data (MI; Fitzmaurice et al., 2012). Although the WGEE and multiple imputation have the marginal interpretation, and are valid under only MAR. These two approaches are benchmark methods under the estimating equation framework for longitudinal missing data, and therefore we also include them for comparisons.

The PQL and the GHQ are carried out using the R functions "glmmPQL" and "glmer", respectively, while the WGEE is obtained by assigning weights to the R function "geem." For the MI approach, we impute missing data following (Fitzmaurice et al., 2012, Chapter

18.2) when the missing pattern is monotone, or apply the R package "MICE" to impute intermittent missing data utilizing the chained equation (Van Buuren, 2007). For WGEE, we tailor the responses to a monotone pattern of missingness in order to apply the weighting strategy.

In addition, we also conduct a chi-square test (Qu et al., 2000) to test the unbiasedness of the fixed-effects estimating equations $\mathrm{E}(\bar{\mathbf{g}}_n^f) = \mathbf{0}$, and the validity of the LCM imputation method indicated in Section 3.3.2.

3.4.1 Study 1: Count Responses under the SPM Assumption

We choose the sample size $n = 150$ and the cluster size $T = 3$. The fixed-effects covariates $\mathbf{x}_{it} = (1, \mathrm{trt}, \mathrm{time}, \mathrm{trt} \times \mathrm{time})'$, where "trt" (treatment) is assigned to 1 if $i \leq n/2$ and 0 otherwise, "time" is the standardized time effect, and "trt $\times$ time" is the interaction effect of treatment and time. The fixed-effects parameter $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)' = (-0.5, 0.5, 0.2, 0.2)'$. The random-effects covariates $\mathbf{z}_{it} = (x_{it1}, x_{it3})'$, and the random effects $b_{ij} \overset{iid}{\sim} \mathrm{Unif}(-0.2, 0.2)$ for $j = 1, 2$. Each $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$ is sampled from a multivariate Poisson distribution with mean $\boldsymbol{\lambda}_i$ satisfying:

$$\log(\lambda_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{b}_i, \ t = 1, \ldots, T;$$

and the correlation structure for the repeated measurements is exchangeable with correlation parameter $\rho = 0, 0.4$ or 0.7. The correlated Poisson data are generated by the R package "corcounts."

For the missing process, let $p_{it}^{\delta} = \mathrm{P}(\delta_{it} = 1)$ and the logistic model of p_{it}^{δ} be

$$\mathrm{logit}(p_{it}^{\delta}) = 0.1b_{i1} - 0.3t/T + 0.1, \ t = 2, \ldots, T,$$

where the assumption (3.1) is satisfied. The term $-t/T$ ensures that the missing rate is higher towards the end of the study, which resembles the real data case. We assume a

monotone pattern of missingness, and $\delta_{it} = \cdots = \delta_{iT} = 0$ if $\delta_{i,t-1} = 0$. The overall missing rate is about 45%.

Table 3.1 provides simulation results based on 200 simulation runs. For the unbiasedness test, we reject the null hypothesis 8, 5 and 10 times out of 200 replications at a significance level of 0.05 when the serial correlation $\rho = 0$, $\rho = 0.4$ and $\rho = 0.7$, respectively. This indicates that the estimating equations are unbiased, and fixed-effects estimates are consistent. For the test of the validity of the LCM, we reject the null hypothesis 101, 63 and 19 times out of 200 replications at a level of 0.05 for $\rho = 0$, 0.4 and 0.7, respectively. This agrees with the theory in that the LCM imputation is only an approximation when the response variable is count data. Nevertheless, the proposed method satisfies the SPM assumption (3.1), and performs the best even when the LCM condition is violated.

Overall the proposed estimators are less biased and have smaller standard errors compared to the WGEE and the MI approaches. In addition, the improvement of the proposed method is more significant when the correlation parameter ρ increases in general. The PQL does not converge due to a small cluster size T, and the GHQ is not applicable here since the dimension of parameters is greater than the number of data points.

3.4.2 Study 2: Binary Responses under the CMAR Assumption

In the second simulation study, we evaluate the performance of the proposed estimator when the assumption (3.1) is violated but the assumption (3.4) is satisfied.

We generate data with sample size $n = 80$ and cluster size $T = 6$. The fixed-effect covariates and the random-effect covariates are the same as in the simulation study 1. The fixed-effect is $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)' = (-0.5, 1, 0.8, 0.8)'$. The random-effect covariate is $\mathbf{z}_{it} = (x_{it1}, x_{it3})'$, and the random effects $b_{ij} \overset{iid}{\sim} \text{Unif}(-0.2, 0.2)$ for $j = 1, 2$. The response $\mathbf{y}_i = (y_{i1}, \ldots, y_{iT})'$ follows a multivariate Bernoulli distribution with mean function μ_{it} satisfying

$$\text{logit}(\mu_{it}) = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{z}_{it}'\mathbf{b}_i, \ t = 1, \ldots, T,$$

and an AR-1 correlation structure with correlation parameter $\rho = 0.2$ or 0.6, generated by the R package "MultiOrd." Note that the correlation structure of $\rho > 0.6$ cannot be generated due to infeasibility of the above mean function (Chaganty and Joe, 2006).

We generate the missing process through a logistic model:

$$\text{logit}(p_{it}^{\delta}) = 0.1y_{i1} + 0.2b_{i1} - 0.5t/T + 0.5,\ t = 2, \ldots, T.$$

In this setting, the assumption (3.4) is satisfied. The missing pattern is intermittent, and the overall missing rate is about 40%.

Table 3.2 provides the simulation results from 200 replications when $\rho = 0.2$ or 0.6. For the unbiasedness test, we reject the null 72 and 80 times out of 200 simulations at a significance level of 0.05 when the serial correlation $\rho = 0.2$ and $\rho = 0.6$, respectively. That is, the unbiasedness of the estimating equations is mildly violated. In addition, since assumption (3.1) is violated, we conduct the chi-square test proposed in Section 3.3.2 to test the validity of the LCM, which leads us to reject the null hypothesis 14 and 37 times out of 200 replications for $\rho = 0.2$ and 0.6, respectively. That is, the LCM is mildly violated in this setting. However, the MEEE still outperforms other approaches and achieves the smallest absolute bias and standard error with its coverage probability around 95%. The improvement of the proposed method is more evident when the correlation parameter $\rho = 0.6$, where the PQL and GHQ deteriorate drastically with higher bias, standard errors, and much lower coverage probability of the confidence interval.

3.5 Application

In this section, we analyze data collected from the 2007-2008 Associated Press-Yahoo! News Poll. This survey intends to evaluate changes in nationwide attitude and opinion towards the presidential election in 2008. It is an eleven-wave survey with the first 9 waves conducted during the year prior to the 2008 general election.

Respondents were invited to participate in all follow-up waves, regardless of their responses to the previous waves, so the missing pattern is non-monotone. However, this survey suffers greatly from data attrition, where only 63% of the first wave respondents still participate in wave 9. To offset the high percentage of the missing rate, this survey recruits new participants as refreshment samples in waves 3, 5, 6 and 9.

We choose one of the survey questions as a response variable: "How much interest do you have in following news about the campaign for president?" Following the Pew Research Center (2010) and Deng et al. (2013) strategy, we dichotomize the 5-level response into a binary variable: 1 for answers "a great deal" or "quite a bit" and 0 otherwise. We analyze all available data collected in the 9 waves before the election, and the total sample size is 4719. The response measurements from the same subject are correlated, with an approximately exchangeable correlation structure and an average correlation around 0.6. The overall missing rate of the response variable is 49.7%. The predictors are all observed, including time, age, education, gender, household income, marital status, whether living in a metropolitan statistical area (MSA Status), and race/ethnicity.

Here the missingness of the response variable is likely to be nonignorable. This is reflected by the left panel of Figure 3.1 showing that respondents are more interested in the presidential election if they stay in the survey longer. That is, the missing probability depends highly on the measurement process. In addition, the missing mechanisms occurring in the refreshment samples are also likely to be MNAR. A two-sample t-test shows that in the last wave before the election, new respondents collected in the last wave have significantly higher interest in the presidential campaign than the respondents recruited in earlier waves, indicating that the earlier measurements from refreshment samples are missing nonignorably. This is also indicated by the left panel of Figure 3.1, in that the responses from subjects with only one observation have a higher average interest in the presidential election, as these subjects are mainly recruited in the last wave.

We assume a random intercept model for the MEEE, the PQL and the GHQ, and compare

them with three marginal approaches: GEE, WGEE, and MI, for which estimations, standard errors and p-values are provided in Table 3.3. We conduct the unbiasedness test of the estimating equations and the validity test of the LCM for the proposed method. Both tests reject the null, indicating that these assumptions are violated. However, the MEEE approach agrees with most of the other methods that as the election time gets closer, older people with higher education level and higher household income are more interested in the presidential election. In addition, except for the MI with monotonized data, the other six methods show that "Black and Non-Hispanic" people are more interested in the presidential election than "White and Non-Hispanic." The most interesting finding here is that methods incorporating refreshment samples such as MEEE, PQL, GHQ, GEE and MI with all available data are able to detect a significant difference in interest between males and females, which coincides with the finding of the Pew Research Center (2010). This implies that refreshment samples may contain important information which should not be ignored. In addition, the MEEE has smaller standard errors for estimators regarding "MSA status" and "Other Non-Hispanic" with more significant p-values.

The right panel of Figure 3.1 plots the average estimated random effects versus the number of observations, which agrees with the left panel in that a large value of random effect implies high interest in the election. Figure 3.2 is a histogram of the estimated random effects given by the MEEE, which shows a bimodal pattern. A Shapiro-Wilk test indicates that the normality assumption for random effects is severely violated (p-value $< 10^{-15}$). Existing approaches that impose the normality assumption may result in estimation bias and misleading inference.

3.6 Discussion

In this chapter, we propose a mixed-effects model to correct estimation bias for nonignorable missing data. Mainly, we construct unbiased estimating equations with unspecified random

effects under a shared-parameter model, and extend it to a more general nonignorable-missing framework. We show that consistency of the fixed-effects parameter estimation can still be achieved under the more general framework. To our knowledge, most existing methods in the shared-parameter model framework require either a parametric distribution assumption or finite support points for the random effects. In contrast, the proposed method allows unspecified random effects which do not have such restrictions. In addition, the proposed method imposes no restriction on the missing pattern, and hence it can be effectively applied to refreshment samples where baseline observations are subject to missing.

For future research, it would be worthwhile to develop a method for handling missing covariates and responses simultaneously (e.g., Lee and Tang, 2006; Chen et al., 2012a). In our framework, since neither the SPM assumption in (3.1) nor the relaxed assumption in (3.4) imposes constraints on covariates, we can treat the covariate with missing values as a new response variable and apply the MEEE.

3.7 Proofs of Theoretical Results

3.7.1 Notation and Regularity Conditions

Define the quadratic inference function:

$$Q_n(\boldsymbol{\beta}|\mathbf{b}) = (\bar{\mathbf{g}}_n^f)'(\bar{C}_n^f)^{-1}(\bar{\mathbf{g}}_n^f),$$

its first partial derivative:

$$\dot{Q}_n(\boldsymbol{\beta}|\mathbf{b}) = \frac{\partial}{\partial\beta}Q_n(\boldsymbol{\beta}|\mathbf{b}) = 2(\dot{\bar{\mathbf{g}}}_n^f)'(\bar{C}_n^f)^{-1}(\bar{\mathbf{g}}_n^f) + o(1),$$

and its second partial derivative:

$$\ddot{Q}_n(\boldsymbol{\beta}|\mathbf{b}) = \frac{\partial^2}{\partial\beta^2} Q_n(\boldsymbol{\beta}|\mathbf{b}) = 2(\dot{\mathbf{g}}_n^f)'(\bar{C}_n^f)^{-1}(\dot{\mathbf{g}}_n^f) + o(1).$$

Define $\dot{\mathbf{g}}_0 = \mathrm{E}(\dot{\mathbf{g}}_i^f|\mathbf{b}_0)$, and $C_0 = \mathrm{Var}(\mathbf{g}_i|\mathbf{b}_0)$.

We here provide the regularity conditions to prove Lemma 2 and Theorem 3.

(i) The response variables $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are i.i.d.

(ii) The fixed effect $\boldsymbol{\beta}$ is identifiable; that is, there exists a unique $\boldsymbol{\beta}_0$, such that $\mathrm{E}\{\mathbf{g}_i^f(\boldsymbol{\beta}_0|\mathbf{b}_0)\} = \mathbf{0}$.

(iii) The estimating function $\mathbf{g}_i(\boldsymbol{\beta}|\mathbf{b})$ is differentiable with respect to both $\boldsymbol{\beta}$ and $\mathbf{b}$, $i = 1, \ldots, n$.

(iv) $\mathrm{Var}(\mathbf{g}_i|\mathbf{b}) < \infty$ in probability, for $i = 1, \ldots, n$.

(v) $\dot{\mathbf{g}}_n^f(\boldsymbol{\beta}|\mathbf{b})$ is uniformly bounded in probability with respect to both $\boldsymbol{\beta}$ and $\mathbf{b}$ in an open bounded space containing $\boldsymbol{\beta}_0$ and $\mathbf{b}_0$, and conditional on $\mathbf{b}_0$, $\dot{\mathbf{g}}_n^f \xrightarrow{a.s.} \dot{\mathbf{g}}_0$ as $n \to \infty$.

(vi) $\bar{C}_n^f(\boldsymbol{\beta}|\mathbf{b})$ is uniformly bounded in probability with respect to both $\boldsymbol{\beta}$ and $\mathbf{b}$ in an open bounded space containing $\boldsymbol{\beta}_0$ and $\mathbf{b}_0$, and conditional on $\mathbf{b}_0$, $\bar{C}_n^f \xrightarrow{a.s.} C_0$ as $n \to \infty$.

(vii) There exists an open bounded parameter space $\mathcal{S} \subseteq \mathbb{R}^p$, such that $\boldsymbol{\beta}_0 \in \mathcal{S}$ and $Q_n(\boldsymbol{\beta}|\mathbf{b}_0)$ is uniformly convergent in probability in $\mathcal{S}$. Define:

$$Q(\boldsymbol{\beta}|\mathbf{b}_0) = \lim_{n\to\infty} Q_n(\boldsymbol{\beta}|\mathbf{b}_0),$$

and thus:

$$\dot{Q}(\boldsymbol{\beta}|\mathbf{b}_0) = \lim_{n\to\infty} \dot{Q}_n(\boldsymbol{\beta}|\mathbf{b}_0).$$

3.7.2 Proofs of Lemma 2 and Theorem 3

Proof of Lemma 2. Solving $\hat{\boldsymbol{\beta}} = \arg\min(\bar{\mathbf{g}}_n^f)'(\bar{C}_n^f)^{-1}(\bar{\mathbf{g}}_n^f)$ is equivalent to solving

$$\dot{Q}_n(\hat{\boldsymbol{\beta}}|\mathbf{b}_0) = \mathbf{0}.$$

By Taylor expansion, we have:

$$\mathbf{0} = \dot{Q}_n(\hat{\boldsymbol{\beta}}|\mathbf{b}_0) = \dot{Q}_n(\boldsymbol{\beta}_0|\mathbf{b}_0) + \ddot{Q}_n(\boldsymbol{\beta}_0|\mathbf{b}_0)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) + o(\frac{1}{\sqrt{n}}),$$

By regularity conditions (ii), (v) and (vi), we have $\mathrm{E}\{\dot{Q}_n(\boldsymbol{\beta}_0|\mathbf{b}_0)\} = \mathbf{0}$. Then by regularity condition (iv) and the central limit theorem, we conclude that:

$$\dot{Q}_n(\boldsymbol{\beta}_0|\mathbf{b}_0) \sim O(\frac{1}{\sqrt{n}}) \text{ and } \sqrt{n}(\dot{Q}_n(\boldsymbol{\beta}_0|\mathbf{b}_0)) \to \mathrm{N}(\mathbf{0}, \Omega_0),$$

where

$$\begin{aligned}
\Omega_0 &= \lim_{n\to\infty} n\mathrm{Var}(\dot{Q}_n(\boldsymbol{\beta}_0|\mathbf{b}_0)) \\
&= 4\lim_{n\to\infty}(\dot{\bar{\mathbf{g}}}_n^f)'(\bar{C}_n^f)^{-1}\{\frac{1}{n}\sum_{i=1}^{n}\mathrm{Var}(\mathbf{g}_i|\mathbf{b}_0)\}(\bar{C}_n^f)^{-1}(\dot{\bar{\mathbf{g}}}_n^f) \\
&= 4(\dot{\mathbf{g}}_0)'(C_0)^{-1}(\dot{\mathbf{g}}_0).
\end{aligned}$$

Since $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = -\ddot{Q}_n^{-1}(\boldsymbol{\beta}_0|\mathbf{b}_0) \cdot \sqrt{n}\dot{Q}_n(\boldsymbol{\beta}_0|\mathbf{b}_0) + o(1)$, we conclude that:

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \to \mathrm{N}(\mathbf{0}, \Sigma_0),$$

where $\Sigma_0 = \lim_{n\to\infty}\{\ddot{Q}_n^{-1}(\boldsymbol{\beta}_0|\mathbf{b}_0)\}\Omega_0\{\ddot{Q}_n^{-1}(\boldsymbol{\beta}_0|\mathbf{b}_0)\}' = \{(\dot{\mathbf{g}}_0)'(C_0)^{-1}(\dot{\mathbf{g}}_0)\}^{-1}$. □

Proof of Theorem 3. Solving $\hat{\boldsymbol{\beta}} = \arg\min(\bar{\mathbf{g}}_n^f)'(\bar{C}_n^f)^{-1}(\bar{\mathbf{g}}_n^f)$ is equivalent to finding $\hat{\boldsymbol{\beta}}$ such that $\dot{Q}_n(\hat{\boldsymbol{\beta}}|\hat{\mathbf{b}}) = \mathbf{0}$.

Based on regularity conditions (ii), (v), (vi) and (vii), we have $Q(\beta_0|\mathbf{b}_0) = 0$ and

$\dot{Q}(\beta_0|\mathbf{b}_0) = \mathbf{0}$. And based on regularity conditions (v) and (vi) and the condition that $\frac{1}{n}\sum_{i=1}^{n}\mathbf{g}_i(\boldsymbol{\beta}_0|\hat{\mathbf{b}}) \to \mathbf{0}$ as $n \to \infty$, we have:

$$\lim_{n\to\infty} \dot{Q}_n(\boldsymbol{\beta}_0|\hat{\mathbf{b}}) = \mathbf{0} = \dot{Q}(\boldsymbol{\beta}_0|\mathbf{b}_0). \tag{3.9}$$

Define the boundary of a ball in $\mathcal{S}$ with center $\boldsymbol{\beta}_0$ and radius $\frac{1}{\sqrt{n}}$ as $\partial B_n(\boldsymbol{\beta}_0) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = \frac{1}{\sqrt{n}}\}$. Then for any $\boldsymbol{\beta} \in \partial B_n(\boldsymbol{\beta}_0)$, we have:

$$\mathbf{0} = Q(\boldsymbol{\beta}_0|\mathbf{b}_0) = Q(\boldsymbol{\beta}|\mathbf{b}_0) + \dot{Q}(\boldsymbol{\beta}|\mathbf{b}_0)(\boldsymbol{\beta}_0 - \boldsymbol{\beta}) + o(\frac{1}{\sqrt{n}}).$$

Since $Q(\boldsymbol{\beta}|\mathbf{b}_0) > 0$ when $\boldsymbol{\beta} \neq \boldsymbol{\beta}_0$, we can find an $\epsilon > 0$, such that:

$$(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\dot{Q}(\boldsymbol{\beta}|\mathbf{b}_0) = Q(\boldsymbol{\beta}|\mathbf{b}_0) + o(\frac{1}{\sqrt{n}}) > \epsilon > 0.$$

Then based on (3.9), for such ϵ, there exists a large N, such that when $n > N$,

$$\begin{aligned} & \|\dot{Q}_n(\boldsymbol{\beta}|\hat{\mathbf{b}}) - \dot{Q}(\boldsymbol{\beta}|\mathbf{b}_0)\| \\ \leq\ & \|\dot{Q}_n(\boldsymbol{\beta}|\hat{\mathbf{b}}) - \dot{Q}_n(\boldsymbol{\beta}_0|\hat{\mathbf{b}})\| + \|\dot{Q}_n(\boldsymbol{\beta}_0|\hat{\mathbf{b}}) - \dot{Q}(\boldsymbol{\beta}_0|\mathbf{b}_0)\| + \|\dot{Q}(\boldsymbol{\beta}_0|\mathbf{b}_0) - \dot{Q}(\boldsymbol{\beta}|\mathbf{b}_0)\| \\ <\ & \epsilon \end{aligned}$$

for $\boldsymbol{\beta} \in \partial B_n(\boldsymbol{\beta}_0)$. This is because $\dot{\bar{\mathbf{g}}}_n^f(\boldsymbol{\beta}|\mathbf{b})$ and $\bar{C}_n^f(\boldsymbol{\beta}|\mathbf{b})$ are uniformly bounded and $\bar{\mathbf{g}}_n^f$ is continuous with respect to $\boldsymbol{\beta}$, so

$$\|\dot{Q}_n(\boldsymbol{\beta}|\hat{\mathbf{b}}) - \dot{Q}_n(\boldsymbol{\beta}_0|\hat{\mathbf{b}})\| < \frac{1}{3}\epsilon, \text{ and } \|\dot{Q}(\boldsymbol{\beta}_0|\mathbf{b}_0) - \dot{Q}(\boldsymbol{\beta}|\mathbf{b}_0)\| < \frac{1}{3}\epsilon$$

for a large N. And because of (3.9),

$$\|\dot{Q}_n(\boldsymbol{\beta}_0|\hat{\mathbf{b}}) - \dot{Q}(\boldsymbol{\beta}_0|\mathbf{b}_0)\| < \frac{1}{3}\epsilon.$$

By the Cauchy-Schwarz Inequality:

$$\begin{aligned} |(\boldsymbol{\beta}-\boldsymbol{\beta}_0)[\dot{Q}_n(\boldsymbol{\beta}|\hat{\mathbf{b}})-\dot{Q}(\boldsymbol{\beta}|\mathbf{b}_0)]| &\leq \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|\cdot\|\dot{Q}_n(\boldsymbol{\beta}|\hat{\mathbf{b}})-\dot{Q}(\boldsymbol{\beta}|\mathbf{b}_0)\| \\ &< \frac{1}{\sqrt{n}}\epsilon. \end{aligned}$$

Therefore,

$$\begin{aligned} (\boldsymbol{\beta}-\boldsymbol{\beta}_0)\dot{Q}_n(\boldsymbol{\beta}|\hat{\mathbf{b}}) &> (\boldsymbol{\beta}-\boldsymbol{\beta}_0)\dot{Q}(\boldsymbol{\beta}|\mathbf{b}_0)-\frac{1}{\sqrt{n}}\epsilon \\ &> (\boldsymbol{\beta}-\boldsymbol{\beta}_0)\dot{Q}(\boldsymbol{\beta}|\mathbf{b}_0)-\epsilon > 0. \end{aligned}$$

Then based on Theorem 6.3.4 of Ortega and Rheinboldt (1970) (p.163), there exists a $\hat{\boldsymbol{\beta}}_n \in B_n(\boldsymbol{\beta}_0)$, such that

$$\dot{Q}_n(\hat{\boldsymbol{\beta}}_n|\hat{\mathbf{b}}) = \mathbf{0}.$$

This is a direct application of the p-dimensional intermediate value theorem. Since $\hat{\boldsymbol{\beta}}_n \in B_n(\boldsymbol{\beta}_0)$, we have $\hat{\boldsymbol{\beta}}_n = O(\frac{1}{\sqrt{n}})$ and $\hat{\boldsymbol{\beta}}_n \to \boldsymbol{\beta}_0$ as $n \to \infty$.

The following part shows the asymptotic normality of $\hat{\boldsymbol{\beta}}_n$.

From Lemma 2, we have:

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_0-\boldsymbol{\beta}_0) = -\ddot{Q}_n^{-1}(\boldsymbol{\beta}_0|\mathbf{b}_0)\cdot\sqrt{n}\dot{Q}_n(\boldsymbol{\beta}_0|\mathbf{b}_0)+O(\frac{1}{\sqrt{n}}), \tag{3.10}$$

where $\hat{\boldsymbol{\beta}}_0$ is the solution of $\hat{\boldsymbol{\beta}} = \arg\min(\bar{\mathbf{g}}_n^f)'(\bar{C}_n^f)^{-1}(\bar{\mathbf{g}}_n^f)$ conditional on $\mathbf{b}_0$.

Since $\hat{\boldsymbol{\beta}}_n \in B_n(\boldsymbol{\beta}_0)$, for any $\epsilon > 0$, we have $\|\dot{Q}_n(\hat{\boldsymbol{\beta}}_n|\hat{\mathbf{b}})-\dot{Q}(\hat{\boldsymbol{\beta}}_n|\mathbf{b}_0)\| < \epsilon$, and hence $\|\dot{Q}_n(\hat{\boldsymbol{\beta}}_n|\hat{\mathbf{b}})-\dot{Q}_n(\hat{\boldsymbol{\beta}}_n|\mathbf{b}_0)\| < \epsilon$ for a large N and $n > N$. In addition,

$$\begin{aligned} \dot{Q}_n(\hat{\boldsymbol{\beta}}_n|\mathbf{b}_0) &= \dot{Q}_n(\hat{\boldsymbol{\beta}}_n|\mathbf{b}_0)-\dot{Q}_n(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0) \\ &= \ddot{Q}_n(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0)(\hat{\boldsymbol{\beta}}_n-\hat{\boldsymbol{\beta}}_0)+O(\frac{1}{n}). \end{aligned}$$

Thus, conditional on $\hat{\mathbf{b}}$,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_0) = \ddot{Q}_n^{-1}(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0) \cdot \sqrt{n}\dot{Q}_n(\hat{\boldsymbol{\beta}}_n|\mathbf{b}_0) + o(1). \tag{3.11}$$

From (3.10) and (3.11), and because $\lim_{n\to\infty} \ddot{Q}_n(\hat{\boldsymbol{\beta}}_0|\mathbf{b}_0) = \lim_{n\to\infty} \ddot{Q}_n(\boldsymbol{\beta}_0|\mathbf{b}_0)$, we have:

$$\begin{aligned}
\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) &= \sqrt{n}(\hat{\boldsymbol{\beta}}_n - \hat{\boldsymbol{\beta}}_0) + \sqrt{n}(\hat{\boldsymbol{\beta}}_0 - \boldsymbol{\beta}_0) \\
&= \ddot{Q}_n^{-1}(\boldsymbol{\beta}_0|\mathbf{b}_0)\{\sqrt{n}\dot{Q}_n(\hat{\boldsymbol{\beta}}_n|\mathbf{b}_0) - \sqrt{n}\dot{Q}_n(\boldsymbol{\beta}_0|\mathbf{b}_0)\} + o(1).
\end{aligned}$$

From the central limit theorem and the consistency of $\hat{\boldsymbol{\beta}}_n$, we know that $\sqrt{n}\dot{Q}_n(\hat{\boldsymbol{\beta}}_n|\mathbf{b}_0)$ and $\sqrt{n}\dot{Q}_n(\boldsymbol{\beta}_0|\mathbf{b}_0)$ are asymptotically normal. Therefore

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to \mathrm{N}(\mathbf{0}, \Sigma),$$

where $\Sigma = \{\ddot{Q}_n^{-1}(\boldsymbol{\beta}_0|\mathbf{b}_0)\}\Omega\{\ddot{Q}_n^{-1}(\boldsymbol{\beta}_0|\mathbf{b}_0)\}'$ and $\Omega = \lim_{n\to\infty} \mathrm{Var}\{\sqrt{n}\dot{Q}_n(\hat{\boldsymbol{\beta}}_n|\mathbf{b}_0) - \sqrt{n}\dot{Q}_n(\boldsymbol{\beta}_0|\mathbf{b}_0)\}$.
□

3.8 Tables and Figures

Table 3.1: The absolute bias, standard error and coverage probability of the fixed-effect estimation from 200 replications for the count responses with an exchangeable correlation structure of parameter ρ.

			MEEE	WGEE	MI
$\rho = 0$	β_1	Abs. Bias	0.106	0.112	0.123
		Std. Error	0.130	0.135	0.148
		CP	0.975	0.930	0.945
	β_2	Abs. Bias	0.139	0.143	0.157
		Std. Error	0.168	0.176	0.191
		CP	0.965	0.930	0.920
	β_3	Abs. Bias	0.122	0.143	0.156
		Std. Error	0.155	0.179	0.198
		CP	0.985	0.950	0.955
	β_4	Abs. Bias	0.158	0.176	0.198
		Std. Error	0.201	0.230	0.249
		CP	0.960	0.935	0.915
$\rho = 0.4$	β_1	Abs. Bias	0.118	0.125	0.137
		Std. Error	0.152	0.166	0.179
		CP	0.980	0.935	0.955
	β_2	Abs. Bias	0.150	0.157	0.179
		Std. Error	0.192	0.203	0.233
		CP	0.965	0.930	0.950
	β_3	Abs. Bias	0.107	0.128	0.156
		Std. Error	0.139	0.163	0.196
		CP	0.955	0.920	0.945
	β_4	Abs. Bias	0.134	0.157	0.205
		Std. Error	0.171	0.202	0.260
		CP	0.955	0.930	0.945
$\rho = 0.7$	β_1	Abs. Bias	0.126	0.128	0.178
		Std. Error	0.157	0.162	0.222
		CP	0.980	0.920	0.975
	β_2	Abs. Bias	0.155	0.167	0.260
		Std. Error	0.205	0.218	0.326
		CP	0.950	0.900	0.935
	β_3	Abs. Bias	0.092	0.101	0.272
		Std. Error	0.117	0.130	0.562
		CP	0.945	0.915	0.990
	β_4	Abs. Bias	0.110	0.141	0.645
		Std. Error	0.138	0.177	1.131
		CP	0.960	0.910	0.985

MEEE: mixed-effects estimating equation; PQL: penalized quasi-likelihood; GHQ: adaptive Gaussian-Hermite quadrature; WGEE: weighted generalized estimating equation; MI: multiple imputation; Abs. Bias: absolute bias; Std. Error: standard error; CP: coverage probability. The PQL does not converge due to a small cluster size T, and the GHQ is not applicable since the dimension of parameters is greater than the number of data points.

Table 3.2: The absolute bias, standard error and coverage probability of the fixed-effect estimation from 200 replications for the binary responses with an AR-1 correlation structure of parameter ρ, where monotonized responses are used for the WGEE.

			MEEE	PQL	GHQ	WGEE	MI
$\rho = 0.2$	β_1	Abs. Bias	0.150	0.214	0.204	0.435	0.172
		Std. Error	0.194	0.272	0.253	0.621	0.173
		CP	0.970	0.933	0.952	0.896	0.935
	β_2	Abs. Bias	0.246	0.401	0.345	0.806	0.413
		Std. Error	0.301	0.483	0.390	1.127	0.225
		CP	0.985	0.867	0.959	0.891	0.715
	β_3	Abs. Bias	0.171	0.311	0.304	0.527	0.148
		Std. Error	0.211	0.341	0.316	0.696	0.181
		CP	0.955	0.860	0.932	0.891	0.980
	β_4	Abs. Bias	0.269	0.379	0.356	0.828	0.405
		Std. Error	0.333	0.471	0.419	1.164	0.230
		CP	0.975	0.927	0.959	0.896	0.790
$\rho = 0.6$	β_1	Abs. Bias	0.217	1.100	5.361	0.515	0.192
		Std. Error	0.280	1.148	7.913	0.658	0.243
		CP	0.935	0.805	0.727	0.873	0.948
	β_2	Abs. Bias	0.333	2.452	8.488	0.750	0.362
		Std. Error	0.424	2.186	11.953	1.050	0.328
		CP	0.960	0.605	0.695	0.923	0.907
	β_3	Abs. Bias	0.177	2.532	7.689	0.577	0.153
		Std. Error	0.234	2.024	10.605	0.737	0.191
		CP	0.910	0.305	0.609	0.845	0.979
	β_4	Abs. Bias	0.306	2.429	4.998	0.862	0.365
		Std. Error	0.402	2.044	6.361	1.246	0.280
		CP	0.975	0.558	0.781	0.901	0.845

MEEE: mixed-effects estimating equation; PQL: penalized quasi-likelihood; GHQ: adaptive Gaussian-Hermite quadrature; WGEE: weighted generalized estimating equation; MI: multiple imputation; Abs. Bias: absolute bias; Std. Error: standard error; CP: coverage probability.

Table 3.3: The estimates, standard errors and p-values of fixed effects on respondents' interest in following news about the presidential campaign

Predictor	Statistics	MEEE	PQL	GHQ	GEE	MI	MI*	WGEE*
Intercept	Estimate	-4.352	-5.370	-7.368	-3.181	-2.395	-1.966	-3.187
	Std. Error	0.210	0.236	0.336	0.152	0.149	0.569	0.211
	p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Time	Estimate	0.112	0.228	0.250	0.111	0.111	0.012	0.110
	Std. Error	0.006	0.007	0.010	0.005	0.005	0.196	0.007
	p-value	0.000	0.000	0.000	0.000	0.000	0.951	0.000
Age	Estimate	0.043	0.047	0.068	0.029	0.021	0.019	0.031
	Std. Error	0.002	0.003	0.004	0.002	0.002	0.003	0.002
	p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Education	Estimate	0.546	0.616	0.869	0.379	0.292	0.283	0.366
	Std. Error	0.036	0.047	0.064	0.029	0.028	0.036	0.041
	p-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Gender	Estimate	0.190	0.220	0.313	0.147	0.113	0.063	0.032
	Std. Error	0.063	0.086	0.114	0.053	0.055	0.065	0.075
	p-value	0.003	0.011	0.006	0.006	0.038	0.333	0.668
Household Income	Estimate	0.047	0.058	0.087	0.036	0.025	0.021	0.031
	Std. Error	0.009	0.012	0.015	0.007	0.008	0.008	0.010
	p-value	0.000	0.000	0.000	0.000	0.001	0.007	0.002
Marital Status	Estimate	-0.011	-0.028	-0.001	-0.018	-0.011	-0.034	-0.019
	Std. Error	0.067	0.093	0.123	0.057	0.056	0.062	0.081
	p-value	0.872	0.760	0.997	0.759	0.846	0.577	0.813
MSA Status	Estimate	0.150	0.133	0.199	0.076	0.051	0.030	0.072
	Std. Error	0.083	0.117	0.156	0.071	0.072	0.082	0.097
	p-value	0.069	0.256	0.200	0.284	0.478	0.713	0.457
Black, Non-Hispanic	Estimate	0.598	0.702	1.000	0.438	0.331	0.117	0.428
	Std. Error	0.137	0.167	0.218	0.101	0.106	0.156	0.134
	p-value	0.000	0.000	0.000	0.000	0.002	0.453	0.001
Other, Non-Hispanic	Estimate	-0.226	-0.261	-0.339	-0.130	-0.105	-0.093	-0.428
	Std. Error	0.124	0.176	0.234	0.112	0.102	0.129	0.160
	p-value	0.068	0.138	0.148	0.244	0.300	0.471	0.007
Hispanic	Estimate	0.053	0.078	0.014	0.053	0.016	-0.028	-0.124
	Std. Error	0.120	0.169	0.222	0.104	0.096	0.112	0.142
	p-value	0.657	0.645	0.950	0.613	0.867	0.802	0.383

*Monotonized responses are used, where all follow-ups are deleted once the first missing datum occurs. MEEE: mixed-effects estimating equation; PQL: penalized quasi-likelihood; GHQ: adaptive Gaussian-Hermite quadrature; GEE: generalized estimating equation; MI: multiple imputation; WGEE: weighted generalized estimating equation; Std: Error, standard error.

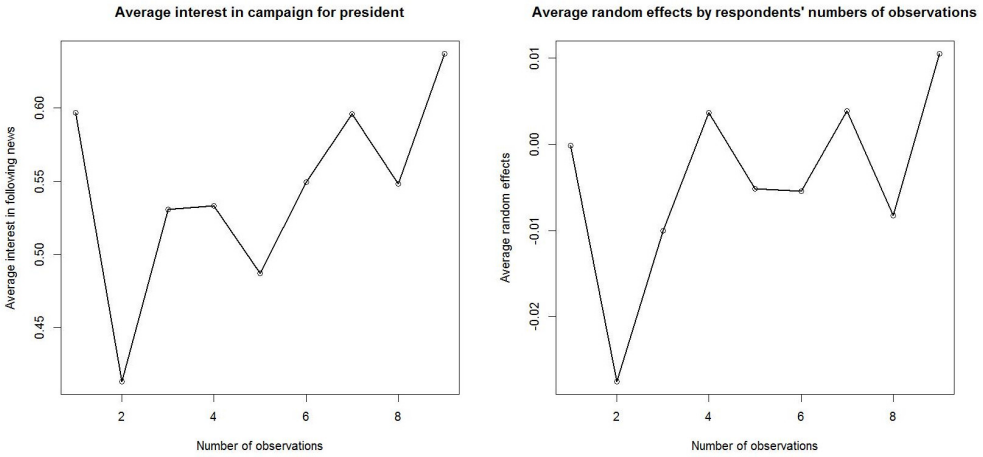

Figure 3.1: A comparison of the respondents' average interest in the presidential campaign and the average of the estimated random effects by MEEE, both plotted against respondents' number of observed occasions; the right panel is plotted using the same model but without time as a predictor.

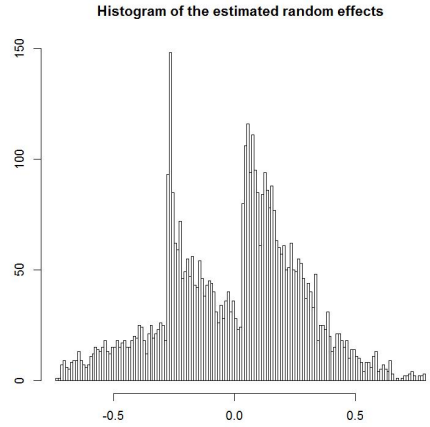

Figure 3.2: Histogram of the estimated random effects by MEEE.

Chapter 4

A Group-Specific Recommender System

4.1 Introduction

Recommender systems have drawn great attention since they can be applied to many areas, such as movies reviews, restaurant and hotel selection, financial services, and even identifying gene therapies. Therefore there is a great demand to develop efficient recommender systems which track users' preferences and recommend potential items of interest to users.

However, developing competitive recommender systems brings new challenges, as information from both users and items could grow exponentially, and the corresponding utility matrix representing users' preferences over items are sparse and high-dimensional. The standard methods and algorithms which are not scalable in practice may suffer from rapid deterioration on recommendation accuracy as the volume of data increases.

In addition, it is important to incorporate dynamic features of data instead of one-time usage only, as data could stream in over time and grow exponentially. For example, in the MovieLens 10M data, 96% of the most recent ratings are either from new users or on new items which did not exist before. This implies that the information collected at an early time may not be representative for future users and items. This phenomenon is also called the "cold-start" problem, where, in the testing set, majority responses are obtained from new users or for new items, and their preference information is not available from the training set. Another important feature of this type of data is that the missing mechanism is likely

nonignorable missing, where the missing mechanism is associated with unobserved responses. For instance, items with fewer and lower rating scores are less likely to attract other users. Existing recommender systems typically assume missing completely at random, which may lead to estimation bias.

Content-based filtering and collaborative filtering are two of the most prevalent approaches for recommender systems. Content-based filtering methods (e.g., Lang, 1995; Mooney and Roy, 2000; Blanco-Fernandez et al., 2008) recommend items by comparing the content of the items with a user's profile, which has the advantage that new items can be recommended upon release. However, domain knowledge is often required to establish a transparent profile for each user (Lops et al., 2011), which entails pre-processing tasks to formulate information vectors for items (Pazzani and Billsus, 2007). In addition, content-based filtering suffers from the "cold-start" problem as well when a new user is recruited (Adomavicius and Tuzhilin, 2005). is mainly compared with collaborative filtering methods.

For collaborative filtering, the key idea is to borrow information from similar users to predict their future actions. One significant advantage is that the domain knowledge for items is not required. Popular collaborative filtering approaches include, but are not limited to, singular value decomposition (SVD; Funk, 2006; Mazumder et al., 2010), restricted Boltzman machines (RBM; Salakhutdinov et al., 2007), and the nearest neighbor methods (kNN; Bell and Koren, 2007). It is well-known that an ensemble of these methods could further enhance prediction accuracy. (See Cacheda et al. (2011) and Feuerverger et al. (2012) for extensive reviews.)

However, most existing collaborative filtering approaches do not effectively solve the "cold-start" problem, although various attempts have been made. For example, Park et al. (2006) suggest adding artificial users or items with pre-defined characteristics, while Goldberg et al. (2001), Melville et al. (2002), and Nguyen et al. (2007) consider imputing "pseudo" ratings. Most recently, a hybrid system incorporating content-based auxiliary information has been proposed (e.g., Agarwal and Chen, 2009; Nguyen and Zhu, 2013; Zhu et al., 2016).

Nevertheless, the "cold-start" problem imposes great challenges, and has not been effectively solved.

In this chapter, we propose a group-specific singular value decomposition method that generalizes the SVD model by incorporating between-subject dependency and utilizes information of missingness. Specifically, we cluster users or items based on their missingness-related characteristics. We assume that individuals within the same cluster are correlated, while individuals from different clusters are independent. The cluster correlation is incorporated through mixed-effects modeling assuming that users or items from the same cluster share the same group effects, along with latent factors modeling using singular value decomposition.

The proposed method has two significant contributions. First, it solves the "cold-start" problem effectively through incorporating group effects. Most collaborative filtering methods rely on subject-specific parameters to predict users' and items' future ratings. However, for a new user or item, the training samples provide no information to estimate such parameters. In contrast, we are able to incorporate additional group information for new users and items to achieve higher prediction accuracy. Second, our clustering strategy takes nonignorable missingness into consideration. In the MovieLens data, we notice that individuals' rating behaviors are highly associated with their missing patterns: movies with higher average rating scores attract more viewers, while frequent viewers tend to be more critical and give low ratings. We cluster individuals into groups based on their non-random missingness, and this allows us to capture individuals' latent characteristics which are not utilized in other approaches.

To implement the proposed method, we propose a new algorithm that embeds a backfitting algorithm into alternating least squares, which avoids large matrices operation and big memory storage, and makes it feasible to achieve scalable computing in practice. Our numerical studies indicate that the proposed method is effective in terms of prediction accuracy. For example, for the MovieLens 1M and 10M data, the proposed method improves

prediction accuracy significantly compared to existing competitive recommender system approaches (e.g., Agarwal and Chen, 2009; Koren et al., 2009; Mazumder et al., 2010; Zhu et al., 2016).

This chapter is organized as follows. Section 4.2 provides the background of the singular value decomposition model and introduces the proposed method. Section 4.3 presents the proposed method, a new algorithm and its implementation. Section 4.4 establishes the theoretical foundation of the proposed method. In Section 4.5 we illustrate the performance and robustness of the proposed method through simulation studies. MovieLens 1M and 10M data are analyzed in Section 4.6. Section 4.7 provides concluding remarks and discussion. All technical details are provided in Section 4.8.

4.2 Background and Model Framework

4.2.1 Background

We provide the background of the singular value decomposition method (Funk, 2006) as follows. Let $\mathbf{R} = (r_{ui})_{n\times m}$ be the utility matrix, where n is the number of users, m is the number of items, and each r_{ui} is an explicit rating from user u for item i ($u = 1, \ldots, n$, $i = 1, \ldots, m$). The SVD method decomposes the utility matrix $\mathbf{R}$ as:

$$\mathbf{R} = \mathbf{P}\mathbf{Q}',$$

where $\mathbf{R}$ is assumed to be low-rank, $\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_n)'$ is an $n \times K$ user preference matrix, $\mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_m)'$ is an $m \times K$ item preference matrix, and K is the pre-specified upper bound of the number of latent factors, which corresponds to the rank of $\mathbf{R}$. Here $\mathbf{q}_i$ and $\mathbf{p}_u$ are K-dimensional latent factors associated with item i and user u, respectively, which explain variability in $\mathbf{R}$.

The predicted value of r_{ui} given by the SVD method is: $\hat{r}_{ui} = \hat{\mathbf{p}}_u'\hat{\mathbf{q}}_i$, where $\hat{\mathbf{q}}_i$ and $\hat{\mathbf{p}}_u$ are

estimated iteratively by:

$$\hat{\mathbf{q}}_i = \underset{q_i}{\operatorname{argmin}} \sum_{u \in U_i} (r_{ui} - \mathbf{p}_u' \mathbf{q}_i)^2 + \lambda \|\mathbf{q}_i\|_2^2,$$

and:

$$\hat{\mathbf{p}}_u = \underset{p_u}{\operatorname{argmin}} \sum_{i \in I_u} (r_{ui} - \mathbf{p}_u' \mathbf{q}_i)^2 + \lambda \|\mathbf{p}_u\|_2^2.$$

Here U_i denotes the set of all users who rate item i, and I_u is the set of all items rated by user u. Different penalty functions can be applied. For example, Zhu et al. (2016) suggest L_0 and L_1 penalties to achieve sparsity of $\mathbf{P}$ and $\mathbf{Q}$. In addition, some SVD methods (e.g., Koren, 2010; Mazumder et al., 2010; Nguyen and Zhu, 2013) are implemented on residuals after a baseline fit, such as linear regression or ANOVA, rather than the raw ratings r_{ui} directly.

The SVD method can be carried out through several algorithms, for example, the alternating least square (ALS; Carroll and Chang, 1970; Harshman, 1970; Koren et al., 2009), gradient descent approaches (Wu, 2007), and one-feature-at-a-time ALS (Funk, 2006).

4.2.2 Model Framework

The general framework of the proposed method is constructed as follows. Suppose $\mathbf{x}_{ui}$ is a covariate vector corresponding to the user u and item i. In the rest of this chapter, we consider $r_{ui} - \mathbf{x}_{ui}' \hat{\boldsymbol{\beta}}$ as the new response, where $\hat{\boldsymbol{\beta}}$ is the linear regression coefficient of $\mathbf{x}_{ui}$ to fit r_{ui}. To simplify our notation, we still use r_{ui} to denote the residual here. In case covariate information is not available, we apply the ANOVA-type model where the grand mean, the user main effects and the item main effects are subtracted and replace r_{ui} by its residual.

Let $\theta_{ui} = E(r_{ui})$. We generalize the SVD model and formulate each θ_{ui} as

$$\theta_{ui} = (\mathbf{p}_u + \mathbf{s}_{v_u})'(\mathbf{q}_i + \mathbf{t}_{j_i}), \tag{4.1}$$

where $\mathbf{s}_{v_u}$ and $\mathbf{t}_{j_i}$ are K-dimensional group effects that are identical across members from

the same cluster. We denote users from the v-th cluster as $V_v = \{u : v_u = v\}$ $(v = 1, \ldots, N)$, and items from the j-th cluster as $J_j = \{i : j_i = j\}$ $(j = 1, \ldots, M)$, where $\sum_{v=1}^{N} |V_v| = n$ and $\sum_{j=1}^{M} |J_j| = m$, $|\cdot|$ is the cardinality of a set, and N and M are the total number of clusters for users and items, respectively. Details about selecting N and M are provided in Section 4.3.3.

In matrix form, we use $\mathbf{S} = (\mathbf{s}_1, \ldots, \mathbf{s}_N)'$ and $\mathbf{T} = (\mathbf{t}_1, \ldots, \mathbf{t}_M)'$ to denote the user and item group-effect matrices, respectively. However, the dimensions of matrix $\mathbf{S}$ and $\mathbf{T}$ are $N \times K$ and $M \times K$, which are not compatible with the dimensions of $\mathbf{P}$ and $\mathbf{Q}$, respectively. Therefore, alternatively we define $\mathbf{S}_c = (\mathbf{s}_1\mathbf{1}'_{|V_1|}, \ldots, \mathbf{s}_N\mathbf{1}'_{|V_N|})'$ and $\mathbf{T}_c = (\mathbf{t}_1\mathbf{1}'_{|J_1|}, \ldots, \mathbf{t}_M\mathbf{1}'_{|J_M|})'$, corresponding to group-effects from users and items, where $\mathbf{1}_k$ is a k-dimensional vector of 1's, and the subscript "c" in $\mathbf{S}_c$ and $\mathbf{T}_c$ denotes the "complete" forms of matrices. Let $\mathbf{\Theta} = (\theta_{ui})_{n\times m}$, then we have

$$\mathbf{\Theta} = (\mathbf{P} + \mathbf{S}_c)(\mathbf{Q} + \mathbf{T}_c)',$$

and if there are no group effects, $\mathbf{\Theta}$ degenerates to $\mathbf{\Theta} = \mathbf{P}\mathbf{Q}'$, which is the same as the SVD model.

Here the users or items can be formed as clusters based on their similar characteristics. For example, we can use missingness-related information such as the number of ratings from each user and each item. Users or items within the same cluster are correlated with each other through the group effects $\mathbf{s}_{v_u}$ or $\mathbf{t}_{j_i}$, while observations from different clusters are assumed to be independent. In Section 4.3, Section 4.4 and Section 4.5.1, we assume N and M are known, and that members in each cluster are correctly labeled.

Remark 1. For easy operation, one could use users' and items' covariate information for clustering. In fact, (4.1) is still a generalization of the SVD method even if $N = M = 1$, because $\mathbf{s}'_{v_u}\mathbf{t}_{j_i}$, $\mathbf{p}'_u\mathbf{t}_{j_i}$, $\mathbf{s}'_{v_u}\mathbf{q}_i$ correspond to the grand mean, the user main effects and the item main effects, analogous to the ANOVA-type of SVD model. Note that covariate

information might not be collected from new users and new items. However, missingness-related information is typically available for clustering, and therefore $\mathbf{s}_{v_u}$ and $\mathbf{t}_{j_i}$ can be utilized for new users and new items. This is crucial to solve the "cold-start" problem.

4.3 The General Method

4.3.1 Parameter Estimation

In this subsection, we illustrate how to obtain estimations of model parameters through training data. In addition, we develop a new algorithm that embeds back-fitting (Breiman and Friedman, 1985) into alternating least squares. This enables us to circumvent large-scale matrix operations through a two-step iteration, and hence significantly improve computational speed and scalability.

Let $\boldsymbol{\gamma}$ be a vectorization of $(\mathbf{P}, \mathbf{Q}, \mathbf{S}, \mathbf{T})$, Ω be a set of user-item pairs associated with observed ratings, and $R^o = \{r_{ui} : (u,i) \in \Omega\}$ be a set of observed ratings. We define the loss function as

$$\mathcal{L}(\boldsymbol{\gamma}|R^o) = \sum_{(u,i)\in\Omega} (r_{ui} - \theta_{ui})^2 + \lambda(\sum_{u=1}^{n} \|\mathbf{p}_u\|_2^2 + \sum_{v=1}^{N} \|\mathbf{s}_v\|_2^2 + \sum_{i=1}^{m} \|\mathbf{q}_i\|_2^2 + \sum_{j=1}^{M} \|\mathbf{t}_j\|_2^2), \quad (4.2)$$

where θ_{ui} is given by (4.1) and λ is a tuning parameter. We can estimate $\boldsymbol{\gamma}$ via

$$\hat{\boldsymbol{\gamma}} = \underset{\gamma}{\operatorname{argmin}}\, \mathcal{L}(\boldsymbol{\gamma}|R^o).$$

Then the predicted value of θ_{ui} can be obtained by $\hat{\theta}_{ui} = (\hat{\mathbf{p}}_u + \hat{\mathbf{s}}_{v_u})'(\hat{\mathbf{q}}_i + \hat{\mathbf{t}}_{j_i})$.

The estimation procedure consists of updating $(\hat{\mathbf{p}}_u + \hat{\mathbf{s}}_{v_u})$ and $(\hat{\mathbf{q}}_i + \hat{\mathbf{t}}_{j_i})$ iteratively. Following the strategy of the alternating least squares, the latent factors and the group effects

associated with item cluster j are estimated by:

$$(\{\hat{\mathbf{q}}_i\}_{i\in J_j}, \hat{\mathbf{t}}_j) = \arg\min_{\{q_i\}_{i\in J_j}, t_j} \sum_{i\in J_j}\sum_{u\in U_i}(r_{ui}-\theta_{ui})^2 + \lambda(\sum_{i\in J_j}\|\mathbf{q}_i\|_2^2 + \|\mathbf{t}_j\|_2^2). \tag{4.3}$$

Similarly, we estimate latent factors and group effects associated with user cluster v:

$$(\{\hat{\mathbf{p}}_u\}_{u\in V_v}, \hat{\mathbf{s}}_v) = \arg\min_{\{p_u\}_{u\in V_v}, s_v} \sum_{u\in V_v}\sum_{i\in I_u}(r_{ui}-\theta_{ui})^2 + \lambda(\sum_{u\in V_v}\|\mathbf{p}_u\|_2^2 + \|\mathbf{s}_v\|_2^2). \tag{4.4}$$

However, directly solving (4.3) and (4.4) by the alternating least square encounters large matrices. In the MovieLens 10M data, it could involve matrices with more than 100,000 rows. We develop a new algorithm which embeds back-fitting into alternating least squares, and minimize each of (4.3) and (4.4) iteratively. Specifically, for each item cluster J_j ($j = 1,\ldots,M$), we fix $\mathbf{P}$ and $\mathbf{S}$, and minimize (4.3) through estimating $\hat{\mathbf{q}}_i$ and $\hat{\mathbf{t}}_j$ iteratively:

$$\hat{\mathbf{q}}_i = \operatorname*{argmin}_{q_i} \sum_{u\in U_i}(r_{ui}-\theta_{ui})^2 + \lambda\|\mathbf{q}_i\|_2^2, i\in J_j, \tag{4.5}$$

$$\hat{\mathbf{t}}_j = \operatorname*{argmin}_{t_j} \sum_{i\in J_j}\sum_{u\in U_i}(r_{ui}-\theta_{ui})^2 + \lambda\|\mathbf{t}_j\|_2^2. \tag{4.6}$$

For each user cluster V_v ($v = 1,\ldots,N$), we fix $\mathbf{Q}$ and $\mathbf{T}$ and minimize (4.4) through estimating $\hat{\mathbf{p}}_u$ and $\hat{\mathbf{s}}_v$ iteratively:

$$\hat{\mathbf{p}}_u = \operatorname*{argmin}_{p_u} \sum_{i\in I_u}(r_{ui}-\theta_{ui})^2 + \lambda\|\mathbf{p}_u\|_2^2, u\in V_v, \tag{4.7}$$

$$\hat{\mathbf{s}}_v = \operatorname*{argmin}_{s_v} \sum_{u\in V_v}\sum_{i\in I_u}(r_{ui}-\theta_{ui})^2 + \lambda\|\mathbf{s}_v\|_2^2. \tag{4.8}$$

The above backfitting is an iterative algorithm for additive models. In contrast, the alternating least squares is an iterative algorithm for multiplicative models. Although they are both blockwise coordinate descent methods under our framework, their convergence properties are different. Ansley and Kohn (1994) show that for penalized least square problems,

the backfitting algorithm converges to the unique optimum solution from any initial values, while the alternating least squares algorithm for two blocks only converges to a stationary point (Chen et al., 2012b).

In addition, the proposed algorithm is also different from the block-wise coordinate descent algorithm which estimates each of $(\mathbf{P}, \mathbf{Q}, \mathbf{S}, \mathbf{T})$ sequentially and iteratively while keeping the other terms as constants. The convergence property of the proposed algorithm is illustrated in Section 4.3.2. Note that the block-wise coordinate descent algorithm does not have such a property.

4.3.2 Algorithm

In this section, we provide the detailed algorithm as follows.

Algorithm 1: Parallel Computing for the Proposed Method

1. *(Initialization)* Set $l = 1$. Set initial values for $(\mathbf{P}^{(0)}, \mathbf{Q}^{(0)}, \mathbf{S}^{(0)}, \mathbf{T}^{(0)})$ and the tuning parameter λ.
2. *(Item Effects)* Estimate $\mathbf{Q}^{(l)}$ and $\mathbf{T}^{(l)}$ iteratively.
   (i) Set $\mathbf{Q}^{(l)} \leftarrow \mathbf{Q}^{(l-1)}$, and set $\mathbf{T}^{(l)} \leftarrow \mathbf{T}^{(l-1)}$.
   (ii) For each item $i = 1, \ldots, m$, calculate $\mathbf{q}_i^{(l)new}$ using (4.5).
   (iii) For each item cluster J_j, $j = 1, \ldots, M$, calculate $\mathbf{t}_j^{(l)new}$ based on (4.6) .
   (iv) Stop iteration if $\frac{1}{mK}\|\mathbf{Q}^{(l)new} - \mathbf{Q}^{(l)}\|_F^2 + \frac{1}{MK}\|\mathbf{T}^{(l)new} - \mathbf{T}^{(l)}\|_F^2 < 10^{-5}$, otherwise assign $\mathbf{Q}^{(l)} \leftarrow \mathbf{Q}^{(l)new}$ and $\mathbf{T}^{(l)} \leftarrow \mathbf{T}^{(l)new}$, and go to step 2(ii).
3. *(User Effects)* Estimate $\mathbf{P}^{(l)}$ and $\mathbf{S}^{(l)}$ iteratively.
   (i) Set $\mathbf{P}^{(l)} \leftarrow \mathbf{P}^{(l-1)}$, and set $\mathbf{S}^{(l)} \leftarrow \mathbf{S}^{(l-1)}$.
   (ii) For each user $u = 1, \ldots, n$, calculate $\mathbf{p}_u^{(l)new}$ using (4.7).
   (iii) For each user cluster V_v, $v = 1, \ldots, N$, calculate $\mathbf{s}_v^{(l)new}$ based on (4.8).
   (iv) Stop iteration if $\frac{1}{nK}\|\mathbf{P}^{(l)new} - \mathbf{P}^{(l)}\|_F^2 + \frac{1}{NK}\|\mathbf{S}^{(l)new} - \mathbf{S}^{(l)}\|_F^2 < 10^{-5}$, otherwise assign $\mathbf{P}^{(l)} \leftarrow \mathbf{P}^{(l)new}$ and $\mathbf{S}^{(l)} \leftarrow \mathbf{S}^{(l)new}$, and go to step 3(ii).
4. *(Stopping criterion)* Stop if $\frac{1}{nK}\|\mathbf{P}^{(l)} + \mathbf{S}_c^{(l)} - \mathbf{P}^{(l-1)} - \mathbf{S}_c^{(l-1)}\|_F^2 + \frac{1}{mK}\|\mathbf{Q}^{(l)} + \mathbf{T}_c^{(l)} - \mathbf{Q}^{(l-1)} - \mathbf{T}_c^{(l-1)}\|_F^2 < 10^{-3}$, otherwise set $l \leftarrow l + 1$ and go to step 2.

Note that the alternating least square is performed by conducting steps 2 and 3 iteratively, while the back-fitting algorithm is carried out within step 2 and step 3. The parallel computing can be implemented in steps 2(ii), (iii) and 3(ii), (iii).

Algorithm 1 does not require large computational and storage cost. We denote I_{B1}, I_{B2} and I_{ALS} as the numbers of iterations for back-fitting in steps 2 and 3, and the ALS, respectively, and C_{Ridge} as the computational complexity of solving the ridge regression with K variables and $\max\{|V_1|, \ldots, |V_N|, |J_1|, \ldots, |J_M|\}$ observations. Then the computational complexity of Algorithm 1 is no greater than $\{(m+M)I_{B1} + (n+N)I_{B2}\}C_{Ridge}I_{ALS}$. Since both ridge regression and Lasso have the same computational complexity as ordinary least squares (Efron et al., 2004), the computational cost of the proposed method is indeed no greater than that of Zhu et al. (2016). For the storage cost, Algorithm 1 requires storages of only item-specific or user-specific information to solve (4.5) or (4.7), and the sizes of items and users information not exceeding $\max\{|J_1|, \ldots, |J_M|\}$ and $\max\{|V_1|, \ldots, |V_N|\}$ to solve (4.6) or (4.8), respectively.

We also establish the convergence property of Algorithm 1 as follows. Let $\boldsymbol{\gamma}^* = \text{vec}(\mathbf{P}^*, \mathbf{Q}^*, \mathbf{S}^*, \mathbf{T}^*)$ be a stationary point of $\mathcal{L}(\boldsymbol{\gamma}|R^o)$ corresponding to two blocks. That is,

$$\text{vec}(\mathbf{P}^*, \mathbf{S}^*) = \underset{P,S}{\text{argmin}}\, \mathcal{L}\left(\text{vec}(\mathbf{P}, \mathbf{Q}^*, \mathbf{S}, \mathbf{T}^*)|R^o\right),$$

and

$$\text{vec}(\mathbf{Q}^*, \mathbf{T}^*) = \underset{Q,T}{\text{argmin}}\, \mathcal{L}\left(\text{vec}(\mathbf{P}^*, \mathbf{Q}, \mathbf{S}^*, \mathbf{T})|R^o\right).$$

The following lemma shows the convergence of Algorithm 1 to a stationary point, which is a local minimum along each block direction. One way to achieve the global minimum is to adopt the branch-and-bound technique, and search all possible local minima (Liu et al., 2005). However, this technique could be computationally intensive.

Lemma 3. *The estimate* $\hat{\boldsymbol{\gamma}} = \text{vec}(\hat{\mathbf{P}}, \hat{\mathbf{Q}}, \hat{\mathbf{S}}, \hat{\mathbf{T}})$ *from Algorithm 1 is a stationary point of the loss function* $\mathcal{L}(\boldsymbol{\gamma}|R^o)$ *in (4.2).*

4.3.3 Implementation

In this subsection we address some implementation issues for the proposed method. Our algorithm is implemented in the R environment, which requires packages "foreach" and "doParallel" for parallel computing and "bigmemory" and "bigalgebra" for big matrix storage and operation. All the reported numerical studies are implemented using the Linux system on cluster computers. We can further enhance computation speed through C++ programming with OpenMP.

To select tuning parameter λ, we search from grid points which minimizes the root mean square error (RMSE) on the validation set. The RMSE on a given set Ω_0 is defined as $\left\{\frac{1}{|\Omega_0|}\sum_{(u,i)\in\Omega_0}(r_{ui}-\hat{\theta}_{ui})^2\right\}^{1/2}$. In selection of the number of latent factors K, we choose K such that it is sufficiently large and leads to stable estimations. In general, K needs to be larger than the rank of the utility matrix $\mathbf{R}$, but not so large as to intensify the computation. Regarding the selection of the number of clusters N and M, Corollary 4 of Section 4.4 provides the lower bound in the order of $O(N)$ and $O(M)$. Note that too small N and M may not have the power to distinguish between the proposed method and the SVD method. In practice, if clustering is based on categorical variables, then we can apply the existing categories, and N and M are known. However, if clustering is based on a continuous variable, we can apply the quantiles of the continuous variable to determine N and M and categorize users and items evenly. We then select the number of clusters through a grid search, similar to the selection of λ and K. See Wang (2010) for a consistent selection of the number of clusters in more general settings.

In particular, for our numerical studies, we split our dataset into 60% training, 15% validation and 25% testing sets based on the time of ratings (timestamps; Zhu et al., 2016). That is, we use historical data to predict future data. If time information is not available, we use a random split to determine training, validation and testing sets instead.

4.4 Theory

In this section, we provide the theoretical foundation of the proposed method in a general setting. That is, we allow r_{ui} to follow a general class of distributions. In particular, we derive an upper bound for the prediction error in probability, and show that existing approaches without utilizing group effects lead to a larger value of the loss function, and therefore are less efficient compared to the proposed method. Furthermore, we establish a lower bound of the number of clusters which guarantees that the group effects can be detected effectively.

Suppose the expected value of each rating is formulated via a known mean function μ. That is,

$$E(r_{ui}) = \mu(\theta_{ui}),$$

and θ_{ui} is defined as in (4.1). For example, if r_{ui} is a continuous variable, then $\mu(\theta_{ui}) = \theta_{ui}$; and if r_{ui}'s are binary, then $\mu(\theta_{ui}) = \frac{\exp(\theta_{ui})}{1+\exp(\theta_{ui})}$.

We let $f_{ui} = f(r_{ui}|\theta_{ui})$ be the probability density function of r_{ui}. Since each r_{ui} is associated with $\boldsymbol{\gamma}$ only through θ_{ui}, we denote $f_{ui}(r, \boldsymbol{\gamma}) = f(r_{ui}|\theta_{ui})$. We define the likelihood-based loss function as:

$$\mathcal{L}(\boldsymbol{\gamma}|R^o) = -\sum_{(u,i)\in\Omega} \log f_{ui} + \lambda_{|\Omega|} D(\boldsymbol{\gamma}),$$

where $\lambda_{|\Omega|}$ is the penalization coefficient, $|\Omega|$ is the total number of observed ratings, and $D(\cdot)$ is a non-negative penalty function of $\boldsymbol{\gamma}$. For example, we have $D(\boldsymbol{\gamma}) = \|\boldsymbol{\gamma}\|_2^2$ for the L_2-penalty.

Since, in practice, the ratings are typically non-negative finite values, it is sensible to assume $\|\boldsymbol{\gamma}\|_\infty \leq L$, where L is a positive constant. We define the parameter vector space as

$$\mathcal{S}(k) = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma}\|_\infty \leq L, D(\boldsymbol{\gamma}) \leq k^2\}.$$

Notice that the dimension of $\boldsymbol{\gamma}$ is $\dim(\boldsymbol{\gamma}) = (n+m+N+M)K$ which goes to infinity as either n or m increases. Therefore, we assume $k \sim O(\sqrt{(n+m+N+M)K})$. Similarly, we define the parameter space for each θ_{ui}: $\mathcal{S}_\Theta(k) = \{\theta : \|\boldsymbol{\gamma}\|_\infty \leq L, D(\boldsymbol{\gamma}) \leq k^2\}$.

Assumption 1. *For some constant $\bar{G} \geq 0$, and $\theta_{ui}, \tilde{\theta}_{ui} \in \mathcal{S}_\Theta(k)$,*

$$\left| f^{1/2}(r_{ui}|\theta_{ui}) - f^{1/2}(r_{ui}|\tilde{\theta}_{ui}) \right| \leq G(r_{ui}) \|\theta_{ui} - \tilde{\theta}_{ui}\|_2,$$

where $EG^2(r_{ui}) \leq \bar{G}^2$ for $u = 1, \ldots, n$, $i = 1, \ldots, m$.

The Hellinger metric $h_\Theta(\cdot, \cdot)$ on $\mathcal{S}_\Theta(k)$ is defined as:

$$h_\Theta(\theta_{ui}, \tilde{\theta}_{ui}) = \left[\int \{ f^{1/2}(r_{ui}|\theta_{ui}) - f^{1/2}(r_{ui}|\tilde{\theta}_{ui}) \}^2 d\nu(r_{ui}) \right]^{1/2},$$

where $\nu(\cdot)$ is a probability measure. Based on Assumption 1, $h_\Theta(\theta_{ui}, \tilde{\theta}_{ui})$ is bounded by $\|\theta_{ui} - \tilde{\theta}_{ui}\|_2$.

We now define the Hellinger metric $h_\mathcal{S}(\cdot, \cdot)$ on $\mathcal{S}(k)$. For $\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}} \in \mathcal{S}(k)$, let

$$h_\mathcal{S}(\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}) = \left\{ \frac{1}{nm} \sum_{i=1}^{m} \sum_{u=1}^{n} h_\Theta^2(\theta_{ui}, \tilde{\theta}_{ui}) \right\}^{1/2}.$$

It is straightforward to show that $h_\mathcal{S}$ is still a metric. In the rest of this chapter, we suppress the subscript and use $h(\cdot, \cdot)$ to denote the Hellinger metric on $\mathcal{S}(k)$. In the following, we show that $h(\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}})$ can be bounded by $\|\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}\|_2$.

Lemma 4. *Under Assumption 1, there exists a constant $d_0 \geq 0$, such that for $\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}} \in \mathcal{S}(k)$,*

$$h(\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}) \leq d_0 \sqrt{\frac{n+m}{nm}} \|\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}\|_2.$$

Suppose $\hat{\boldsymbol{\gamma}} = \arg\min_{\gamma \in \mathcal{S}(k)} \mathcal{L}(\boldsymbol{\gamma}|R^o)$ is a penalized maximum likelihood estimator of $\boldsymbol{\gamma}$. Theorem 4 indicates that $\hat{\boldsymbol{\gamma}}$ converges to $\boldsymbol{\gamma}$ exponentially in probability, with a convergence rate of $\epsilon_{|\Omega|}$.

Theorem 4. *Under Assumption 1 and suppose* $\lambda_{|\Omega|} < \frac{1}{2k}\epsilon_{|\Omega|}^2$*, the best possible convergence rate of* $\hat{\boldsymbol{\gamma}}$ *is*

$$\epsilon_{|\Omega|} \sim \frac{\sqrt{(n+m)K}}{|\Omega|^{1/2}} \left\{ \log\left(\frac{|\Omega|}{\sqrt{nmK}} \right) \right\}^{1/2},$$

and there exists a constant $c > 0$*, such that*

$$P\left(h(\hat{\boldsymbol{\gamma}}, \boldsymbol{\gamma}) \geq \epsilon_{|\Omega|} \right) \leq 7 \exp(-c|\Omega|\epsilon_{|\Omega|}^2).$$

Remark 2. Theorem 4 can be generalized to achieve the convergence property measured by the L_2 distance as a special case of Corollary 2 in Shen (1998). However, the convergence under the L_2 distance is more restrictive than the convergence under the Hellinger distance. We adopt the Hellinger distance because of the following advantages. First, the convergence rate of $\hat{\boldsymbol{\gamma}}$ depends only on the size of the parameter space $\mathcal{S}(k)$ and the penalization coefficient $\lambda_{|\Omega|}$ (Shen, 1998). In contrast, the convergence rate based on the L_2 distance depends on additional local and global behavior of $\mathrm{Var}\{\mathcal{L}(\hat{\boldsymbol{\gamma}}|R^o) - \mathcal{L}(\boldsymbol{\gamma}|R^o)\}$. Second, the exponential bound under the Hellinger distance does not rely on the existence of the moment generating function of $G(\cdot)$, which is needed for the exponential bound under the L_2 distance.

Remark 3. Theorem 4 is quite general in terms of the rates of n and m. If we assume $O(n) = O(m) = O(n+m)$ such as in the MovieLens data, then $\epsilon_{|\Omega|}$ converges faster than $\epsilon_{|\Omega|}^{SAJ}$, where $\epsilon_{|\Omega|}^{SAJ} \sim \frac{\sqrt{(n+m)K}}{|\Omega|^{1/2}} \left\{ \log\left(\frac{|\Omega|}{(n+m)k} \right) \right\}^{1/2} \left\{ \log\left(\frac{m}{k} \right) \right\}^{1/2}$ is the convergence rate provided by the collaborative prediction method with binary ratings (Srebro et al., 2005). The exact rate comparison is not available here.

Remark 4. The definition of $\mathcal{S}(k)$ is for the purpose of achieving the best possible convergence rate. Specifically, let $\mathcal{S} \in \mathbb{R}^{(n+m+N+M)K}$ be the true underlying parameter space. Since $\mathcal{S}$ is in an infinite dimensional space when n or m goes to infinity, $\hat{\boldsymbol{\gamma}}$ obtained by optimizing over $\mathcal{S}$ may not achieve the best possible convergence rate (Shen and Wong, 1994). Instead, we adopt the idea of *sieve* MLE (Grenander, 1981), and approximate $\mathcal{S}$ by $\mathcal{S}(k)$ which grows as the sample size increases. This ensures that the penalized MLE $\hat{\boldsymbol{\gamma}}$ on $\mathcal{S}(k)$ is capable of

achieving the best possible convergence rate (Shen, 1998).

Remark 5. If we impose $\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}\|_2 \leq d_{n,m}$ with radius $d_{n,m} = \sqrt{\frac{2nm}{d_0^2(n+m)}}\epsilon_{|\Omega|}$, then the entropy of $\mathcal{S}(k)$ under Assumption 1 also satisfies the condition of local entropy (Wong and Shen, 1995). That is,

$$\mathcal{S}(k) = \mathcal{S}(k) \cap \left\{ \frac{1}{nm} \sum_{i=1}^{m} \sum_{u=1}^{n} \|f^{1/2}(r_{ui}, \boldsymbol{\gamma}) - f^{1/2}(r_{ui}, \hat{\boldsymbol{\gamma}})\|_2^2 \leq 2s^2 \right\}, \text{ for all } s \geq \epsilon_{|\Omega|}.$$

Consequently, the convergence rate of $\epsilon_{|\Omega|}$ is $\log(|\Omega|)$ times faster than the convergence rate calculated by using global entropy.

We now assume that the density function f_{ui} is a member of the exponential family in its canonical form. That is,

$$f(r_{ui}|\theta_{ui}) = H(r_{ui}) \exp\{\theta_{ui} T(r_{ui}) - A(\theta_{ui})\}.$$

In fact, the following results still hold if f is in the over-dispersed exponential family.

Suppose $\boldsymbol{\gamma} \in \mathcal{S}(k)$ and $\theta_{ui} \in \mathcal{S}_\Theta(k)$ are the true parameters. Then Theorem 5 indicates that if misspecified $\tilde{\theta}_{ui}$'s are not close to θ_{ui}'s, then the loss function of the corresponding $\tilde{\boldsymbol{\gamma}}$ cannot be closer to the loss function of $\boldsymbol{\gamma}$ than a given threshold in probability.

Theorem 5. *Under Assumption 1 and* $\lambda_{|\Omega|} < \frac{1}{2k}\epsilon_{|\Omega|}^2$, *there exist* $c_i > 0$, $i = 1, 2$, *such that for* $\epsilon_{|\Omega|} > 0$, *there exists* $\delta_{|\Omega|} > 0$, *and* $\min_{1 \leq u \leq n, 1 \leq i \leq m} |\tilde{\theta}_{ui} - \theta_{ui}| > \delta_{|\Omega|}$ *implies that*

$$P^*\left(\frac{1}{|\Omega|} \{\mathcal{L}(\tilde{\boldsymbol{\gamma}}|R^o) - \mathcal{L}(\boldsymbol{\gamma}|R^o)\} > c_1 \epsilon_{|\Omega|}^2 \right) \geq 1 - 7\exp(-c_2|\Omega|\epsilon_{|\Omega|}^2),$$

where P^* *denotes the outer measure (Pollard, 2012).*

Remark 6. Theorem 4 and Theorem 5 still hold if the loss function $\mathcal{L}(\cdot|\cdot)$ is not likelihood-based, but is a general criterion function. For such $\mathcal{L}(\cdot|\cdot)$, we can replace $h(\cdot,\cdot)$ by $\rho(\cdot,\cdot) = K^{1/2}(\cdot,\cdot)$ as the new measure of convergence, where $K(\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}) = E\{\mathcal{L}(\boldsymbol{\gamma}|\mathbf{R}) - \mathcal{L}(\tilde{\boldsymbol{\gamma}}|\mathbf{R})\}$. Note

that $K(\cdot,\cdot)$ is the Kullback-Leiber information if $\mathcal{L}(\cdot|\cdot)$ is a log-likelihood, which dominates the Hellinger distance $h(\cdot,\cdot)$, and hence the convergence is stronger under $K(\cdot,\cdot)$. See Shen (1998) for more details about regularity conditions for a more general criterion function.

Suppose $\boldsymbol{\gamma}_0 \in \mathcal{S}(k)$ is a vectorization of $(\mathbf{P}, \mathbf{Q}, \mathbf{0}, \mathbf{0})$, which corresponds to models with no group effects. The following Corollary 3 shows that if the true group effects are not close to 0, then existing methods ignoring group effects such as the SVD model ($\theta^0_{ui} = \mathbf{p}'_u\mathbf{q}_i$) lead to a larger loss in probability than the proposed method.

Corollary 3. *Under Assumption 1 and* $\lambda_{|\Omega|} < \frac{1}{2k}\epsilon^2_{|\Omega|}$, *there exists* $c_i > 0$, $i = 1, 2$, *and a constant* $\phi \in (0, 1]$, *such that for* $\frac{1}{\sqrt{\phi}}\epsilon_{|\Omega|} > 0$, *there exists* $\delta_{|\Omega|} > 0$. *Assume that at least* (ϕnm) *pairs of* (u, i) *satisfy* $|\theta^0_{ui} - \theta_{ui}| > \delta_{|\Omega|}$. *Then*

$$P^*\left(\frac{1}{|\Omega|}\{\mathcal{L}(\boldsymbol{\gamma}_0|R^o) - \mathcal{L}(\boldsymbol{\gamma}|R^o)\} > c_1\epsilon^2_{|\Omega|}\right) \geq 1 - 7\exp(-c_2|\Omega|\epsilon^2_{|\Omega|}).$$

The following corollary provides the minimal rate of N and M, in terms of n, m, K and $|\Omega|$. This implies that the number of clusters should be sufficiently large so that the group effects can be detected.

Corollary 4. *Under assumptions in Theorem 4, the rate of* N *and* M *satisfies*

$$O(N + M) \succeq \frac{nm}{|\Omega|}\log\left(\frac{|\Omega|}{\sqrt{nmK}}\right).$$

If we further assume that the number of ratings is proportional to the size of the utility matrix, that is, $O(|\Omega|) = O(nm)$, then $O(N + M) \succeq \log(\frac{|\Omega|^{1/2}}{K^{1/2}})$. The lower bound of $O(N + M)$ is useful in determining the minimal number of clusters. For example, for the MovieLens 10M data where $|\Omega| = 10,000,000$, we have the lower bound $\log(\frac{|\Omega|^{1/2}}{K^{1/2}}) \approx 7$ if $K \leq 10$.

4.5 Simulation Studies

In this section, we provide simulation studies to investigate the numerical performance of the proposed method in finite samples. Specifically, we compare the proposed method with four matrix factorization methods in Section 4.5.1 under a dynamic setting where new users and new items appear at later times. In Section 4.5.2, we test the robustness of the proposed model under various degrees of cluster misspecification.

4.5.1 Comparison under the "Cold-Start" Problem

In this simulation studies, we simulate the "cold-start" problem where new users' and new items' information is not available in the training set. In addition, we simulate that users' behavior is affected by other users' behavior, and therefore the missingness is not missing completely at random. Here users and items from the same group are generated to be dependent from each other.

We set $n = 650$ and $m = 660$ and generate $\mathbf{p}_u, \mathbf{q}_i \overset{iid}{\sim} \mathrm{N}(0, \mathbf{I}_K)$ for $u = 1, \ldots, n$, $i = 1, \ldots, m$, where $\mathbf{I}_K$ is a K-dimensional identity matrix with $K = 3$ or 6. To simulate group effects, we let $\mathbf{s}_v = (-3.5 + 0.5v)\mathbf{1}_K, v = 1, \ldots, N$, and $\mathbf{t}_j = (-3.6 + 0.6j)\mathbf{1}_K, j = 1, \ldots, M$, where $N = 13$, $M = 11$. We set cluster size $|V_1| = \cdots = |V_N| = 50$, and $|J_1| = \cdots = |J_M| = 60$. Without loss of generality, we assume that covariate information is not available for this simulation.

In contrast to other simulation studies, we do not generate the entire utility matrix $\mathbf{R}$. Instead, we mimic the real data case where only a small percentage of ratings is collected. We choose the total number of ratings to be $|\Omega| = (1 - \bar{\pi})nm$, where $\bar{\pi} = 0.7, 0.8, 0.9$ or 0.95 is the missing rate. The following procedure is used to generate these ratings.

We first select the l-th user-item pair (u_l, i_l), where $l = 1, \ldots, |\Omega|$ indicates the sequence of ratings from the earliest to the latest. If item i_l's current average rating is greater than 0.5, then for user u_l, we assign a rating $r_{u_l i_l}$ with probability 0.85; otherwise we assign $r_{u_l i_l}$ with

probability 0.2. The rating $r_{u_l i_l}$ is generated by $(\mathbf{p}_{u_l}+\mathbf{s}_{v_{u_l}})'(\mathbf{q}_{i_l}+\mathbf{t}_{j_{i_l}})/3+\varepsilon$, where $\varepsilon \overset{iid}{\sim} \mathrm{N}(0,1)$. That is, we simulate a setting where users tend to rate highly-rated items. Here u_l and i_l are sampled from $1,\ldots,n$ and $1,\ldots,m$ independently, but with weights proportional to the density of normal distributions $\mathrm{N}(nl/|\Omega|,(0.2n)^2)$ and $\mathrm{N}(ml/|\Omega|,(0.2m)^2)$, respectively. That is, ratings appearing at a later time are more likely corresponding to newer users or to newer items. If we fail to assign $r_{u_l i_l}$ a value, we re-draw (u_l, i_l) and restart this procedure. The selection of $r_{u_l i_l}$ is based on observed information, so the missing mechanism is missing at random (Rubin, 1976).

We compare the performance of the proposed method with four competitive matrix factorization models, namely, the regularized singular value decomposition method solved by the alternating least square algorithm (RSVD; Funk, 2006; Koren et al., 2009), a regression-based latent factor model (Agarwal and Chen, 2009), a nuclear-norm matrix completion method (Soft-Impute; Mazumder et al., 2010), and a latent factor model with sparsity pursuit (Zhu et al., 2016). For the last three methods, we apply the codes in `https://github.com/beechung/latent-factor-models`, the R package "softImpute", and that of Zhu et al. (2016), respectively.

For the proposed method, we apply the loss function (4.2). The tuning parameter λ for the proposed method and the RSVD is selected from grid points ranging from 1 to 29 to minimize the RMSEs on the validation set. For Agarwal and Chen (2009), we use the default of 10 iterations, while for Mazumder et al. (2010), the default $\lambda = 0$ is chosen to achieve convergence for the local minimum; and for Zhu et al. (2016), the tuning parameter selection is integrated in their programming coding. We generate simulation settings when the number of latent factors $K = 3$ and 6, and the missing rate $\bar{\pi} = 0.7, 0.8, 0.9, 0.95$. The means and standard errors of RMSEs on the testing set are reported in Table 4.1. The simulation results are based on 500 replications.

Table 4.1 indicates that the proposed method performs the best across all settings. Overall, the proposed method is relatively robust against different missing rates or different num-

bers of latent factors, and has the smallest standard error in most settings. In the most extreme case with $K = 6$ and $\bar{\pi} = 0.95$, the proposed method is still more than 100% better than the best of the four existing methods in terms of the RMSEs. The RSVD method performs well when both $\bar{\pi}$ and K are small, but performs poorly when either $\bar{\pi}$ or K increases. By contrast, Agarwal and Chen (2009), Mazumder et al. (2010) and Zhu et al. (2016) are able to provide small standard errors when $K = 6$ and $\bar{\pi} = 0.95$, but have large RMSEs across all settings. Mazumder et al. (2010) occasionally provides outlying results due to a convergence problem when $\bar{\pi}$ is 0.9 or 0.95. We remove these extreme results in our simulations.

4.5.2 Robustness against Cluster Misspecification

In this simulation study, we test the robustness of the proposed method when the clusters are misspecified.

We follow the same data-generating process as in the previous study, but allow the cluster assignment to be misspecified. Specifically, we misassign users and items to adjacent clusters with 10%, 30% and 50% chance. Here adjacent clusters are defined as the clusters with the closest group effects. This definition of adjacent clusters reflects the real data situation. For example, a horror movie might be misclassified as a thriller movie, but less likely a romantic movie.

The simulation results based on 500 replications are summarized in Table 4.2. In general, the proposed method is robust against the misspecification of clusters. In comparison with the previous results from Table 4.1, the proposed method performs better than the other four methods in all settings even when 50% of the cluster members are misclassified. On the other hand, the misspecification rate affects the performance of the proposed method to different degrees for various settings of $\bar{\pi}$ and K. For example, the proposed method below the 50% misspecification rate is 2.7% worse than the proposed method when there is no misspecification, in terms of the RMSE under $K = 3$ and $\bar{\pi} = 0.7$; and becomes 18.8% worse

than the one with no misspecification under $K = 6$ and $\bar{\pi} = 0.95$.

4.6 MovieLens Data

We apply the proposed method to MovieLens 1M and 10M data. The two datasets are collected by GroupLens Research and are available at `http://grouplens.org/datasets/movielens`. The MovieLens 1M data contains 1,000,209 ratings of 3,883 movies by 6,040 users, and rating scores range from 1 to 5. In addition, the 1M dataset provides demographic information for the users (age, gender, occupation, zipcode), and genres and release dates of the movies. In the MovieLens 10M data, we have 10,000,054 ratings collected from 71,567 users over 10,681 items, and 99% of the movie ratings are actually missing. Rating scores range from $0.5, 1, \ldots, 5$, but no user information is available.

Figure 4.1 illustrates the missing pattern of MovieLens 1M data. Both graphs indicate that the missing mechanism is possibly missing not at random. In the left figure, the right-skewed distribution from users indicates that only a few users rated a large number of movies. While the median number of ratings is 96, the maximum can reach up to 2,314. The right figure shows that popular movies attract more viewers. That is, the number of ratings for each movie is positively associated with its average rating score, indicating nonignorable missingness.

For the proposed method, we take advantage of missingness information from each user and item for clustering. We observe that users who give a large number of ratings tend to assign low rating scores; therefore we classify users based on the quantiles of the number of their ratings. For items, we notice that old movies being rated are usually classical and have high average rating scores. Therefore, the items are clustered based on their release dates. We use $N = 12$ and $M = 10$ as the number of clusters for users and items in both data sets. The means of ratings from different clusters are significantly different based on their pairwise two-sample t-tests. In addition, we also try a large range of N's and M's, but they

do not affect the results very much.

The proposed method is compared with the four matrix factorization methods described in Section 4.5.1. Tuning parameters for each method are selected from grid points to minimize the RMSEs on the validation set. For the proposed method, we apply the loss function (4.2) and select $K = 2$ and $\lambda = 12$ for the 1M data, and $K = 6$ and $\lambda = 16$ for the 10M data. For Agarwal and Chen (2009), we select $K = 1$ for both the 1M and 10M data, which requires 25 and 10 iterations of the EM algorithm to guarantee convergence, respectively. For Mazumder et al. (2010), $K = 4$ and $K = 9$ are selected for the 1M and 10M data, and while using different λ's does not influence the RMSE very much, we apply $\lambda = 0$ to estimate the theoretical local minimum. For Zhu et al. (2016), the tuning and the selection of K are provided in their coding automatically, and the L_0-penalty function is applied. For the RSVD, $K = 4$ and $\lambda = 7.5$ are selected for the 1M data, and $K = 4$ and $\lambda = 6$ are selected for the 10M data. In addition, we also compare the proposed method with the "grand mean imputation" approach, which predicts each rating by the mean of the training set and the validation set, and the "linear regression" approach using ratings from the training and the validation sets against all available covariates from users and items.

Table 4.3 provides the prediction results on the testing set, which indicates that the proposed method outperforms the other methods quite significantly. For example, for the 1M data, the RMSE of the proposed method is 8.7% less than the RSVD, 19.5% less than Agarwal and Chen (2009), 10.3% less than Mazumder et al. (2010), 9.4% less than Zhu et al. (2016), and 13.2% and 11.6% less than grand mean imputation and linear regression, respectively. For the 10M data, the proposed method improves on grand mean imputation, linear regression, the RSVD, Agarwal and Chen (2009), Mazumder et al. (2010) and Zhu et al. (2016) by 8.7%, 7.1%, 6.7%, 4.5%, 8.7% and 8.0% in terms of the RMSE, respectively. In addition, while some of the matrix factorization methods are worse than the linear regression method, the proposed method always beats the linear regression method.

The numerical studies are run on Dell C8220 computing sleds each with two 10-core Intel

Xeon E5-2670V2 processors and 64GB RAM. The proposed method uses 27 minutes for 1M data ($K = 2$ and $\lambda = 12$), and 10.9 hours for 10M data ($K = 6$ and $\lambda = 16$). The RSVD uses 6.4 minutes for 1M data ($K = 4$ and $\lambda = 7.5$), and 7.1 hours for 10M data ($K = 4$ and $\lambda = 6$). The Agarwal and Chen (2009) method requires 18.1 minutes for 1M data ($K = 1$ with 25 iterations), and 1.1 hours for 10M data ($K = 1$ with 10 iterations), while Mazumder et al. (2010) method uses 20.8 seconds for 1M data ($K = 4$), and 11.6 minutes for 10M data ($K = 9$), and Zhu et al. (2016) uses 1.1 minutes for 1M data, and 18.5 minutes for 10M data. The proposed method requires 5-10 more iterations to converge than its counterpart which does not incorporate group effects. As we discussed in Section 4.3.2, the proposed method has the same computational complexity as the Zhu et al. (2016) method, and is expected to be significantly faster if it is programmed in C and implemented through OpenMP.

We also investigate the "cold-start" problem in the MovieLens 10M data, where 96% of the ratings in the testing set are either from new users or on new items which are not available in the training set. We name these ratings "new ratings", in contrast to the "old ratings" given by existing users to existing items. In Table 4.4, we compare the proposed method with the four competitive methods on the "old ratings", the "new ratings", and the entire testing set. On the one hand, the RSVD, Mazumder et al. (2010), Zhu et al. (2016) and the proposed method have similar RMSE for the "old ratings" set, indicating similar performances on prediction accuracy for existing users and items. On the other hand, the proposed method has the smallest RMSE compared to the other methods for the "new ratings" and the entire testing sets, indicating the superior performance of the proposed method for the "cold-start" problem.

4.7 Discussion

We propose a new recommender system which improves prediction accuracy through incorporating dependency among users and items, in addition to utilizing information from the

non-random missingness.

In most collaborative filtering methods, training data may not have sufficient information to estimate subject-specific parameters for new users and items. Therefore, only baseline models such as ANOVA or linear regression are applied. For example, for a new user u, $\hat{\mathbf{p}}_u = \mathbf{0}$, and a method without specifying the group effects has $\hat{\theta}_{ui} = \mathbf{x}'_{ui}\hat{\boldsymbol{\beta}}$. In contrast, the proposed method provides a prediction through $\hat{\theta}_{ui} = \mathbf{x}'_{ui}\hat{\boldsymbol{\beta}} + \hat{\mathbf{s}}'_{v_u}(\hat{\mathbf{q}}_i + \hat{\mathbf{t}}_{j_i})$. The interaction term $\hat{\mathbf{s}}'_{v_u}\hat{\mathbf{q}}_i$ provides the average rating of the v_u-th cluster on the i-th item, which guarantees that $\hat{\theta}_{ui}$ is item-specific. The same property also holds for new items. The group effects $\mathbf{s}_{v_u}$ and $\mathbf{t}_{j_i}$ allow us to borrow information from existing users and items, and provide more accurate recommendations to new subjects.

The proposed model also takes advantage of missingness information as users or items may have missing patterns associated with their rating behaviors. Therefore, we propose clustering users and items based on the numbers of their ratings or other variables associated with the missingness. Thus the group effects $(\mathbf{s}_{v_u}, \mathbf{t}_{j_i})$ could provide unique latent information which are not available in $\mathbf{x}_{ui}$, $\mathbf{p}_u$ or $\mathbf{q}_i$. Note that if the group effects $(\mathbf{s}_{v_u}, \mathbf{t}_{j_i})$ are the only factors that are associated with the missing process, then the proposed method captures the entire missing-not-at-random mechanism. In other words, correctly estimating $(\mathbf{s}_{v_u}, \mathbf{t}_{j_i})$ enables us to achieve consistent and efficient estimation of θ_{ui}, regardless of the missing mechanism.

4.8 Proofs of Theoretical Results

4.8.1 Proof of Lemma 3

By Theorem 2.1 of Ansley and Kohn (1994), each of (4.3) and (4.4) has a unique solution, and the back-fitting algorithms for (4.5) and (4.6) can be applied in (4.3), and (4.7) and (4.8) can be applied in (4.4). This guarantees convergence to the unique solution given any initial value. Therefore, Algorithm 1 is equivalent to minimizing (4.3) and (4.4) iteratively.

Note that minimizing (4.3) and (4.4) iteratively is a special case of Algorithm MBI (Chen et al., 2012b) with two blocks. Therefore, following Theorem 3.1 of Chen et al. (2012b), Algorithm 1 converges to a stationary point. This completes the proof.

4.8.2 Proof of Lemma 4

Since $\theta_{ui} = (\mathbf{p}_u + \mathbf{s}_{v_u})'(\mathbf{q}_i + \mathbf{t}_{j_i})$ is a quadratic function of $(\mathbf{p}'_u, \mathbf{q}'_i, \mathbf{s}'_{v_u}, \mathbf{t}'_{j_i})'$, and given that $\|(\mathbf{p}'_u, \mathbf{q}'_i, \mathbf{s}'_{v_u}, \mathbf{t}'_{j_i})'\|_\infty$ and $\|(\tilde{\mathbf{p}}'_u, \tilde{\mathbf{q}}'_i, \tilde{\mathbf{s}}'_{v_u}, \tilde{\mathbf{t}}'_{j_i})'\|_\infty$ are bounded by L, there exists a constant $C_1 \geq 0$, such that

$$\|\theta_{ui} - \tilde{\theta}_{ui}\|_2 \leq C_1 \|(\mathbf{p}'_u, \mathbf{q}'_i, \mathbf{s}'_{v_u}, \mathbf{t}'_{j_i})' - (\tilde{\mathbf{p}}'_u, \tilde{\mathbf{q}}'_i, \tilde{\mathbf{s}}'_{v_u}, \tilde{\mathbf{t}}'_{j_i})'\|_2.$$

Recall that $f_{ui}(r, \boldsymbol{\gamma}) = f(r_{ui}|\theta_{ui})$, then based on Assumption 1:

$$\begin{aligned}
h^2(\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}) &= \frac{1}{nm} \sum_{u=1}^{n} \sum_{i=1}^{m} \int |f_{ui}^{1/2}(r, \boldsymbol{\gamma}) - f_{ui}^{1/2}(r, \tilde{\boldsymbol{\gamma}})|^2 d\nu(r) \\
&\leq \frac{1}{nm} \sum_{u=1}^{n} \sum_{i=1}^{m} \bar{G}^2 C_1^2 \|(\mathbf{p}'_u, \mathbf{q}'_i, \mathbf{s}'_{v_u}, \mathbf{t}'_{j_i})' - (\tilde{\mathbf{p}}'_u, \tilde{\mathbf{q}}'_i, \tilde{\mathbf{s}}'_{v_u}, \tilde{\mathbf{t}}'_{j_i})'\|_2^2 \\
&\leq \frac{1}{nm} \bar{G}^2 C_1^2 (\|\mathbf{P} - \tilde{\mathbf{P}}\|_F^2 + \|\mathbf{Q} - \tilde{\mathbf{Q}}\|_F^2 + \|\mathbf{S}_c - \tilde{\mathbf{S}}_c\|_F^2 + \|\mathbf{T}_c - \tilde{\mathbf{T}}_c\|_F^2) \\
&\leq \frac{1}{nm} \bar{G}^2 C_1^2 \{\|\mathbf{P} - \tilde{\mathbf{P}}\|_F^2 + \|\mathbf{Q} - \tilde{\mathbf{Q}}\|_F^2 + (n+m)(\|\mathbf{S} - \tilde{\mathbf{S}}\|_F^2 + \|\mathbf{T} - \tilde{\mathbf{T}}\|_F^2)\} \\
&\leq \frac{n+m}{nm} \bar{G}^2 C_1^2 \|\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}\|_2^2.
\end{aligned}$$

The second-to-last inequality results from the fact that

$$\begin{aligned}
\|\mathbf{S}_c - \tilde{\mathbf{S}}_c\|_F^2 &= \sum_{v=1}^{N} |V_v| \cdot \|\mathbf{s}_v - \tilde{\mathbf{s}}_v\|_2^2 \\
&\leq \max_{v=1,\ldots,N} \{|V_v|\} \|\mathbf{S} - \tilde{\mathbf{S}}\|_F^2 \\
&\leq (n+m) \|\mathbf{S} - \tilde{\mathbf{S}}\|_F^2,
\end{aligned}$$

and similarly $\|\mathbf{T}_c - \tilde{\mathbf{T}}_c\|_F^2 \leq (n+m)\|\mathbf{T} - \tilde{\mathbf{T}}\|_F^2$. The last inequality results from the fact that

$$\|\boldsymbol{\gamma} - \tilde{\boldsymbol{\gamma}}\|_2^2 = \|\mathbf{P} - \tilde{\mathbf{P}}\|_F^2 + \|\mathbf{Q} - \tilde{\mathbf{Q}}\|_F^2 + \|\mathbf{S} - \tilde{\mathbf{S}}\|_F^2 + \|\mathbf{T} - \tilde{\mathbf{T}}\|_F^2.$$

Define $d_0 = \bar{G}C_1$, and the result then follows. This completes the proof.

4.8.3 Proof of Theorem 4

We first verify the condition of Lemma 2.1 of Ossiander (1987). Based on Lemma 4,

$$\begin{aligned}
& \left\{ \frac{1}{nm} \sum_{u=1}^{n} \sum_{i=1}^{m} E(\sup_{\hat{\gamma} \in B_d(\gamma)} |f_{ui}^{1/2}(r, \hat{\boldsymbol{\gamma}}) - f_{ui}^{1/2}(r, \boldsymbol{\gamma})|^2) \right\}^{1/2} \\
= & \left\{ \frac{1}{nm} \sum_{u=1}^{n} \sum_{i=1}^{m} \int \sup_{\hat{\gamma} \in B_d(\gamma)} |f_{ui}^{1/2}(r, \hat{\boldsymbol{\gamma}}) - f_{ui}^{1/2}(r, \boldsymbol{\gamma})|^2 d\nu(r) \right\}^{1/2} \\
\leq & \left\{ \frac{n+m}{nm} \bar{G}^2 C_1^2 \sup_{\hat{\gamma} \in B_d(\gamma)} \|\hat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}\|_2^2 \right\}^{1/2} \\
\leq & \sqrt{\frac{n+m}{nm}} d_0 d \\
:= & \; g(d)
\end{aligned}$$

Hence for $u > 0$,

$$H^B(u, \mathcal{S}(k), \rho) \leq H(g^{-1}(u/2), \mathcal{S}(k), \rho),$$

where H^B is the metric entropy of $\mathcal{S}(k)$ with bracketing of $f^{1/2}$, H is the ordinary metric entropy of $\mathcal{S}(k)$, and ρ is the L_2-norm.

Next we provide an upper bound for $H(g^{-1}(u/2), \mathcal{S}(k), \rho)$. Since $g^{-1}(u/2) = \frac{\sqrt{nm}}{2d_0\sqrt{n+m}} u$,

$N \leq n$, $M \leq m$, and $\|\boldsymbol{\gamma}\|_\infty \leq L$, we have

$$\begin{aligned}
0 &\leq H^B(u, \mathcal{S}(k), \rho) \\
&\leq H(g^{-1}(u/2), \mathcal{S}(k), \rho) \\
&\leq \log\left[\max\left\{\left(\frac{L\sqrt{(n+m+N+M)K}}{\frac{\sqrt{nm}}{2d_0\sqrt{n+m}}u}\right)^{(n+m+N+M)K}, 1\right\}\right] \\
&\leq \max\left\{(n+m+N+M)K\log\left(\frac{2\sqrt{2K}d_0L(n+m)}{\sqrt{nm}u}\right), 0\right\} \\
&= \max\left\{(n+m+N+M)K\log\left(\frac{\sqrt{K}C(n+m)}{\sqrt{nm}u}\right), 0\right\}
\end{aligned}$$

for $u \geq \epsilon^2_{|\Omega|}$ and $C = 2\sqrt{2}d_0L$.

We now find the convergence rate $\epsilon_{|\Omega|}$, which is the smallest ϵ that satisfies the conditions of Theorem 1 of Shen (1998). That is,

$$\sup_{k\geq k_0}\psi_1(\epsilon, k) \leq c_2|\Omega|^{1/2}$$

for a constant k_0, where $\psi_1(\epsilon, k) = \int_x^{x^{1/2}} \left\{H^B(u, \mathcal{F}(k))\right\}^{1/2} du/x$ with $x = (c_1\epsilon^2 + \lambda_{|\Omega|}(k-k_0))$, and $\mathcal{F}(k) = \{f^{1/2}(r, \boldsymbol{\gamma}) : \boldsymbol{\gamma} \in \mathcal{S}(k)\}$.

Note that $\psi_1 \leq 0 \leq c_2|\Omega|^{1/2}$ when $x \geq 1$, so we only consider the case when $0 < x < 1$. Notice that $K \geq 1$ and $n + m \geq \sqrt{nm}$. Then with a sufficiently large L, we have $\max\left\{\log\left(\frac{\sqrt{K}C(n+m)}{\sqrt{nm}u}\right), 0\right\} = \log\left(\frac{\sqrt{K}C(n+m)}{\sqrt{nm}u}\right)$ for $u \in [x, x^{1/2}]$. Then:

$$\begin{aligned}
\psi_1(\epsilon, k) &= \int_x^{x^{1/2}} \left\{H^B(u, \mathcal{F}(k))\right\}^{1/2} du/x \\
&\leq \big((n+m+N+M)K\big)^{1/2} \int_x^{x^{1/2}} \left\{\log\left(\frac{\sqrt{K}C(n+m)}{\sqrt{nm}}\right) - \log u\right\}^{1/2} du/x \\
&\leq \big((n+m+N+M)K\big)^{1/2}(x^{-1/2}-1)\left\{\log\left(\frac{\sqrt{K}C(n+m)}{\sqrt{nm}}\right) + \log(x^{-1})\right\}^{1/2}.
\end{aligned}$$

Since $\lambda_{|\Omega|} < \frac{1}{2k}\epsilon_{|\Omega|}^2$ and $k \sim O\big(\sqrt{(n+m+N+M)K}\big)$, we have $\lambda_{|\Omega|} = o(\epsilon_{|\Omega|}^2)$. Therefore, we solve

$$\begin{aligned}
\sup_{k\geq k_0}\psi_1(\epsilon,k) &= \psi_1(\epsilon,k_0) \\
&\sim \sqrt{(n+m+N+M)K}\frac{1}{\epsilon_{|\Omega|}}\left\{\log\left(\frac{\sqrt{K}(n+m)}{\epsilon_{|\Omega|}^2\sqrt{nm}}\right)\right\}^{1/2} \\
&= c_2|\Omega|^{1/2}.
\end{aligned}$$

Then the smallest rate $\epsilon_{|\Omega|}$ is determined by

$$\frac{1}{\epsilon_{|\Omega|}}\left\{\log\left(\frac{\sqrt{K}(n+m)}{\epsilon_{|\Omega|}^2\sqrt{nm}}\right)\right\}^{1/2} \sim \frac{|\Omega|^{1/2}}{\sqrt{(n+m+N+M)K}}.$$

Note that $N \leq n$ and $M \leq m$, then we have

$$\epsilon_{|\Omega|} \sim \frac{\sqrt{(n+m)K}}{|\Omega|^{1/2}}\left\{\log\left(\frac{|\Omega|}{\sqrt{nmK}}\right)\right\}^{1/2}.$$

For $\epsilon_{|\Omega|}$ and $\lambda_{|\Omega|}$, the conditions of Corollary 1 of Shen (1998) are now satisfied. The result then follows.

This completes the proof.

4.8.4 Proof of Theorem 5

Based on Theorem 4 and Theorem 1 of Shen (1998), there exists $c_i > 0$, $i = 1, 2$, such that:

$$P^*\left(\sup_{\{\tilde{\gamma}\in\mathcal{S}(k),h(\gamma,\tilde{\gamma})\geq\epsilon\}}\{\mathcal{L}(\boldsymbol{\gamma}|R^o) - \mathcal{L}(\tilde{\boldsymbol{\gamma}}|R^o)\} \geq -c_1|\Omega|\epsilon^2\right) \leq 7\exp(-c_2|\Omega|\epsilon^2).$$

Therefore, if there exists $\tilde{\boldsymbol{\gamma}} \in \mathcal{S}(k)$ such that $h(\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}) \geq \epsilon$, then

$$P^*\left(\{\mathcal{L}(\boldsymbol{\gamma}|R^o) - \mathcal{L}(\tilde{\boldsymbol{\gamma}}|R^o)\} \geq -c_1|\Omega|\epsilon^2\right) \leq 7\exp(-c_2|\Omega|\epsilon^2).$$

We suppress the subscript, write $h_\Theta(\theta_{ui}, \tilde{\theta}_{ui})$ as $h_\Theta(\theta, \tilde{\theta})$, and write $f(r_{ui}|\theta_{ui})$ as $f(r|\theta)$. We now lower-bound $h(\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}})$ by a function of $|\theta_{ui} - \tilde{\theta}_{ui}|$:

$$\begin{aligned} h^2_\Theta(\theta, \tilde{\theta}) &= E\left\{f^{1/2}(r|\theta) - f^{1/2}(r|\tilde{\theta})\right\}^2 \\ &= \left(\int_{\{f(r|\theta)>f(r|\tilde{\theta})\}} + \int_{\{f(r|\tilde{\theta})\geq f(r|\theta)\}}\right)\left\{f^{1/2}(r|\theta) - f^{1/2}(r|\tilde{\theta})\right\}^2 d\nu(r) \\ &:= I_1 + I_2, \end{aligned}$$

where $I_1 = \int_{\{f(r|\theta)>f(r|\tilde{\theta})\}} f(r|\theta)\left(1 - \exp\left[\frac{1}{2}\left\{(\tilde{\theta}-\theta)T(r) - (A(\tilde{\theta}) - A(\theta))\right\}\right]\right)^2 d\nu(r)$, and $I_2 = \int_{\{f(r|\tilde{\theta})\geq f(r|\theta)\}} f(r|\tilde{\theta})\left(1 - \exp\left[\frac{1}{2}\left\{(\theta-\tilde{\theta})T(r) - (A(\theta) - A(\tilde{\theta}))\right\}\right]\right)^2 d\nu(r)$.

For I_1, since $f(r|\theta) > f(r|\tilde{\theta})$, we have $Z := (\tilde{\theta}-\theta)T(r) - (A(\tilde{\theta}) - A(\theta)) \leq 0$. Since $\|\boldsymbol{\gamma}\|_\infty \leq L$, we have θ bounded in a closed set, and hence $A'(\theta) = E_\theta[T(r)]$ is bounded. Let $L_A = \sup_\theta |E_\theta T(r)|$, then

$$|A(\tilde{\theta}) - A(\theta)| \leq L_A|\tilde{\theta} - \theta|.$$

Then $-Z = |Z| \geq (|T(r)| - L_A)|\tilde{\theta} - \theta|$. That is,

$$1 - \exp\{-\frac{1}{2}|Z|\} \geq \max\left\{1 - \exp\left[\frac{1}{2}(L_A - |T(r)|)|\tilde{\theta} - \theta|\right], 0\right\},$$

and

$$\begin{aligned} I_1 &= \int_{\{f(r|\theta)>f(r|\tilde{\theta})\}} f(r|\theta)\left(1 - \exp\{-\frac{1}{2}|Z|\}\right)^2 d\nu(r) \\ &\geq \int_{\{f(r|\theta)>f(r|\tilde{\theta})\}} f(r|\theta)\max\left\{1 - \exp\left[\frac{1}{2}(L_A - |T(r)|)|\tilde{\theta} - \theta|\right], 0\right\}^2 d\nu(r). \end{aligned}$$

In a similar way,

$$\begin{aligned} I_2 &\geq \int_{\{f(r|\tilde{\theta})\geq f(r|\theta)\}} f(r|\tilde{\theta})\max\left\{1 - \exp\left[\frac{1}{2}(L_A - |T(r)|)|\tilde{\theta} - \theta|\right], 0\right\}^2 d\nu(r) \\ &\geq \int_{\{f(r|\tilde{\theta})\geq f(r|\theta)\}} f(r|\theta)\max\left\{1 - \exp\left[\frac{1}{2}(L_A - |T(r)|)|\tilde{\theta} - \theta|\right], 0\right\}^2 d\nu(r). \end{aligned}$$

Notice that $1-\exp\left\{\frac{1}{2}(L_A-|T(r)|)|\tilde{\theta}-\theta|\right\}\geq 0$ if and only if $|T(r)|\geq L_A$. Hence,

$$\begin{aligned} h^2_{\Theta}(\theta,\tilde{\theta}) &= I_1+I_2 \\ &\geq \int_{\{|T(r)|\geq L_A\}} f(r|\theta)\left[1-\exp\left\{\frac{1}{2}(L_A-|T(r)|)|\tilde{\theta}-\theta|\right\}\right]^2 d\nu(r), \end{aligned}$$

which is a non-decreasing function of $|\tilde{\theta}-\theta|$.

Therefore, for each θ_{ui}, and given the $\epsilon_{|\Omega|}$ in Theorem 4, there exists a $\delta_{|\Omega|}(\theta_{ui})$, such that $|\tilde{\theta}_{ui}-\theta_{ui}|>\delta_{|\Omega|}(\theta_{ui})$ implies $h_{\Theta}(\theta_{ui},\tilde{\theta}_{ui})\geq\epsilon_{|\Omega|}$.

Let $\delta_{|\Omega|}=\max\limits_{1\leq u\leq n,1\leq i\leq m}\sup\limits_{\theta_{ui}}\delta_{|\Omega|}(\theta_{ui})$, then $|\tilde{\theta}_{ui}-\theta_{ui}|>\delta_{|\Omega|}$ for each (u,i) implies $h(\boldsymbol{\gamma},\tilde{\boldsymbol{\gamma}})\geq\epsilon_{|\Omega|}$, and the result follows. This completes the proof.

4.8.5 Proof of Corollary 3

Define $\Phi=\{(u,i):|\theta^0_{ui}-\theta_{ui}|>\delta_{|\Omega|}\}$. Then the cardinality of Φ satisfies $|\Phi|\geq\phi nm$. From Theorem 5, for $(u,i)\in\Phi$, we have $h_{\Theta}(\theta^0_{ui},\theta_{ui})\geq\frac{1}{\sqrt{\phi}}\epsilon_{|\Omega|}$. Hence by the definition of $\boldsymbol{\gamma}_0$, we have:

$$h(\boldsymbol{\gamma},\boldsymbol{\gamma}_0)=\left\{\frac{1}{nm}\sum_{i=1}^{m}\sum_{u=1}^{n}h^2_{\Theta}(\theta^0_{ui},\theta_{ui})\right\}^{1/2}\geq\left\{\frac{1}{nm}(\phi nm)\frac{1}{\phi}\epsilon^2_{|\Omega|}\right\}^{1/2}=\epsilon_{|\Omega|}.$$

This completes the proof.

4.8.6 Proof of Corollary 4

In the proof of Corollary 3, we verify that $h(\boldsymbol{\gamma},\boldsymbol{\gamma}_0)\geq\epsilon_{|\Omega|}$. Then by Lemma 4, we have:

$$\epsilon^2_{|\Omega|}\leq h^2(\boldsymbol{\gamma},\boldsymbol{\gamma}_0)\leq\frac{d_0^2(n+m)}{nm}\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|_2^2.$$

Then $\|\mathbf{S}\|_F^2+\|\mathbf{T}\|_F^2=\|\boldsymbol{\gamma}-\boldsymbol{\gamma}_0\|^2\geq\frac{nm}{d_0^2(n+m)}\epsilon^2_{|\Omega|}$.

Meanwhile, we have each entry of $\mathbf{S}$ and $\mathbf{T}$ bounded by L, and hence $\|\mathbf{S}\|_F^2+\|\mathbf{T}\|_F^2\leq$

$(N+M)KL^2$.

Therefore, $N+M \geq \frac{nm}{d_0^2 L^2 K(n+m)} \epsilon_{|\Omega|}^2$. By the rate of $\epsilon_{|\Omega|}$ provided in Theorem 4, we have:

$$O(N+M) \succeq \frac{nm}{|\Omega|} \log\left(\frac{|\Omega|}{\sqrt{nmK}}\right).$$

This completes the proof.

4.9 Tables and Figures

Table 4.1: RMSE (standard error) of the proposed method compared with four existing methods, with the missing rate $\bar{\pi} = 70\%$, 80%, 90% and 95%, and the number of latent factors $K = 3$ or 6, where RSVD, AC, MHT and ZSY stand for regularized singular value decomposition, the regression-based latent factor model (Agarwal and Chen, 2009), Soft-Impute (Mazumder et al., 2010), and the latent factor model with sparsity pursuit (Zhu et al., 2016), respectively.

No. of latent factors	Missing Rate	The Proposed Method	RSVD	AC	MHT	ZSY
$K = 3$	70%	1.232 (0.029)	1.823 (0.324)	4.218 (0.089)	3.591 (0.178)	2.384 (0.077)
	80%	1.329 (0.042)	2.574 (0.506)	4.190 (0.091)	4.064 (0.140)	2.574 (0.085)
	90%	1.521 (0.070)	4.002 (0.689)	4.109 (0.095)	4.581 (0.116)	2.982 (0.095)
	95%	1.800 (0.103)	4.526 (0.172)	4.087 (0.096)	4.774 (0.123)	3.288 (0.100)
$K = 6$	70%	1.461 (0.035)	3.728 (0.188)	7.164 (0.132)	7.126 (0.294)	5.844 (0.656)
	80%	1.634 (0.058)	4.926 (0.274)	6.962 (0.134)	8.038 (0.267)	5.885 (0.145)
	90%	2.032 (0.136)	7.048 (0.270)	6.805 (0.136)	8.931 (0.172)	6.019 (0.420)
	95%	2.839 (0.388)	8.316 (0.270)	6.846 (0.149)	9.142 (0.176)	6.207 (0.151)

Table 4.2: RMSE (standard error) of the proposed method when the missing rate is 70%, 80%, 90% or 95%, and the number of latent factors $K = 3$ or 6, under 0%, 10%, 30% and 50% cluster misspecification rate.

		Misspecification Rate			
No. of latent factors	Missing Rate	0%	10%	30%	50%
$K = 3$	70%	1.232 (0.029)	1.237 (0.029)	1.250 (0.032)	1.265 (0.038)
	80%	1.329 (0.042)	1.340 (0.052)	1.359 (0.051)	1.380 (0.049)
	90%	1.521 (0.070)	1.544 (0.180)	1.591 (0.162)	1.626 (0.255)
	95%	1.800 (0.103)	1.810 (0.116)	1.869 (0.102)	1.920 (0.093)
$K = 6$	70%	1.461 (0.035)	1.502 (0.049)	1.560 (0.048)	1.623 (0.059)
	80%	1.634 (0.058)	1.698 (0.070)	1.815 (0.074)	1.911 (0.092)
	90%	2.032 (0.136)	2.229 (0.198)	2.428 (0.146)	2.648 (0.150)
	95%	2.839 (0.388)	3.041 (0.302)	3.245 (0.238)	3.373 (0.178)

Table 4.3: RMSE of the proposed method compared with six existing methods for MovieLens 1M and 10M data, where RSVD, AC, MHT and ZSY stand for regularized singular value decomposition, the regression-based latent factor model (Agarwal and Chen, 2009), Soft-Impute (Mazumder et al., 2010), and the latent factor model with sparsity pursuit (Zhu et al., 2016), respectively.

	MovieLens 1M	MovieLens 10M
Grand Mean Imputation	1.1112	1.0185
Linear Regression	1.0905	1.0007
The Proposed Method	0.9635	0.9295
RSVD	1.0552	0.9966
AC	1.1974	0.9737
MHT	1.0737	1.0177
ZSY	1.0635	1.0108

Table 4.4: RMSE of the proposed method compared with four existing methods on the MovieLens 10M data to study the "cold-start" problem: "old ratings" and "new ratings" stand for ratings in the testing sets given by existing users to existing items, and by new users or to new items. Here RSVD, AC, MHT and ZSY stand for regularized singular value decomposition, the regression-based latent factor model (Agarwal and Chen, 2009), Soft-Impute (Mazumder et al., 2010), and the latent factor model with sparsity pursuit (Zhu et al., 2016), respectively.

	The proposed method	RSVD	AC	MHT	ZSY
"old ratings"	0.7971	0.8062	1.3324	0.8160	0.8018
"new ratings"	0.9348	1.0039	0.9553	1.0252	1.0189
the entire testing set	0.9295	0.9966	0.9737	1.0177	1.0108

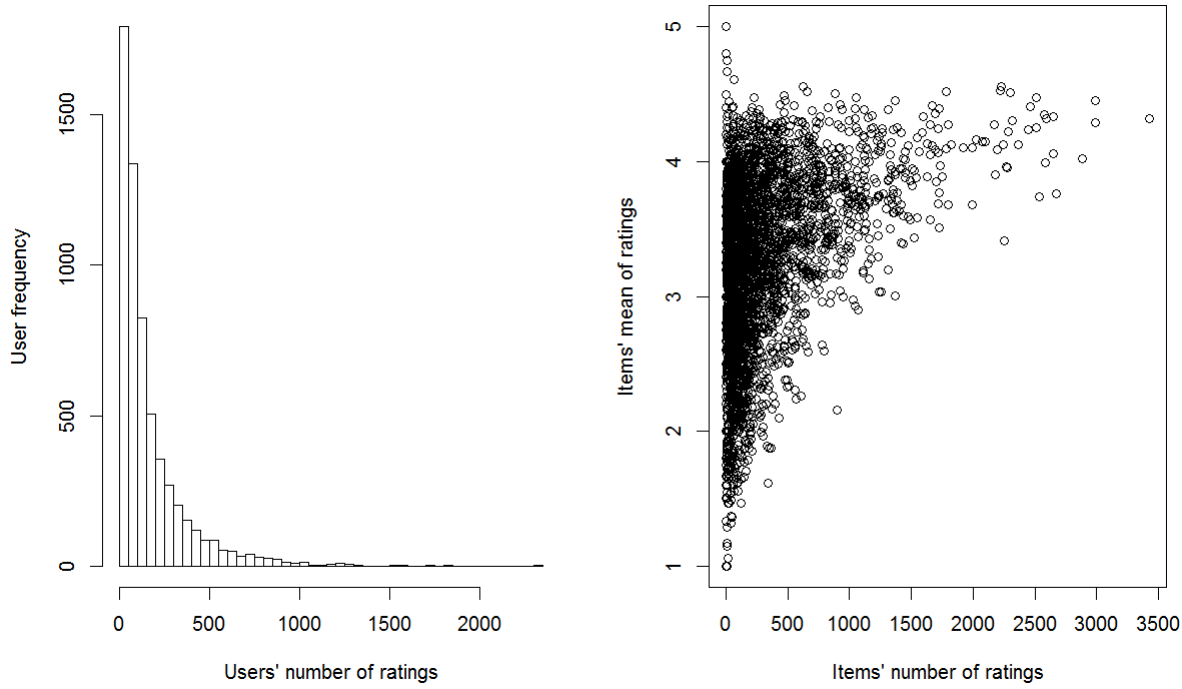

Figure 4.1: Missing pattern analysis for the MovieLens 1M data. Left: Most users rated a small number of movies, while few users rated a large number of movies. Right: Movies with a high average rating attract more users.

References

Adomavicius, G. and Tuzhilin, A. (2005). Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749.

Agarwal, D. and Chen, B.-C. (2009). Regression-based latent factor models. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 19–28. ACM.

Anderson, D. A. and Aitkin, M. (1985). Variance component models with binary response: interviewer variability. *Journal of the Royal Statistical Society. Series B (Methodological)*, 203–210.

Ansley, C. F. and Kohn, R. (1994). Convergence of the backfitting algorithm for additive models. *Journal of the Australian Mathematical Society (Series A)*, 57(03):316–329.

Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. *Studies in item analysis and prediction*, 6:158–168.

Beckenback, E. F. and Bellman, R. E. (1965). *Inequalities*. Springer Verlag.

Bell, R. M. and Koren, Y. (2007). Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In *Proceedings of the 2007 7th IEEE International Conference on Data Mining*, 43–52. IEEE.

Blanco-Fernandez, Y., Pazos-Arias, J. J., Gil-Solla, A., Ramos-Cabrer, M., and Lopez-Nores, M. (2008). Providing entertainment by content-based filtering and semantic reasoning in intelligent recommender systems. *IEEE Transactions on Consumer Electronics*, 54(2):727–735.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 80(391):580–598.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American statistical Association*, 88(421):9–25.

Cacheda, F., Carneiro, V., Fernández, D., and Formoso, V. (2011). Comparison of collaborative filtering algorithms: Limitations of current techniques and proposals for scalable, high-performance recommender systems. *ACM Transactions on the Web (TWEB)*, 5(1):2.

Carroll, J. D. and Chang, J.-J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3):283–319.

Chaganty, N. R. and Joe, H. (2006). Range of correlation matrices for dependent bernoulli random variables. *Biometrika*, 93(1):197–206.

Chen, B., Grace, Y. Y., and Cook, R. J. (2012a). Weighted generalized estimating functions for longitudinal response and covariate data that are missing at random. *Journal of the American Statistical Association*.

Chen, B., He, S., Li, Z., and Zhang, S. (2012b). Maximum block improvement and polynomial optimization. *SIAM Journal on Optimization*, 22(1):87–107.

Cho, H., Wang, P., and Qu, A. (2016). Personalized treatment for longitudinal data using unspecified random-effects model. *Statistica Sinica*, In press.

Cook, R. D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, 91(435):983–992.

Cook, R. D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics. 1998*. Wiley, New York.

Cook, R. D. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Annals of Statistics*, 455–474.

Cook, R. D. and Ni, L. (2005). Sufficient dimension reduction via inverse regression: A minimum discrepancy approach. *Journal of the American Statistical Association*, 100(470):410–428.

Cook, R. D. and Ni, L. (2006). Using intraslice covariances for improved estimation of the central subspace in regression. *Biometrika*, 93(1):65–74.

Cook, R. D. and Weisberg, S. (1991). Discussion of "sliced inverse regression for dimension reduction," by k.c. li. *Journal of the American Statistical Association*, 86:328–332.

Cook, R. D. and Weisberg, S. (1994). Transforming a response variable for linearity. *Biometrika*, 81(4):731–737.

Cook, R. D. and Yin, X. (2001). Dimension reduction and visualization in discriminant analysis (with discussion). *Australian & New Zealand Journal of Statistics*, 43(2):147–199.

Deng, Y., Hillygus, D. S., Reiter, J. P., Si, Y., Zheng, S., et al. (2013). Handling attrition in longitudinal studies: The case for refreshment samples. *Statistical Science*, 28(2):238–256.

Diggle, P. and Kenward, M. G. (1994). Informative drop-out in longitudinal data analysis. *Applied statistics*, 49–93.

Dong, Y. and Li, B. (2010). Dimension reduction for non-elliptically distributed predictors: second-order methods. *Biometrika*, 97(2):279–294.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Feuerverger, A., He, Y., and Khatri, S. (2012). Statistical significance of the Netflix challenge. *Statistical Science*, 27(2):202–231.

Fitzmaurice, G. M., Laird, N. M., and Ware, J. H. (2012). *Applied longitudinal analysis*, volume 998. John Wiley & Sons.

Follmann, D. A. and Wu, M. C. (1995). An approximate generalized linear model with random effects for informative missing data. *Biometrics*, 51(1):151–168.

Fu, W. J. (2003). Penalized estimating equations. *Biometrics*, 59(1):126–132.

Fung, W. K., He, X., Liu, L., and Shi, P. (2002). Dimension reduction based on canonical correlation. *Statistica Sinica*, 1093–1113.

Funk, S. (2006). Netflix update: Try this at home. *URL* ***http://sifter.org/~simon/journal/20061211.html***.

Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. (2001). Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151.

Grenander, U. (1981). *Abstract Inference*. Wiley, New York.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, 1029–1054.

Harshman, R. A. (1970). Foundations of the parafac procedure: Models and conditions for an "explanatory" multi-modal factor analysis. *UCLA Working Papers in Phonetics*, 16:1–84.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Henderson, R., Diggle, P., and Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, 1(4):465–480.

Hirano, K., Imbens, G. W., Ridder, G., and Rubin, D. B. (2001). Combining panel data sets with attrition and refreshment samples. *Econometrica*, 69(6):1645–1659.

Hogan, J. W. and Laird, N. M. (1997). Mixture models for the joint distribution of repeated measures and event times. *Statistics in medicine*, 16(3):239–257.

Ibrahim, J. G., Chen, M.-H., and Lipsitz, S. R. (2001). Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable. *Biometrika*, 88(2):551–564.

Kim, J. K. and Yu, C. L. (2012). A semiparametric estimation of mean functionals with nonignorable missing data. *Journal of the American Statistical Association*.

Koren, Y. (2010). Factor in the neighbors: Scalable and accurate collaborative filtering. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 4(1):1.

Koren, Y., Bell, R., and Volinsky, C. (2009). Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Kosorok, M. R. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

Lang, K. (1995). Newsweeder: Learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning*, 331–339.

Lee, S.-Y. and Tang, N.-S. (2006). Analysis of nonlinear structural equation models with nonignorable missing covariates and ordered categorical data. *Statistica Sinica*, 1117–1141.

Li, B., Cook, R. D., and Chiaromonte, F. (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. *Annals of statistics*, 1636–1668.

Li, B. and Dong, Y. (2009). Dimension reduction for nonelliptically distributed predictors. *The Annals of Statistics*, 1272–1298.

Li, B. and Wang, S. (2007). On directional regression for dimension reduction. *Journal of the American Statistical Association*, 102(479):997–1008.

Li, B., Zha, H., and Chiaromonte, F. (2005). Contour regression: a general approach to dimension reduction. *Annals of statistics*, 1580–1616.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction (with discussion). *Journal of the American Statistical Association*, 86(414):316–327.

Li, K.-C. (1992). On principal hessian directions for data visualization and dimension reduction: another application of stein's lemma. *Journal of the American Statistical Association*, 87(420):1025–1039.

Li, K.-C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics*, 1009–1052.

Li, L. and Yin, X. (2009). Longitudinal data analysis using sufficient dimension reduction method. *Computational Statistics & Data Analysis*, 53(12):4106–4115.

Li, N., Elashoff, R. M., Li, G., and Tseng, C.-H. (2012). Joint analysis of bivariate longitudinal ordinal outcomes and competing risks survival times with nonparametric distributions for random effects. *Statistics in medicine*, 31(16):1707–1721.

Liang, K.-Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Lin, H., McCulloch, C. E., and Rosenheck, R. A. (2004). Latent pattern mixture models for informative intermittent missing data in longitudinal studies. *Biometrics*, 60(2):295–305.

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Sinha, D., Parzen, M., and Lipshultz, S. (2009). Joint generalized estimating equations for multivariate longitudinal binary outcomes with missing data: an application to acquired immune deficiency syndrome data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):3–20.

Little, R. J. (1994). A class of pattern-mixture models for normal incomplete data. *Biometrika*, 81(3):471–483.

Little, R. J. (1995). Modeling the drop-out mechanism in repeated-measures studies. *Journal of the American Statistical Association*, 90(431):1112–1121.

Little, R. J. (2008). Selection and pattern-mixture models. *Longitudinal data analysis*, 409–431.

Liu, S., Shen, X., and Wong, W. H. (2005). Computational developments of ψ-learning. In *Proceedings 5th SIAM International Conference on Data Mining, Newport Beach, CA*, 1–12. SIAM.

Lops, P., De Gemmis, M., and Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In *Recommender Systems Handbook*, 73–105. Springer.

Luo, R., Wang, H., and Tsai, C.-L. (2009). Contour projected dimension reduction. *The Annals of Statistics*, 37(6B):3743–3778.

Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association*, 107(497):168–179.

Ma, Y. and Zhu, L. (2013a). Efficiency loss and the linearity condition in dimension reduction. *Biometrika*, 100(2):371–383.

Ma, Y. and Zhu, L. (2013b). Efficient estimation in sufficient dimension reduction. *Annals of statistics*, 41(1):250.

Ma, Y. and Zhu, L. (2013c). A review on dimension reduction. *International Statistical Review*, 81(1):134–150.

Ma, Y. and Zhu, L. (2014). On estimation efficiency of the central mean subspace. *Journal of the Royal Statistical Society: Series B*, 76(5):885–901.

Maruotti, A. (2015). Handling non-ignorable dropouts in longitudinal data: a conditional model based on a latent markov heterogeneity structure. *TEST*, 24(1):84–109.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *The Journal of Machine Learning Research*, 11:2287–2322.

McCullagh, P. (1983). Quasi-likelihood functions. *The Annals of Statistics*, 59–67.

McLeish, D. L. et al. (1975). A maximal inequality and dependent strong laws. *The Annals of probability*, 3(5):829–839.

Melville, P., Mooney, R. J., and Nagarajan, R. (2002). Content-boosted collaborative filtering for improved recommendations. In *Proceedings of the 18th National Conference on Artificial Intelligence*, 187–192.

Molenberghs, G., Kenward, M. G., and Lesaffre, E. (1997). The analysis of longitudinal ordinal data with nonrandom drop-out. *Biometrika*, 84(1):33–44.

Mooney, R. J. and Roy, L. (2000). Content-based book recommending using learning for text categorization. In *Proceedings of the 5th ACM Conference on Digital Libraries*, 195–204. ACM.

Nguyen, A.-T., Denos, N., and Berrut, C. (2007). Improving new user recommendations with rule-based induction on cold user data. In *Proceedings of the 2007 ACM Conference on Recommender Systems*, 121–128. ACM.

Nguyen, J. and Zhu, M. (2013). Content-boosted matrix factorization techniques for recommender systems. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 6(4):286–301.

Ortega, J. M. and Rheinboldt, W. C. (1970). *Iterative solution of nonlinear equations in several variables*, volume 30. Siam.

Ossiander, M. (1987). A central limit theorem under metric entropy with L2 bracketing. *The Annals of Probability*, 15(3):897–919.

Paik, M. C. (1997). The generalized estimating equation approach when data are not missing completely at random. *Journal of the American Statistical Association*, 92(440):1320–1329.

Pardoe, I., Yin, X., and Cook, R. D. (2007). Graphical tools for quadratic discriminant analysis. *Technometrics*, 49(2):172–183.

Park, S.-T., Pennock, D., Madani, O., Good, N., and DeCoste, D. (2006). Naïve filterbots for robust cold-start recommendations. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 699–705. ACM.

Pazzani, M. J. and Billsus, D. (2007). Content-based recommendation systems. In *The Adaptive Web*, 325–341. Springer.

Pew Research Center (2010). Four years later republicans faring better with men, whites, independents, and seniors (press release). *URL* ***http://www.people-press.org/files/legacy-pdf/643.pdf***.

Pfeiffer, R. M., Forzani, L., and Bura, E. (2012). Sufficient dimension reduction for longitudinally measured predictors. *Statistics in medicine*, 31(22):2414–2427.

Pollard, D. (2012). *Convergence of Stochastic Processes*. Springer Science & Business Media.

Qaqish, B. F. (2003). A family of multivariate binary distributions for simulating correlated binary variables with specified marginal means and correlations. *Biometrika*, 90(2):455–463.

Qu, A., Lindsay, B. G., and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87(4):823–836.

Qu, A., Lindsay, B. G., and Lu, L. (2010). Highly efficient aggregate unbiased estimating functions approach for correlated data with missing at random. *Journal of the American Statistical Association*, 105(489):194–204.

Qu, A., Yi, G., Song, P.-K., and Wang, P. (2011). Assessing the validity of weighted generalized estimating equations. *Biometrika*, 98(1):215–224.

Ridder, G. (1992). An empirical evaluation of some models for non-random attrition in panel data. *Structural Change and Economic Dynamics*, 3(2):337–355.

Rizopoulos, D., Verbeke, G., and Molenberghs, G. (2008). Shared parameter models under random effects misspecification. *Biometrika*, 95(1):63–74.

Robins, J. M., Rotnitzky, A., and Scharfstein, D. O. (2000). Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In *Statistical models in epidemiology, the environment, and clinical trials*, 1–94. Springer.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association*, 89(427):846–866.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *Journal of the American Statistical Association*, 90(429):106–121.

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. (1998). Semiparametric regression for repeated outcomes with nonignorable nonresponse. *Journal of the American Statistical Association*, 93(444):1321–1339.

Roy, J. (2003). Modeling longitudinal data with nonignorable dropouts using a latent dropout class model. *Biometrics*, 59(4):829–836.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Salakhutdinov, R., Mnih, A., and Hinton, G. (2007). Restricted Boltzmann machines for collaborative filtering. In *Proceedings of the 24th International Conference on Machine Learning*, 791–798. ACM.

Seaman, S. and Copas, A. (2009). Doubly robust generalized estimating equations for longitudinal data. *Statistics in medicine*, 28(6):937–955.

Shao, J. and Zhang, J. (2015). A transformation approach in linear mixed-effects models with informative missing responses. *Biometrika*, 102(1):107–119.

Shen, X. (1998). On the method of penalization. *Statistica Sinica*, 8(2):337–357.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Annals of Statistics*, 22(2):580–615.

Spagnoli, A., Henderson, R., Boys, R. J., and Houwing-Duistermaat, J. J. (2011). A hidden markov model for informative dropout in longitudinal response data with crisis states. *Statistics & Probability Letters*, 81(7):730–738.

Srebro, N., Alon, N., and Jaakkola, T. S. (2005). Generalization error bounds for collaborative prediction with low-rank matrices. In *In Advances In Neural Information Processing Systems 17*, 5–27.

Stubbendick, A. L. and Ibrahim, J. G. (2003). Maximum likelihood methods for nonignorable missing responses and covariates in random effects models. *Biometrics*, 59(4):1140–1150.

Stubbendick, A. L. and Ibrahim, J. G. (2006). Likelihood-based inference with nonignorable missing responses and covariates in models for discrete longitudinal data. *Statistica Sinica*, 1143–1167.

Tsonaka, R., Rizopoulos, D., Verbeke, G., and Lesaffre, E. (2010). Nonignorable models for intermittently missing categorical longitudinal responses. *Biometrics*, 66(3):834–844.

Tsonaka, R., Verbeke, G., and Lesaffre, E. (2009). A semi-parametric shared parameter model to handle nonmonotone nonignorable missingness. *Biometrics*, 65(1):81–87.

Van Buuren, S. (2007). Multiple imputation of discrete and continuous data by fully conditional specification. *Statistical methods in medical research*, 16(3):219–242.

Van Buuren, S. (2012). *Flexible imputation of missing data*. CRC press.

Vansteelandt, S., Rotnitzky, A., and Robins, J. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika*, 94(4):841–860.

Wang, H. and Xia, Y. (2008). Sliced regression for dimension reduction. *Journal of the American Statistical Association*, 103(482):811–821.

Wang, J. (2010). Consistent selection of the number of clusters via crossvalidation. *Biometrika*, 97(4):893–904.

Wang, P., Tsai, G.-f., and Qu, A. (2012). Conditional inference functions for mixed-effects models with unspecified random-effects distribution. *Journal of the American Statistical Association*, 107(498):725–736.

Wedderburn, R. W. (1974). Quasi-likelihood functions, generalized linear models, and the gauss?newton method. *Biometrika*, 61(3):439–447.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics*, 23(2):339–362.

Wu, M. (2007). Collaborative filtering via ensembles of matrix factorizations. In *Proceedings of KDD Cup and Workshop*.

Wu, M. C. and Bailey, K. R. (1989). Estimation and comparison of changes in the presence of informative right censoring: conditional linear model. *Biometrics*, 939–955.

Wu, M. C. and Carroll, R. J. (1988). Estimation and comparison of changes in the presence of informative right censoring by modeling the censoring process. *Biometrics*, 175–188.

Xia, Y., Tong, H., Li, W., and Zhu, L.-X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B*, 64(3):363–410.

Ye, Z. and Weiss, R. E. (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Journal of the American Statistical Association*, 98(464):968–979.

Yin, X. and Cook, R. D. (2002). Dimension reduction for the conditional kth moment in regression. *Journal of the Royal Statistical Society: Series B*, 64(2):159–175.

Yin, X. and Li, B. (2011). Sufficient dimension reduction based on an ensemble of minimum average variance estimators. *The Annals of Statistics*, 3392–3416.

Yuan, Y. and Little, R. J. (2009). Mixed-effect hybrid models for longitudinal data with nonignorable dropout. *Biometrics*, 65(2):478–486.

Zhou, J. (2009). Robust dimension reduction based on canonical correlation. *Journal of Multivariate Analysis*, 100(1):195–209.

Zhou, J. and He, X. (2008). Dimension reduction based on constrained canonical correlation and variable filtering. *The Annals of Statistics*, 1649–1668.

Zhou, J. and Qu, A. (2012). Informative estimation and selection of correlation structure for longitudinal data. *Journal of the American Statistical Association*, 107(498):701–710.

Zhou, Y., Little, R. J., Kalbfleisch, J. D., et al. (2010). Block-conditional missing at random models for missing data. *Statistical Science*, 25(4):517–532.

Zhu, L.-P. and Zhu, L.-X. (2009). Dimension reduction for conditional variance in regressions. *Statistica Sinica*, 869–883.

Zhu, Y., Shen, X., and Ye, C. (2016). Personalized prediction and sparsity pursuit in latent factor models. *Journal of the American Statistical Association*, 111(513):241–252.