Development of Joint Estimating Equation Approaches to Merging Clustered or Longitudinal Datasets from Multiple Biomedical Studies

by

Fei Wang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Biostatistics)
in The University of Michigan
2012

Doctoral Committee:

Professor Peter X.-K. Song, Co-Chair
Assistant Professor Lu Wang, Co-Chair
Associate Professor Timothy D. Johnson
Professor Ji Zhu

To my parents

ACKNOWLEDGEMENTS

It is a pleasure to thank the many people who made this thesis possible. It is difficult to overstate my gratitude to my Ph.D. advisors, Dr. Peter X.-K. Song and Dr. Lu Wang, for their support and advice throughout the course of my graduate studies. Without their encouragement, sound advice, good teaching, and lots of good ideas, I would have been lost.

I am also deeply indebted to my committee members Dr. Timothy D. Johnson and Dr. Ji Zhu. Their valuable comments, inputs and discussions greatly improve my thesis.

My appreciation also goes to the Department of Biostatistics which has provided a strongly supportive environment for study and research. I wish to thank all of the knowledgeable faculty members for their expertise in biostatistics.

Finally, I would like to give my special thank to my parents and my sister for their faith in me. It was under their watchful eye that I gained so much drive and an ability to tackle challenges head on.

TABLE OF CONTENTS

CHAPTER

LIST OF TABLES

Table

LIST OF FIGURES

Figure

LIST OF APPENDICES

Appendix

ABSTRACT

Development of Joint Estimating Equation Approaches to Merging Clustered or
Longitudinal Datasets from Multiple Biomedical Studies

by

Fei Wang

Co-Chairs: Peter X.-K. Song and Lu Wang

Jointly analyzing multiple datasets arising from similar studies has drawn increasing attention in recent years. In this dissertation, we investigate three primary problems pertinent to merging clustered or longitudinal datasets from multiple biomedical studies.

The first project concerns the development of a rigorous hypothesis testing procedure to assess the validity of data merging and a joint estimation approach to obtaining regression coefficient estimates when merging data is permitted. The proposed methods account for different within-subject correlations and follow-up schedules in different longitudinal studies. We establish large sample properties for the proposed test statistics and estimation. It is shown through simulations that our proposed test statistics are desirable in controlling test size even if within-subject correlation structures are misspecified. It is also shown that our joint estimation method improves estimation efficiency on all regression coefficients with merged data. For illustration, we apply the proposed methods to analyze data from a randomized interventional trial of asthma care.

The second project aims to generalize the work developed in the first project by allowing study-specific missing covariates. In particular, the presence of study-specific missing covariates (e.g. covariates observed in some studies but completely missing in other studies) gives rise to a great challenge in data merging and analysis. We propose a joint estimating function approach to addressing this key challenge, in which a novel nonparametric estimating function constructed via splines-based sieve approximation is utilized to bridge estimating equations from studies with missing covariates to those with fully observed covariates. Under mild regularity conditions, we show that the proposed estimator is consistent and asymptotically normal. We evaluate finite sample performances of the proposed method through simulation studies. In comparison to the conventional multiple imputation approach, our method exhibits smaller estimation bias. We provide an illustrative data analysis using longitudinal cohorts collected in Mexico City to assess the effect of lead exposure on children's somatic growth.

The third project is devoted to the development of a screening procedure for parameter homogeneity, which is the key feature to reduce model complexity in the process of data merging. We consider the longitudinal marginal model for merged studies, in which the classical hypothesis testing approach to evaluating all possible subsets of common regression parameters can be combinatorially complex and computationally prohibitive. We develop a regularization method that can overcome this difficulty by applying the idea of adaptive fused lasso in that restrictions are imposed on differences of pairs of parameters between studies. The selection procedure will automatically detect common parameters across all or subsets of studies. Through simulation studies we show that the proposed method performs well to consistently identify common parameters. We illustrate our method through merging HIV surveillance cohorts collected in China to assess if common effects are present over five geographic regions when these cohorts are monitored.

CHAPTER I

Introduction

1.1 Motivating Data Sets

This dissertation is motivated by three data, Physician Asthma Care Education (PACE), Lead exposures in cohorts of mother-infant pairs living in Mexico City, and HIV/AIDS surveillance data for drug users. They are described in detail in the next three sections.

1.1.1 Physician Asthma Care Education (PACE)

PACE data is collected from a community-based clinical trial aiming to evaluate the effectiveness of a continuing medical education program, Physician Asthma Care Education (PACE) (*Cabana et al.*, 2006), in the hope to improve pediatricians' asthma therapeutic and communication skills.

The randomized intervention trial was conducted in 10 locations in U.S.A, including Corpus Christi, TX; Fresno/Bakersfield, CA; Nashville, TN; Jacksonville, FL; Omaha, NE; St Paul, MN; Kent County, MI; New Castle County, DE; Columbus, OH; and Indianapolis, IN. To recruit physicians, using yellow-page listings and membership lists from local professional societies and asthma coalitions, the study investigators found 1219 primary care providers who agreed to participate into the research. All those selected providers had complete lists of their asthma patients.

To evaluate the effectiveness of continuous medical training on physicians' medical and communication skills, PACE study first matched 10 sites into 5 pairs according to population, asthma prevalence, percentage of the population that is Hispanic and/or black, climate, and other important factors. Each matched pair contained two similar sites and one was randomly selected as the control and the other as the intervention. Only physicians in the intervention sites accepted continuous medical training. Given that asthma symptoms are environmentally sensitive, such as meteorological conditions, it is natural to separate these 5 pairs of regions into two regions, "south" and "north", to control potential confounding factors specific to the regions. Study sites in each region are listed in Table 1.1. Due to potential regional differences in asthma prevalences, it is of interest to examine if the two regional datasets share common effects in some covariates.

Table 1.1: Study sites in "south" and "north" regions.

south	north
Fresno/Bakersfield (CA)	St. Paul (MN)
Corpus Christi (TX)	Omaha (NE)
Jacksonville (FL)	Columbus (OH)
Nashville (TN)	Indianapolis (IN)
	Castle County (DE)
	Kent County (MI)

Patient samples in PACE data were chosen from patient lists provided by selected physicians according to the following criteria: a diagnosis of asthma; between 2 and 12 years of age; a patient of a study physician; and no other diseases associated with pulmonary complications. Each patient was interviewed at baseline , year 1 and year 2 follow-ups. In total, there were 870 asthma patients interviewed at baseline, and at year 1 and year 2 follow-ups. The number of pediatricians associated with patient samples was 101, including 48 controls and 53 interventions.

One of primary outcomes is the number of nights with trouble sleeping, along with other demographic variables and some risk factors, such as time, self-evaluated asthma severity on patient's night-time asthma symptoms, symptom frequency, baseline outcome, age and gender. This study had average age of 2.7 years, 69% male patients, and 40.4% patients in the intervention group. Figure 1.1 shows longitudinal plots of the number of trouble-sleeping nights for asthma patients within "south" and "north" regions. Analyzing the datasets, we aim to examine if they may be combined before the entire data is used for the analysis.

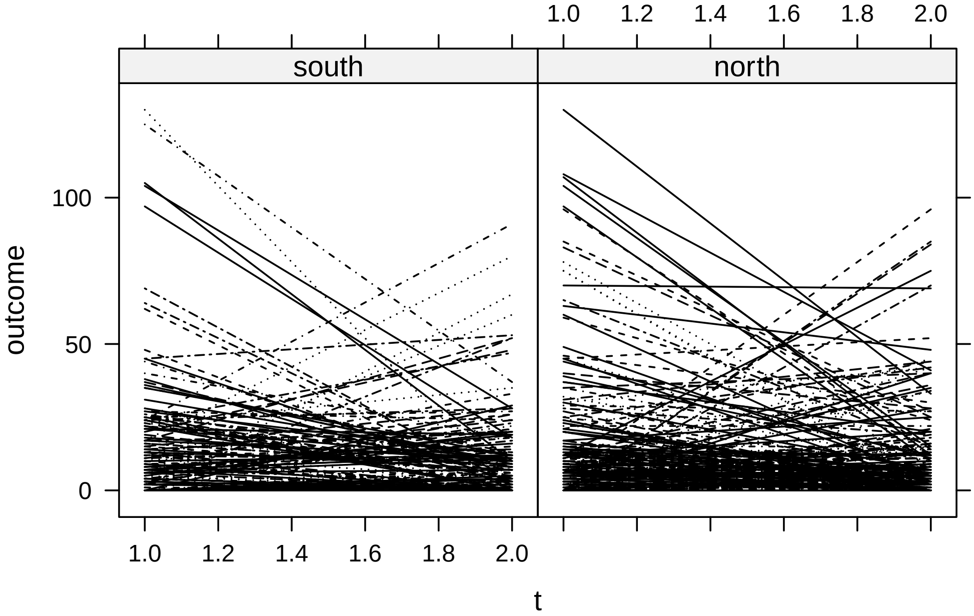

Figure 1.1: Longitudinal plot of outcomes for asthma patients over 2 years

1.1.2 Lead Exposure of Mother-infant Pairs in Mexico City

To evaluate the effect of children's lead exposures on their somatic growth, investigators established two birth cohorts from two hospitals in Mexico City. The two cohorts are termed as cohort B and cohort C in the thesis, which contain 88 and 427 mother-infant pairs, respectively. Children participated in the study were followed within 5 years, and the investigators collected the information about mother's blood lead exposure (PBL) and child's cord blood lead exposure (CBL). Exposure PBL was observed completely in both cohorts but exposure CBL was only fully observed in cohort C and was significantly missing for 54% children in cohort B. The number of repeated measurements of children in the study varies from 2 to 11 with its distribution presented in Figure 1.2. Children's weight trajectories are displayed in Figure 1.3.

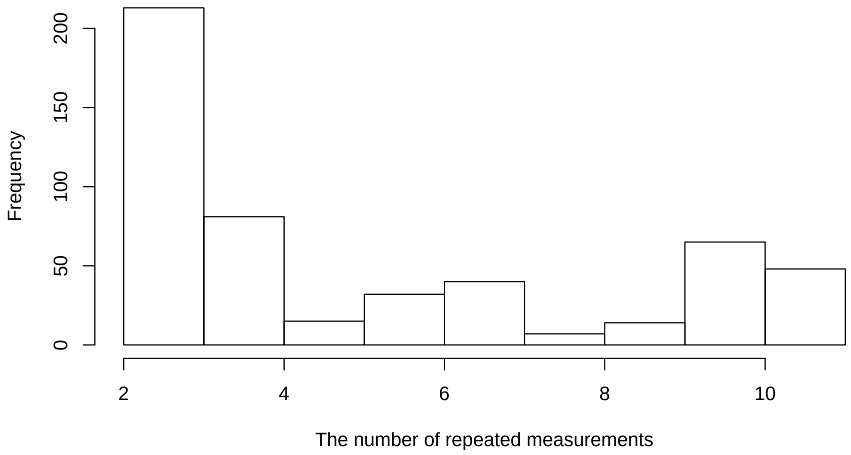

Figure 1.2: Histogram of the number of repeated measurements of lead exposure data in Mexico City.

The study objective is to evaluate the association between exposure CBL and children's weight growth, adjusting for other observed covariates in both cohorts. A key challenge in the analysis arises from the missingness of CBL in cohort B. But the fact that CBL and PBL were both observed in cohort C makes it possible for us to borrow information from cohort C to recover missing measurements of CBL in cohort B.

1.1.3 HIV/AIDS Surveillance Data for Drug Users

In recent years, HIV/AIDS has become an urgent public health issue in China. It is paramount to identify risk factors on HIV infection among injection drug users so that the government can monitor and control HIV/ADIS pandemics. For this purpose, annual surveys were conducted by Chinese CDC in 2006, 2007, 2008 and

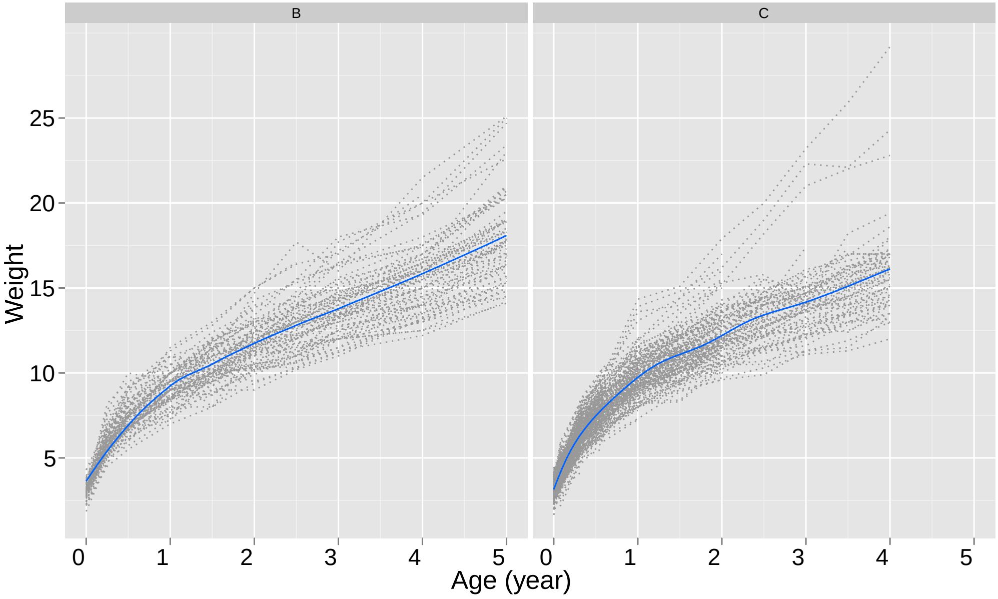

Figure 1.3: Trajectories of children's weights vs children's ages over two cohorts.

2009 in a southwestern province of China, aiming to identify and evaluate drug users' risk behaviors related to HIV infection. Some of risk factors related to drug users' behavioral characteristics are injection drug use, needle sharing, and unsafe sex.

The annual survey is based on a multiple stage stratified sampling with counties/cities as sampling units in the first stage and surveillance centers of countries/cities, such as centers of disease control, hospitals, or drug rehabilitation, as primary sampling units (PSU) in the second stage. Within each selected surveillance center, blood samples of all drug users were gathered and tested for HIV positive. The stratified sampling units are illustrated in Figure 1.4.

The number of selected surveillance centers in the 2006 to 2009 survey were 20, 12, 4 and 37, respectively. The selected surveillance centers are sampled from 5 regions termed as A, B, C, D, and E, which are different in terms their population sizes and socioeconomic status. Cluster sizes, namely, the number of injection drug users in a surveillance center, varies unevenly, ranging from 11 to 440. Figure 1.5 shows the distribution of cluster sizes.

The primary outcome measured in the study is HIV infection defined by a binary

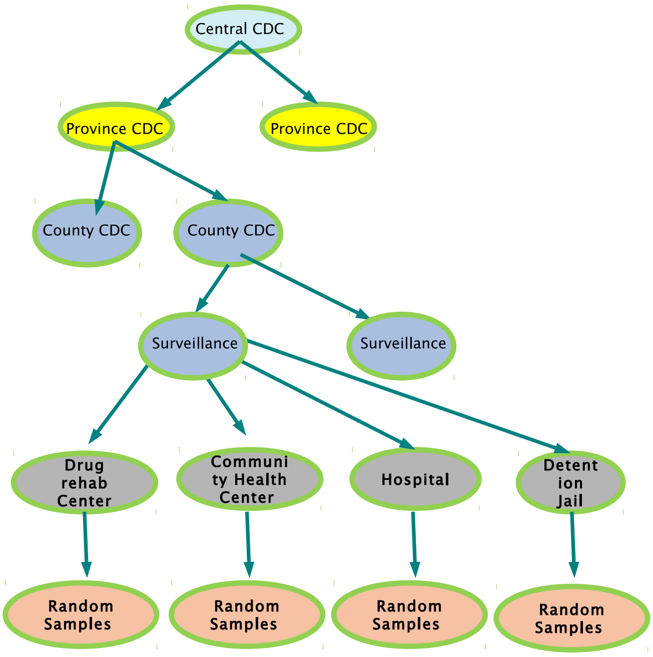

Figure 1.4: Diagram of sampling design

variable with 1 for being infected and 0 otherwise. To investigate drug users' risk behaviors, the survey collected information about injection drug use, unsafe sex, needle sharing as well as drug users' demographic variables including gender and marital status. As shown in Table 1.2, the majority of the drug users were adult male and almost at least 40% of drug users were not in marriage.

Our objective of the data analysis is to assess the common effects of behavioral activities across the five regions on HIV infection. The approach of hypothesis testing for such purpose is infeasible due to the large number of tests when either the number of covariates or the number of studies is large. But the idea of regularized estimation can provide an alternative fast screening procedure to solve the problem.

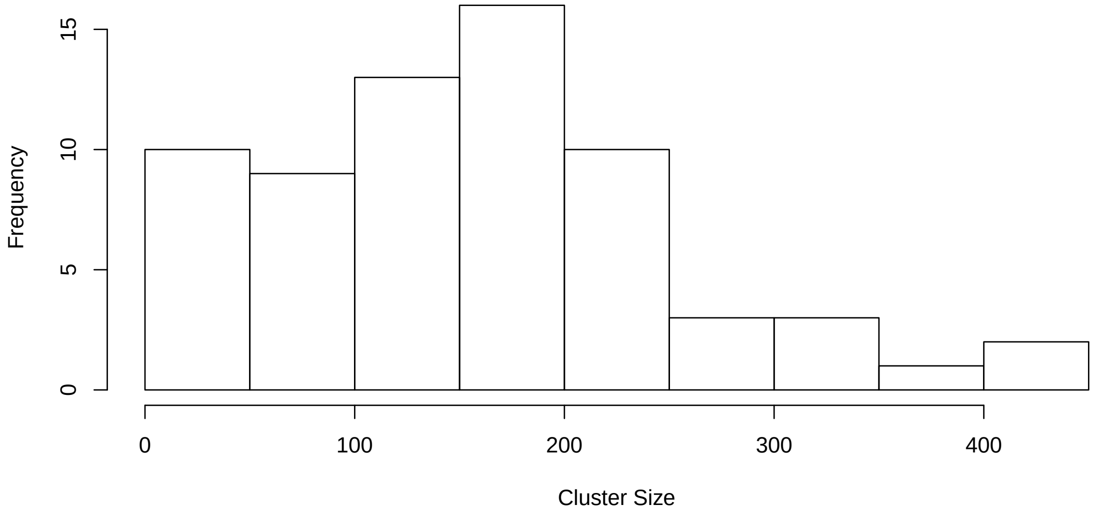

Figure 1.5: Histogram of cluster sizes for 2006 to 2009 HIV/AIDS surveillance data.

Table 1.2: Summary statistics for the outcome and demographic variables.

Variables	2006	2007	2008	2009
The number of clusters	20	12	4	37
Male (%)	85.8	82.3	88.1	79.9
Single (%)	51.1	47.6	47.5	40.3
Positive HIV test (%)	14.4	13.7	22.1	6.7

1.2 Meta Analysis

Meta analysis has long been known as one of the most important statistical approaches in biomedical and public health studies. The key of such study is to combine datasets from several similar studies, which allows investigators to achieve study objectives that, otherwise, cannot be achieved using any single study. For example *Davis et al.* (2010) proposed a model-based approach that combines information from two surveys. There are many other published work of similar types, which clearly indicates the popularity of meta analysis in practice.

One primary motivation of meta analysis is to increase sample size, which will hopefully result in more powerful analysis, such as smaller p-values. However, this idea of combining data sets is not a free lunch and often controversial. According to a case study in *Mendenhall et al.* (2008) , Edmonton company CT Technologies conducted two randomized clinical trials to test whether their proprietary ginseng extract would reduce the incidence of respiratory illnesses. Combining the two trials, they concluded that their results indicted a 89% reduction in laboratory-confirmed respiratory illness. But two professors from University of British Columbia criticized their claim and accused the article's authors of doing a form of data-mining by taking two studies that don't show a benefit and then adding them together to get a positive result. An important lesson we learned from this example is that there is a clear need of validation procedure that can either approve or disapprove the data merging before the combined data can be used to draw scientific conclusions.

1.3 Specific Objectives

The need of validation procedures as a pre-meta analysis "door-keeper" has been advocated by some thoughtful statisticians. One classical example is the well-known Brewslow and Day's test (*Breslow and Day*, 1980) for a common odds ratio across different strata in contingency table analysis. This test is performed to validate the existence of a common association before data from different strata can be combined to estimate such common association. In the current literature, there are very few results available to address statistical issues related to merging clustered or longitudinal data, in particular statistical tools for validation. As a matter of fact, in practice many studies have applied the strategy of data merging without considering the validity of this operation.

In this dissertation, we aim to develop methodologies for the following analytic objectives:

(i) To establish a data-driven validation procedure verifying data merging. In particular, we will develop a tool of hypothesis test to approve or disapprove common effects of covariates in longitudinal regression models;

(ii) To obtain estimates of regression coefficients when data merging is approved;

(iii) To jointly analyze multiple studies when study-specific missing covariates exist;

(iv) To propose a fast screening procedure for analyzing multiple longitudinal studies when the hypothesis testing approach is challenged by a huge number of tests.

When multiple studies are combined, it is important to account for between-study heterogeneity. This is a difficult task because such heterogeneity is attributed to multiple factors that would confound each other and can not be easily disentangled. In the case of clustered or longitudinal data, such heterogeneity may arise from, for example, different within-subject correlations, different study follow-up schedules, or study-specific missing covariates (covariates is completely missing in one study but fully observed in other studies). Thus, most of the existing meta analysis methods developed for cross-sectional data are no longer applicable for clustered or longitudinal data. We will develop several novel strategies to validate clustered or longitudinal data merging, and then to establish estimating function approaches to analyzing combined data after the validity of data merging is granted.

1.4 Organization of the Dissertation

In Chapter 2, we consider a standard longitudinal framework in that all variables are collected from different studies under a common study protocol without missing covariates. We first develop a rigorous quadratic inference function (*Qu et al.*, 2000) strategy to validate longitudinal data merging by testing for the unbiasedness of the generalised estimating functions under a common set of regression coefficients but with possibly different covariance structures. We establish large sample properties for the proposed test and estimation procedure. Using simulation studies, we compare our method with several popular methods, including meta analysis and generalized estimating equations. It is shown that our test gives a robust control of type I error against the misspecification of working correlation structures. In addition, our joint estimation procedure leads to an improvement in estimation efficiency on the all regression coefficients after data merging is validated. We illustrate the proposed methods through a randomized interventional trial of asthma care introduced in Section 1.1.1.

In Chapter 3, we relax the setting considered in Chapter 2 by allowing study-specific missing covariates (e.g. covariates observed in some studies but completely missing in other studies). We propose a joint estimating function approach to analyzing merged studies with study-specific missing covariates, in which we first construct a nonparametric estimating function via splines-based sieve approximation, and then utilize it to bridge estimating equations from studies with missing covariates to those with fully observed covariates. Under some mild regularity conditions, we establish consistency and asymptotic normality for the proposed estimator. We use simulation studies to evaluate finite sample performances of the proposed method. We also compare our method to the conventional multiple imputation approach, and our method exhibits smaller estimation bias. We analyze the lead exposure data introduced in Section 1.1.2 to evaluate the effect of cord blood lead exposure on children's somatic growth.

Chapter 4 focuses on the challenge that the classical hypothesis testing approach becomes computationally prohibitive when it is used to evaluate all possible subsets of common regression parameters across different studies. We address this challenge by applying the method of adaptive fused lasso in which penalties are imposed on paired differences of parameters between studies. The regularization procedure will automatically detect common parameters across all or subsets of studies and provide estimates of distinctive parameters. Through simulation studies we show that the proposed method performs well to consistently identify common and distinctive parameters. We illustrate our method through HIV surveillance cohorts collected in China to examine the presence of common covariate effects over five geographic regions where these cohorts are sampled.

The thesis concludes with the perspective of future work. Appendices are included to provide the detailed technical proofs.

CHAPTER II

Quadratic Inference Function Approach to Merging Longitudinal Studies: Validation Test and Joint Estimation

2.1 Introduction

Merging data from clinical trials or longitudinal cohort studies with identical or similar protocols can offer a powerful way to better understand effects of treatment and exposure on patient outcomes (e.g. *Localio et al.*, 2001; *Xie and Ahn*, 2010). Appropriate data merging achieves a desired statistical power. A well-known approach for this practice is meta analysis (e.g. *Becker*, 2007; *Hartung et al.*, 2008). But meta analysis often utilises summary statistics from individual analysis with no or little justification provided on the validity of data merging. When the original datasets are fully available, a statistical model incorporating interaction terms between studies and covariates of interest may be used to characterise different effect sizes of covariates across studies. However, in such analysis most existing approaches use a common correlation structure and a common dispersion parameter for different studies. According to *Crowder* (1995), misspecification of working correlation structures, particularly for multiple longitudinal studies, may inflate type I errors and distort power.

Another widely used approach is to model cross-study heterogeneity of regression coefficients through a mixed-effects model (e.g. *Laird and Ware*, 1982; *Zhang et al.*, 2009). This requires a relatively large number of studies and correct distribution assumptions in order to adequately estimate the cross-study variability (*Follmann and Proschan*, 1999). In addition, the general theory regarding tests for nonzero variance components is difficult to apply (*Stram and Lee*, 1994; *Crainiceanu and Ruppert*, 2004), especially for non-normal data (*Zhang and Lin*, 2008).

Breslow and Day's (1980) test for homogeneity of conditional odds ratios is a classical example of validation prior to the calculation of the common odds ratio for multiple strata. In this chapter, we consider longitudinal studies that collect the same types of variables under similar protocols. We develop a novel quadratic inference function (*Qu et al.*, 2000) strategy to validate longitudinal data merging by testing for the unbiasedness of the generalised estimating functions under a common set of regression coefficients but with possibly different covariance structures. The unbiasedness implies that study-specific estimating functions are compatible with a shared regression mean model, so that the resulting analysis of merged data would lead to consistent estimators of regression coefficients and robust control of type I error against covariance misspecification.

We organize the chapter as follows. After presenting the formulation in Section 2.2, in Section 2.3 we propose two testing statistics and establish their asymptotic distributions under the null hypothesis of homogeneity of regression parameter. In Section 2.4, we provide theoretical justifications for the efficiency improvement in estimation of both common and study-specific regression coefficients using merged data. Section 2.5 concerns simulation studies, in which the proposed methodologies are compared with several current popular methods. A data analysis is illustrated in Section 2.6, and a few concluding remarks are presented in Section 2.7. All technical details are listed in the Appendix.

2.2 Formulation

We consider $K \geq 2$ longitudinal studies that collect a common set of variables under similar study protocols. Let $Y_{k,ij}$ be the outcome for jth observation of subject i in study k, and let $\boldsymbol{X}_{k,ij}$ be the corresponding covariate vector for $i = 1, \ldots, n_k$, $j = 1, \ldots, m_k$ and $k = 1, \ldots, K$, where n_k and m_k are the numbers of subjects and the numbers of observations on each subject in study k, respectively. We assume a marginal model for outcome $Y_{k,ij}$, consisting of conditional mean model $E(Y_{k,ij} \mid \boldsymbol{X}_{k,ij}) = \mu_{k,ij} = h(\boldsymbol{X}_{k,ij}^T \boldsymbol{\beta}_k)$ and conditional variance $\text{var}(Y_{k,ij} \mid \boldsymbol{X}_{k,ij}) = \phi_k v(\mu_{k,ij})$, where $h(\cdot)$ is a known link function, $\boldsymbol{\beta}_k$ is a p-dimensional regression parameter, $v(\cdot)$ is a known unit variance function, and ϕ_k is a dispersion parameter. The within-subject correlation is accommodated via a working correlation matrix $\boldsymbol{R}_k(\alpha_k)$, as suggested by *Zeger et al.* (1988), where α_k is the correlation parameter of study k.

For study k, $k = 1, \cdots, K$, an estimator of $\boldsymbol{\beta}_k$ from generalised estimating equations solves

$$n_k^{-1} \sum_{i=1}^{n_k} \boldsymbol{D}_{k,i}^T \boldsymbol{A}_{k,i}^{-1/2} \boldsymbol{R}_k^{-1}(\alpha_k) \boldsymbol{A}_{k,i}^{-1/2} (\boldsymbol{Y}_{k,i} - \boldsymbol{\mu}_{k,i}) = \boldsymbol{0}, \tag{2.1}$$

where $\boldsymbol{Y}_{k,i} = (Y_{k,i1}, \ldots, Y_{k,im_k})^T$, $\boldsymbol{\mu}_{k,i} = (\mu_{k,i1}, \ldots, \mu_{k,im_k})^T$, $\boldsymbol{D}_{k,i} = \partial \boldsymbol{\mu}_{k,i} / \partial \boldsymbol{\beta}_k^T$ and $\boldsymbol{A}_{k,i} = \text{diag}\{v(\mu_{k,i1}), \ldots, v(\mu_{k,im_k})\}$. To deal with merged data, we propose to utilise the quadratic inference function method (*Qu et al.*, 2000) to join the above study-specific estimating functions. A quadratic inference function is derived via approximating the inverse working correlation matrix by $\boldsymbol{R}_k^{-1}(\alpha_k) \approx \sum_{s=1}^{s_k} a_{k,s} \boldsymbol{M}_{k,s}$ for $k = 1, \ldots, K$, where $a_{k,1}, \ldots, a_{k,s_k}$ are constants possibly dependent on the correlation parameter α_k and $\boldsymbol{M}_{k,1}, \ldots, \boldsymbol{M}_{k,s_k}$ are known basis matrices with elements 0 and 1, which are determined by a given correlation matrix $\boldsymbol{R}_k(\alpha_k)$. *Qu et al.* (2000) give more details on the forms of basis matrices in some widely used correlation matrices.

Plugging the expansion of $\boldsymbol{R}_k^{-1}(\alpha_k)$ into (2.1) leads to

$$n_k{}^{-1}\sum_{i=1}^{n_k}\sum_{j=1}^{s_k} a_{k,j}\boldsymbol{D}_{k,i}^T\boldsymbol{A}_{k,i}^{-1/2}\boldsymbol{M}_{k,j}\boldsymbol{A}_{k,i}^{-1/2}(\boldsymbol{Y}_{k,i}-\boldsymbol{\mu}_{k,i}) = \boldsymbol{0} \quad (k=1,\ldots,K),$$

which may be regarded as a combination of elements of the extended score vector

$$\bar{\boldsymbol{g}}_k(\boldsymbol{\beta}_k) = n_k{}^{-1}\sum_{i=1}^{n_k}\boldsymbol{g}_{k,i}(\boldsymbol{\beta}_k) = n_k{}^{-1}\sum_{i=1}^{n_k}\begin{pmatrix}\boldsymbol{D}_{k,i}^T\boldsymbol{A}_{k,i}^{-1/2}\boldsymbol{M}_{k,1}\boldsymbol{A}_{k,i}^{-1/2}(\boldsymbol{Y}_{k,i}-\boldsymbol{\mu}_{k,i})\\ \vdots \\ \boldsymbol{D}_{k,i}^T\boldsymbol{A}_{k,i}^{-1/2}\boldsymbol{M}_{k,s_k}\boldsymbol{A}_{k,i}^{-1/2}(\boldsymbol{Y}_{k,i}-\boldsymbol{\mu}_{k,i})\end{pmatrix}. \tag{2.2}$$

Unlike generalised estimating equations, a quadratic inference function does not need to estimate nuisance coefficients $a_{k,1},\ldots,a_{k,s_k}$ in order to estimate parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_K^T)^T$ of interest.

Define study indicator $\delta_i(k)$, with 1 indicating that subject i belongs to study k and 0 otherwise. For the merged longitudinal data, $\boldsymbol{\beta}$ can be estimated by $\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} Q(\boldsymbol{\beta})$ where

$$Q(\boldsymbol{\beta}) = n\bar{\boldsymbol{g}}(\boldsymbol{\beta})^T\boldsymbol{C}^{-}(\boldsymbol{\beta})\bar{\boldsymbol{g}}(\boldsymbol{\beta}), \tag{2.3}$$

with $n=\sum_{k=1}^K n_k$, and

$$\bar{\boldsymbol{g}}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}(\ \delta_i(1)\boldsymbol{g}_{1,i}(\boldsymbol{\beta}_1)^T \quad ,\ldots,\quad \delta_i(K)\boldsymbol{g}_{K,i}(\boldsymbol{\beta}_K)^T\ )^T = n^{-1}\sum_{i=1}^{n}\boldsymbol{g}_i(\boldsymbol{\beta}),$$
$$\boldsymbol{C}(\boldsymbol{\beta}) = n^{-1}\sum_{i=1}^{n}\mathrm{diag}\{\delta_i(1)\boldsymbol{g}_{1,i}(\boldsymbol{\beta}_1)\boldsymbol{g}_{1,i}(\boldsymbol{\beta}_1)^T,\ldots,\delta_i(K)\boldsymbol{g}_{K,i}(\boldsymbol{\beta}_K)\boldsymbol{g}_{K,i}(\boldsymbol{\beta}_K)^T\}.$$

Here $\boldsymbol{C}(\boldsymbol{\beta})$ is a block-diagonal matrix under the assumption of mutually independent study cohorts, which however may be relaxed in the case of related cohorts. We adopt the unique Moore Penrose generalised inverse in equation (2.3) to enhance numerical stability, as matrix $\boldsymbol{C}(\boldsymbol{\beta})$ may become singular in some practical cases (*Hu and Song*, 2012). See Lemma 9.2.6 in (*Harville*, 2008) for the construction of such inverse operation.

2.3 Homogeneity Test

We develop methods of hypothesis test for global and partial homogeneity of regression parameters across multiple studies. By homogeneity, we mean the equality of regression parameters across all studies, including global homogeneity $\boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_K$ or equality on a subset of coefficients for partial homogeneity. To derive asymptotic distributions of the proposed test statistics, we assume the study-specific mean models are correctly specified, so $\boldsymbol{\beta}_k$ can be consistently estimated in the corresponding individual study k.

Let $\mathcal{M} \subset \{1, \ldots, p\}$ denote an index set, and then $|\mathcal{M}|$ denotes the number of elements in $\mathcal{M}$, or the cardinality of $\mathcal{M}$. Accordingly, $\boldsymbol{\beta}_k(\mathcal{M})$ and $\boldsymbol{\beta}_k(\mathcal{M}^c)$ are subsets of parameters indexed by $\mathcal{M}$ and its complementary set $\mathcal{M}^c$, respectively. Clearly, set $\mathcal{M} = \{1, \ldots, p\}$ corresponds to a global homogeneity, while a partial homogeneity is given by $\mathcal{M}$ being a certain subset of $\{1, \ldots, p\}$.

To test hypotheses $H_0 : \boldsymbol{\beta}_1(\mathcal{M}) = \cdots = \boldsymbol{\beta}_K(\mathcal{M})$ versus $H_a : \boldsymbol{\beta}_i(\mathcal{M}) \neq \boldsymbol{\beta}_j(\mathcal{M})$ for some $i \neq j$ and $i, j \in \{1, \ldots, K\}$, let $\Omega_0(\mathcal{M}) = \{(\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_K^T)^T : \boldsymbol{\beta}_1(\mathcal{M}) = \cdots = \boldsymbol{\beta}_K(\mathcal{M}), \boldsymbol{\beta}_k \in R^p, k = 1, \ldots, K\}$ be the null parameter space under H_0, and let Ω be the whole parameter space. Estimators of $\boldsymbol{\beta}$ under $\Omega_0(\mathcal{M})$ and under Ω are, respectively

$$\hat{\boldsymbol{\beta}}_{\Omega_0(\mathcal{M})} = \underset{\boldsymbol{\beta} \in \Omega_0(\mathcal{M})}{\arg\min}\, Q(\boldsymbol{\beta}), \text{ and } \hat{\boldsymbol{\beta}}_{\Omega} = \underset{\boldsymbol{\beta} \in \Omega}{\arg\min}\, Q(\boldsymbol{\beta}), \tag{2.4}$$

where $Q(\boldsymbol{\beta})$ is given by (2.3) with the corresponding parametrisation. Under H_0, the following Theorem II.1 establishes the asymptotic distribution of $Q(\hat{\boldsymbol{\beta}}_{\Omega_0(\mathcal{M})})$ provided any root-n consistent estimator $\hat{\boldsymbol{\beta}}_{\Omega_0(\mathcal{M})}$.

Theorem II.1. *Let $\hat{\boldsymbol{\beta}}_{\Omega_0(\mathcal{M})}$ be a root-n consistent estimator of true parameter $\boldsymbol{\beta}_0$*

under H_0. Suppose the following regularity conditions hold: (a) $\boldsymbol{\beta}_0 \in$ interior of $\mathcal{B} \subset \mathcal{R}^{Kp-(K-1)|\mathcal{M}|}$, and $\mathcal{B}$ is compact; (b) $\boldsymbol{g}_i(\boldsymbol{\beta})$ is continuously differentiable in a neighborhood $\mathcal{N}$ of true $\boldsymbol{\beta}_0$; (c) $E\{\boldsymbol{g}_i(\boldsymbol{\beta})\} = 0$ if and only if $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $E\{||\boldsymbol{g}_i(\boldsymbol{\beta}_0)||^2\}$ is finite; (d) $E\{\sup_{\boldsymbol{\beta}\in\mathcal{N}} ||\partial \boldsymbol{g}_i(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^T||\} < \infty$; (e) $n^{1/2}\bar{g}(\boldsymbol{\beta}_0)$ converges to $N(\mathbf{0}, \boldsymbol{\Sigma})$ in distribution, where $\boldsymbol{\Sigma}$ is a block-diagonal matrix, $\boldsymbol{\Sigma} = \mathrm{diag}(\rho_1^{-1}\boldsymbol{\Sigma}_1, \ldots, \rho_K^{-1}\boldsymbol{\Sigma}_K)$, with $\boldsymbol{\Sigma}_k = \mathrm{cov}\{\boldsymbol{g}_{k,i}(\boldsymbol{\beta}_0)\}$ and $\rho_k = \lim_{n\to\infty} n_k/n$ for $k = 1, \ldots, K$; (f) $\boldsymbol{G}^T\boldsymbol{\Sigma}^-\boldsymbol{G}$ is nonsingular, where $\boldsymbol{G} = E\{\partial\boldsymbol{g}_i(\boldsymbol{\beta}_0)/\partial\boldsymbol{\beta}^T\}$; and (g) $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^-\boldsymbol{G} = \boldsymbol{G}$. Then, $Q(\hat{\boldsymbol{\beta}}_{\Omega_0(\mathcal{M})})$ converges in distribution to $\chi^2_{\mathrm{rank}(\boldsymbol{\Sigma})-Kp+(K-1)|\mathcal{M}|}$.

Here $||\cdot||$ denotes the Euclidean norm. The proof of Theorem II.1 is given in the appendix. Since the matrix $\boldsymbol{\Sigma}$ may not be of full rank, the degrees of freedom of $Q(\hat{\boldsymbol{\beta}}_{\Omega_0(\mathcal{M})})$ take the form of $\mathrm{rank}(\boldsymbol{\Sigma}) - Kp + (K-1)|\mathcal{M}|$, where $\mathrm{rank}(\boldsymbol{\Sigma})$ can be estimated from orthogonal triangularisation of an estimated $\boldsymbol{\Sigma}$.

When all study-specific mean models are correctly specified, $\hat{\boldsymbol{\beta}}_\Omega$ is a root-n consistent estimator of $\boldsymbol{\beta}_0$. Under the null hypothesis, $\bar{\boldsymbol{g}}(\boldsymbol{\beta})$ is an unbiased estimating function for $\boldsymbol{\beta} \in \Omega_0(\mathcal{M})$, so we can obtain another root-n consistent estimator $\hat{\boldsymbol{\beta}}_{\Omega_0(\mathcal{M})}$ of $\boldsymbol{\beta}_0$. Therefore, we propose two test statistics. The first is $Q(\hat{\boldsymbol{\beta}}_{\Omega_0(\mathcal{M})})$, mimicking the score type test statistic, denoted as $\hat{Q}_S$. Its asymptotic chi-square distribution under H_0 is shown in Theorem II.1. The second is $Q(\hat{\boldsymbol{\beta}}_{\Omega_0(\mathcal{M})}) - Q(\hat{\boldsymbol{\beta}}_\Omega)$, which resembles the likelihood ratio type test statistic, denoted as $\hat{Q}_{LR}$. The asymptotic distribution of $\hat{Q}_{LR}$ is given as follows.

Corollary II.2. *Under the regularity conditions in Theorem II.1, test statistic $\hat{Q}_{LR}$ converges in distribution to $\chi^2_{(K-1)|\mathcal{M}|}$.*

2.4 Joint Estimation with Merged Data

When either global or partial homogeneity is established, the merged data will lead to efficiency improvement in estimation for both common and study-specific regression coefficients. To elucidate, without loss of generality, we consider a case of partial homogeneity. Let $\boldsymbol{\zeta}$ denote a vector of common coefficients for covariates $\boldsymbol{X}_{k,ij}$ shared by the studies and let $\boldsymbol{\gamma}_k$ denote study-specific parameters associated with covariates $\boldsymbol{Z}_{k,ij}$. Then $\boldsymbol{\beta} = (\boldsymbol{\zeta}^T, \boldsymbol{\gamma}_1^T, \ldots, \boldsymbol{\gamma}_K^T)^T$ represents all the parameters and $\boldsymbol{\beta}_k = (\boldsymbol{\zeta}^T, \boldsymbol{\gamma}_k^T)^T$ contains parameters in study k only. Accordingly, the mean model is rewritten as $E(Y_{k,ij} \mid \boldsymbol{X}_{k,ij}, \boldsymbol{Z}_{k,ij}) = \mu_{k,ij} = h(\boldsymbol{X}_{k,ij}^T\boldsymbol{\zeta} + \boldsymbol{Z}_{k,ij}^T\boldsymbol{\gamma}_k)$ $(k = 1, \ldots, K)$. Consequently, we obtain a consistent estimate $\hat{\boldsymbol{\beta}}$ by minimising function in (2.3) with the merged data. Under Assumptions (a)-(f) of Theorem II.1, as shown in the appendix, $n^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ converges in distribution to $\mathrm{N}\{0, (\boldsymbol{G}^T\boldsymbol{\Sigma}^-\boldsymbol{G})^{-1}\}$ with $\boldsymbol{\Sigma}$ defined in Theorem II.1 and

$$\boldsymbol{G} = E\left\{\frac{\partial \boldsymbol{g}_i(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\beta}^T}\right\} = \begin{pmatrix} \boldsymbol{G}_1 \\ \vdots \\ \boldsymbol{G}_K \end{pmatrix}$$
$$= \begin{bmatrix} E\{\frac{\partial \boldsymbol{g}_{1,i}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\zeta}^T}\} & E\{\frac{\partial \boldsymbol{g}_{1,i}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\gamma}_1^T}\} & \ldots & \ldots & \boldsymbol{0} \\ E\{\frac{\partial \boldsymbol{g}_{2,i}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\zeta}^T}\} & \boldsymbol{0} & E\{\frac{\partial \boldsymbol{g}_{2,i}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\gamma}_2^T}\} & \ldots & \boldsymbol{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ E\{\frac{\partial \boldsymbol{g}_{K,i}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\zeta}^T}\} & \boldsymbol{0} & \ldots & \ldots & E\{\frac{\partial \boldsymbol{g}_{K,i}(\boldsymbol{\beta}_0)}{\partial \boldsymbol{\gamma}_K^T}\} \end{bmatrix}.$$

To explicate the efficiency gain in the merged data analysis, we focus on parameter $\boldsymbol{\beta}_k$. Let $\tilde{\boldsymbol{\beta}}_k$ be an estimator obtained by minimising function in (2.3) only using the kth study data and $\hat{\boldsymbol{\beta}}_k$ be the subvector of $\hat{\boldsymbol{\beta}}$, obtained with the merged data. The root-n asymptotic variance for $\tilde{\boldsymbol{\beta}}_k$ is $\{\rho_k(\boldsymbol{G}_k^T\boldsymbol{\Sigma}_k^-\boldsymbol{G}_k)_{[\boldsymbol{\beta}_k,\boldsymbol{\beta}_k]}\}^{-1}$, where $\boldsymbol{G}_k$ is the kth block-row of matrix $\boldsymbol{G}$ above, $\boldsymbol{\Sigma}_k$ is defined in Theorem II.1, and $\boldsymbol{B}_{[\boldsymbol{\beta}_k,\boldsymbol{\beta}_k]}$ denotes the sub-block matrix of $\boldsymbol{B}$ with rows and columns selected by those elements corresponding to $\boldsymbol{\beta}_k$. The asymptotic variance for $\hat{\boldsymbol{\beta}}_k$ is $\{(\boldsymbol{G}^T\boldsymbol{\Sigma}^-\boldsymbol{G})^{-1}\}_{[\boldsymbol{\beta}_k,\boldsymbol{\beta}_k]}$. Theorem II.3 below establishes the efficiency improvement achieved through the joint estimation with merged data.

Theorem II.3. *Suppose for study $k = 1, \ldots, K$, $(\boldsymbol{G}_k^T \boldsymbol{\Sigma}_k^- \boldsymbol{G}_k)_{[\boldsymbol{\beta}_k, \boldsymbol{\beta}_k]}$ is positive definite. Then the asymptotic variances of $\hat{\boldsymbol{\beta}}_k$ and $\tilde{\boldsymbol{\beta}}_k$ satisfy*

$$\{(\boldsymbol{G}^T \boldsymbol{\Sigma}^- \boldsymbol{G})^{-1}\}_{[\boldsymbol{\beta}_k, \boldsymbol{\beta}_k]} \preceq \frac{1}{\rho_k} \{(\boldsymbol{G}_k^T \boldsymbol{\Sigma}_k^- \boldsymbol{G}_k)_{[\boldsymbol{\beta}_k, \boldsymbol{\beta}_k]}\}^{-1} \quad (k = 1, \ldots, K),$$

where $\preceq$ is in the sense of Löwner's partial ordering in the space of non-negative definite matrices.

The proof of Theorem II.3 is given in the appendix. Theorem II.3 implies that the asymptotic variance of $\hat{\boldsymbol{\beta}}_k$ is smaller than that of $\tilde{\boldsymbol{\beta}}_k$. This suggests that the estimation with the merged data is not only flexible to accommodate different study-specific correlations and follow-up schedules but also leads to estimation efficiency gain on the regression coefficients. This efficiency benefit is not easily achieved by meta analysis, in which the effective sample size is not really increased from combining individual analyses. Moreover, when additional nuisance parameters are introduced into the joint estimation procedure in generalised estimating equations to account for study-specific covariance parameters, the efficiency gain is not guaranteed for the estimation of parameters of interest. This is because even though the merged data have more samples, the number of nuisance parameters gets increased too, which can offset the benefit from increased sample sizes.

2.5 Simulation Study

Two simulation studies were conducted to investigate the finite sample performance of our proposed tests and to compare with Wald-type tests using the method of generalised estimating equations. We consider several versions of Wald-type test statistics, denoted by W_{zla}, W_{p}, W_{md} and W_{wl}. They are computed by utilising different sandwich variance estimators proposed by *Zeger et al.* (1988), *Pan* (2001), *Mancl and DeRouen* (2001) and *Wang and Long* (2011), respectively. Wald-type tests are applied to test for no interactions between study dummy covariates and covariates of interest under a common correlation structure for multiple studies. Technically speaking, these approaches may be modified to accommodate study-specific covariance matrices, but the resulting methods require iteratively updating regression parameters and study-specific covariance nuisance parameters, so their performances will be affected by the estimation of nuisance parameters. In this chapter we do not implement such extended estimating equation approaches but focus on using robust sandwich variance estimators to account for the covariance heterogeneity across multiple studies.

For meta analysis, we adopt the Cochran's test for partial homogeneity. According to *Hartung et al.* (2008), the Cochran's test is approximately distributed as χ^2_{K-1} under the null hypothesis of homogeneous coefficients across all K studies. Similarly as for the Wald-type tests, we use T_{zla}, T_{p}, T_{md} and T_{wl} to denote the Cochran's test with the corresponding robust sandwich variance estimators.

Besides meta analysis and Wald-type tests, we also consider homogeneity test using mixed-effects models by testing zero variance components of random slopes. In the linear mixed-effects model, the asymptotic distribution of a likelihood ratio test for one zero variance component is $0.5\chi^2_0 + 0.5\chi^2_1$ (*Stram and Lee*, 1994), while in a generalised linear mixed model with nonidentity link functions, such mixtures of chi-squares for likelihood ratio tests are hard to obtain (*Fitzmaurice et al.*, 2007; *Sinha*, 2009). In our simulation studies, because data are generated by the population-average model with some pre-fixed correlation structures, tests for zero variance components cannot control type I error at all. Thus, we do not include results from the mixed-effects models in the comparison.

The first simulation study is generated by a population-average linear model $Y_{k,ij} = \beta_{k,0} + \beta_{k,1}X_{k,ij} + \beta_{k,2}Z_{k,ij} + \varepsilon_{k,ij}$ for $j = 1, \ldots, m_k$, $i = 1, \ldots, n_k$ with $m_k = 8$, $n_k = 100$ and $k = 1, \ldots, K$. Covariate vector $\boldsymbol{Z}_{k,i} = (Z_{k,i1}, \ldots, Z_{k,im_k})^T$

is a time dependent variable simulated from multivariate normal distribution with mean vector $(1,\ldots,m_k)^T$ and the identity covariance matrix $\boldsymbol{I}_{m_k}$. Covariate vector $\boldsymbol{X}_{k,i} = (X_{k,i1},\ldots,X_{k,im_k})^T$ is a time independent (baseline) covariate generated from exponential distribution with rate parameter 4. The error terms $\boldsymbol{\varepsilon}_{k,i} = (\varepsilon_{k,i1},\ldots,\varepsilon_{k,im_k})^T$ follow $\mathrm{N}\{\boldsymbol{0},\phi_k\boldsymbol{R}_k(\alpha_k)\}$ with correlation matrix $\boldsymbol{R}_k(\alpha_k)$. Denote all correlation parameters and dispersion parameters by $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_K)^T$ and $\boldsymbol{\phi} = (\phi_1,\ldots,\phi_K)^T$, respectively. And denote the order-1 autoregressive and compound symmetric correlations by $\boldsymbol{R}_{ar}$ and $\boldsymbol{R}_{cs}$ respectively. We consider three cases with varying number of studies and covariances: (i) $K = 4$, $\boldsymbol{\phi} = (50,10,4,1)^T$, $\boldsymbol{\alpha} = (0.7,0.4,0.2,0.1)^T$, and $\{\boldsymbol{R}_1(\cdot),\boldsymbol{R}_2(\cdot),\boldsymbol{R}_3(\cdot),\boldsymbol{R}_4(\cdot)\} = \{\boldsymbol{R}_{ar},\ \boldsymbol{R}_{cs},\ \boldsymbol{R}_{cs},\ \boldsymbol{R}_{ar}\}$; (ii) $K = 3$, $\boldsymbol{\phi} = (10,4,1)^T$, $\boldsymbol{\alpha} = (0.7,0.2,0.1)^T$, $\{\boldsymbol{R}_1(\cdot),\boldsymbol{R}_2(\cdot),\boldsymbol{R}_3(\cdot)\} = \{\boldsymbol{R}_{ar},\ \boldsymbol{R}_{cs},\ \boldsymbol{R}_{ar}\}$; (iii) $K = 2$, $\boldsymbol{\phi} = (10,1)^T$, $\boldsymbol{\alpha} = (0.7,0.2)^T$, and $\{\boldsymbol{R}_1(\cdot),\boldsymbol{R}_2(\cdot)\} = \{\boldsymbol{R}_{ar},\ \boldsymbol{R}_{cs}\}$. Let $\boldsymbol{\beta}_k = (\beta_{k,0},\beta_{k,1},\beta_{k,2})^T$ for $k = 1,\ldots,K$. We are interested in a global test $H_0 : \boldsymbol{\beta}_1 = \cdots = \boldsymbol{\beta}_K$ and a partial test concerning only the coefficients of $\boldsymbol{X}_{k,ij}$, $H_0 : \beta_{1,1} = \cdots = \beta_{K,1}$. Type I errors are computed with $\boldsymbol{\beta}_k = (-1,-2,3)^T$ for $k = 1,\ldots,K$, while power is calculated under $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_3 = (-1,-2,3)^T$ and $\boldsymbol{\beta}_2 = (-1,-1.85,3)^T$. In the use of Wald-type tests for zero interaction effects between covariates and study indicators, only coefficients of interaction terms will be involved in the test.

Table 2.1 summarises type I errors and power of all test statistics at a significance level of 0.05 over 4000 replications. For a fair comparison, we compute Wald-type tests, meta analyses and our proposed tests under a common correlation structure of $\boldsymbol{R}_{ar}$ or $\boldsymbol{R}_{cs}$. The results clearly shows that no matter which working correlation structure is used, our proposed tests, $\hat{Q}_{LR}$ and $\hat{Q}_S$, can properly control type I error rates. In contrast, Wald-type tests and meta analyses cannot, particularly for global homogeneity tests and for $K > 2$. Wald-type tests have inflated type I error rates, mostly because the modified robust variance estimators still underestimate variances

Table 2.1: Average type I error rates and power of test statistics of the proposed (Q's), Wald-type (W's) and meta-based (T's) versions over 4000 replications for continuous outcomes from K studies. Upper and lower panels correspond to the global and partial homogeneity tests respectively. Two correlations are used: order-1 autoregression, $\boldsymbol{R}_{ar}$, and compound symmetry, $\boldsymbol{R}_{cs}$.

	$K = 4$				$K = 3$				$K = 2$			
	Size %		Power %		Size %		Power %		Size %		Power %	
Test	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$
$\hat{Q}_S$	4.1	4.1	27.0	38.7	4.8	4.6	85.4	95.2	4.5	4.0	50.5	65.7
$\hat{Q}_{LR}$	4.0	4.2	42.4	43.4	4.4	4.1	95.9	96.6	4.0	4.3	73.6	72.3
W_{zla}	9.6	9.6	79.0	80.2	8.8	8.3	99.9	100	7.1	7.4	91.3	90.1
W_{p}	11.4	12.1	42.5	40.0	8.5	9.2	95.2	94.8	5.8	5.9	90.8	89.4
W_{md}	6.7	7.0	71.2	73.3	6.6	6.6	99.8	99.9	5.7	6.0	88.0	87.0
W_{wl}	10.2	11.0	39.4	36.8	7.5	8.2	94.6	94.0	5.4	5.2	89.4	88.8
$\hat{Q}_{LR}$	4.7	4.6	71.1	72.2	5.0	4.7	98.8	98.8	5.3	5.3	87.4	87.0
W_{zla}	8.4	8.2	88.8	90.0	7.4	7.3	100	100	7.2	7.1	96.5	95.5
W_{p}	9.0	8.8	62.1	60.4	7.2	7.6	99.4	99.5	6.1	6.0	96.4	95.6
W_{md}	6.2	6.4	85.0	86.6	5.4	5.6	100	100	5.5	5.8	95.0	94.1
W_{wl}	8.5	8.4	60.5	58.7	6.9	7.2	99.3	99.5	5.5	5.4	96.1	95.4
T_{zla}	9.5	9.8	65.0	65.7	8.2	8.2	99.2	99.1	7.6	7.6	78.7	76.9
T_{p}	5.7	5.4	60.2	60.6	5.7	5.8	99.2	99.2	5.4	5.3	77.0	74.9
T_{md}	6.2	6.5	56.4	57.0	5.6	6.0	98.6	98.6	5.7	5.8	72.7	71.8
T_{wl}	4.9	4.8	57.9	58.6	5.1	5.0	99.0	99.0	4.8	4.7	75.2	73.5

of regression coefficients and cannot fully account for differences among covariance structures across studies. Meta analyses appear to have less inflated type I errors than Wald-type tests, but since meta analyses cannot sufficiently utilise all data information, they tend to have lower power.

The second simulation study concerns binary outcome $Y_{k,ij}$, which is generated from a population-average logistic model $\text{logit}\{E(Y_{k,ij} \mid Z_{k,ij})\} = \beta_{k,0} + \beta_{k,1} Z_{k,ij}$ for $j = 1, \ldots, 8$, $i = 1, \ldots, 100$ and $k = 1, \ldots, K$. $Z_{k,ij}$ is generated from Unif$(0, 1)$ distribution. We consider the same global and partial homogeneity hypotheses as those in the first simulation study. Within-subject correlations have three cases: (i) $K = 4$, $\boldsymbol{\alpha} = (0.7, 0.4, 0.2, 0.1)^T$, and $\{\boldsymbol{R}_1(\cdot), \boldsymbol{R}_2(\cdot), \boldsymbol{R}_3(\cdot), \boldsymbol{R}_4(\cdot)\} = \{\boldsymbol{R}_{ar}, \boldsymbol{R}_{cs}, \boldsymbol{R}_{cs}, \boldsymbol{R}_{ar}\}$; (ii) $K = 3$, $\boldsymbol{\alpha} = (0.7, 0.2, 0.1)^T$, and $\{\boldsymbol{R}_1(\cdot), \boldsymbol{R}_2(\cdot), \boldsymbol{R}_3(\cdot)\} = \{\boldsymbol{R}_{ar}, \boldsymbol{R}_{cs}, \boldsymbol{R}_{ar}\}$; (iii) $K = 2$, $\boldsymbol{\alpha} = (0.7, 0.2)^T$, and $\{\boldsymbol{R}_1(\cdot), \boldsymbol{R}_2(\cdot)\} = \{\boldsymbol{R}_{ar}, \boldsymbol{R}_{cs}\}$. Type I errors are computed with $\boldsymbol{\beta}_k = (-0.2, 1.5)^T$ for all $k = 1, \ldots, K$, while the power is calculated under $\boldsymbol{\beta}_2 = (-0.2, 2.5)^T$ and $\boldsymbol{\beta}_k = (-0.2, 1.5)^T$ for $k \neq 2$.

Table 2.2 presents results summarised over 4000 replications at significance level 0.05. Similar conclusions are drawn to those obtained in the case of continuous outcomes. Wald-type tests and meta analyses both produce inflated type I errors. For instance, type I error of W_{p} appears above 7% when 4 studies are considered regardless of working correlation structure used. Among all Wald-type tests, the one, W_{md}, based on Mancl & DeRouen's (2001) sandwich variance estimator seems to have a reasonable control of type I error. To deal with the violation of a common correlation structure, their method strives to reduce the bias in estimation of the covariance matrix of data while the other methods focus on improving correlation matrix estimation. To compare power of Mancl & DeRouen's test to our test, a ratio, (Power of W_{md})/(Power of $\hat{Q}_{LR}$), decreases as the number of studies increases, dropping from 98.6% in the case of two studies to 88.8% in the case of four studies for the global homogeneity test under $\boldsymbol{R}_{ar}$ working correlation. This implies that although Mancl & DeRouen's method can correct for the bias in the covariance estimation, it is inferior to our $\hat{Q}_{LR}$ test in terms of power. Since meta analysis does not utilise data from all individual studies efficiently, it has lower power than our methods even when its type I error is properly controlled. Finally, since the degrees of freedom of a chi-square test statistic increase when the number of studies increases, our test statistics may lose power in the setting of many studies. In this scenario, we recommend using mixed-effects models to handle merged data if distribution assumptions for multiple studies can be properly pre-specified.

2.6 Application

We illustrate the proposed methodology by a community-based clinical trial that aims to evaluate the effectiveness of a continuing medical education program, Physi-

Table 2.2: Average type I error rates and power of test statistics of the proposed (Q's), Wald-type (W's) and meta-based (T's) versions over 4000 replications for binary outcomes from K studies. Upper and lower panels correspond to the global and partial homogeneity tests respectively. Two correlations are used: order-1 autoregression, $\boldsymbol{R}_{ar}$, and compound symmetry, $\boldsymbol{R}_{cs}$.

	$K=4$				$K=3$				$K=2$			
	Size %		Power %		Size %		Power %		Size %		Power %	
Test	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$	$\boldsymbol{R}_{ar}$	$\boldsymbol{R}_{cs}$
$\hat{Q}_S$	4.8	4.9	65.2	66.9	5.0	4.9	79.8	79.6	5.1	5.4	57.0	51.0
$\hat{Q}_{LR}$	5.2	5.1	86.0	82.6	4.8	4.9	92.0	91.3	5.1	5.5	78.7	71.0
W_{zla}	5.2	5.5	73.6	76.6	5.5	4.9	89.3	90.5	6.0	5.9	78.6	76.5
W_{p}	7.8	8.1	81.6	81.6	7.6	6.1	90.1	89.4	6.3	6.6	79.5	77.5
W_{md}	4.5	4.8	71.6	75.2	5.0	4.6	88.6	89.7	5.5	5.4	77.6	75.7
W_{wl}	7.4	7.8	80.5	80.7	7.1	5.9	89.5	88.9	6.0	6.3	79.1	77.0
$\hat{Q}_{LR}$	5.3	5.0	88.2	89.4	5.2	5.3	95.6	95.4	5.0	4.9	86.0	80.8
W_{zla}	5.1	4.9	82.5	85.9	5.4	5.2	94.4	95.2	5.5	4.8	85.2	83.4
W_{p}	6.4	6.8	87.9	89.6	6.6	6.6	94.6	95.2	5.6	5.8	85.9	84.3
W_{md}	4.7	4.6	81.5	85.0	5.2	4.9	94.1	94.8	5.2	4.3	84.8	82.5
W_{wl}	6.1	6.7	87.5	89.2	6.2	6.4	94.4	95.1	5.4	5.5	85.7	84.0
T_{zla}	5.1	5.4	58.8	65.4	5.4	5.4	70.8	69.3	5.0	5.0	63.2	52.2
T_{p}	10.4	10.2	58.7	64.6	5.7	4.8	70.4	69.2	4.8	4.7	63.2	52.0
T_{md}	4.8	4.8	56.8	63.8	4.9	5.0	69.5	68.3	4.6	4.7	62.4	51.1
T_{wl}	10.3	10.0	58.2	64.4	5.5	4.7	70.1	68.8	4.7	4.6	62.9	51.8

cian Asthma Care Education (PACE) (*Cabana et al.*, 2006), in improving pediatricians' asthma therapeutic and communication skills. This randomized trial was conducted in 10 sites in U.S.A. Four of them are from "southern" states (Texas, California, Tennessee and Florida), while the others are from "northern" states (Nebraska, Minnesota, Michigan, Delaware, Ohio and Indiana). A total of 101 pediatricians and a random sample of 870 asthma patients with age 2 to 12 years old participated in the study. Every patient is observed at baseline, and at one year and two year follow-up periods. The number of nights with trouble sleeping was the primary outcome of interest. The primary scientific objective is to evaluate the effectiveness of intervention, as well as the effects of time, self-evaluated asthma severity on patient's night-time asthma symptoms, and frequency, adjusting for age, gender and baseline asthma symptoms. Self-evaluation of asthma severity contains three categories including 1 (mild), 2 (moderate) and 3 (severe). Covariate frequency is measured as how often in the past 12 months asthma symptoms occur, where symptoms include coughing, wheezing, chest tightness, or shortness of breath interfaced with children's sleep. It is coded in 6 categories (1,2,3,4,5 and 6) in a decreasing order of symptom frequency.

Chronic asthma is known to be affected by multiple environmental and cultural factors. Therefore, patients from different regions may experience different effects of the intervention. To illustrate, we fit the following regression model, separately, using data from the southern states only, data from the northern states only, and the merged data from both southern and northern states,

$$y_{it} = \beta_0 + \beta_1 x_{it}^A + \beta_2 x_{it}^G + \beta_3 t + \beta_4 x_{it}^I + \beta_5 x_{it}^S + \beta_6 x_{it}^F + \beta_7 x_i^B + \epsilon_{it} \quad t = 1, 2,$$

where covariates are age(x_{it}^A), gender(x_{it}^G, 1 for male, 0 for female), time(t=1,2), intervention(x_{it}^I, 1 for receiving PACE, 0 for not receiving PACE), severity(x_{it}^S), frequency(x_{it}^F) and baseline(x_i^B). Table 2.3 reports the estimated effect of intervention as well as those of other covariates across the two regions.

Table 2.3: Regression coefficients estimated by QIF based on data from the southern, northern, combined two regions, as well as meta analysis.

	South		North		Combined		Meta	
	Est.	Pvalue	Est.	Pvalue	Est.	Pvalue	Est.	Pvalue
Intercept	59.86	0.00	74.30	0.00	67.12	0.00	69.10	0.00
Age	-0.12	0.57	-0.15	0.43	-0.14	0.31	-0.14	0.34
Gender	-0.43	0.77	-0.25	0.85	-0.17	0.86	-0.34	0.74
Time	-3.32	0.09	-1.96	0.13	-2.54	0.02	-2.37	0.03
Intervention	-0.61	0.67	1.26	0.29	0.62	0.50	0.48	0.60
Severity	1.07	0.47	-2.26	0.06	-0.74	0.42	-1.02	0.27
Frequency	-9.57	0.00	-11.98	0.00	-10.80	0.00	-11.26	0.00
Baseline	0.22	0.00	0.19	0.00	0.20	0.00	0.21	0.00

As expected, the estimation results in Table 2.3 do not appear to be consistent between southern and northern states. To evaluate if the data merging is proper, we begin by applying our partial homogeneity test proposed in Section 2.3 to identify if there exists a subset of covariates that appear to have the same effects. This proceeds in two steps: (i) to test homogeneity on each regression coefficient across two regions; and (ii) to choose a subset of homogeneous coefficients identified in the first step and then to test whether the chosen set of coefficients is homogeneous across two regions.

At significance level 0.1, two covariates, severity (x_{it}^S) and frequency (x_{it}^F), are chosen as potentially having different effects across the two regions, because their p-values are 0.08 and 0.09(below 0.1), respectively, in the individual homogeneity test. In a test for the homogeneity of severity and frequency jointly, a smaller p-value (0.06) is given by our test. This indicates a marginally accepted homogeneity for severity and frequency between northern and southern states. In contrast, a much larger p-value (0.53) is given when jointly testing the homogeneity of the other covariates: age, gender, time, intervention and baseline. Therefore, in the joint estimation we specify common region coefficients of age, gender, time and intervention, but leave coefficients for intercept, severity and frequency to be different. Table 2.4 lists the results obtained by the proposed joint estimation method.

We compare the region-specific analyses to the naive "combined" analysis, in which all coefficients are naively assumed to be the same, and to the meta analysis. The effect of time is worth noting. Given the increased sample size for this common time effect, time covariate becomes significant in the merged data analysis. This means that the longer in time the intervention of education program is in use, the lower the expected number of nights with trouble sleeping, adjusted by age, gender, and baseline symptom. It is also interesting to notice the differences in the estimation for the effect of self-evaluation of asthma severity. As shown in Table 2.4, self-evaluation of asthma severity is not significantly associated with the number of nights with trouble sleeping for children in the southern region, but significantly for children in the northern region. In the naive "combined" data analysis, as well as the meta analysis reported in Table 2.3, such a significant effect in the northern region has been masked (p-value 0.42 and 0.27, respectively). This means that it is not always valid to perform combined data analysis without recognizing a certain important feature present in a subcohort of subjects. In the north region-specific analysis, this effect is marginally significant (p-value 0.06). Using our proposed joint estimation, with the utility of data from both regions, we are able to establish evidence for significant effect (p-value 0.04) of asthma severity in the northern region.

2.7 Concluding Remarks

Practitioners often combine multiple similar small datasets to hopefully achieve a more powerful statistical analysis. This data merging practice should be cautiously considered because some influential heterogeneous features of individual studies can cause misleading results in a combined data analysis. We developed a data-driven approach to addressing this validation issue in data merging, which is useful in practical studies. By comparing to other popular methods based on meta analysis, generalised estimating equations and mixed-effects model, our methods have shown to control type I error satisfactorily and achieve larger power for the homogeneity test. In addition, our joint estimation procedure provides more efficient estimation of regression coefficients with merged data. When the number of studies becomes large, our methods may be affected with reduced power. In a separate publication, we will provide some procedural guidelines on how to conduct the homogeneity test with multiple covariates.

Table 2.4: Estimated regression coefficients by the joint estimation method using the merged data.

	Covariates	Est.	Pvalue
common in both regions	Age	-0.09	0.47
	Gender	-0.53	0.58
	Time	-2.23	0.02
	Intervention	0.47	0.56
	Baseline	0.20	0.00
only in south	Intercept	57.71	0.00
	Severity	1.45	0.24
	Frequency	-9.67	0.00
only in north	Intercept	74.07	0.00
	Severity	-2.25	0.04
	Frequency	-11.88	0.00

CHAPTER III

Merging Multiple Longitudinal Studies with Study-Specific Missing Covariates: A Joint Estimating Function Approach

3.1 Introduction

Analyzing combined datasets collected from multiple similar studies has been popular in practice in order to achieve greater power in statistical analysis. The increased power is obtained when parameters across multiple study populations are common and therefore can be estimated using more observations with the combined datasets than using each dataset separately. The larger sample size, if properly utilized, will lead to improved performances in both statistical estimation and inference. In addition, combined data potentially provide richer information to answer some questions that otherwise may not be answered using data from each individual study.

Such potential power gain from combined data is subject to additional complexities in study design, data collection as well as data structures, and it is not a free benefit in data analysis. For example, misaligned missing covariates across different studies (e.g. covariates observed in some studies but completely missing in other studies) are difficult to handle. This chapter is motivated by a health study involving multiple longitudinal cohorts gathered in Mexico City, whose aim is to evaluate the effect of children's lead exposures on their somatic growth. This study consists of two birth cohorts established by the same study team from two hospitals in Mexico City, termed as cohort B and cohort C throughout the chapter, respectively. Two lead exposure measures recorded in the study include mother's blood lead exposure (PBL) and child's cord blood lead exposure (CBL), where the former is fully recorded in both cohorts but the latter is only fully measured in cohort C. One of the primary objectives was to assess the association between CBL and child's weight growth, adjusting for other covariates available in both cohorts. Apparently, a key challenge in the analysis of merged data from cohorts B and C pertains to the fact that CBL measurements in cohort B are very heavily missing.

There are several other challenges that also need to be handled properly in the analysis of merged data. Inter-study heterogeneity often gives rise to some complicating factors that may impair the popular working correlation strategy for the modeling covariances of longitudinal data. For instance, data collected from hospitals located in urban areas might be more volatile than those collected from hospitals located in rural areas because hospitals in cities tend to have more diversified patient population. Also, in the above motivating example, children in cohort B were followed repeatedly at months of 0,3,6, 12, 18, 24, 30, 36 and 60 after birth, while children in cohort C were seen according to different times of visits at 0, 1, 4, 7, 12, 18, 24, 30, 36, 42 and 48 months after birth. Ignoring such heterogeneities and using a common covariance structure could lead to inappropriate results (*Wang et al.*, 2012) in the analysis of combined data. Similarly, multi-center clinical trials, even administrated by a common protocol, may still vary in actual operations for data collection, due, for example, to study coordinator's personal effort on retaining patient's follow-up visits. Joint modeling of mean and covariance has been discussed in the literature to account for covariance heterogeneity. For instance, *Pourahmadi* (1999) and *Leng et al.* (2010) proposed models to address covariate-adjusted covariances. However, when these approaches are applied to merged data from multiple studies, both model building and computation may become demanding, where the number of parameters in covariance models can escalate along the increased number of studies. In addition, diagnostic tools for covariance models have been little considered in the literature and misspecified covariance models can lead to incorrect statistical inference and misleading data analysis. All of these, as a result, may offset the benefit of estimation efficiency from merged data.

Meta analysis (e.g. *Hedges and Olkin*, 1985) is one of the oldest topics in statistics and may be the most widely used method to deal with integrated data from multiple studies. In Bayesian meta analysis, a hierarchical model is postulated to derive combined effects through a weighted average under certain prior distributions (e.g. *Lopes et al.*, 2003, *Müller et al.*, 2004, *Inoue et al.*, 2004, among others). For longitudinal data analysis, *Ishak et al.* (2007) and *Ishak et al.* (2008) investigated meta analysis using mixed-effects models. Most existing meta analysis techniques are developed to either combine summary statistics of individual studies or to use original datasets under strong distributional assumptions.

Wang et al. (2012) proposed an estimating equation approach to assessing the validity of data merging and to analyzing the merged longitudinal dataset. It is shown that their approach is flexible to handle covariance heterogeneity (e.g. different within-subject correlations across studies) and provides proper control of type I error in hypothesis testing. However, their method is limited only to the case of fully observed data and is not applicable to the aforementioned study of lead exposures on weight growth where measurements of covariate CBL in cohort B are substantially missing. Although the popular imputation technique may be a simple and direct solution to the problem, as shown in our simulation studies in Section 3.7, such a strategy may fail to work properly when longitudinal studies are highly heterogeneous.

The current literature of missing data has provided many approaches to handling missing covariates in a single study. For example, *Reilly and Pepe* (1995) proposed a mean score method to impute the score function with missing covariates. *Robins et al.* (1994b) and *Robins and Rotnitzky* (1995), among others, developed various versions of inverse probability weighted (IPW) estimating equation approaches to analyzing incomplete longitudinal data with an assumption on full distributions. *Scharfstein et al.* (1999) extended IPW methods to data with non-ignorable dropouts. To improve robustness against model misspecification and estimation efficiency of IPW methods, several authors (e.g. *Robins et al.*, 1994a; *Davidian et al.*, 2005) proposed augmented IPW (AIPW) estimators, which are sometimes also called doubly robust estimators. An alternative strategy to the idea of IPW is multiple imputation, which has been extensively studied under parametric models (e.g. *Rubin*, 1987, 1996) or under nonparametric models (e.g. *Lipsitz et al.*, 1998; *Kim and Fuller*, 2004). So far, IPW, AIPW and multiple imputation approaches have been mainly developed for a single study with missing data. Applying them to the setting of combined studies requires nontrivial analytic work, especially when the merged dataset involves misaligned or study-specific missing covariates. Although *Qu et al.*'s (2010) aggregated unbiased estimating function approach does not require estimating the probability of missingness or imputing the missing response, it is still developed for one study and only for missing responses.

In this article we propose a new estimating function approach to analyzing merged data from multiple studies with study-specific missing covariates. The novelty of our method lies in the idea of joining study-specific estimating functions, instead of directly joining multiple datasets. In this way, we allow great flexibility to accommodate different covariance structures and other properties across studies. Given that it is not feasible to evaluate estimating functions of studies with missing covariates, integrating these estimating functions with respect to missing covariates is inevitable. The resulting estimating functions are then evaluated nonparametrically without assuming any specific distributions. We show that if the full-data mean models are correctly specified in all individual studies under a valid missing data mechanism, our proposed joint estimating functions are asymptotically unbiased, leading to valid estimation and inference.

We organize the rest of the chapter as follows. Section 3.2 introduces notation and models of interest in the chapter. Section 3.3 presents our estimating procedures and Section 3.4 derives relevant asymptotic properties of the proposed estimator. We then discuss issues concerning implementation of our method in Section 3.5. Some important extensions are given in Section 3.6. After presenting simulation results in Section 3.7, we illustrate our method by analyzing the motivating data to assess the effect of lead exposures on child's weight growth in Section 3.8. All technical details, including proofs and assumptions, are included in the Appendix.

3.2 Model

We consider subjects collected from $K \geq 2$ longitudinal studies with n_k number of subjects in study k, $k = 1, \dots, K$, and the total number of subjects is $n = \sum_{k=1}^{K} n_k$. Let $D_i \in \{1, \dots, K\}$ be the study indicator of subject i, and Y_{ij} be the outcome measured for subject i at visit time j, $j = 1, \dots, m_{D_i}$, and m_{D_i} denotes the number of visits in study D_i for subject $i = 1, \dots, n$. For the ease of exposition we assume that subjects in the same study have the same number of repeated measurements in the rest of the chapter. Let $\boldsymbol{X}_{ij}$ denote a p-dimensional vector of covariates fully observed in all K studies and let $\boldsymbol{Z}_{ij}$ denote a q-dimensional vector of covariates completely observed only in study $k \in \mathcal{S}_o \subset \{1, \dots, K\}$ and missing in study $k \in \mathcal{S}_m \subset \{1, \dots, K\}$, where $\mathcal{S}_m \cup \mathcal{S}_o = \{1, \dots, K\}$. Correspondingly we let $n_o = \sum_{l \in \mathcal{S}_o} n_l$ denote the number of subjects in studies belonging to $\mathcal{S}_o$.

To facilitate our discussion, here we consider two groups of covariates, denoted by $\boldsymbol{Z}^1$ and $\boldsymbol{Z}^2$, and list all scenarios of their missing covariate patterns. Without loss of

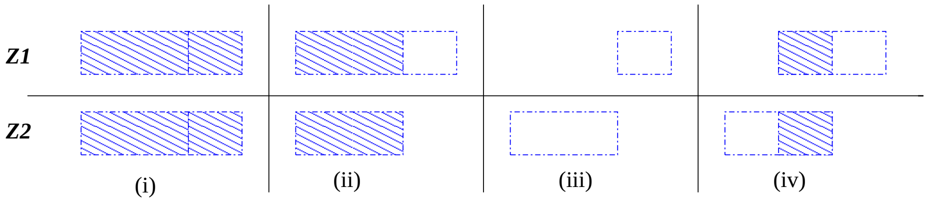

Figure 3.1: Illustration of four scenarios of missing covariate patterns. The shadowed bar indicates a subset of studies with overlapped missing covariates and the blank bar represents a subset of studies with non-overlapped missing covariates.

generality, we write covariates as $[\boldsymbol{X}_{ij}, (\boldsymbol{Z}_{ij}^1, \boldsymbol{Z}_{ij}^2)]$, where $\boldsymbol{X}_{ij}$ is fully observed in all K studies but $\boldsymbol{Z}_{ij}^1$ and $\boldsymbol{Z}_{ij}^2$ are two subvectors of $\boldsymbol{Z}_{ij}$ missing in a set of studies indexed by ${\mathcal{S}_m}^1$ and another set of studies indexed by ${\mathcal{S}_m}^2$, respectively, with $\mathcal{S}_m = {\mathcal{S}_m}^1 \cup {\mathcal{S}_m}^2$ and ${\mathcal{S}_m}^1, {\mathcal{S}_m}^2 \subset \{1, \ldots, K\}$. As shown in Figure 3.1, there are four possible scenarios: (i) Aligned missingness. $\boldsymbol{Z}_{ij}^1$ and $\boldsymbol{Z}_{ij}^2$ are missing in the common set of studies, i.e. ${\mathcal{S}_m}^1 = {\mathcal{S}_m}^2$. In this case, there is no need to distinguish those two covariate subvectors, and conveniently we denote $\boldsymbol{Z}_{ij}^1$ and $\boldsymbol{Z}_{ij}^2$ as $\boldsymbol{Z}_{ij}$. (ii) Nested missingness. A set of studies with missing $\boldsymbol{Z}_{ij}^1$ contains the subset of studies with missing $\boldsymbol{Z}_{ij}^2$, ${\mathcal{S}_m}^1 \subset {\mathcal{S}_m}^2$ or vice versa ${\mathcal{S}_m}^2 \subset {\mathcal{S}_m}^1$. (iii) Completely misaligned missingness. There are no common covariates missing in a common set of studies, i.e. ${\mathcal{S}_m}^1 \cap {\mathcal{S}_m}^2 = \emptyset$. (iv) Partially misaligned missingness. There exists a subset of studies in which a subvector of missing covariates are aligned.

We begin with the simplest scenario of aligned missingness, namely case (i) in Figure 3.1, in which both model and estimation procedure will be discussed in detail. The other three scenarios will be discussed in Section 3.6 as extensions from the case of aligned missingness.

Suppose that the mean of Y_{ij}, given all covariates $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_{ij}$ in study k, satisfies

$$\mu_{k,ij} = E(Y_{ij} \mid \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, D_i = k) = h(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}_{0,k} + \boldsymbol{Z}_{ij}^T \boldsymbol{\lambda}_{0,k}), \quad k = 1, \ldots, K, \tag{3.1}$$

where $h(\cdot)$ is a known link function and $\boldsymbol{\theta}_{0,k} = (\boldsymbol{\beta}_{0,k}^T, \boldsymbol{\lambda}_{0,k}^T)^T$ are the true regression parameters defined in a compact set $\mathcal{B} \subseteq R^{p+q}$. Here we assume that the true parameters are fully or partially shared across studies. The conditional variance of Y_{ij} takes the form: $\text{var}(Y_{ij} \mid \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}, D_i = k) = \phi_k v(\mu_{k,ij})$, where $v(\cdot)$ is a known variance function and ϕ_k is the dispersion parameter.

Note that parameter $\boldsymbol{\theta}_{0,k}$, $k \in \mathcal{S}_m$, cannot be estimated only using data from study k. But it is possible to estimate the parameter by borrowing information from other studies with fully or partially observed data. In matrix notation, let $\boldsymbol{Y}_i = (Y_{i1}, \ldots, Y_{im_{D_i}})^T$, $\boldsymbol{X}_i = (\boldsymbol{X}_{i1}, \ldots, \boldsymbol{X}_{im_{D_i}})^T$ and similarly $\boldsymbol{Z}_i$, $i = 1, \ldots, n$. Our approach relies on the following assumption of missing data mechanism: the study indicator D_i is independent of missing covariates $\boldsymbol{Z}_i$ given $\boldsymbol{X}_i$, denoted by

$$D_i \perp \boldsymbol{Z}_i \mid \boldsymbol{X}_i, \text{ for all } i. \tag{3.2}$$

This assumption is slightly stronger than the typical missing at random (MAR) assumption, in which missing data mechanism may also depend on observed outcomes. The reason is that if (3.2) were allowed to depend on outcome $\boldsymbol{Y}_i$, it would contradict to the conditions assumed in model (3.1), where the regression parameters are allowed to be (partially) different across studies. Assumption (3.2) implies that

$$\begin{aligned} P(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, D_i = k \in \mathcal{S}_m) &= E\{P(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, D_i = k \in \mathcal{S}_m) \mid \boldsymbol{X}_i, D_i = k \in \mathcal{S}_m\} \\ &= E\{P(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, D_i = k \in \mathcal{S}_m) \mid \boldsymbol{X}_i, D_i \in \mathcal{S}_o\}, \end{aligned}$$

where the expectation is taken with respect to the conditional distribution of $\boldsymbol{Z}_i$ given $\boldsymbol{X}_i$ and D_i. The above expressions clearly indicate that under assumption (3.2) the missingness of $\boldsymbol{Z}_i$ in study $k \in \mathcal{S}_m$ can be overcome via data from studies with fully observed data, and hence parameter $\boldsymbol{\theta}_{0,k}$ in study $k \in \mathcal{S}_m$ becomes estimable.

3.3 Estimation

In this section, we propose an estimating function approach to estimating all unknown regression parameters $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{0,1}^T, \ldots, \boldsymbol{\theta}_{0,K}^T)^T$. Before presenting the detail, we like to comment that although our approach is discussed for the scenario of $\boldsymbol{Z}_i$ being fully missing in study $k \in \mathcal{S}_m$, it is also applicable when $\boldsymbol{Z}_i$ is partially missing (i.e. $\boldsymbol{Z}_i$ observed on some subjects) in study $k \in \mathcal{S}_m$ as long as assumption (3.2) holds.

3.3.1 Conditional Moments

Firstly note that $\boldsymbol{\theta}_{0,k}$ in study $k \in \mathcal{S}_o$ can be routinely estimated by generalized estimating equations (GEE) method. On the other hand, for $\boldsymbol{\theta}_{0,k}$ in study $k \in \mathcal{S}_m$, we need to integrate the full data model (3.1) with respect to missing covariates $\boldsymbol{Z}_i$. Precisely, let $\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_k)$ denote the conditional expectation of $h(\boldsymbol{X}_k, \boldsymbol{Z}_{ij}, \boldsymbol{\theta}_k) = h(\boldsymbol{X}_{ij}^T\boldsymbol{\beta}_k + \boldsymbol{Z}_{ij}^T\boldsymbol{\lambda}_k)$ in (3.1) with respect to $\boldsymbol{Z}_{ij}$ conditioning on $\boldsymbol{X}_{ij}$ in study $k \in \mathcal{S}_m$, and assume that the resulting marginal estimating function $\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_k)$ is a smooth function satisfying $\eta_k(\boldsymbol{X}_i, \boldsymbol{\theta}_{0,k}) = E(Y_{ij} \mid \boldsymbol{X}_{ij}, D_i = k)$ uniquely at $\boldsymbol{\theta}_{0,k}$ under the conditional distribution of $\boldsymbol{Y}_{ij}$ given $\boldsymbol{X}_{ij}$ and $D_i = k$. In this case assumption (3.2) implies that

$$\begin{aligned}\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k}) &= E\{h(\boldsymbol{X}_k, \boldsymbol{Z}_{ij}, \boldsymbol{\theta}_{0,k}) \mid \boldsymbol{X}_{ij}, D_i = k \in \mathcal{S}_m\} \\ &= E\{h(\boldsymbol{X}_k, \boldsymbol{Z}_{ij}, \boldsymbol{\theta}_{0,k}) \mid \boldsymbol{X}_{ij}, D_i \in \mathcal{S}_o\}.\end{aligned}$$

This suggests that estimating function $\eta_k(\cdot, \boldsymbol{\theta}_{0,k})$ can be estimated by using data from studies in $\mathcal{S}_o$. Similarly, the variance of Y_{ij} conditioning on $\boldsymbol{X}_{ij}$ in study $k \in \mathcal{S}_m$, denoted by $\nu_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})$, is given by

$$\begin{aligned}\nu_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k}) =& \mathrm{Var}(Y_{ij} \mid \boldsymbol{X}_{ij}, D_i = k \in \mathcal{S}_m) \\ =& \phi_k E[v\{h(\boldsymbol{X}_k, \boldsymbol{Z}_{ij}, \boldsymbol{\theta}_{0,k})\} \mid \boldsymbol{X}_{ij}, D_i = k \in \mathcal{S}_m] \\ &+ E\{h(\boldsymbol{X}_k, \boldsymbol{Z}_{ij}, \boldsymbol{\theta}_{0,k})^2 \mid \boldsymbol{X}_{ij}, D_i \in \mathcal{S}_o\} - \{\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})\}^2.\end{aligned}$$

In particular, when Y_{ij} follows a normal distribution, $\nu_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})$ is equal to $\phi_k + Var(\boldsymbol{Z}_{ij} \mid \boldsymbol{X}_{ij}, D_i \in \mathcal{S}_o)\boldsymbol{\lambda}_{0,k}^2$. When Y_{ij} is a binary response, $\nu_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})$ becomes $\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})\{1 - \eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})\}$. In both cases, which are frequently encountered in practice, this conditional variance $\nu_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k})$ can be estimated using data from studies in $\mathcal{S}_o$.

3.3.2 Estimation with Missing Covariates

We consider nonparametric estimation for $\eta_k(\boldsymbol{x}, \boldsymbol{\theta}_k)$ based on the following argument of feasibility. According to *Newey* (1994), we can show that our proposed estimators of the regression coefficients in this section are consistent and asymptotically normal, as long as the plug-in estimator of $\eta_k(\boldsymbol{x}, \boldsymbol{\theta}_k)$ satisfies a convergence rate faster than $n^{-1/4}$. This rate is achievable when $\eta_k(\boldsymbol{x}, \boldsymbol{\theta}_k)$ is assumed to be sufficiently smooth with respect to $\boldsymbol{x}$.

In this chapter, we adopt the sieve least square method (e.g. *Newey*, 1997; *Shen*, 1997; *Shen and Wong*, 1994) to estimate $\eta_k(\boldsymbol{x}, \boldsymbol{\theta}_k)$ using basis functions $\{b_l(\boldsymbol{x})\}_{l=1}^{j_{n_k}}$ that enable us to approximate a square-integrable smooth function on a compact support. The number of basis functions, j_{n_k}, increases along the increase of sample size n. Sieve estimation is regarded as being of both analytical and computational convenience due to the fact that a sieve estimator has explicit analytic expressions, which simplifies the proposed estimation procedure.

For the ease of exposition, we suppress covariates in the following short-handed notation. In study k, we denote $\eta_{k,ij}(\boldsymbol{\theta}_k) = \eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_k)$, $h_{k,ij}(\boldsymbol{\theta}_k) = h(\boldsymbol{X}_k, \boldsymbol{Z}_{ij}, \boldsymbol{\theta}_k)$ and $\nu_{k,ij}(\boldsymbol{\theta}_k) = \nu_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_k)$ and so forth. The corresponding vectors for subject i are denoted by $\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$, $\boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)$, and $\boldsymbol{\nu}_{k,i}(\boldsymbol{\theta}_k)$, respectively. A sieve estimator of $\eta_{k,ij}(\boldsymbol{\theta}_k), k \in \mathcal{S}_m$, takes the following form:

$$\hat{\eta}_{k,ij}(\boldsymbol{\theta}_k) = \sum_{l=1}^{j_{n_k}} a_{k,l}(\boldsymbol{\theta}_k) b_l(\boldsymbol{x}_{ij}) = \boldsymbol{b}(\boldsymbol{x}_{ij})^T \boldsymbol{a}_k(\boldsymbol{\theta}_k),$$

where $\boldsymbol{a}_k(\boldsymbol{\theta}_k) = (a_{k,1}(\boldsymbol{\theta}_k), \ldots, a_{k,j_{n_k}}(\boldsymbol{\theta}_k))^T$ is the vector of unknown coefficients to be estimated, and $\boldsymbol{b}(\boldsymbol{x}_{ij}) = (b_1(\boldsymbol{x}_{ij}), \ldots, b_{j_{n_k}}(\boldsymbol{x}_{ij}))^T$ is the vector of basis functions. Estimation of $\boldsymbol{a}_k(\boldsymbol{\theta}_k)$ is carried out by minimizing the following objective function using all studies from $\mathcal{S}_o$:

$$\hat{\boldsymbol{a}}_k(\boldsymbol{\theta}_k) = \underset{\boldsymbol{a}_k(\boldsymbol{\theta}_k)}{\arg\min} \sum_{i=1}^{n} \sum_{j=1}^{m_i} I[D_i \in \mathcal{S}_o] \left\{ h_{k,ij}(\boldsymbol{\theta}_k) - \boldsymbol{b}(\boldsymbol{X}_{ij})^T \boldsymbol{a}_k(\boldsymbol{\theta}_k) \right\}^2, \ k \in \mathcal{S}_m,$$

where $I[A]$ is the indicator function of set A. For subject i in study $D_i = l$, we define notation: an $m_l \times j_{n_k}$ matrix $\boldsymbol{W}_i = (\boldsymbol{b}(\boldsymbol{X}_{i1}), \ldots, \boldsymbol{b}(\boldsymbol{X}_{im_l}))^T$, a $j_{n_k} \times \sum_{l \in \mathcal{S}_o} n_l m_l$ matrix $\boldsymbol{U}^T = (\boldsymbol{W}_i^T)_{D_i = l \in \mathcal{S}_o}$, and $(\sum_{l \in \mathcal{S}_o} n_l m_l)$-dimensional vector $\boldsymbol{H}_k(\boldsymbol{\theta}_k) = (\boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)^T)^T_{D_i = l \in \mathcal{S}_o}$. It is easy to see that

$$\hat{\boldsymbol{a}}_k(\boldsymbol{\theta}_k) = (\boldsymbol{U}^T \boldsymbol{U})^{-1} \boldsymbol{U}^T \boldsymbol{H}_k(\boldsymbol{\theta}_k),$$

and hence $\hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_k) = \boldsymbol{W}_i \hat{\boldsymbol{a}}_k(\boldsymbol{\theta}_k)$. Correspondingly the estimated $\partial \hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_k) / \partial \boldsymbol{\theta}_k$ is

$$\nabla_{\boldsymbol{\theta}_k} \hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_k) = \boldsymbol{W}_i \nabla_{\boldsymbol{\theta}_k} \hat{\boldsymbol{a}}_k(\boldsymbol{\theta}_k),$$

where $\nabla_a f(a)$ denotes a gradient vector of function f with respect to a. In Section 3.5.2 we will discuss the selection of the number of basis functions to balance between the goodness-of-fit and model parsimony.

With the availability of estimated $\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$, we are now ready to derive the estimating function for the regression parameter $\boldsymbol{\theta}_k$ of interest. Following *Qu et al.* (2000), we propose to join the estimating functions from individual studies using the method of quadratic inference function (QIF). Pointed out by *Wang et al.* (2012), the QIF approach provides great flexibility to account for inter-study heterogeneities. Briefly, QIF begins with an expansion on the inverse of a working correlation matrix for study k of the form: $\boldsymbol{R}_k^{-1}(\alpha_k) \approx \sum_{s=1}^{s_k} \rho_{k,s}\boldsymbol{M}_{k,s}$, where $\rho_{k,1}, \ldots, \rho_{k,s_k}$ are constants possibly dependent on nuisance correlation parameter α_k, and $\boldsymbol{M}_{k,1}, \ldots, \boldsymbol{M}_{k,s_k}$ are known basis matrices with elements 0 and 1 determined by the given working correlation matrix $\boldsymbol{R}_k(\alpha_k)$. Refer to *Qu et al.* (2000) for more details concerning the forms of basis matrices corresponding to different working correlation structures such as exchangeable and AR-1. See also *Song* (2007, Chapter 5) for the selection of working correlation matrix in the application of QIF.

First let us focus on the estimation of regression coefficients in a regression model for study $k \in \mathcal{S}_m$, in which we need to borrow data from all studies in $\mathcal{S}_o$ in order to estimate coefficients of the missing covariates. Denote the estimating function for subject i in study $k \in \mathcal{S}_m$ by $\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i})$, which is expressed with an explicit involvement of $\hat{\boldsymbol{\eta}}_{k,i}$. The same treatment is given to other notation whenever applicable. The extended score vector $\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k)$ takes the form:

$$\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k) = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i}) \overset{def.}{=} \frac{1}{n_k}\sum_{i=1}^{n_k} \begin{pmatrix} \boldsymbol{g}_{k,i,1}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i}) \\ \vdots \\ \boldsymbol{g}_{k,i,s_k}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i}) \end{pmatrix}, \tag{3.3}$$

where for $s = 1, \ldots, s_k$,

$$\boldsymbol{g}_{k,i,s}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i}) = \nabla_{\boldsymbol{\theta}_k}\hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_k)^T \boldsymbol{V}_{k,i,s}(\boldsymbol{\theta}_k)\{\boldsymbol{Y}_i - \hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_k)\},$$

with $\boldsymbol{V}_{k,i,s}(\boldsymbol{\theta}_k) = {\boldsymbol{A}_{k,i}}^{-1/2}\boldsymbol{M}_{k,s}{\boldsymbol{A}_{k,i}}^{-1/2}$ and $\boldsymbol{A}_{k,i} = \mathrm{diag}\{\nu_{k,i1}(\boldsymbol{\theta}_k), \ldots, \nu_{k,im_k}(\boldsymbol{\theta}_k)\}$.

Minimizing a quadratic function

$$Q_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k) = n_k \bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k)^T \boldsymbol{C}_k^{-}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k) \bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k) \tag{3.4}$$

gives an estimator of $\boldsymbol{\theta}_k$, that is,

$$\hat{\boldsymbol{\theta}}_k = \underset{\boldsymbol{\theta}_k}{\arg\min} Q_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k), \tag{3.5}$$

where $\boldsymbol{C}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k)$ is given by $\boldsymbol{C}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k) = \frac{1}{n_k} \sum_{i=1}^{n_k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i}) \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i})^T$. Note that as discussed above we do not need to consider another nonparametric estimate for $\nu_{k,ij}(\boldsymbol{\theta}_k)$ in (3.5) separately in the linear model or in the logistic model. Even if $\nu_{k,ij}(\boldsymbol{\theta}_k)$ has to be estimated separately, for instance in the log-linear model, the large sample properties of the consistency and asymptotic normality given in Section 3.4 for $\hat{\boldsymbol{\theta}}_k$ still hold under certain regularity conditions, as long as $\nu_{k,ij}(\boldsymbol{\theta}_k)$ is replaced by a root-n consistent estimator. This plug-in exercise has been well studied by many authors in the literature of semiparametric models and estimation (e.g. *Powell*, 1986; *Ichimura and Lee*, 2010; *Andrews*, 1994).

3.3.3 Joint Estimation with Complete and Incomplete Datasets

An advantage of performing joint analysis of merged data is to improve estimation efficiency on the regression coefficients across studies (*Wang et al.*, 2012). This property is expected to prevail even when some covariates are not observed in some studies, a situation considered in this chapter. Let $\mathcal{M}_l \subset \{1, \ldots, K\}$, $l = 1, \ldots, p+q$, be the subset of studies within which the l^{th} covariate has a common effect size. The parameter space constrained by all $\mathcal{M}_l$, $l = 1, \ldots, p+q$, is denoted by Ω with $\Omega = \{(\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T : \theta_{kl} = \theta_{k'l} \text{ for } \forall\ k \neq k' \in \mathcal{M}_l, l = 1, \ldots, p+q\}$ representing the subspace of parameters restricted under all conditions of common regression coefficients.

In study $k \in \mathcal{S}_o$, similar to equation (3.3), the extended score vector $\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \boldsymbol{h}_k)$ is given by

$$\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \boldsymbol{h}_k) = \frac{1}{n_k}\sum_{i=1}^{n_k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_k, \boldsymbol{h}_{k,i}) \overset{def.}{=} \frac{1}{n_k}\sum_{i=1}^{n_k} \begin{pmatrix} \nabla_{\boldsymbol{\theta}_k}\boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)^T \boldsymbol{V}_{k,i,1}(\boldsymbol{\theta}_k)\{\boldsymbol{Y}_i - \boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)\} \\ \vdots \\ \nabla_{\boldsymbol{\theta}_k}\boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)^T \boldsymbol{V}_{k,i,s_k}(\boldsymbol{\theta}_k)\{\boldsymbol{Y}_i - \boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)\} \end{pmatrix},$$

where $\boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)$ is defined in (3.1) and $\nabla_{\boldsymbol{\theta}_k}\boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k) = \partial \boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)/\partial \boldsymbol{\theta}_k^T$. Now we are ready to form a joint quadratic inference function to simultaneously estimate all regression coefficients using all K studies. This objective function is $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}})$ defined as

$$Q(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) = n\bar{\boldsymbol{g}}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}})^T \boldsymbol{C}^-(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}})\bar{\boldsymbol{g}}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}),$$

where

$$\bar{\boldsymbol{g}}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{g}_i(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}_i) \overset{def.}{=} \frac{1}{n}\sum_{i=1}^{n} \Big(I[D_i = 1]\boldsymbol{g}_{1,i}^T, \ldots, I[D_i = K]\boldsymbol{g}_{K,i}^T\Big)^T, \qquad (3.6)$$

with $\boldsymbol{g}_{k,i} = I[D_i = k \in \mathcal{S}_o]\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_k, \boldsymbol{h}_{k,i}) + I[D_i = k \in \mathcal{S}_m]\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i})$ for $k = 1, \ldots, K$, and $\boldsymbol{C}(\boldsymbol{\theta}, \boldsymbol{\eta})$ is a block-diagonal matrix, $\boldsymbol{C}(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{diag}\{\boldsymbol{g}_{1,i}\boldsymbol{g}_{1,i}^T, \ldots, \boldsymbol{g}_{K,i}\boldsymbol{g}_{K,i}^T\}$. Parameter vector $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^T, \ldots, \boldsymbol{\theta}_K^T)^T$ is then estimated by minimizing $Q(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}})$ over Ω, that is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta} \in \Omega}{\arg\min} Q(\boldsymbol{\theta}, \hat{\boldsymbol{\eta}}). \qquad (3.7)$$

When there are no missing covariates, *Wang et al.* (2012) showed that the above QIF estimator of $\boldsymbol{\theta}$ based on the merged data is more efficient than that based on each individual study. In this chapter we show that such efficiency gain remains in the presence of study-specific missing covariates, where unknown $\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$ for $k \in \mathcal{S}_m$ is replaced by a sieve nonparametric estimate $\hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_k)$. The detail is presented in Section 3.4.

3.4 Asymptotic Properties

This section concerns asymptotic properties of both estimator $\hat{\boldsymbol{\theta}}_k$ obtained in individual study $k \in \mathcal{S}_m$ according to (3.5), and the joint estimator $\hat{\boldsymbol{\theta}}$ given in (3.7). For convenience, the study-specific expectation under the distribution generating data of study k is denoted by $E_k(\cdot) = E(\cdot \mid D_i = k)$, $k = 1, \ldots, K$. Denote the Euclidean norm of a vector $\boldsymbol{b}$ by $\|\boldsymbol{b}\|$, the induced norm of a matrix $\boldsymbol{A}$ by $\|\boldsymbol{A}\| = \sup_{\|\boldsymbol{b}\|=1} \|\boldsymbol{A}\boldsymbol{b}\|$, the sup-norm of a function $f(\boldsymbol{x})$ by $\|f\|_\infty = \sup_{\boldsymbol{x}} \|f(\boldsymbol{x})\|$, and the L_2 norm of a random vector $\boldsymbol{X}$ by $\|\boldsymbol{X}\|_2$. For any p-dimensional vector $\boldsymbol{a} = (a_1, \ldots, a_p)^T$ of nonnegative integers, write $|\boldsymbol{a}| = \sum_{i=1}^p a_i$. The $|\boldsymbol{a}|$-th order derivative of an analytic function $f(\boldsymbol{x})$, $f(\boldsymbol{x}) : \mathcal{R}^p \to \mathcal{R}$, with respect to $\boldsymbol{x}$, is represented by $\nabla^{\boldsymbol{a}} f(\boldsymbol{x}) = \frac{\partial^{|\boldsymbol{a}|}}{\partial x_1^{a_1} \ldots \partial x_p^{a_p}} f(\boldsymbol{x})$. For some $\gamma > 0$, let $[\gamma]$ be the largest integer smaller than γ. We also let $\Lambda^\gamma(\mathcal{X})$ denote the space of functions $f : \mathcal{X} \to \mathcal{R}$ that have up to $[\gamma]$-th order continuous derivatives and the highest $[\gamma]$-th order derivative is Hölder continuous with the exponent $\gamma - [\gamma] \in (0, 1]$ (see *Chen et al.* (2003) for the detail).

To establish large-sample properties for the two proposed estimators, we impose some regularity conditions with the details listed in the Appendix. It is worth noting that among those conditions, Assumption B.1.1 is a smoothness condition for $\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$ regularizing the order of an approximation error in the sieve expansion with respect to a set of basis functions. According to Theorem 12.8 of *Schumaker* (1981), the uniform approximation error to $\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$ is $O(j_{n_k}^{|\gamma|/p})$ for any $\boldsymbol{\theta}_k$. Thus, to achieve *Stone*'s (1982) optimal convergence rate, it requires $|\boldsymbol{\gamma}| > p/2$ for a p-dimensional covariate $\boldsymbol{X}_i$. Other conditions in Assumption B.1 pertain to behaviors of estimating functions for the regression parameters to ensure that the estimating functions are asymptotically unbiased when a nonparametric estimator of $\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$ is plugged in.

Theorem III.1. *Let $n_o = \sum_{l \in \mathcal{S}_o} n_l$. Suppose that (i) the mean model* (3.1) *is correctly specified, and that (ii) missing mechanism assumption* (3.2) *holds. Under As-*

sumption B.1 and $j_{n_k} = o(n_o)$*, estimator* $\hat{\boldsymbol{\theta}}_k$ *for* $k \in \mathcal{S}_m$ *given in* (3.5) *is consistent, namely,* $\hat{\boldsymbol{\theta}}_k \xrightarrow{p} \boldsymbol{\theta}_{0,k}$ *as* $n_o \to \infty$*.*

To prove Theorem III.1, we need to first establish the consistency for the nonparametric estimator $\|\hat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k\|_\infty = o_p(1)$, which is the focus of Lemma B.3 in the Appendix, by applying similar arguments to those given in *Chen et al.* (2005). Consequently, we are able to obtain the uniform consistency of the score functions, $\sup_{\boldsymbol{\theta}_k \in \mathcal{B}} \|\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k) - \bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \boldsymbol{\eta}_k)\| = o_p(1)$. Moreover, we achieve the consistency for $\hat{\boldsymbol{\theta}}_k$ according to Glivenko-Cantelli Theorem and Lemma 5.2 of *Newey* (1994).

Theorem III.2. *Consider* $\hat{\boldsymbol{\theta}}_k$ *given by* (3.5) *for* $k \in \mathcal{S}_m$*. Then under Assumptions B.1 and B.2 stated in the Appendix as well as* $j_{n_k} = o(n_o)$*, the estimated score function* $\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_k)$ *can be represented by*

$$n_k^{1/2} \bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_k) = n_k^{-1/2} \sum_{D_i = k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i}) + \tau_k^{1/2} n_o^{-1/2} \sum_{D_i \in \mathcal{S}_o} \boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i}) + o_p(1),$$

where $\frac{n_k}{n_o} \to \tau_k$ *as* $n_k \to \infty, n_o \to \infty$ *and*

$$\boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i}) = \Big(\boldsymbol{q}_{k,i,1}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})^T, \ldots, \boldsymbol{q}_{k,i,s_k}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})^T \Big)^T$$

consists of elements $\boldsymbol{q}_{k,i,s}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})$*, which is defined as for* $s = 1, \cdots, s_k$

$$\boldsymbol{q}_{k,i,s}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i}) = \frac{f(\boldsymbol{X}_i \mid D_i = k)}{f(\boldsymbol{X}_i \mid D_i \in \mathcal{S}_o)} \nabla \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})^T \boldsymbol{V}_{k,i,s} \{\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k}) - \boldsymbol{h}_{k,i}(\boldsymbol{\theta}_{0,k})\}.$$

Moreover, the asymptotic distribution of $\hat{\boldsymbol{\theta}}_k$ *is given by*

$$\sqrt{n}_k(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{0,k}) \xrightarrow{d} N\{\boldsymbol{0}, (\boldsymbol{G}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{G}_k)^{-1}\},$$

where $\boldsymbol{G}_k = E_k\{\nabla \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i})\}$, *and* $\boldsymbol{\Sigma}_k = \boldsymbol{\Sigma}_{k,1} + \tau_k \boldsymbol{\Sigma}_{k,2}$ *with*

$$\boldsymbol{\Sigma}_{k,1} = E_k\{\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i})\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i})^T\},$$

and

$$\boldsymbol{\Sigma}_{k,2} = E_o\{\boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})\boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})^T\}.$$

From Theorem III.2, we see that the representation of $n_k^{1/2}\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_{k,i})$ constitutes two components:

$$n_k^{-1/2}\sum_{D_i=k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i}) \quad \text{and} \quad \tau_k^{1/2} n_o^{-1/2} \sum_{D_i \in \mathcal{S}_o} \boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i}).$$

It is interesting to note that the second component $\tau_k^{1/2} n_o^{-1/2} \sum_{D_i \in \mathcal{S}_o} \boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})$ is related to the weighted likelihood (e.g. *Hu and Zidek*, 2002; *Wang and Zidek*, 2005). Since covariate $\boldsymbol{Z}_i$ is not recorded in study $k \in \mathcal{S}_m$, $\tau_k^{1/2} n_o^{-1/2} \sum_{D_i \in \mathcal{S}_o} \boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})$ presents an inference function using the observed data on $\boldsymbol{Z}_i$ from other studies in $\mathcal{S}_o$ weighted by the measure of relevance defined by $f(\boldsymbol{X}_i \mid D_i = k)/f(\boldsymbol{X}_i \mid D_i \in \mathcal{S}_o)$. Thus, it becomes natural to yield the asymptotic variance of $\hat{\boldsymbol{\theta}}_k$ that consists of two pieces, $\boldsymbol{\Sigma}_{k,1}$ and $\boldsymbol{\Sigma}_{k,2}$, where $\boldsymbol{\Sigma}_{k,1}$ gives the asymptotic variance of $\hat{\boldsymbol{\theta}}_k$ when $\boldsymbol{\eta}_{k,i}$ were known, while $\boldsymbol{\Sigma}_{k,2}$ characterizes the additional variance incurred by the nonparametric sieve estimation of $\boldsymbol{\eta}_{k,i}$. The extra contribution by $\boldsymbol{\Sigma}_{k,2}$ towards the total variance of $\boldsymbol{\Sigma}_k$ is weighted according to a rate of τ_k; when n_o exceeds n_k in the sense of $\frac{n_k}{n_o} \to 0$, the contribution from studies in $\mathcal{S}_m$ will vanish and may be ignored. To evaluate $(\boldsymbol{G}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{G}_k)^{-1}$, we need to replace $\boldsymbol{G}_k$ and $\boldsymbol{\Sigma}_k$ by their consistent estimates respectively. This step involves estimating an unknown density ratio between $f(\boldsymbol{X}_i \mid D_i = k)$ and $f(\boldsymbol{X}_i \mid D_i \in \mathcal{S}_o)$. Note that we may rewrite this ratio as $\frac{f(D_i=k|\boldsymbol{X}_i)f(D_i\in\mathcal{S}_o)}{f(D_i\in\mathcal{S}_o|\boldsymbol{X}_i)f(D_i=k)}$, in a spirit similar to the strategy of inverse probability weighting, where $\frac{f(D_i=k|\boldsymbol{X}_i)}{f(D_i\in\mathcal{S}_o|\boldsymbol{X}_i)}$ may be estimated by a multinomial logistic model and

$\frac{f(D_i \in \mathcal{S}_o)}{f(D_i = k)}$ by $\frac{n_o}{n_k}$. Obviously this approach needs some additional model assumptions which may not be easily checked in practice. An alternative way is to perform a bootstrap variance estimation, which avoids making extra model assumptions in the above ratio estimation, and hence is recommended and implemented in Section 3.5.

Now we turn to the estimator $\hat{\boldsymbol{\theta}}$ given in (3.7). Using similar arguments, we obtain the following representations for the extended scores $n^{1/2}\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\eta}})$: for $k \in \mathcal{S}_o$

$$n^{-1/2} \sum_{D_i=k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i}) = \Big(\frac{\tau_k}{1+\tau_{\mathcal{S}_m}}\Big)^{1/2} n_k^{-1/2} \sum_{D_i=k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i}) + o_p(1),$$

and for $k \in \mathcal{S}_m$

$$\begin{aligned} & n^{-1/2} \sum_{D_i=k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_{k,i}) \\ = & \Big(\frac{\tau_k}{1+\tau_{\mathcal{S}_m}}\Big)^{1/2} \Big\{ n_k^{-1/2} \sum_{D_i=k} \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i}) + \tau_k^{1/2} n_o^{-1/2} \sum_{D_i \in \mathcal{S}_o} \boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i}) \Big\} + o_p(1), \end{aligned}$$

where $\tau_{\mathcal{S}_m} = \sum_{k \in \mathcal{S}_m} \tau_k$. Thus, the asymptotic variance of $n^{1/2}\bar{\boldsymbol{g}}(\boldsymbol{\theta}_0, \hat{\boldsymbol{\eta}})$, $\boldsymbol{\Sigma}$, is a block-diagonal matrix whose k-th element is given as follows:

$$\frac{\tau_k}{1+\tau_{\mathcal{S}_m}} \boldsymbol{\Sigma}_k I[k \in \mathcal{S}_o] + \frac{\tau_k}{1+\tau_{\mathcal{S}_m}} \boldsymbol{\Sigma}_{k,1} I[k \in \mathcal{S}_m] + \frac{\tau_k^2}{1+\tau_{\mathcal{S}_m}} \boldsymbol{\Sigma}_{k,2} I[k \in \mathcal{S}_m], \tag{3.8}$$

where $\boldsymbol{\Sigma}_k = E_k\{\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})^T\}$, and the other two covariances, $\boldsymbol{\Sigma}_{k,1}$ and $\boldsymbol{\Sigma}_{k,2}$, are given in Theorem III.2. The block-diagonal structure for $\boldsymbol{\Sigma}$ is due to the fact that $\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})$ and $\boldsymbol{q}_{l,i}(\boldsymbol{\theta}_{0,l}, \boldsymbol{h}_{l,i})$ for study k and l, $k \neq l$, are uncorrelated. When there exist shared parameters, namely $dim(\Omega) < (p+q)K$, the joint estimation given in (3.7) can improve efficiency for all regression coefficients by using similar arguments in *Wang et al.* (2012). We summarize the above discussion concerning asymptotic properties of $\hat{\boldsymbol{\theta}}$ into the following Theorem.

Theorem III.3. *Under Assumptions B.1 and B.2 given in the Appendix, the joint* *estimator* $\hat{\boldsymbol{\theta}}$ *given in* (3.7) *is asymptotically normally distributed with mean* $\mathbf{0}$ *and asymptotic variance* $(\boldsymbol{G}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{G})^{-1}$*, namely*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} N\{\mathbf{0}, (\boldsymbol{G}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{G})^{-1}\}, \text{ as } n \to \infty$$

where $\boldsymbol{\Sigma}$ *is given in* (3.8) *and* $\boldsymbol{G} = (\boldsymbol{G}_1^T, \ldots, \boldsymbol{G}_k^T)^T$ *with the* k*-th matrix* $\boldsymbol{G}_k$ *given by*

$$\boldsymbol{G}_k = \begin{cases} E_k\{\nabla_{\boldsymbol{\theta}_k}\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})\}, & k \in \mathcal{S}_o; \\ E_k\{\nabla_{\boldsymbol{\theta}_k}\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i})\}, & k \in \mathcal{S}_m. \end{cases}$$

When there exist shared parameters, $\hat{\boldsymbol{\theta}}$ *has a smaller asymptotic variance than any* $\hat{\boldsymbol{\theta}}_k$*,* $k = 1, \ldots, K$*, obtained by* (3.5) *using data from individual studies.*

3.5 Implementation

This section focuses on two key elements in the implementation of our method: (i) bootstrap variance estimation and (ii) selection of the number of basis functions in the nonparametric estimation of $\boldsymbol{\eta}(\boldsymbol{x}, \boldsymbol{\theta}_k)$.

3.5.1 Bootstrap Variance Estimation

As noted above, directly estimating the asymptotic variance of $\hat{\boldsymbol{\theta}}_k$ in study $k \in \mathcal{S}_m$ is challenging because it involves an unknown measure of relevance, $f(\boldsymbol{X}_i \mid D_i = k \in \mathcal{S}_m)/f(\boldsymbol{X}_i \mid D_i \in \mathcal{S}_o)$. Thus we consider estimating the asymptotic variance by using bootstrap resampling techniques. We follow *Chen et al.* (2003) and *Hall and Horowitz* (1996) to establish our bootstrap procedure. Let $\{\boldsymbol{Y}_i^*, \boldsymbol{X}_i^*, \boldsymbol{Z}_i^*, D_i^*\}_{i=1}^n$ be a bootstrap sample, which is generated by the scheme of stratified sampling with individual studies as strata, so that the resulting bootstrap sample constitutes the same proportions of subjects from K studies and preserves the same within-subject correlation as that of the original sample. According to *Hall and Horowitz* (1996), a bootstrap version of extended score $\bar{\boldsymbol{g}}_k^*(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i})$ needs to be centered, given by

$$\bar{\boldsymbol{g}}_k^c(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k^*) = \bar{\boldsymbol{g}}_k^*(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k^*) - \bar{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\eta}}_k),$$

where $\hat{\boldsymbol{\theta}}_k$ and $\hat{\boldsymbol{\eta}}_k$ are estimated from the original sample and $\hat{\boldsymbol{\eta}}_k^*$ is estimated from the bootstrap sample. The reason for the need of centering is that the QIF estimator is obtained as a minimizer of an objective function, and the resulting estimated moments of the extended scores are not necessarily equal to $\mathbf{0}$. It is imperative to subtract $\bar{\boldsymbol{g}}_k(\hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\eta}}_k)$ from $\bar{\boldsymbol{g}}_k^*(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k^*)$ to obtain asymptotically unbiased estimating functions, which is critical to ensure consistent estimation. Consequently the bootstrap estimator $\hat{\boldsymbol{\theta}}_k^*$ is defined as the minimizer of $Q_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k^*)$ given in (3.5) where $\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k)$ is replaced by its bootstrap version $\bar{\boldsymbol{g}}_k^c(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_k^*)$.

Repeating the bootstrap procedure a certain number of times, we yield a set of bootstrap estimates of $\boldsymbol{\theta}_k$, which are then used to calculate the bootstrap variances. The same procedure can be established for the joint estimation of $\boldsymbol{\theta}$.

3.5.2 Selection of the Number of Basis Functions

Another critical issue in the implementation of the proposed method is to determine the number of basis functions for the estimation of $\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$. Since a nonparametric regression is used to estimate the conditional mean model instead of estimating regression coefficients, selecting the number of basis functions is more relevant to estimation of $\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$ than estimation of $\boldsymbol{\theta}_k$. There are several criteria potentially useful to serve for such a selection purpose, including *Schwarz*'s (1978) Bayesian information criterion (BIC) and *Craven and Wahba*'s (1979) generalized cross validation (GCV). In the context of longitudinal data, *Wang and Qu* (2009) proposed QIF-based BIC, termed as BIQIF, to perform model selection in parametric regression. Note that BIQIF cannot be directly applied in our semiparametric model setting because the penalty term $\frac{1}{2}\log(n)(p+k)$ appears always to dominate BIQIF when the mean model structure in (3.1) is correctly specified. As a result, the BIQIF tends to select underfitting models. As a remedy, we follow the work of *He et al.* (2002) and propose a new BIC-type model selection criterion:

$$\text{BIC}(j_{n_k}) = Q(\hat{\boldsymbol{\theta}}_k^{(j_{n_k})}, \hat{\boldsymbol{\eta}}_k) + \frac{\log n}{2n}(p + j_{n_k}),\ k = 1, \cdots, K,$$

where p is the number of regression parameters, j_{n_k} is the number of basis functions and $\hat{\boldsymbol{\theta}}_k^{(j_{n_k})}$ is the estimate of $\boldsymbol{\theta}_k$ when j_{n_k} basis functions are used. The number j_{n_k} is chosen by searching j_{n_k} within a sufficiently wide range of candidate values and the best j_{n_k} is the one with the smallest $\text{BIC}(j_{n_k})$. Performance of $\text{BIC}(j_{n_k})$ is examined through simulation studies in Section 3.7.3.

3.6 Missing Covariates in Other Scenarios

Now let us extend the above development in the first case of aligned missing covariates to the other three scenarios. The nested missingness (i.e. scenario (ii) in Figure 3.1) might be handled using the same framework under the same assumption (3.2) of missing data mechanism. This is because if we naively treat observed covariates contained in $\boldsymbol{Z}_2$ in Figure 3.1 for case (ii) as "missing" covariates, we may simply turn nested missingness into aligned missingness. But this approach is not desirable and can be improved by explicitly using the nested missing pattern. Without loss of generality, we assume ${\mathcal{S}_m}^1 \subset {\mathcal{S}_m}^2$. To handle scenario (ii), we make the following assumptions regarding mechanisms of missing covariates:

$$D_i \perp \boldsymbol{Z}_i^1 \mid \boldsymbol{X}_i,\ \text{and } D_i \perp \boldsymbol{Z}_i^2 \mid \boldsymbol{X}_i,\ \boldsymbol{Z}_i^1,\ \text{for all } i. \tag{3.9}$$

This implies that for study $k \in \mathcal{S}_m{}^1$ when both $\boldsymbol{Z}_{ij}^1$ and $\boldsymbol{Z}_{ij}^2$ are missing,

$$\begin{aligned}\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{\theta}_{0,k}) &= E\{h(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^1, \boldsymbol{Z}_{ij}^2, \boldsymbol{\theta}_{0,k}) \mid \boldsymbol{X}_{ij}, D_i = k \in \mathcal{S}_m{}^1\} \\ &= E\{h(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^1, \boldsymbol{Z}_{ij}^2, \boldsymbol{\theta}_{0,k}) \mid \boldsymbol{X}_{ij}, D_i \in \mathcal{S}_o{}^2\},\end{aligned}$$

where the expectation is under the distribution of $\boldsymbol{Z}_{ij}^1$ and $\boldsymbol{Z}_{ij}^2$ given $\boldsymbol{X}_{ij}$ and D_i. On the other hand for study $k \in \mathcal{S}_m{}^2 \backslash \mathcal{S}_m{}^1$, when only $\boldsymbol{Z}_{ij}^2$ is missing

$$\begin{aligned}\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^1, \boldsymbol{\theta}_{0,k}) &= E\{h(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^1, \boldsymbol{Z}_{ij}^2, \boldsymbol{\theta}_{0,k}) \mid \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^1, D_i = k \in \mathcal{S}_m{}^2 \backslash \mathcal{S}_m{}^1\} \\ &= E\{h(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^1, \boldsymbol{Z}_{ij}^2, \boldsymbol{\theta}_{0,k}) \mid \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^1, D_i \in \mathcal{S}_o{}^2\},\end{aligned}$$

where the expectation is under the distribution of $\boldsymbol{Z}_{ij}^2$ given $\boldsymbol{X}_{ij}$ and $\boldsymbol{Z}_{ij}^1$ and D_i. Following similar procedures developed above for aligned missing covariates, we can estimate study-specific parameters in $\mathcal{S}_m{}^1$ and in $\mathcal{S}_m{}^2 \backslash \mathcal{S}_m{}^1$ by incorporating the corresponding mean models induced from the above nested missing patterns with the respective $\boldsymbol{\eta}_k(\cdot)$ functions.

The situation of misaligned missingness (i.e. scenario (iii) in Figure 3.1) may also occur in practice. This is a very challenging situation, because different studies collect exclusive sources of covariates on subjects and no studies collect complete data. Using the same notation in the nested missingness above, we now have $\mathcal{S}_m{}^1 \cap \mathcal{S}_m{}^2 = \emptyset$. Model (3.1) is now rewritten as follows:

$$\mu_{k,ij} = E(Y_{ij} \mid \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^1, \boldsymbol{Z}_{ij}^2, D_i = k) = h(\boldsymbol{X}_{ij}^T \boldsymbol{\beta}_{0,k} + \boldsymbol{Z}_{ij}^1{}^T \boldsymbol{\lambda}_{0,k}^1 + \boldsymbol{Z}_{ij}^2{}^T \boldsymbol{\lambda}_{0,k}^2),\ k = 1, \ldots, K.$$

Assumption (3.2) implies that in study $k \in \mathcal{S}_m{}^1$, $D_i \perp \boldsymbol{Z}_i^1 \mid \boldsymbol{X}_i, \boldsymbol{Z}_i^2$, while in study $k \in \mathcal{S}_m{}^2$, $D_i \perp \boldsymbol{Z}_i^2 \mid \boldsymbol{X}_i, \boldsymbol{Z}_i^1$. This assumption is no longer sufficient to deal with the case when $\boldsymbol{Z}_i^1$ and $\boldsymbol{Z}_i^2$ are completely misaligned. To overcome, we postulate the following extra assumption to facilitate the estimation procedure:

$$\boldsymbol{Z}_i^1 \perp \boldsymbol{Z}_i^2 \mid \boldsymbol{X}_i. \tag{3.10}$$

Under the assumptions of (3.2), (3.9) and (3.10), we know $D_i \perp \boldsymbol{Z}_i^2 \mid \boldsymbol{X}_i$. Taking a conditional expectation on $h(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^1, \boldsymbol{Z}_{ij}^2, \boldsymbol{\theta}_{0,k})$ with respect to $\boldsymbol{Z}_{ij}^1$ in study $k \in {\mathcal{S}_m}^1$, we obtain

$$\begin{aligned}\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^2, \boldsymbol{\theta}_{0,k}) &= E\{h(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^1, \boldsymbol{Z}_{ij}^2, \boldsymbol{\theta}_{0,k}) \mid \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^2, D_i = k \in {\mathcal{S}_m}^1\} \\ &= E\{h(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^1, \boldsymbol{Z}_{ij}^2, \boldsymbol{\theta}_{0,k}) \mid \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^2, D_i \in {\mathcal{S}_m}^2\} \\ &= E\{h(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^1, \boldsymbol{Z}_{ij}^2, \boldsymbol{\theta}_{0,k}) \mid \boldsymbol{X}_{ij}, D_i \in {\mathcal{S}_m}^2\}.\end{aligned}$$

Similar derivations hold for study $k \in {\mathcal{S}_m}^2$. It is interesting to note that when Y_{ij} follows linear or log-linear model, $\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^2, \boldsymbol{\theta}_{0,k})$ can be estimated by similar procedures discussed in Section 3.3. For instance, in the linear model, we have

$$\begin{aligned}\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^2, \boldsymbol{\theta}_{0,k}) &= E(\boldsymbol{X}_{ij}\boldsymbol{\beta}_{0,k} + {\boldsymbol{Z}_{ij}^1}^T \boldsymbol{\lambda}_{0,k}^1 + {\boldsymbol{Z}_{ij}^2}^T \boldsymbol{\lambda}_{0,k}^2 \mid \boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^2, D_i = k \in {\mathcal{S}_m}^1) \\ &= \boldsymbol{X}_{ij}\boldsymbol{\beta}_{0,k} + {\boldsymbol{Z}_{ij}^2}^T \boldsymbol{\lambda}_{0,k}^2 + E(\boldsymbol{Z}_{ij}^1 \mid \boldsymbol{X}_{ij}, D_i \in {\mathcal{S}_m}^2)^T \boldsymbol{\lambda}_{0,k}^1,\end{aligned}$$

in which $E(\boldsymbol{Z}_{ij}^1 \mid \boldsymbol{X}_{ij}, D_i \in {\mathcal{S}_m}^2)$ may be estimated parametrically by a linear model or nonparametrically through the spline-based sieve method. In the log-linear model, similar derivations can be carried out because factorization leads to a separable structure multiplicatively so that $\eta_k(\boldsymbol{X}_{ij}, \boldsymbol{Z}_{ij}^2, \boldsymbol{\theta}_{0,k})$ can be directly estimated. When Y_{ij} follows a logistic model, this separation property no longer holds, and thus it is necessary to include some studies with both fully observed $\boldsymbol{Z}_{ij}^1$ and $\boldsymbol{Z}_{ij}^2$ in order to apply the proposed method in this chapter.

The situation of partially misaligned missingness (i.e. scenario (iv) in Figure 3.1) can be handled using similar strategies to those proposed for the case of misaligned missingness and nested missingness. Studies in ${\mathcal{S}_m}^1 \backslash {\mathcal{S}_m}^2$ and ${\mathcal{S}_m}^2 \backslash {\mathcal{S}_m}^1$ are exactly the case of misaligned missingness, while to deal with studies in ${\mathcal{S}_m}^1 \cap {\mathcal{S}_m}^2$ we need studies with fully observed $\boldsymbol{Z}^1$ and $\boldsymbol{Z}^2$, assuming the complementary set $({\mathcal{S}_m}^1 \cup {\mathcal{S}_m}^2)^c$ containing the other studies is not empty.

3.7 Simulation Studies

We conduct several simulation studies to evaluate performances of the proposed methods.

3.7.1 Simulation Study I: Comparison to Existing Methods

We run a simulation study to compare our proposed method with two existing methods, GEE and QIF, using complete data, imputed data by either parametric multiple imputation or nonparametric hot-deck multiple imputation (*Little and Rubin*, 1987). Summary statistics are drawn based on 4000 datasets generated from the following model:

$$Y_{ij} = \begin{cases} \beta_{0,1} + \beta_{1,1}X_{ij} + \lambda_{1,1}Z_{ij} + \epsilon^1_{ij}, & D_i = 1 \\ \beta_{0,2} + \beta_{1,2}X_{ij} + \lambda_{1,2}Z_{ij} + \epsilon^2_{ij}, & D_i = 2 \end{cases}, \quad j = 1, \ldots, m, i = 1, \ldots, n,$$

where the intercepts are set for the same, $\beta_{0,1} = \beta_{0,2} = 1$, simply denoted by β_0, and the true regression coefficients are $\boldsymbol{\theta}_0 = (\beta_0, \beta_{1,1}, \lambda_{1,1}, \beta_{1,2}, \lambda_{1,2})^T = (1, 1, -0.5, 2, 0.5)^T$. Also, we set $n = 200$ subjects and $m = 4$ repeated measurements. Covariate X_{ij} is generated from Unif$(0, 1)$, and covariates Z_{ij} is generated from a conditional model given X_{ij} of the form: $Z_{ij} = sin(4\pi X_{ij}) + \zeta_{ij}$, where $\zeta_{ij} \overset{iid}{\sim} N(0, 0.5)$. Here Z_{ij} is treated as a study-specific missing covariate whose state of missingness, D_i, is determined by a logistic model on X_{i1}, logit$\{P(D_i = 2 \mid \boldsymbol{X}_i)\} = 0.5 + 0.4X_{i1}$. As a result, 39% of subjects are sampled from study 2 and treated as missing subjects. The above specification implies that $E(Y_{ij} \mid X_{ij}, D_i = 2) = \beta_0 + \beta_{1,2}X_{ij} + \lambda_{1,2}sin(4\pi X_{ij})$. Error terms,

$\boldsymbol{\epsilon}_i^k = (\epsilon_{i1}^k, \ldots, \epsilon_{im}^k)^T, k = 1, 2,$ are independently generated from $N_m(\mathbf{0}, \phi_k \boldsymbol{R}_k(\alpha_k))$, $k = 1, 2$, where the covariance matrix $\phi_k \boldsymbol{R}_k(\alpha_k)$ is specified in the following two cases:

Case I. correlation matrices $\boldsymbol{R}_1(\cdot)$ and $\boldsymbol{R}_2(\cdot)$ in two studies are both AR-1 correlation with $(\alpha_1, \alpha_2) = (0.4, 0.4)$, and variance parameters are $(\phi_1, \phi_2) = (1, 1)$;

Case II. correlation matrix $\boldsymbol{R}_1(\cdot)$ in study 1 is AR-1 with $\alpha_1 = 0.7$ while correlation matrix $\boldsymbol{R}_2(\cdot)$ in study 2 is compound symmetry with $\alpha_2 = 0.2$; variance parameters are different, $(\phi_1, \phi_2) = (10, 1)$.

The imputed datasets for study 2 are created according to the true conditional distribution of $\boldsymbol{Z}_i$ given $\boldsymbol{X}_i$ to avoid potential uncertainty in the estimation of this conditional distribution. Here we use $f(\boldsymbol{Z}_i \mid \boldsymbol{X}_i)$ for imputation instead of $f(\boldsymbol{Z}_i \mid \boldsymbol{X}_i, \boldsymbol{Y}_i)$ because two studies are governed by two different regression models, and therefore $f(\boldsymbol{Z}_i \mid \boldsymbol{X}_i, \boldsymbol{Y}_i)$ in study 2 is not estimable using observed data in study 1 (see a detailed explanation provided in a paragraph below). Likewise, in the implementation of hot-deck imputation we select a set of observed data that are similar to the missing $\boldsymbol{Z}_i$ in terms of small Euclidean distances in their $\boldsymbol{X}_i$ values, in which we randomly generate 10 imputed datasets.

The conditional mean function, $E(Y_{ij} \mid X_{ij}, D_i = 2) = \beta_0 + \beta_{1,2} X_{ij} + \lambda_{1,2} sin(4\pi X_{ij})$, is estimated using the B-spline regression with 6 basis functions. We postpone our discussion about the basis function selection to Section 3.7.3, and now focus on the comparison of our method with the imputation methods. Simulation results for the true model of case I and II above are reported in Tables 3.1, 3.2, 3.3 and 3.4 under two different working correlation structures. In the ideal case where the complete data are used, both QIF and GEE have shown little biases and reached desirable 95% nominal coverage for both working correlation scenarios. When covariate $\boldsymbol{Z}_i$ is missing in study 2, both parametric and hot-deck multiple imputation methods produce noticeable estimation biases in GEE and QIF, particularly for those parameters exclusively belonging to study 2, where severe undercoverage is evident for for $\beta_{1,2}$ and $\lambda_{1,2}$ (substantially lower than 95% nominal level). Similar results are drawn from Tables 3.3 and 3.4.

The failure of both parametric multiple imputation and hot-deck imputation may be attributed to the validity of the imputation methods, which have been justified only under the selection model. Note that in a selection model regression parameters are present in the distribution $f(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, \boldsymbol{Z}_i, \theta)$, which however is not the case in this simulation model where regression parameters are different across two studies. Thus the imputation is in general not applicable to multiple studies that are governed by models with different parameters.

In effect $f(\boldsymbol{Z}_i \mid \boldsymbol{Y}_i, \boldsymbol{X}_i, D_i = 2)$ cannot be estimated using the observed data from study 1. Even if here the true conditional distribution $f(\boldsymbol{Z}_i \mid \boldsymbol{X}_i)$ is used in the imputation, imputed values for missing $\boldsymbol{Z}_i$ may still violate unbiasedness of $E\{Y_{ij} - h(\beta_0 + \beta_{1,2}X_{ij} + \lambda_{1,2}Z_{ij}) \mid X_{ij}, Z_{ij}, D_i = 2\} = 0$. Therefore, both GEE and QIF with the imputed data are impaired and yield significant estimation biases. *Molenberghs and Kenward* (2007, Chap. 2) examine the performance of GEE with multiple imputation for missing responses. By comparing IPW GEE with imputation-based GEE under the selection model, they show that imputation-based GEE produces significantly larger bias as well as mean squared error (MSE) than IPW GEE in various longitudinal data settings. Our findings from the above simulation study are in agreement with theirs.

In contrast to the imputation methods, our proposed method demonstrates satisfactory performances in terms of bias and coverage. For example, the coverage of $\lambda_{1,2}$ is close to the nominal 95% level in various settings. This is because our method uses asymptotically unbiased estimating functions derived by plugging in a consistent nonparametric estimation of $E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, D_i = 2)$. As shown in Tables 3.1 and 3.2 for case I, and Tables 3.3 and 3.4 for case II, the price paid to gain the benefit of desirable coverage is larger standard deviations than the ideal QIF and GEE using the complete data. This is not surprising because $E(\boldsymbol{Y}_i \mid \boldsymbol{X}_i, D_i = 2)$ is estimated nonparametrically in our method. This further confirms the theoretical results given in Theorem III.2 and Theorem III.3 regarding the asymptotic covariances, in which, as explained already, the uncertainty from the nonparametric estimation is to be accounted for.

3.7.2 Simulation Study II: Comparison under Different Missing Percentages

We now explore the performance of our proposed methods with various missing data percentages, namely different proportions of subjects between the two study cohorts. We consider three missing data percentages, 31%, 50% and 62%, which are yielded by varying coefficients in the model $\text{logit}\{P(D_i = 2 \mid X_{ij})\} = \vartheta_0 + \vartheta_1 X_{i1}$. The corresponding parameter estimates and coverage probabilities are listed in Table 3.5 under AR-1 working correlation. When the missing data percentage increases, the amount of estimation bias increases, which is resulted from the reduction of fully observed data in study 1. Consequently, the mean square errors increase for all parameters β_0, $\beta_{1,1}$ and $\lambda_{1,1}$ in study 1, whereas the mean square errors drop for parameters $\beta_{1,2}$ and $\lambda_{1,2}$ in study 2 due to the increased size of observed data in study 2. It is noticeable that the estimated coverage rates of the regression parameters remain close to 95% nominal level in all these settings.

3.7.3 Simulation Study III: Basis Function Selection

We also examine how the proposed BIC criterion behaves in the selection of basis functions. Under the same settings of the first simulation study, we increase the number of basis functions from 4 to 12 in the estimation of $E(Y_{ij} \mid \boldsymbol{X}_{ij}, D_i = 2)$, and summarize the results in Figure 3.2. This figure indicates that BIC criterion is minimized at 6, after which the MSE cannot be improved significantly with more basis functions being used. This evidence implies that our criterion tends to chose a parsimonious nonparametric model with small MSE.

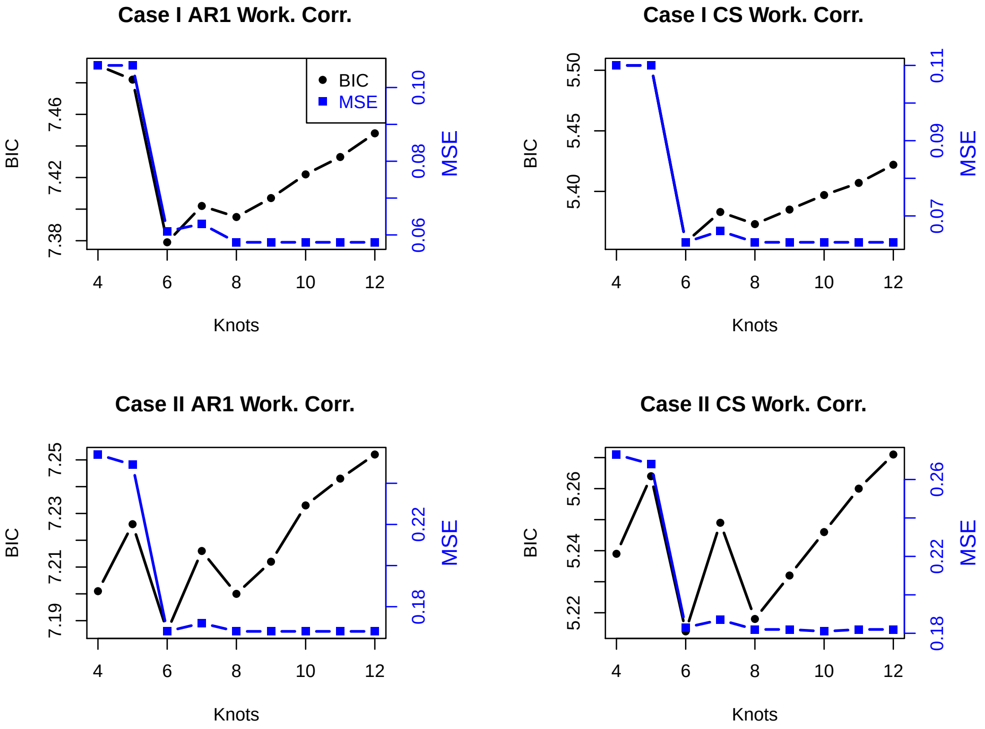

Figure 3.2: BIC and MSE profile curves for knot selection.

3.7.4 Simulation Study IV: Comparison of Joint Analysis and Individual analysis

To illustrate the efficiency gain in the joint analysis, we run a simulation study to compare the standard errors obtained from the joint analysis and those obtained from the individual analysis. This is to confirm the theoretical result given in Theorem III.3. The data is generated in the same way as in case I of the first simulation in Section 3.7.1. The joint analysis utilizes the fact that two studies have a common intercept parameter, while the individual analysis ignores this fact and includes different intercepts in the respective models. The standard errors are calculated by the bootstrap method discussed in Section 3.5. Summarized results over 100 replications in Table 3.6 clearly show that the joint analysis has given smaller standard errors for all regression coefficients. This efficiency improvement appears very substantial for study 2 where missing covariates are present. The individual analysis only uses 61% of the sample size to obtain parameter estimation. In conclusion, it is clearly beneficial to borrow data information from study 1 to improve inference for the parameters in study 2.

3.8 Application

We apply our method to analyze the lead exposure data collected from two longitudinal cohorts of infants in Mexico City. Between 1994 and 2005 the study recruited 89 mother-infant pairs in cohort B and 492 mother-infant pairs in cohort C at two maternity hospitals serving low-to-moderate income populations (*Afeiche et al.*, 2011). We are interested in studying the effect of cord blood lead exposure on child's weight growth.

Child's weight was measured repeatedly at every 0, 3, 6, 12, 18, 24, 30, 36, 48 and 60 months after birth in cohort B, while at 0, 1, 4, 7, 12, 18, 24, 30, 36, 42 and 48 months in cohort C. Two lead exposure measures, mother's blood lead (PBL) and child's cord blood lead (CBL), are recorded at baseline. PBL was measured for all mothers in both cohorts while CBL was collected for all infants in cohort C and approximately 46% of infants in cohort B due to children's or maternal refusal, inability to give blood or because a blood lead measure was not scheduled.

Figure 3.3 displays trajectories of child's weights vs child's ages across two cohorts, and Figure 3.4 includes two scatter-plots of child's weights vs child's CBL in log scale. Adjusting child's gender and age, we estimate the effect of CBL on weight growth via the following model:

$$E\{Y_{k,ij} \mid X_{k,ij}, Z_{k,ij}\} = \beta_{0,k} + \beta_{1,k}X_{k,ij} + \beta_{2,k}G_{k,i} + \beta_{3,k}B_1(t_{k,ij}) + \beta_{4,k}B_2(t_{k,ij}) + \beta_{5,k}B_3(t_{k,ij}) + \lambda_{1,k}B_1(Z_{k,ij}) + \lambda_{2,k}B_2(Z_{k,ij}), \quad k = 1, 2, \tag{3.11}$$

where cohorts C and B correspond to $k = 1$ and $k = 2$, respectively. For subject i at jth visit, variable $Y_{k,ij}$, $t_{k,ij}$, $X_{k,ij}$, $G_{k,i}$ and $Z_{k,ij}$ are log(weight), child's ages (year), log(PBL), child's gender (1 for male and 0 for female), and log(CBL), respectively. We apply log-transformation on weight, PBL and CBL to reduce skewness. Effects on time $t_{k,ij}$ and $Z_{k,ij}$ are captured by linear splines with three basis functions, $B_1(t_{k,ij})$, $B_2(t_{k,ij})$ and $B_3(t_{k,ij})$, for covariate time $t_{k,ij}$ at knots 0.5 and 2, and two basis functions, $B_1(Z_{k,ij})$ and $B_2(Z_{k,ij})$, for $Z_{k,ij}$ at knot 2.3 in log scale. The piecewise linear trend of child's weight versus child's age can be observed in Figure 3.3. Given that 46% of CBL measurements are missing in cohort B, we estimate the effect of covariate CBL by merging the two cohorts. Through a routine model screening process using

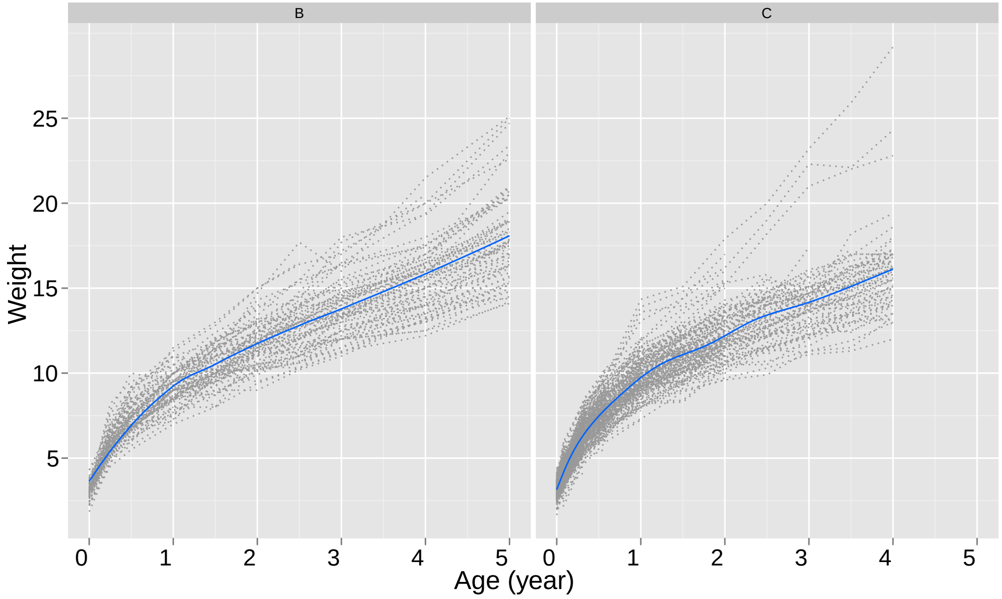

Figure 3.3: Trajectories of children's weights vs children's ages over two cohorts.

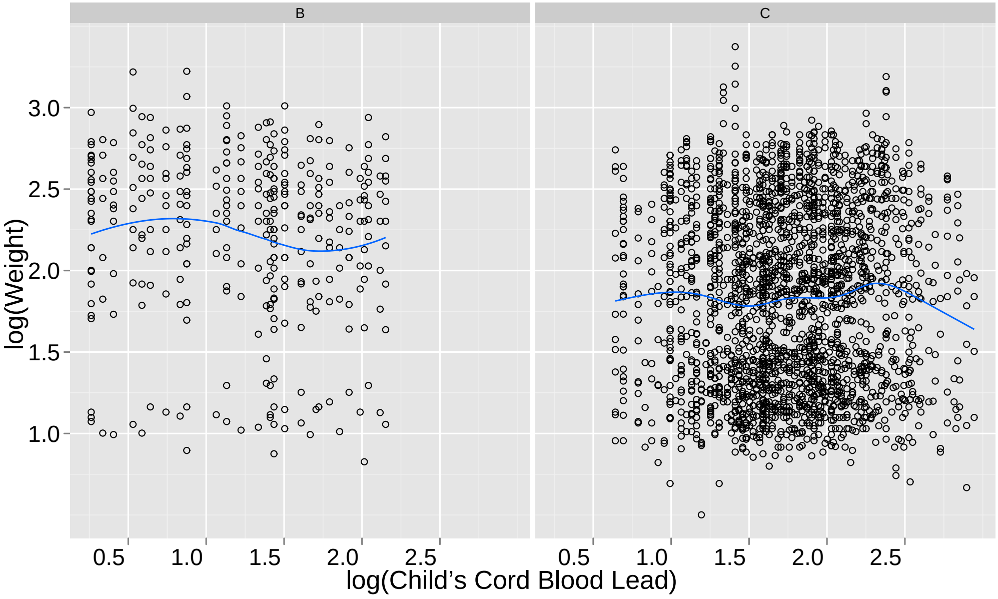

Figure 3.4: Scatterplots of log-transformed children's weights vs log-transformed children's cord blood lead exposure across two cohorts.

interactions between covariates and cohort dummy variates, we finally reach a model with common coefficients for $X_{k,ij}$, $G_{k,i}$, $B_1(t_{k,ij})$ and $B_2(t_{k,ij})$ across two cohorts.

Results in Table 3.7 indicate that gender and age both are strongly associated with weight growth of children. For children age 2 or younger in two cohorts, they have similar weight growth on average. Children older than 2 years in cohort B grow faster than their peers in cohort C. As for the effect of lead exposure in child's cord blood the effect of $\log(CBL)$ on weight growth in cohort C appears to be nearly significant when $\log(CBL)$ is greater than 2.3, or equivalently CBL concentration larger than $10\mu g/L$.

3.9 Concluding Remarks

We have developed a novel estimating function approach to assessing covariate effects through merging datasets from multiple longitudinal studies. The proposed method accounts for various aspects of heterogeneity across studies so the resulting estimation and inference are not only synthenized with integrated data and but also adaptive to individual study features. The innovation of our method lies in the strategies of handling multiple datasets with study-specific missing covariates, which often occur in data merging. In this setting, different missing patterns cannot be handled properly by traditional imputation approach which requires different datasets to be generated under the same distribution. When datasets of multiple studies are collected respectively from different subpopulations, it is problematic to use studies with fully observed data either to impute study-specific missing covariates or to adjust the chance of missingness by the method of inverse probability weighting. Our approach features a sieve nonparametric estimation of a marginalized mean model which is resulted from integrating the set of missing covariates out the original full-data mean model. Under appropriate missing mechanism assumptions, the marginalized mean model can be estimated properly by using studies with fully observed covariates and hence the resulting estimation for regression coefficients is consistent and asymptotically normal.

In addition, we evaluate bootstrap variance estimation and BIC criterion for the selection of basis functions in nonparametric estimation. More importantly the implementation of our method is numerically straightforward. Both theoretical and numerical evidence is provided to show the large-sample properties and finite-sample performances of the proposed methods. In conclusion, our method works well to handle study-specific missing covariates as long as the full-data mean model and missing data mechanisms are both appropriately specified. Since our method relies on the nonparametric estimation of the marginalized estimating functions, it could be challenged when the number of observed covariates is large. Also when the number of studies is large, it would be computationally demanding to use traditional hypothesis testing method to determine shared parameters across studies in the joint analysis. Providing a flexible and efficient way to detect common parameters in multiple studies in the presence of missing covariates is worth future exploration.

Table 3.1: Summary of regression parameter estimates for data generated in Case I under AR-1 working correlation. Complete, Par-MI and Hot-deck represent complete data, data imputed by parametric multiple imputation and data imputed by hot-deck multiple imputation, respectively. E.S.E. is the empirical standard error computed from 4000 simulated datasets. A.S.E. is the asymptotic standard error. For our method, A.S.E. is the bootstrap standard error computed using 400 bootstrap samples. The coverage probability, C.P. is computed by using the asymptotic standard error.

Data	Methods	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\theta}}$	Bias	E.S.E	A.S.E.	MSE	C.P.
Complete	QIF	β_0	1.000	0.000	0.075	0.075	0.006	0.952
		$\beta_{1,1}$	1.001	0.001	0.124	0.126	0.015	0.948
		$\lambda_{1,1}$	-0.501	-0.001	0.040	0.040	0.002	0.946
		$\beta_{1,2}$	2.006	0.006	0.158	0.160	0.025	0.956
		$\lambda_{1,2}$	0.499	-0.001	0.060	0.058	0.004	0.941
	GEE	β_0	1.000	0.000	0.072	0.072	0.005	0.945
		$\beta_{1,1}$	1.002	0.002	0.122	0.122	0.015	0.948
		$\lambda_{1,1}$	-0.501	-0.001	0.040	0.039	0.002	0.939
		$\beta_{1,2}$	2.003	0.003	0.151	0.152	0.023	0.949
		$\lambda_{1,2}$	0.499	-0.001	0.057	0.055	0.003	0.951
Par-MI	QIF	β_0	1.035	0.035	0.077	0.079	0.007	0.927
		$\beta_{1,1}$	0.961	-0.039	0.126	0.129	0.017	0.939
		$\lambda_{1,1}$	-0.504	-0.004	0.040	0.040	0.002	0.944
		$\beta_{1,2}$	1.865	-0.135	0.169	0.180	0.047	0.894
		$\lambda_{1,2}$	0.241	-0.259	0.049	0.083	0.069	0.040
	GEE	β_0	1.042	0.042	0.074	0.076	0.007	0.917
		$\beta_{1,1}$	0.951	-0.049	0.123	0.125	0.018	0.923
		$\lambda_{1,1}$	-0.505	-0.005	0.040	0.039	0.002	0.939
		$\beta_{1,2}$	1.851	-0.149	0.162	0.169	0.049	0.879
		$\lambda_{1,2}$	0.239	-0.261	0.047	0.078	0.070	0.017
Hot-deck	QIF	β_0	1.034	0.034	0.076	0.079	0.007	0.931
		$\beta_{1,1}$	0.961	-0.039	0.126	0.129	0.017	0.937
		$\lambda_{1,1}$	-0.504	-0.004	0.040	0.040	0.002	0.944
		$\beta_{1,2}$	1.866	-0.134	0.170	0.181	0.047	0.893
		$\lambda_{1,2}$	0.237	-0.263	0.049	0.083	0.072	0.039
	GEE	β_0	1.041	0.041	0.073	0.076	0.007	0.923
		$\beta_{1,1}$	0.952	-0.048	0.123	0.125	0.018	0.924
		$\lambda_{1,1}$	-0.504	-0.004	0.040	0.039	0.002	0.937
		$\beta_{1,2}$	1.852	-0.148	0.162	0.170	0.048	0.879
		$\lambda_{1,2}$	0.235	-0.265	0.047	0.078	0.073	0.015
	Our method	β_0	1.002	0.002	0.080	0.081	0.006	0.944
		$\beta_{1,1}$	1.000	0.000	0.129	0.130	0.017	0.943
		$\lambda_{1,1}$	-0.501	-0.001	0.040	0.040	0.002	0.949
		$\beta_{1,2}$	1.992	-0.008	0.206	0.207	0.042	0.948
		$\lambda_{1,2}$	0.483	-0.017	0.211	0.210	0.045	0.936

Table 3.2: Summary of regression parameter estimates for data generated in Case I under compound symmetry working correlation. Complete, Par-MI and Hot-deck represent complete data, data imputed by parametric multiple imputation and data imputed by hot-deck multiple imputation, respectively. E.S.E. is the empirical standard error computed from 4000 simulated datasets. A.S.E. is the asymptotic standard error. For our method, A.S.E. is the bootstrap standard error computed using 400 bootstrap samples. The coverage probability, C.P. is computed by using the asymptotic standard error.

Data	Methods	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\theta}}$	Bias	E.S.E	A.S.E.	MSE	C.P.
Complete	QIF	β_0	1.001	0.001	0.078	0.077	0.006	0.946
		$\beta_{1,1}$	1.000	0.000	0.131	0.131	0.017	0.951
		$\lambda_{1,1}$	-0.501	-0.001	0.043	0.042	0.002	0.944
		$\beta_{1,2}$	2.002	0.002	0.165	0.164	0.027	0.952
		$\lambda_{1,2}$	0.498	-0.002	0.062	0.060	0.004	0.946
	GEE	β_0	1.001	0.001	0.076	0.075	0.006	0.947
		$\beta_{1,1}$	1.000	0.000	0.130	0.128	0.017	0.948
		$\lambda_{1,1}$	-0.501	-0.001	0.042	0.041	0.002	0.940
		$\beta_{1,2}$	2.002	0.002	0.160	0.158	0.026	0.952
		$\lambda_{1,2}$	0.498	-0.002	0.060	0.058	0.004	0.940
Par-MI	QIF	β_0	1.036	0.036	0.079	0.081	0.008	0.931
		$\beta_{1,1}$	0.958	-0.042	0.133	0.134	0.019	0.929
		$\lambda_{1,1}$	-0.504	-0.004	0.043	0.042	0.002	0.942
		$\beta_{1,2}$	1.863	-0.137	0.172	0.183	0.048	0.894
		$\lambda_{1,2}$	0.240	-0.260	0.050	0.084	0.070	0.047
	GEE	β_0	1.042	0.042	0.077	0.078	0.008	0.914
		$\beta_{1,1}$	0.949	-0.051	0.130	0.131	0.020	0.926
		$\lambda_{1,1}$	-0.505	-0.005	0.042	0.041	0.002	0.934
		$\beta_{1,2}$	1.853	-0.147	0.169	0.174	0.050	0.872
		$\lambda_{1,2}$	0.239	-0.261	0.049	0.080	0.071	0.031
Hot-deck	QIF	β_0	1.035	0.035	0.079	0.081	0.007	0.935
		$\beta_{1,1}$	0.959	-0.041	0.132	0.134	0.019	0.933
		$\lambda_{1,1}$	-0.504	-0.004	0.043	0.042	0.002	0.944
		$\beta_{1,2}$	1.864	-0.136	0.173	0.183	0.048	0.893
		$\lambda_{1,2}$	0.236	-0.264	0.050	0.084	0.072	0.038
	GEE	β_0	1.042	0.042	0.076	0.078	0.008	0.917
		$\beta_{1,1}$	0.949	-0.051	0.130	0.131	0.019	0.927
		$\lambda_{1,1}$	-0.505	-0.005	0.042	0.041	0.002	0.936
		$\beta_{1,2}$	1.855	-0.145	0.169	0.174	0.050	0.871
		$\lambda_{1,2}$	0.235	-0.265	0.049	0.080	0.073	0.022
	Our Method	β_0	1.002	0.002	0.083	0.084	0.007	0.942
		$\beta_{1,1}$	0.999	-0.001	0.136	0.136	0.018	0.943
		$\lambda_{1,1}$	-0.501	-0.001	0.043	0.042	0.002	0.940
		$\beta_{1,2}$	1.992	-0.008	0.210	0.211	0.044	0.960
		$\lambda_{1,2}$	0.483	-0.017	0.210	0.214	0.044	0.943

Table 3.3: Summary of regression parameter estimates for data generated in Case II under AR-1 working correlation. Complete, Par-MI and Hot-deck represent complete data, data imputed by parametric multiple imputation and data imputed by hot-deck multiple imputation, respectively. E.S.E. is the empirical standard error computed from 4000 simulated datasets. A.S.E. is the asymptotic standard error. For our method, A.S.E. is the bootstrap standard error computed using 400 bootstrap samples. The coverage probability, C.P. is computed by using the asymptotic standard error.

Data	Methods	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\theta}}$	Bias	E.S.E	A.S.E.	MSE	C.P.
Complete	QIF	β_0	1.002	0.002	0.132	0.124	0.017	0.917
		$\beta_{1,1}$	0.992	-0.008	0.287	0.282	0.082	0.938
		$\lambda_{1,1}$	-0.497	0.003	0.093	0.096	0.009	0.952
		$\beta_{1,2}$	2.002	0.002	0.219	0.210	0.048	0.936
		$\lambda_{1,2}$	0.503	0.003	0.063	0.064	0.004	0.947
	GEE	β_0	1.009	0.009	0.201	0.186	0.041	0.923
		$\beta_{1,1}$	0.984	-0.016	0.290	0.283	0.084	0.944
		$\lambda_{1,1}$	-0.498	0.002	0.089	0.090	0.008	0.957
		$\beta_{1,2}$	1.995	-0.005	0.243	0.229	0.059	0.929
		$\lambda_{1,2}$	0.502	0.002	0.069	0.069	0.005	0.943
Par-MI	QIF	β_0	1.102	0.102	0.140	0.136	0.030	0.884
		$\beta_{1,1}$	0.921	-0.079	0.289	0.285	0.090	0.928
		$\lambda_{1,1}$	-0.502	-0.002	0.093	0.096	0.009	0.954
		$\beta_{1,2}$	1.783	-0.217	0.231	0.237	0.100	0.866
		$\lambda_{1,2}$	0.235	-0.265	0.049	0.089	0.073	0.049
	GEE	β_0	1.052	0.052	0.203	0.188	0.044	0.916
		$\beta_{1,1}$	0.952	-0.048	0.290	0.283	0.086	0.941
		$\lambda_{1,1}$	-0.501	-0.001	0.089	0.090	0.008	0.957
		$\beta_{1,2}$	1.800	-0.200	0.251	0.251	0.103	0.882
		$\lambda_{1,2}$	0.237	-0.263	0.055	0.096	0.072	0.093
Hot-deck	QIF	β_0	1.101	0.101	0.141	0.137	0.030	0.886
		$\beta_{1,1}$	0.922	-0.078	0.289	0.285	0.090	0.935
		$\lambda_{1,1}$	-0.502	-0.002	0.093	0.096	0.009	0.952
		$\beta_{1,2}$	1.786	-0.214	0.232	0.238	0.099	0.856
		$\lambda_{1,2}$	0.232	-0.268	0.050	0.089	0.074	0.056
	GEE	β_0	1.052	0.052	0.203	0.188	0.044	0.916
		$\beta_{1,1}$	0.952	-0.048	0.290	0.283	0.086	0.942
		$\lambda_{1,1}$	-0.500	0.000	0.090	0.090	0.008	0.957
		$\beta_{1,2}$	1.801	-0.199	0.251	0.252	0.102	0.881
		$\lambda_{1,2}$	0.233	-0.267	0.057	0.096	0.074	0.101
	Our method	β_0	1.007	0.007	0.179	0.163	0.032	0.931
		$\beta_{1,1}$	0.989	-0.011	0.298	0.288	0.089	0.937
		$\lambda_{1,1}$	-0.497	0.003	0.093	0.095	0.009	0.953
		$\beta_{1,2}$	1.990	-0.010	0.326	0.309	0.106	0.941
		$\lambda_{1,2}$	0.487	-0.013	0.259	0.255	0.067	0.956

Table 3.4: Summary of regression parameter estimates for data generated in Case II under compound symmetry working correlation. Complete, Par-MI and Hot-deck represent complete data, data imputed by parametric multiple imputation and data imputed by hot-deck multiple imputation, respectively. E.S.E. is the empirical standard error computed from 4000 simulated datasets. A.S.E. is the asymptotic standard error. For our method, A.S.E. is the bootstrap standard error computed using 400 bootstrap samples. The coverage probability, C.P. is computed by using asymptotic standard error.

Data	Method	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\theta}}$	Bias	E.S.E	A.S.E.	MSE	C.P.
Complete	QIF	β_0	1.001	0.001	0.129	0.123	0.017	0.929
		$\beta_{1,1}$	0.988	-0.012	0.306	0.301	0.094	0.947
		$\lambda_{1,1}$	-0.497	0.003	0.104	0.105	0.011	0.958
		$\beta_{1,2}$	2.004	0.004	0.212	0.205	0.045	0.940
		$\lambda_{1,2}$	0.503	0.003	0.061	0.062	0.004	0.941
	GEE	β_0	1.011	0.011	0.211	0.197	0.045	0.935
		$\beta_{1,1}$	0.979	-0.021	0.330	0.319	0.109	0.934
		$\lambda_{1,1}$	-0.498	0.002	0.105	0.104	0.011	0.952
		$\beta_{1,2}$	1.998	-0.002	0.241	0.228	0.058	0.928
		$\lambda_{1,2}$	0.502	0.002	0.062	0.063	0.004	0.947
Par-MI	QIF	β_0	1.105	0.105	0.137	0.135	0.030	0.881
		$\beta_{1,1}$	0.906	-0.094	0.308	0.305	0.103	0.935
		$\lambda_{1,1}$	-0.504	-0.004	0.104	0.105	0.011	0.954
		$\beta_{1,2}$	1.779	-0.221	0.222	0.232	0.098	0.866
		$\lambda_{1,2}$	0.234	-0.266	0.048	0.086	0.073	0.034
	GEE	β_0	1.054	0.054	0.212	0.198	0.048	0.918
		$\beta_{1,1}$	0.943	-0.057	0.330	0.319	0.112	0.935
		$\lambda_{1,1}$	-0.501	-0.001	0.105	0.104	0.011	0.952
		$\beta_{1,2}$	1.810	-0.190	0.246	0.246	0.097	0.892
		$\lambda_{1,2}$	0.236	-0.264	0.049	0.087	0.072	0.036
Hot-deck	QIF	β_0	1.104	0.104	0.137	0.136	0.030	0.882
		$\beta_{1,1}$	0.908	-0.092	0.308	0.305	0.103	0.941
		$\lambda_{1,1}$	-0.504	-0.004	0.104	0.105	0.011	0.952
		$\beta_{1,2}$	1.782	-0.218	0.223	0.233	0.098	0.864
		$\lambda_{1,2}$	0.231	-0.269	0.049	0.087	0.075	0.044
	GEE	β_0	1.054	0.054	0.212	0.198	0.048	0.920
		$\beta_{1,1}$	0.944	-0.056	0.330	0.319	0.112	0.936
		$\lambda_{1,1}$	-0.500	0.000	0.105	0.104	0.011	0.953
		$\beta_{1,2}$	1.811	-0.189	0.247	0.247	0.096	0.891
		$\lambda_{1,2}$	0.232	-0.268	0.052	0.087	0.074	0.045
	Our method	β_0	1.008	0.008	0.176	0.164	0.031	0.929
		$\beta_{1,1}$	0.982	-0.018	0.319	0.311	0.102	0.944
		$\lambda_{1,1}$	-0.498	0.002	0.104	0.104	0.011	0.955
		$\beta_{1,2}$	1.988	-0.012	0.316	0.309	0.100	0.952
		$\lambda_{1,2}$	0.485	-0.015	0.246	0.252	0.061	0.956

Table 3.5: Summary of regression parameter estimates by the proposed method in Case I and II under AR-1 working correlation and different missing percentages. M.P. is the missing percentage. E.S.E. is the empirical standard error computed from 4000 simulated datasets. A.S.E. is the asymptotic standard error. For our method, A.S.E. is the bootstrap standard error computed using 400 bootstrap samples. The coverage probability, C.P., is computed by using the asymptotic standard error.

Case	M.P.	$\boldsymbol{\theta}$	$\hat{\boldsymbol{\theta}}$	Bias	E.S.E	A.S.E.	MSE	C.P.
I	31%	β_0	0.999	-0.001	0.084	0.081	0.007	0.933
		$\beta_{1,1}$	0.997	-0.003	0.136	0.129	0.018	0.937
		$\lambda_{1,1}$	-0.501	-0.001	0.039	0.039	0.002	0.944
		$\beta_{1,2}$	2.001	0.001	0.220	0.210	0.049	0.941
		$\lambda_{1,2}$	0.489	-0.011	0.222	0.217	0.050	0.947
	50%	β_0	1.004	0.004	0.091	0.088	0.008	0.939
		$\beta_{1,1}$	1.002	0.002	0.148	0.145	0.022	0.943
		$\lambda_{1,1}$	-0.501	-0.001	0.046	0.046	0.002	0.947
		$\beta_{1,2}$	1.992	-0.008	0.194	0.192	0.038	0.950
		$\lambda_{1,2}$	0.483	-0.017	0.190	0.186	0.037	0.934
	62%	β_0	1.011	0.011	0.093	0.096	0.009	0.950
		$\beta_{1,1}$	0.985	-0.015	0.169	0.164	0.029	0.938
		$\lambda_{1,1}$	-0.497	0.003	0.055	0.054	0.003	0.946
		$\beta_{1,2}$	1.973	-0.027	0.192	0.195	0.037	0.944
		$\lambda_{1,2}$	0.475	-0.025	0.185	0.178	0.035	0.929
II	31%	β_0	1.013	0.013	0.159	0.165	0.025	0.956
		$\beta_{1,1}$	0.992	-0.008	0.285	0.286	0.081	0.949
		$\lambda_{1,1}$	-0.503	-0.003	0.091	0.093	0.008	0.960
		$\beta_{1,2}$	1.978	-0.022	0.309	0.315	0.096	0.950
		$\lambda_{1,2}$	0.485	-0.015	0.249	0.260	0.062	0.954
	50%	β_0	1.004	0.004	0.153	0.151	0.023	0.942
		$\beta_{1,1}$	0.994	-0.006	0.334	0.325	0.111	0.940
		$\lambda_{1,1}$	-0.501	-0.001	0.113	0.109	0.013	0.934
		$\beta_{1,2}$	1.993	-0.007	0.280	0.280	0.078	0.956
		$\lambda_{1,2}$	0.479	-0.021	0.231	0.224	0.054	0.943
	62%	β_0	1.023	0.023	0.145	0.145	0.022	0.954
		$\beta_{1,1}$	0.988	-0.012	0.374	0.370	0.140	0.945
		$\lambda_{1,1}$	-0.492	0.008	0.125	0.127	0.016	0.945
		$\beta_{1,2}$	1.959	-0.041	0.265	0.266	0.072	0.945
		$\lambda_{1,2}$	0.460	-0.040	0.204	0.213	0.043	0.939

Table 3.6: Comparison of standard errors from joint estimation and individual estimation in Case I under AR-1 and CS working correlations. For our method, standard error is the bootstrap standard error computed using 200 bootstrap samples.

		Standard Error			
		AR-1		CS	
Study	$\hat{\beta}$	Joint	Individual	Joint	Individual
	β_0	0.076	0.089	0.079	0.092
I	$\beta_{1,1}$	0.125	0.136	0.131	0.142
	$\lambda_{1,1}$	0.039	0.040	0.041	0.042
	β_0	0.076	0.258	0.079	0.255
II	$\beta_{1,2}$	0.174	0.448	0.178	0.442
	$\lambda_{1,2}$	0.097	0.228	0.100	0.225

Table 3.7: Estimates of regression parameters and p-values for the analysis of children's growth on lead exposures. Intercept, log(PBL), Gender, B_1(age) and B_2(age) have common coefficients across two cohorts, while B_3(age) and log(CBL) have different effect sizes.

	Cohort C		Cohort B	
Covariates	Estimate	P-values	Estimate	P-values
Intercept	1.156	<0.001	1.156	<0.001
log(PBL)	0.006	0.730	0.006	0.730
Gender	0.036	<0.001	0.036	<0.001
B_1(age)	0.910	<0.001	0.910	<0.001
B_2(age)	1.303	<0.001	1.303	<0.001
B_3(age)	1.561	<0.001	1.726	<0.001
B_1(log(CBL))	0.035	0.330	1.561	0.980
B_2(log(CBL))	-0.111	0.052	-0.017	0.993

CHAPTER IV

Adaptive Fused Lasso in Meta Longitudinal Studies

4.1 Introduction

In biomedical research, when the sample size is not large enough to achieve adequate statistical power, it is a common practice to merge data from multiple studies (*Zhang et al.*, 2007; *Thase et al.*, 2009). For instance, in a quantitative trait loci (QTL) analysis of exploring risk factors of a certain disease, a single QTL study dataset often has been limited with a small sample size. To overcome, geneticists attempt to routinely combine datasets from multiple similar studies in order to improve statistical power (*Aschard et al.*, 2011).

The increased sample size from the merged data may not always lead to an improvement in estimation efficiency or higher testing power if datasets are collected from inhomogeneous subpopulations, such as study populations subject to major population-specific confounders. In this case, merging data can bring in much noise and additional complications in data distributions, which may offset the power gain from the increased sample size. Thus, the fundamental task in the use of data integration strategy is to check homogeneity across multiple studies. In a regression model, which is the setting considered in this chapter, finding a set of homogeneous (or common) regression coefficients across multiple studies enables us to better understand certain common mean structures, which will then be utilized to improve statistical power. In contrast, it is hard to interpret estimated common effects of covariates when it is blindly assumed that data from different subpopulations share common parameters without a prior data evidence. Unfortunately this is the current practice widely accepted in the meta analysis. It is hoped that the work described in this chapter can provide some useful guidelines to practitioners in the aspect of checking parameter homogeneity using available data.

Hypothesis testing would naturally be the first choice of method to examine certain underlying homogeneity of parameters by setting the null hypothesis to be the equality of parameters of interest. In general, a desirable test statistic used in the situation of merged data should be the one that is robust against heterogeneous characteristics among longitudinal studies, such as discrepancies of within-subject correlation, dispersion and follow-up schedule. Various versions of modified sandwich covariance estimators have been proposed to provide robust within-subject covariance estimators, see for example *Zeger et al.* (1988), *Wang and Long* (2011), *Mancl and DeRouen* (2001), *Pan* (2001). As pointed out by *Wang et al.* (2012), Wald testing statistics constructed using the modified sandwich covariance estimators may suffer severely from inflated type I errors when multiple longitudinal datasets are inhomogeneous.

Recently, *Wang et al.* (2012) proposed two new test statistics that have been shown robust and therefore can perform valid test for homogeneity of regression coefficients. However, their approach is challenged by a large number of hypotheses needed to be tested in order to reach a full understanding of coefficient homogeneity. *Wang et al.*'s (2012) method is not computationally feasible to deal with many tests, and hence a fast screening technique is indeed needed rather than performing all possible individual hypothesis tests. As a matter of fact, the number of tests required in the case of p covariates in K studies is of order $\binom{K}{2}^p$. When either K, or p, or both are large, the number of required tests will easily exceed available computational power of modern computers, which hinders the application of hypothesis test based approach.

In our proposed fast screening procedure, we pay much attention to the idea of regularized estimation, which has become a popular technique for variable selection in the recent literature. Among many existing regularization procedures, fused lasso is of most relevance to our need. Fused lasso was developed by *Tibshirani et al.* (2005) for ordered parameters that penalizes differences between coefficients of adjacent continuous covariates. For categorical covariates, *Bondell and Reich* (2008a) proposed a form of penalty for unordered categorical covariates. *Yuan et al.* (2006) and *Zou and Yuan* (2008), among others, have proposed methods to regularize a group of covariates. *Bondell and Reich* (2008b) considered an approach to regularizing selected variables while clustering selected variables into predictive groups.

As an application of adaptive fused lasso (*Zou*, 2006; *Tibshirani et al.*, 2005), our efficient screening procedure developed in this chapter can uncover sets of common coefficients across studies. We consider regularization on differences for pairs of parameters, and when a difference is zero the two corresponding coefficients will be fused into a common one. As a result, the number of distinctive parameters will be greatly reduced. Furthermore, improved statistical power may be achieved from the increased sample size as a result of the reduced number of distinctive parameters. More importantly, the resulting statistical interpretations are more appealing and meaningful. The chapter is organized as follows. Section 4.2 introduces notation, models and the penalized objective function in this chapter. Section 4.3 discusses the algorithm for the optimization of the proposed method. After evaluating our method using two simulation studies in Section 4.4, we apply our method in Section 4.5 to an analysis of HIV monitoring cohort data to assess effects of needle sharing on HIV positive among injection drug users in China.

4.2 Formulation

We consider K longitudinal studies with outcome $Y_{k,ij}$ and p-dimensional covariate vector $\boldsymbol{X}_{k,ij}$, where $i = 1, \ldots, n_k$, $j = 1, \ldots, m$ and $k = 1, \ldots, K$. For the ease of exposition, we assume all studies have the same numbers of repeated measurements m. We consider the longitudinal marginal model in this chapter (*Song*, 2007, Chapter 5). For study k the conditional mean of $Y_{k,ij}$ satisfies $E(Y_{k,ij} \mid \boldsymbol{X}_{k,ij}) = \mu_{k,ij} = h(\boldsymbol{X}_{k,ij}^T \boldsymbol{\beta}_k)$, and the conditional variance of $Y_{k,ij}$ satisfies $\text{var}(Y_{k,ij} \mid \boldsymbol{X}_{k,ij}) = \phi_k v(\mu_{k,ij})$, where ϕ_k is the dispersion parameter, $h(\cdot)$ and $v(\cdot)$ are respectively known link and variance functions, and $\boldsymbol{\beta}_k = (\beta_{k,1}, \ldots, \beta_{k,p})^T$ is the vector of regression coefficients associated with $\boldsymbol{X}_{k,ij}$. We denote $\boldsymbol{Y}_{k,i} = (Y_{k,i1}, \ldots, Y_{k,im})^T$, $\boldsymbol{\mu}_{k,i} = (\mu_{k,i1}, \ldots, \mu_{k,im})^T$, and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \cdots, \boldsymbol{\beta}_K^T)^T$.

To describe the underlying parameter configuration, an index set of studies, $\mathcal{A}_l \subseteq \{1, \cdots, K\}$, is defined for the l-th covariate, over which their corresponding regression coefficients are the same, namely with respect to $\boldsymbol{X}_{k,ijl}$, $l = 1, \cdots, p$, such that

$$\beta_{k,l} = \beta_{k',l} \overset{def}{=} \beta_{\mathcal{A}_l,l}, \forall\ k \neq k' \in \mathcal{A}_l.$$

Note that the set $\mathcal{A}_l$ is unknown. Our objective in this chapter is twofold: to determine the set $\mathcal{A}_l$ and to estimate the homogeneous coefficients $\beta_{\mathcal{A}_l,l}$ with respect to $\mathcal{A}_l$, $l = 1, \ldots, p$. These two tasks can be achieved simultaneously by using the regularization technique described in this section.

For each longitudinal study, an estimating function may be formed to estimate $\boldsymbol{\beta}_k$, and then according to *Wang et al.* (2012) all such individual study-specific estimating functions may be combined by the means of quadratic inference function (QIF) (*Qu et al.*, 2000). To apply QIF, we need to first approximate the inverse of working correlation matrix by $\boldsymbol{R}_k^{-1}(\alpha_k) \approx \sum_{s=1}^{s_k} \rho_s \boldsymbol{M}_{k,s}$, where $\rho_1, \ldots, \rho_{s_k}$ are constants and possibly dependent on α_k and $\boldsymbol{M}_{k,1}, \ldots, \boldsymbol{M}_{k,s_k}$ are known basis matrices with elements 0 and 1 completely determined by a given working correlation matrix $\boldsymbol{R}_k(\alpha_k)$. Refer to *Qu et al.* (2000) for more details concerning basis matrices corresponding to different working correlation structures such as exchangeable and AR-1. Using this expansion of $\boldsymbol{R}_k^{-1}$, following *Wang et al.* (2012), we can form a quadratic objective function as follows:

$$Q(\boldsymbol{\beta}) = n\bar{\boldsymbol{g}}(\boldsymbol{\beta})^T\boldsymbol{C}^-(\boldsymbol{\beta})\bar{\boldsymbol{g}}(\boldsymbol{\beta}), \tag{4.1}$$

with $n = \sum_{k=1}^K n_k$, and

$$\begin{aligned}
\bar{\boldsymbol{g}}(\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^n (\ \delta_i(1)\boldsymbol{g}_{1,i}(\boldsymbol{\beta}_1)^T \quad , \dots , \quad \delta_i(K)\boldsymbol{g}_{K,i}(\boldsymbol{\beta}_K)^T\ )^T = n^{-1}\sum_{i=1}^n \boldsymbol{g}_i(\boldsymbol{\beta}),\\
\boldsymbol{C}(\boldsymbol{\beta}) &= n^{-1}\sum_{i=1}^n \mathrm{diag}\{\delta_i(1)\boldsymbol{g}_{1,i}(\boldsymbol{\beta}_1)\boldsymbol{g}_{1,i}(\boldsymbol{\beta}_1)^T, \dots, \delta_i(K)\boldsymbol{g}_{K,i}(\boldsymbol{\beta}_K)\boldsymbol{g}_{K,i}(\boldsymbol{\beta}_K)^T\},\\
\bar{\boldsymbol{g}}_k(\boldsymbol{\beta}_k) &= {n_k}^{-1}\sum_{i=1}^{n_k}\boldsymbol{g}_{k,i}(\boldsymbol{\beta}_k) = {n_k}^{-1}\sum_{i=1}^{n_k}\begin{pmatrix} \dot{\boldsymbol{\mu}}_{k,i}^T\boldsymbol{A}_{k,i}^{-1/2}\boldsymbol{M}_{k,1}\boldsymbol{A}_{k,i}^{-1/2}(\boldsymbol{Y}_{k,i} - \boldsymbol{\mu}_{k,i}) \\ \vdots \\ \dot{\boldsymbol{\mu}}_{k,i}^T\boldsymbol{A}_{k,i}^{-1/2}\boldsymbol{M}_{k,s_k}\boldsymbol{A}_{k,i}^{-1/2}(\boldsymbol{Y}_{k,i} - \boldsymbol{\mu}_{k,i}) \end{pmatrix},
\end{aligned}$$

where $\delta_i(k)$ is the study indicator, with 1 indicating that subject i belongs to study k and 0 otherwise, $\dot{\boldsymbol{\mu}}_{k,i} = \partial\boldsymbol{\mu}_{k,i}/\partial\boldsymbol{\beta}_k^T$ and $\boldsymbol{A}_{k,i} = \mathrm{diag}\{v(\mu_{k,i1}), \dots, v(\mu_{k,im})\}$.

To identify elements in the set $\mathcal{A}_l$ for $l = 1, \dots, p$, we regularize the objective function $Q(\boldsymbol{\beta})$ in (4.1) with a penalty function that is constructed in a way similar to the adaptive fused lasso (*Zou*, 2006; *Tibshirani et al.*, 2005). Our adaptive fused penalty function takes the following form:

$$P(\boldsymbol{\beta}) = \sum_{l=1}^p\sum_{k=1}^K\sum_{k'>j}^K w_{kk',l}|\beta_{k,l} - \beta_{k',l}|, \tag{4.2}$$

where weight $w_{kk',l} \geq 0$ is specified by $w_{kk',l} = 1/|\beta_{k,l}^* - \beta_{k',l}^*|^\gamma$ using a positive constant γ, $\gamma > 0$, and initial consistent estimates $\beta_{k,l}^*$ and $\beta_{k',l}^*$ for $\beta_{k,l}$ and $\beta_{k',l}$, respectively, obtained from individual study based analysis. *Ueki* (2009) and *Ueki and Kawasaki*

(2011) consider the similar variable grouping problem for a single cross-sectional study where L_2 norm is used. Equivalently, (4.2) can be expressed in the matrix notation:

$$P(\boldsymbol{\beta}) = \|\boldsymbol{D}\boldsymbol{\beta}\|_1,$$

where $\boldsymbol{D}$ is a $\binom{K}{2}p' \times Kp$ matrix defining $\binom{K}{2}p'$ restrictions for p', $p' \leq p$, covariates. Entries of $\boldsymbol{D}$ corresponding to $\beta_{k,l}$ and $\beta_{k',l}$ are $w_{kk',l}$ and $-w_{kk',l}$, respectively.

The regularized estimation for parameters in all K studies is carried out by minimizing the following penalized objective function $PQ(\boldsymbol{\beta}, \lambda)$, namely,

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} PQ(\boldsymbol{\beta}, \lambda), \tag{4.3}$$

where $PQ(\boldsymbol{\beta}, \lambda) = Q(\boldsymbol{\beta}) + \lambda\|\boldsymbol{D}\boldsymbol{\beta}\|_1$ and $\lambda > 0$ is a tuning parameter that controls the size of the set $\mathcal{A}_l$.

When $\boldsymbol{D}$ is an identity matrix, $\|\boldsymbol{D}\boldsymbol{\beta}\|_1$ becomes the popular lasso penalty, and in this case minimizing $PQ(\boldsymbol{\beta}, \lambda)$ can be carried out by several algorithms; for instance, *Friedman et al.*'s (2010) coordinate descent algorithm, *Efron et al.*'s (2004) least angle regression, and *Fan and Li*'s (2001) algorithm that approximates the L_1 norm penalty by a quadratic function around initial parameter estimates. However our penalty function in (4.2) is more complex, which requires a different strategy to carry out the optimization of $PQ(\boldsymbol{\beta}, \lambda)$. The related details are given in Section 4.3.

4.3 Dual Optimization

We convert the optimization problem in (4.3) into a problem with simpler restrictions to facilitate numerical calculation. We first consider a dual optimization problem for (4.3). This is established by the second order Taylor approximation of the $Q(\boldsymbol{\beta})$ function at an initial estimate $\boldsymbol{\beta}^{(0)}$. The initial estimates of regression coefficients may be obtained by running GEE analysis in individual studies, where the estimation consistency holds when their mean models are correctly specified. The second order approximation to the objective function $PQ(\boldsymbol{\beta}, \lambda)$ around $\boldsymbol{\beta}^0$ is given by

$$PQ(\boldsymbol{\beta}, \lambda) \approx Q_0 + \dot{\boldsymbol{Q}}_0^T(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \ddot{\boldsymbol{Q}}_0(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \lambda\|\boldsymbol{D}\boldsymbol{\beta}\|_1, \tag{4.4}$$

where Q_0, $\dot{\boldsymbol{Q}}_0$ and $\ddot{\boldsymbol{Q}}_0$ are $Q(\boldsymbol{\beta})$, the first and second derivatives of $Q(\boldsymbol{\beta})$ evaluated at $\boldsymbol{\beta}^0$, respectively. Following the argument of *Kim et al.* (2009), we let $\boldsymbol{z} = \boldsymbol{D}\boldsymbol{\beta}$ and rewrite the minimization of (4.4) as follows:

$$\begin{aligned} &\min_{\boldsymbol{\beta}, \boldsymbol{z}} Q_0 + \dot{\boldsymbol{Q}}_0^T(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^0)^T \ddot{\boldsymbol{Q}}_0(\boldsymbol{\beta} - \boldsymbol{\beta}^0) + \lambda\|\boldsymbol{z}\|_1, \\ &\text{subject to } \boldsymbol{D}\boldsymbol{\beta} = \boldsymbol{z}. \end{aligned}$$

Then, the Lagrangian takes the form

$$L(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\tau}) = Q_0 + \dot{\boldsymbol{Q}}_0^T(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) + \frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)})^T \ddot{\boldsymbol{Q}}_0(\boldsymbol{\beta} - \boldsymbol{\beta}^{(0)}) + \lambda\|\boldsymbol{z}\|_1 + \tau^T(\boldsymbol{D}\boldsymbol{\beta} - \boldsymbol{z}),$$

where $\boldsymbol{\tau} \in R_+^m$ is the Lagrangian multiplier. Being a function of $\boldsymbol{\beta}$, the objective function $L(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\tau})$ is actually minimized at $\boldsymbol{\beta} = \boldsymbol{\beta}^{(0)} - \ddot{\boldsymbol{Q}}_0^{-1}(\dot{\boldsymbol{Q}}_0 + \boldsymbol{D}^T\boldsymbol{\tau}_\lambda)$ with the minimum given by, up to a constant,

$$\min_{\boldsymbol{\beta} \in R^{Kp}} L(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\tau}_\lambda) = \boldsymbol{\tau}_\lambda^T \boldsymbol{D}\boldsymbol{\beta}^{(0)} \; - \frac{1}{2}(\dot{\boldsymbol{Q}}_0 + \boldsymbol{D}^T\boldsymbol{\tau}_\lambda)^T \ddot{\boldsymbol{Q}}_0^{-1}(\dot{\boldsymbol{Q}}_0 + \boldsymbol{D}^T\boldsymbol{\tau}_\lambda).$$

On the other hand, being a function of $\boldsymbol{z}$, the objective function $L(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\tau}_\lambda)$ is minimized with the minimum

$$\min_{\boldsymbol{z}} L(\boldsymbol{\beta}, \boldsymbol{z}, \boldsymbol{\tau}_\lambda) = \begin{cases} 0, & \text{if } \|\boldsymbol{\tau}\|_\infty < \lambda, \\ -\infty, & \text{otherwise,} \end{cases}$$

where $\|\cdot\|_\infty$ is the supremum norm for a vector. Therefore the dual optimization with regard to $\boldsymbol{\tau}$ is

$$\begin{aligned} &\min_{\boldsymbol{\tau}\in R_+^m} -\boldsymbol{\tau}^T\boldsymbol{D}\boldsymbol{\beta}^{(0)} + \frac{1}{2}(\dot{\boldsymbol{Q}}_0 + \boldsymbol{D}^T\boldsymbol{\tau})^T\ddot{\boldsymbol{Q}}_0^{-1}(\dot{\boldsymbol{Q}}_0 + \boldsymbol{D}^T\boldsymbol{\tau}), \\ &\text{subject to } \|\boldsymbol{\tau}_\lambda\|_\infty < \lambda. \end{aligned} \tag{4.5}$$

Given the solution $\hat{\boldsymbol{\tau}}_\lambda$ of (4.5), we can update $\boldsymbol{\beta}$ via

$$\hat{\boldsymbol{\beta}}_\lambda = \boldsymbol{\beta}^{(0)} - \ddot{\boldsymbol{Q}}_0^{-1}(\dot{\boldsymbol{Q}}_0 + \boldsymbol{D}^T\hat{\boldsymbol{\tau}}_\lambda).$$

In effect, the optimization required in (4.5) is a quadratic programming problem with boundedness restrictions, $\|\boldsymbol{\tau}_\lambda\|_\infty < \lambda$, which can be solved by applying standard convex optimization algorithms, e.g. the interior-point methods.

4.4 Numerical Examples

4.4.1 Simulation Study I

The first simulation study is to illustrate the performance of our method to determine the underlying homogeneity of parameters for continuous outcomes. There are 8 longitudinal studies with 4 repeated measurements considered under the following true linear models:

$$Y_{k,ij} = \beta_{k,0} + \beta_{k,1}X_{k,ij} + \beta_{k,2}Z_{k,ij} + \epsilon_{k,ij} \quad k = 1,\ldots,8,\ i = 1,\ldots,n_k,\ j = 1,\ldots,4,$$

where $\boldsymbol{\beta}_k = (\beta_{k,0}, \beta_{k,1}, \beta_{k,2})^T$ is the vector of true regression parameters and the error term $\boldsymbol{\epsilon}_k = (\epsilon_{k,1},\ldots,\epsilon_{k,4})^T$ follows $N\{\mathbf{0}, \phi_k\boldsymbol{R}_k(\alpha_k)\}$ for $k = 1,\ldots,8$. Covariate $X_{k,ij}$ is a baseline covariate generated from standard normal distribution. Covariate $\boldsymbol{Z}_{k,i} = (Z_{k,i1},\cdots,Z_{k,i8})$ is dependent on time and simulated from standard multivariate normal distribution. Inhomogeneous covariance structures are specified to mimic the reality that the ten studies are sampled from different subpopulations. In particular, we set $\boldsymbol{R}_k(\cdot)$ for $k = 1, 4, 6, 7, 8$ as AR-1, and $\boldsymbol{R}_k(\cdot)$ for $k = 2, 3, 5$ as compound symmetry, respectively. The dispersion parameters $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_8)^T = (1, 2, 2, 1, 2, 1, 1, 1)^T$, and the correlation parameters $\alpha_1 = \cdots = \alpha_8 = 0.5$. We consider the following two scenarios of the underlying homogeneity for the regression parameters:

Case I. $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = (-1, 2, 0)^T$, $\boldsymbol{\beta}_3 = \cdots = \boldsymbol{\beta}_6 = (-1, 3, 0)^T$, and $\boldsymbol{\beta}_7 = \boldsymbol{\beta}_8 = (-1, 3.5, 0)^T$;

Case II. $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = (-1, 2, 2)^T$, $\boldsymbol{\beta}_3 = \cdots = \boldsymbol{\beta}_6 = (-1, 3, 2)^T$, and $\boldsymbol{\beta}_7 = \boldsymbol{\beta}_8 = (-1, 3.5, 3)^T$.

The regularization matrix $\boldsymbol{D}$ in equation (4.2), which contains all needed pairwise restrictions on covariate $\beta_{k,1}$ and $\beta_{k,2}$. In case I, the regression model contains 16 parameters in all 8 studies and matrix $\boldsymbol{D}$ is a $\binom{8}{2} \times 16$ matrix, while in case II matrix $\boldsymbol{D}$ takes the following form:

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D}_1 \\ \boldsymbol{D}_2 \end{pmatrix}, \tag{4.6}$$

where $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ are two $\binom{8}{2} \times 24$ matrices which are constructed in the same way as the following two matrices given in the case of three studies illustrated as follows:

$$\boldsymbol{D}_1 = \begin{pmatrix} 0 & w_{12,1} & 0 & 0 & -w_{12,1} & 0 & 0 & 0 & 0 \\ 0 & w_{13,1} & 0 & 0 & 0 & 0 & 0 & -w_{13,1} & 0 \\ 0 & 0 & 0 & 0 & w_{23,1} & 0 & 0 & -w_{23,1} & 0 \end{pmatrix},$$

$$\boldsymbol{D}_2 = \begin{pmatrix} 0 & 0 & w_{12,2} & 0 & 0 & -w_{12,2} & 0 & 0 & 0 \\ 0 & 0 & w_{13,2} & 0 & 0 & 0 & 0 & 0 & -w_{13,2} \\ 0 & 0 & 0 & 0 & 0 & w_{23,2} & 0 & 0 & -w_{23,2} \end{pmatrix}.$$

For convenience, we arrange the parameter order in the columns of $\boldsymbol{D}$ as the same order of all parameters arranged in $\boldsymbol{\beta}$.

Tunning parameter λ is chosen according to the smallest BIC criterion given as follows:

$$BIC(\lambda) = Q(\hat{\boldsymbol{\beta}}_\lambda) + df(\hat{\boldsymbol{\beta}}_\lambda)log(n),$$

where $\hat{\boldsymbol{\beta}}_\lambda$ is the resulting parameter estimate at a given value of λ and $df(\hat{\boldsymbol{\beta}}_\lambda)$ is the total number of distinctive parameter estimates. Figure 4.1 shows two BIC curves computed from a simulated dataset, and their shapes are representative in our simulation study.

Sensitivity, specificity and model size are reported in Table 4.1 for Case I and Case II under AR-1 and compound symmetry working correlations based on 200 replications. Sensitivity is estimated as the proportion of equal coefficient pairs that are correctly identified, while specificity is estimated in the same way for unequal coefficients pairs over 200 rounds of simulations. Model size is the number of distinctive parameters estimated by the proposed method for covariates $X_{k,ij}$ and $Z_{k,ij}$. The true number of distinctive parameters are 3 for case I and 5 for case II, respectively. The parameter γ in the construction of the penalty function in (4.2) is fixed at 2.8 in this simulation study.

As show in Table 4.1 our method can identify the shared parameters across studies in all designed cases. The performance in terms of sensitivities and specificities is increasing along the increase of sample sizes in each study, indicating that the proposed method can consistently identify the underlying parameter structures. We notice that specificities under different scenarios are all in a moderate or high level no matter which case is considered. Besides the model performance assessed by sensitivities, specificities and model sizes, we learn that the choice of working correlation structures does not have a significant impact on our method's performance.

Table 4.1: Simulation Results for the Linear Model

		AR-1			CS		
Case	n_k	Sensitivity	Specificity	Model Size (Std)	Sensitivity	Specificity	Model Size (Std)
	100	0.245	0.290	3.505 (1.037)	0.230	0.290	3.505 (0.951)
I	400	0.555	0.795	3.415 (0.718)	0.555	0.795	3.410 (0.666)
	800	0.755	0.960	3.405 (0.875)	0.750	0.960	3.420 (0.887)
	100	0.200	0.355	6.595 (2.490)	0.205	0.340	6.155 (2.217)
II	400	0.320	0.840	6.655 (2.563)	0.345	0.840	6.470 (2.460)
	800	0.450	0.985	7.005 (3.176)	0.445	0.980	6.595 (2.786)

4.4.2 Simulation Study II

The second simulation study is designed for binary outcomes with five longitudinal studies generated under the following logistic models:

$$\text{logit}\{E(Y_{k,ij} \mid X_{k,ij}, Z_{k,ij})\} = \beta_{k,0} + \beta_{k,1}X_{k,ij} + \beta_{k,2}Z_{k,ij}$$
$$k = 1, \ldots, 5, i = 1, \ldots, n_k, j = 1, \ldots, 4,$$

where baseline covariate $\boldsymbol{X}_{k,i} = (X_{k,i1}, \ldots, X_{k,i8})$ is generated from $Binomial(0.5)$ and time-dependent covariate $\boldsymbol{Z}_{k,i} = (Z_{k,i1}, \ldots, Z_{k,i8})$ is simulated from $Unif(0, 1)$. The five studies are different in terms of their study-specific covariance and correlation parameters. We set $\boldsymbol{R}_k(\cdot)$ for $k = 1, 4$ as AR-1, and $\boldsymbol{R}_k(\cdot)$ for $k = 2, 3, 5$ as compound symmetry, respectively. The correlation parameters $\alpha_1 = \cdots = \alpha_5 = 0.5$. We consider the following two cases of underlying homogeneity for the regression coefficients:

1) $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \beta_3 = (-1, -1, 0)$ and $\boldsymbol{\beta}_4 = \boldsymbol{\beta}_5 = (-1, 1, 0)^T$.

2) $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = (-1, -2, 3)$, $\boldsymbol{\beta}_3 = (-1, -2, 4)$, and $\boldsymbol{\beta}_4 = \boldsymbol{\beta}_5 = (-1, 2, 4)$.

Matrix $\boldsymbol{D}$ is constructed in the same way as the first numerical example. Based on 200 rounds of simulation, we summarize results in table 4.2. It shows that our method's performance increases along the increase of the sample size. For example the sensitivity computed under AR-1 working correlation for case 2 increases from 0.560 to 0.81 when the sample size increases from 100 to 800. At the same time, the model size approaches to the true model sizes, 5 for case I and 4 for case II, respectively. The parameter γ is 3 in this simulation study.

4.5 Application

We now apply the proposed method to analyze clustered datasets collected by national HIV surveillance project on injection drug users (IDU) in a southwestern

Table 4.2: Simulation Results for the Logistic Model

			AR-1			CS	
Case	n_k	Sensitivity	Specificity	Model Size (Std)	Sensitivity	Specificity	Model Size (Std)
I	100	0.760	0.990	2.260 (0.504)	0.750	0.980	2.28 (0.541)
	400	0.925	1.000	2.080 (0.290)	0.880	1.000	2.14 (0.376)
	800	0.975	1.000	2.025 (0.157)	0.940	1.000	2.07 (0.256)
II	100	0.560	0.170	3.885 (1.216)	0.52	0.175	4.085 (1.374)
	400	0.620	0.665	4.100 (0.657)	0.59	0.640	4.175 (0.805)
	800	0.810	0.950	4.160 (0.464)	0.75	0.935	4.285 (0.613)

province of China. By the end of 2006, China has established 393 national and 370 provincial monitoring sites reporting HIV incidences to the national centre for AIDS/sexually transmitted disease control and prevention (*Sun et al.*, 2007). Provincial HIV sentinel surveillance program involved community health center, hospitals and drug addiction treatment centres conducting surveys among high risk groups such as IDUs.

The HIV surveillance cohort data were collected between 2006 and 2009 using stratified sampling from 67 hospitals, community health center, and drug addiction treatment centers as primary sample units (PSU) to monitor incidences of HIV infection among IDUs in the study area. All IDUs sampled in the survey were tested for HIV positive and interviewed for their behavioral characteristics related to drug usage, such as if they inhale drugs, if they share needles with other IDUs, and if they are infected by syphilis virus and so on. Cluster sizes of PSUs varied greatly from 11 to 440 IDUs. To reduce within the PSU heterogeneity, we further divide IDUs within a PSU into 3 classes according to their martial status (single, marriage, divorce). As a result we create 194 smaller but more homogeneous clusters.

The study contained five regions termed as A, B, C, D, and E, which are very different in many aspects, such as the population size and socioeconomic status. For example, region A is the largest metropolitan city in the province, whereas region E is primarily dominated by minorities living in mountain villages. Thus, it is expected that such highly diversified backgrounds of IDUs across these regions could possibly lead to different trends and covariate effects on HIV positive.

In our application, we are particularly interested in assessing the effects of behavioral activities on HIV positive, among which needle sharing is the central variable that has been proved as a critical factor for the infection of HIV. Here we focus on investigating the common effects of behavioral activities among the five regions on HIV positive. Based on such clustered data, we fit a marginal logistic regression model of the following form:

$$\mathrm{logit}\{E(Y_{k,ij} \mid X_{k,i1}, X_{k,i2}, X_{k,i3}, X_{k,i4})\} = \beta_0 + \beta_1 X_{k,i1} + \beta_2 X_{k,i2} + \beta_3 X_{k,i3} + \beta_4 X_{k,i4}, \tag{4.7}$$

where response Y_{ij} is a binary outcome of HIV positive for the jth subject in the ith cluster, and covariates $X_{k,i1}$ to $X_{k,i4}$ are gender (1 for male, 0 for female), time (0 to 4 years), needle-sharing (1 for needle-sharing, 0 otherwise) and syphilis (1 for syphilis, 0 otherwise), respectively.

First the clustered surveillance cohort data from the five regions are analyzed separately by the QIF (*Qu et al.*, 2000) using the compound symmetric correlation matrix. All covariates are standardized and estimated coefficients associated with covariates are reported in Table 4.3. It seems that the effect sizes of needle sharing in the five regions fall in approximately two groups: the first group contains A, B and C with effect sizes between 0.8 and 0.6, while the second group includes D and E with much smaller effect sizes. Effects of syphilis in the five regions are not as large as needle sharing and probably no patterns appear.

We apply the proposed approach to finding common parameters of needle sharing and syphilis shared by the five regions. The BIC curves and solution paths of needle sharing and syphilis are showed in Figure 4.2. Regression coefficient estimates summarized in Table 4.4 are chosen according the minimum of BIC at which $\lambda = 2.375$ when γ is fixed at 1.

Needle sharing's solution path displayed in the middle panel of Figure 4.2 indicates region A, B, and C share a common effect of needle sharing on HIV positive, which is significantly higher than the needle sharing's effect in region D and E. It shows that IDUs sharing needles with other IDUs in region A, B and C tend to have larger chance to be infected with HIV than IDUs in region D and E. On the other hand, our method shrinks all parameters associated with syphilis to 0.158, indicating there is no regional discrepancy in the effect of syphilis on HIV positive.

4.6 Concluding Remarks

In this chapter, we apply the adaptive fused lasso approach to merging multiple longitudinal studies when the underlying homogeneous parameter configuration is complex. If either the number of studies or the number of parameters is large, the conventional approach of hypothesis testing is computationally prohibited due to the considerably large number of possible parameter configurations across studies. Our approach based on the adaptive fused lasso can overcome this difficulty and efficiently detect the pattern of homogeneous parameters shared by studies and estimate distinctive parameters at the same time. However, our method may be affected with the increased complexity concerning the underlying pattern of homogeneous parameters. We conjecture that this limitation may be caused by the inflexibility of using one tuning parameter in the regularization procedure. When the configuration of common parameters is complex, one tuning parameter may not work flexibly and properly for grouped restrictions on regression parameters. For example, if $\beta_1 = \beta_2$ and $\beta_2 = \beta_3$ then $\beta_1 = \beta_3$ must happen. In our current version of penalty function, this type of nested conditions are not accounted for. Another issue of our method is the selection of the parameter γ, which is still an open problem. For the future research, we plan to extend the proposed approach to make it more flexible and adaptive to more complex parameter configurations, and provide more informative guidelines for the selection of γ in the construction of penalty functions.

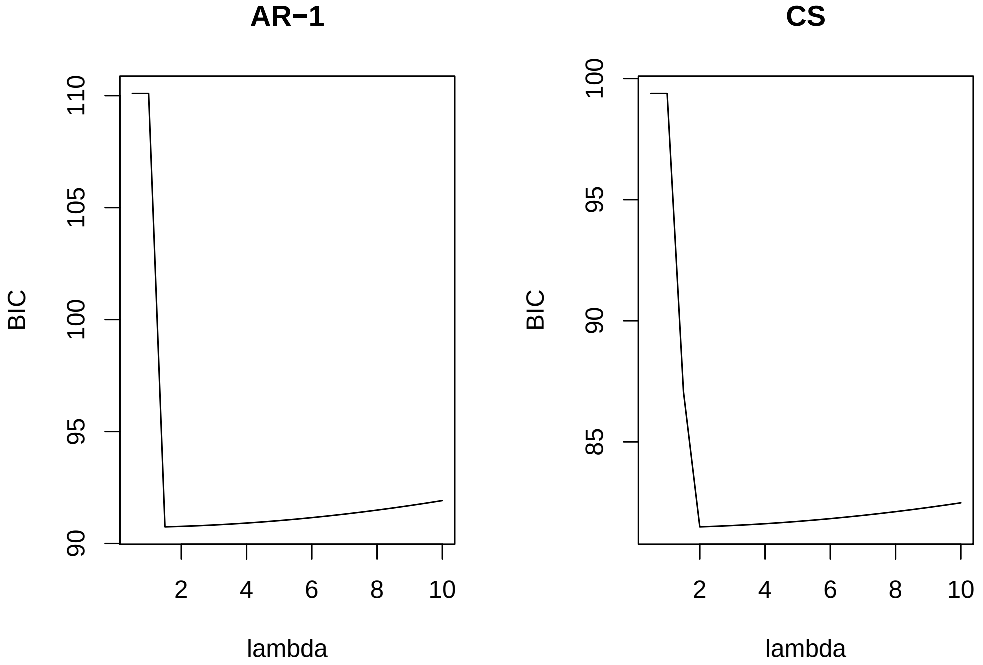

Figure 4.1: BIC curves under two different working correlations.

Table 4.3: Estimates of regression coefficients obtained from individual analyses using data from the five areas. All covariates are standardized with mean 0 and variance 1.

	A	B	C	D	E
Intercept	-3.102	-3.531	-2.312	-1.864	-0.810
Gender	0.145	-0.106	-0.037	0.507	0.411
Time	-0.214	0.320	0.601	-0.611	-0.218
Needle sharing	0.828	0.808	0.604	0.012	0.360
Syphilis	0.180	-0.096	0.056	-0.433	0.147

Table 4.4: Parameter estimates obtained from penalized QIF using data from five areas. All covariates are standardized with mean 0 and variance 1.

	A	B	C	D	E
Intercept	-2.985	-3.487	-2.423	-1.915	-0.796
Gender	0.229	-0.140	-0.050	1.005	0.432
Time	-0.336	0.091	0.658	-0.427	-0.254
Needle sharing	0.579	0.579	0.579	0.006	0.434
Syphilis	0.158	0.158	0.158	0.158	0.158

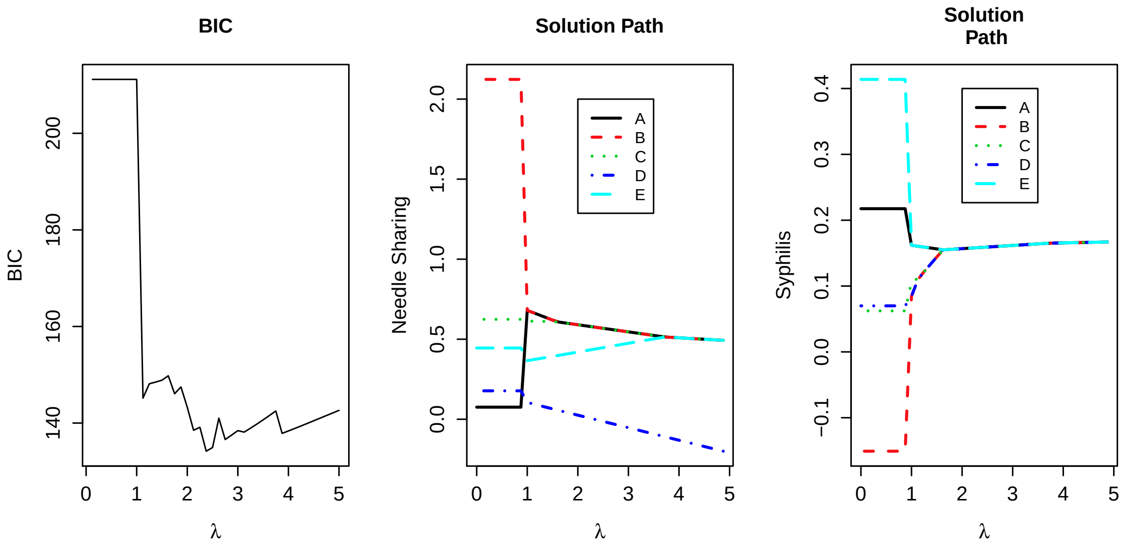

Figure 4.2: BIC and solution paths for needle sharing and syphilis.

CHAPTER V

Future Work

Most methods in meta analysis and my first two topics in analyzing multiple longitudinal or clustered studies are all in the track of testing whether data can be merged or be analyzed jointly by proposed testing statistics. But considering the nature of meta analysis, quantitatively integrating several empirical research studies related to a common topic, we can achieve the aim by the method of regularization as well.

Compared to the approach of testing hypotheses, the method of regularization has several benefits: (i) testing statistics and their null distributions have to be derived under the correct model specifications. Thus very often it is not easy to unify hypothesis testing and model selection into a single procedure. In contrast, the method of regularization may achieve this purpose by estimating parameters of unimportant variables exactly 0 while detecting common distinctive parameters; (ii) datasets may be collected from different protocols in different experiments, so that types of outcomes may not be the same. The method of regularization can join estimation procedures for different types of outcomes as long as a common objective function can be derived. In contrast, it is not that flexible to develop tests for different types of outcomes; (iii) it is common that several covariates of interest may exhibit different types of common features, such as a common effect size and a common effect sign.

Using appropriate penalty functions, the method of regularization may help us to solve such complex problems.

Along this direction, there are several potential problems, including data merging with different types of outcomes and with high dimensional covariates. I plan to extend the work in Chapter IV to make it more flexible and adaptive to address those challenges mentioned above. Also, it seems very useful to develop a method for the validation of common effect signs. For example it maybe meaningfulless to prove the same gene measured in different experiments to have the same effect size, but it is meaningful to examine if the same gene could have the same sign of effect in multiple experiments. Another future topic pertains to jointly analyzing two different types of outcomes such as longitudinal outcomes and survival outcomes, or longitudinal outcomes and brain image measurements. For the joint analysis of longitudinal and survival data, if longitudinal data is of primary interest, using the event time data will improve parameter estimates of longitudinal models by allowing adjustment for informative censoring of repeated measures by the disease process. I will explore how to jointly analyze the two types of outcomes in the framework of generalized method of moments.

APPENDICES

APPENDIX A

Appendices for Chapter II

Proof of Theorem II.1. Let $\hat{\boldsymbol{\beta}}$ be a root-n consistent estimator for $\boldsymbol{\beta}_0$. A Taylor expansion of $\bar{\boldsymbol{g}}(\hat{\boldsymbol{\beta}})$ about $\boldsymbol{\beta}_0$ gives $\bar{\boldsymbol{g}}(\hat{\boldsymbol{\beta}}) = \bar{\boldsymbol{g}}(\boldsymbol{\beta}_0) + \hat{\boldsymbol{G}}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$, where $\boldsymbol{\beta}^*$ is between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{G}}(\boldsymbol{\beta}^*) = \partial\bar{\boldsymbol{g}}(\boldsymbol{\beta}^*)/\partial\boldsymbol{\beta}^T$. Substituting this expression for $g(\hat{\boldsymbol{\beta}})$ in $Q(\hat{\boldsymbol{\beta}})$, we represent $Q(\hat{\boldsymbol{\beta}})$ as

$$Q(\hat{\boldsymbol{\beta}}) = \left\| n^{1/2}\{\boldsymbol{C}^{-}(\hat{\boldsymbol{\beta}})\}^{1/2}\left\{\bar{g}(\boldsymbol{\beta}_0) + \hat{\boldsymbol{G}}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right\}\right\|^2, \qquad \text{(A.1)}$$

where $\|\cdot\|$ is the Euclidean norm. Another Taylor expansion of $\bar{\boldsymbol{g}}(\hat{\boldsymbol{\beta}})$ about $\boldsymbol{\beta}_0$ in the first order condition of $\hat{\boldsymbol{\beta}}$, $\partial Q(\hat{\boldsymbol{\beta}})/\partial\boldsymbol{\beta}^T = 0$, gives

$$\hat{\boldsymbol{G}}(\hat{\boldsymbol{\beta}})^T\boldsymbol{C}^{-}(\hat{\boldsymbol{\beta}})\left\{\bar{\boldsymbol{g}}(\boldsymbol{\beta}_0) + \hat{\boldsymbol{G}}(\boldsymbol{\beta}^{**})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)\right\} + o_p(1) = 0$$

, where $\boldsymbol{\beta}^{**}$ is between $\boldsymbol{\beta}_0$ and $\hat{\boldsymbol{\beta}}$. Provided that $\hat{\boldsymbol{G}}(\hat{\boldsymbol{\beta}})^T\boldsymbol{C}^{-}(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{G}}(\boldsymbol{\beta}^{**})$ is nonsingular,

$$(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = -\left\{\hat{\boldsymbol{G}}(\hat{\boldsymbol{\beta}})^T\boldsymbol{C}^{-}(\hat{\boldsymbol{\beta}})\hat{\boldsymbol{G}}(\boldsymbol{\beta}^{**})\right\}^{-1} a_n(\hat{\boldsymbol{\beta}})\bar{\boldsymbol{g}}(\boldsymbol{\beta}_0), \qquad \text{(A.2)}$$

where $a_n(\hat{\boldsymbol{\beta}}) = \hat{\boldsymbol{G}}(\hat{\boldsymbol{\beta}})^T \boldsymbol{C}^-(\hat{\boldsymbol{\beta}})$. Substituting (A.2) for $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$ into (A.1) yields

$$\begin{aligned}
Q(\hat{\boldsymbol{\beta}}) &= \left\| n^{1/2}\{\boldsymbol{C}^-(\hat{\boldsymbol{\beta}})\}^{1/2} \left\{ \bar{g}(\boldsymbol{\beta}_0) + \hat{\boldsymbol{G}}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \right\} \right\|^2 \\
&= \left\| n^{1/2}\{\boldsymbol{C}^-(\hat{\boldsymbol{\beta}})\}^{1/2} \left[\boldsymbol{I} - \hat{\boldsymbol{G}}(\boldsymbol{\beta}^*) \left\{ \hat{\boldsymbol{G}}(\hat{\boldsymbol{\beta}})^T \boldsymbol{C}^-(\hat{\boldsymbol{\beta}}) \hat{\boldsymbol{G}}(\boldsymbol{\beta}^{**}) \right\}^{-1} a_n(\hat{\boldsymbol{\beta}}) \right] \bar{g}(\boldsymbol{\beta}_0) \right\|^2 .
\end{aligned}$$

By assumptions (b) and (d) stated in Theorem II.1 and *Davidson* (2001, Theorem 21.6), we obtain

$$\begin{aligned}
\frac{\partial \bar{\boldsymbol{g}}(\hat{\boldsymbol{\beta}})}{\partial \boldsymbol{\beta}^T} &= \boldsymbol{G} + o_p(1), \\
\frac{\partial \bar{\boldsymbol{g}}(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^T} &= \boldsymbol{G} + o_p(1), \\
\frac{\partial \bar{\boldsymbol{g}}(\boldsymbol{\beta}^{**})}{\partial \boldsymbol{\beta}^T} &= \boldsymbol{G} + o_p(1), \\
\text{and } \boldsymbol{C}^-(\hat{\boldsymbol{\beta}}) &= \boldsymbol{\Sigma}^- + o_p(1).
\end{aligned}$$

Assumptions (c) and (e) give $n^{1/2}\bar{\boldsymbol{g}}(\boldsymbol{\beta}_0) \to \boldsymbol{Y} \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ in distribution, where $\boldsymbol{\Sigma}$ could be singular. The extended definition for multivariate normal distribution with singular covariance matrix is given by Definition 2.4.1 (*Anderson*, 2003). Then Slutsky's Theorem implies

$$\begin{aligned}
& n^{1/2}\{\boldsymbol{C}^-(\hat{\boldsymbol{\beta}})\}^{1/2} \left[\boldsymbol{I} - \hat{\boldsymbol{G}}(\boldsymbol{\beta}^{**}) \left\{ \hat{\boldsymbol{G}}(\hat{\boldsymbol{\beta}})^T \boldsymbol{C}^-(\hat{\boldsymbol{\beta}}) \hat{\boldsymbol{G}}(\boldsymbol{\beta}^{**}) \right\}^{-1} a_n(\hat{\boldsymbol{\beta}}) \right] \bar{\boldsymbol{g}}(\boldsymbol{\beta}_0) \\
& \to (\boldsymbol{\Sigma}^-)^{1/2}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y}
\end{aligned}$$

in distribution, where $\boldsymbol{P} = \boldsymbol{G}(\boldsymbol{G}^T\boldsymbol{\Sigma}^-\boldsymbol{G})^{-1}\boldsymbol{G}^T\boldsymbol{\Sigma}^-$ and $\boldsymbol{Y}$ follows a $N(\boldsymbol{0}, \boldsymbol{\Sigma})$. Let $\boldsymbol{S} = (\boldsymbol{\Sigma}^-)^{1/2}(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{\Sigma}(\boldsymbol{I} - \boldsymbol{P})^T(\boldsymbol{\Sigma}^-)^{1/2}$. Since $\boldsymbol{P}$ is idempotent, so is $\boldsymbol{S}$. Thus $Q(\hat{\boldsymbol{\beta}})$ converges in distribution to

$$\boldsymbol{Y}^T(\boldsymbol{I} - \boldsymbol{P})^T\boldsymbol{\Sigma}^-(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{Y} \sim \chi^2_{\text{rank}(\boldsymbol{S})},$$

where $\text{rank}(\boldsymbol{S}) = \text{trace}(\boldsymbol{S}) = \text{rank}(\boldsymbol{\Sigma}) - Kp + (K-1)|\mathcal{M}|$. □

Proof of Theorem II.3. Note that $\boldsymbol{G}^T = (\boldsymbol{G}_1^T, \boldsymbol{G}_2^T, \ldots, \boldsymbol{G}_K^T)$ and

$$\boldsymbol{\Sigma}^- = \text{diag}\{\rho_1 \boldsymbol{\Sigma}_1^-, \ldots, \rho_K \boldsymbol{\Sigma}_K^-\}$$

where $\boldsymbol{G}_k$ $(k = 1, \ldots, K)$ and $\boldsymbol{\Sigma}$ are defined in Theorem II.1. We have

$$\boldsymbol{G}^T \boldsymbol{\Sigma}^- \boldsymbol{G} = \rho_1 \boldsymbol{G}_1^T \boldsymbol{\Sigma}_1^- \boldsymbol{G}_1 + \cdots + \rho_K \boldsymbol{G}_K^T \boldsymbol{\Sigma}_K^- \boldsymbol{G}_K.$$

Denote $\boldsymbol{B} = \boldsymbol{G}^T \boldsymbol{\Sigma}^- \boldsymbol{G}$ and $\boldsymbol{B}_k = \rho_k \boldsymbol{G}_k^T \boldsymbol{\Sigma}_k^- \boldsymbol{G}_k$ for $k = 1, \ldots, K$. $\boldsymbol{B}$ can be partitioned as

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{B}_{[\boldsymbol{\beta}_1,\boldsymbol{\beta}_1]} & \boldsymbol{B}_{[\boldsymbol{\beta}_1,-\boldsymbol{\beta}_1]} \\ \boldsymbol{B}_{[-\boldsymbol{\beta}_1,\boldsymbol{\beta}_1]} & \boldsymbol{B}_{[-\boldsymbol{\beta}_1,-\boldsymbol{\beta}_1]} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^K \boldsymbol{B}_{k[\boldsymbol{\beta}_1,\boldsymbol{\beta}_1]} & \sum_{k=1}^K \boldsymbol{B}_{k[\boldsymbol{\beta}_1,-\boldsymbol{\beta}_1]} \\ \sum_{k=1}^K \boldsymbol{B}_{k[-\boldsymbol{\beta}_1,\boldsymbol{\beta}_1]} & \sum_{k=1}^K \boldsymbol{B}_{k[-\boldsymbol{\beta}_1,-\boldsymbol{\beta}_1]} \end{pmatrix},$$

where $-\boldsymbol{\beta}_1$ means not corresponding to $\boldsymbol{\beta}_1$, block-diagonal matrix

$$\sum_{k=1}^K \boldsymbol{B}_{k[-\boldsymbol{\beta}_1,-\boldsymbol{\beta}_1]} = \text{diag}\{\boldsymbol{B}_{2[\boldsymbol{\gamma}_2,\boldsymbol{\gamma}_2]}, \ldots, \boldsymbol{B}_{K[\boldsymbol{\gamma}_K,\boldsymbol{\gamma}_K]}\},$$

$$\sum_{k=1}^K \boldsymbol{B}_{k[\boldsymbol{\beta}_1,-\boldsymbol{\beta}_1]} = \{\boldsymbol{B}_{2[\boldsymbol{\beta}_1,\boldsymbol{\gamma}_2]}, \ldots, \boldsymbol{B}_{K[\boldsymbol{\beta}_1,\boldsymbol{\gamma}_K]}\},$$

and

$$\sum_{k=1}^K (\boldsymbol{B}_{k[-\boldsymbol{\beta}_1,\boldsymbol{\beta}_1]})^T = \{(\boldsymbol{B}_{2[\boldsymbol{\gamma}_2,\boldsymbol{\beta}_1]})^T, \ldots, (\boldsymbol{B}_{K[\boldsymbol{\gamma}_K,\boldsymbol{\beta}_1]})^T\}.$$

Following *Horn and Johnson* (1990, page 18), one can easily derive the inverse of partitioned matrix $\boldsymbol{B}$ and

$$(\boldsymbol{B}^{-1})_{[\boldsymbol{\beta}_1,\boldsymbol{\beta}_1]} = \left\{ \sum_{k=1}^K \boldsymbol{B}_{k[\boldsymbol{\beta}_1,\boldsymbol{\beta}_1]} - \sum_{k=2}^K \boldsymbol{B}_{k[\boldsymbol{\beta}_1,\boldsymbol{\gamma}_k]} (\boldsymbol{B}_{k[\boldsymbol{\gamma}_k,\boldsymbol{\gamma}_k]})^{-1} \boldsymbol{B}_{k[\boldsymbol{\gamma}_k,\boldsymbol{\beta}_1]} \right\}^{-1}.$$

Since $\boldsymbol{B}_{k[\boldsymbol{\beta}_k,\boldsymbol{\beta}_k]}$ is positive definite, so is $\boldsymbol{B}_{k[\boldsymbol{\zeta},\boldsymbol{\zeta}]} - \boldsymbol{B}_{k[\boldsymbol{\zeta},\boldsymbol{\gamma}_k]}(\boldsymbol{B}_{k[\boldsymbol{\gamma}_k,\boldsymbol{\gamma}_k]})^{-1}\boldsymbol{B}_{k[\boldsymbol{\gamma}_k,\boldsymbol{\zeta}]}$. By the fact that $\boldsymbol{B}_{k[\boldsymbol{\beta}_1,\boldsymbol{\beta}_1]} - \boldsymbol{B}_{k[\boldsymbol{\beta}_1,\boldsymbol{\gamma}_k]}(\boldsymbol{B}_{k[\boldsymbol{\gamma}_k,\boldsymbol{\gamma}_k]})^{-1}\boldsymbol{B}_{k[\boldsymbol{\gamma}_k,\boldsymbol{\beta}_1]}$ is a block-diagonal matrix with diagonal components $\boldsymbol{B}_{k[\boldsymbol{\zeta},\boldsymbol{\zeta}]} - \boldsymbol{B}_{k[\boldsymbol{\zeta},\boldsymbol{\gamma}_k]}(\boldsymbol{B}_{k[\boldsymbol{\gamma}_k,\boldsymbol{\gamma}_k]})^{-1}\boldsymbol{B}_{k[\boldsymbol{\gamma}_k,\boldsymbol{\zeta}]}$ and a zero matrix, we show that $\sum_{k=2}^{K}\boldsymbol{B}_{k[\boldsymbol{\beta}_1,\boldsymbol{\beta}_1]} - \sum_{k=2}^{K}\boldsymbol{B}_{k[\boldsymbol{\beta}_1,\boldsymbol{\gamma}_k]}(\boldsymbol{B}_{k[\boldsymbol{\gamma}_k,\boldsymbol{\gamma}_k]})^{-1}\boldsymbol{B}_{k[\boldsymbol{\gamma}_k,\boldsymbol{\beta}_1]}$ is nonnegative definite. Applying *Horn and Johnson* (1990, Theorem 7.7.4), we obtain

$$(\boldsymbol{B}^{-1})_{[\boldsymbol{\beta}_1,\boldsymbol{\beta}_1]} \preceq (\boldsymbol{B}_{1[\boldsymbol{\beta}_1,\boldsymbol{\beta}_1]})^{-1},$$

where $(\boldsymbol{B}^{-1})_{[\boldsymbol{\beta}_1,\boldsymbol{\beta}_1]}$ and $(\boldsymbol{B}_{1[\boldsymbol{\beta}_1,\boldsymbol{\beta}_1]})^{-1}$ are root-n asymptotic variances for $\hat{\boldsymbol{\beta}}_1$ and $\tilde{\boldsymbol{\beta}}_1$, respectively. Rearranging the order of $\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_K$ in parameter $\boldsymbol{\beta}$, we can prove $(\boldsymbol{B}^{-1})_{[\boldsymbol{\beta}_k,\boldsymbol{\beta}_k]} \preceq (\boldsymbol{B}_{k[\boldsymbol{\beta}_k,\boldsymbol{\beta}_k]})^{-1}$ for all $k = 1,\ldots,K$. □

APPENDIX B

Appendices for Chapter III

We first introduce notation used in the following proofs. For a function f of a random variable U with expectation $Ef(U)$, we define two empirical measures $\mathbb{P}_n$ and $\mathbb{G}_n$:

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(U_i), \quad \mathbb{G}_n f = n^{-1/2}\sum_{i=1}^{n}\{f(U_i) - Ef(U_i)\}.$$

We suppress the argument in $\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$ for notational convenience: for instance we let $(\boldsymbol{\theta}_k, \boldsymbol{X}_i, \boldsymbol{\eta}_{k,i})$ denote $(\boldsymbol{\theta}_k, \boldsymbol{X}_i, \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k))$. When no ambiguity exists, the index i of $\boldsymbol{\eta}_{k,i}$ can be dropped further, for example $\boldsymbol{\eta}_k$ for $\boldsymbol{\eta}_{k,i}$. Similarly we use a simplier notation $\nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$ to denote $\nabla_{\boldsymbol{\theta}_k}\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)$ throughout this appendix. In our proofs $\boldsymbol{\eta}_{k,i}$ always denotes the true function and $\boldsymbol{\eta}'_{k,i}$ denotes a function different from $\boldsymbol{\eta}_{k,i}$. We also define the following residuals for study $k \in \mathcal{S}_m$ and $s = 1, \ldots, s_k$:

$$\begin{aligned}
\boldsymbol{q}_{k,i,s}(\boldsymbol{\theta}_k, \boldsymbol{h}_{k,i}) &= \frac{f(\boldsymbol{X}_i \mid D_i = k)}{f(\boldsymbol{X}_i \mid D_i \in \mathcal{S}_o)}\nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)^T \boldsymbol{V}_{k,i,s}\{\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k) - \boldsymbol{h}_{k,i}(\boldsymbol{\theta}_k)\}\\
\boldsymbol{g}_{k,i,s}(\boldsymbol{\theta}_k, \hat{\boldsymbol{\eta}}_{k,i}) &= \nabla\hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_k)^T \boldsymbol{V}_{k,i,s}\{\boldsymbol{Y}_i - \hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_k)\},\\
\boldsymbol{r}_{k,i,s}(\boldsymbol{\theta}_k) &= \frac{f(\boldsymbol{X}_i \mid D_i = k)}{f(\boldsymbol{X}_i \mid D_i \in \mathcal{S}_o)}\nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)^T \boldsymbol{V}_{k,i,s},
\end{aligned}$$

where s_k is the number of basis matrices used to approximate $\boldsymbol{R}_k^{-1}(\alpha_k)$. For $f(\boldsymbol{x})$:

$R^p \to R$, we let $\tilde{f}(\boldsymbol{x})$ be the orthogonal projection of f onto the linear span of j_{n_k} basis functions under the norm $\|\cdot\|_2$.

Assumption B.1.

(B.1.1) For all $k \in \mathcal{S}_m$ and $\boldsymbol{\theta}_k \in \mathcal{B}$, and $\boldsymbol{X}_{ij}$ in a compact subset $\mathcal{X} \subset R^p$, $\eta_{k,ij}(\boldsymbol{\theta}_k)$ belongs to the set $\Lambda^\gamma(\mathcal{X})$ for some $\gamma > p/2$. $\eta_{k,ij}(\boldsymbol{\theta}_k)$ is second order continuously differentiable with respect to $\boldsymbol{\theta}_k$.

(B.1.2) $\|E_k[\{\boldsymbol{Y}_i - \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})\}\{\boldsymbol{Y}_i - \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})\}^T \mid \boldsymbol{X}_i]\| < \infty$.

(B.1.3) There exists $c > 0$ such that $\inf\limits_{i,j} v(\mu_{k,ij}) \geq c$. Also both $v(\cdot)$ and $h(\cdot)$ have bounded second order derivatives.

(B.1.4) There exists a function $\Pi_n \eta_{k,ij}(\boldsymbol{\theta}_k)$ in sieve space defined by

$$\mathcal{G}_n = \{\Pi_n \eta_{k,ij}(\boldsymbol{\theta}_k) = \boldsymbol{b}(\boldsymbol{x}_{k,ij})^T \boldsymbol{a}(\boldsymbol{\theta}_k)\}$$

such that $\|\eta_{k,ij}(\boldsymbol{\theta}_k) - \Pi_n \eta_{k,ij}(\boldsymbol{\theta}_k)\|_\infty = o_p(1)$.

(B.1.5) $E_k\Big\{\sup\limits_{\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{0,k}\| < \delta} \|\nabla^2 \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)\|^2\Big\} < \infty$ for some small $\delta > 0$.

(B.1.6) $\boldsymbol{C}_k(\hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{\eta}}_k)^- = \boldsymbol{\Sigma}_k^- + o_p(1)$ where $\boldsymbol{\Sigma}_k^-$ is a positive definite.

(B.1.7) As $n_k \to \infty$ and $n_o \to \infty$, $\frac{n_k}{n_o} \to \tau_k \in (0, \infty)$ for $k \in \mathcal{S}_m$.

Assumptions below are required to establish asymptotic distribution of the estimator in (3.5).

Assumption B.2. *(B.2.1) Let $\boldsymbol{G}_k = E_k\{\nabla \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i})\}$, $\boldsymbol{G}_k^T \boldsymbol{\Sigma}_k^{-1} \boldsymbol{G}_k$ is positive definite.*

(B.2.2) $j_{n_k} = O(n_o^{\frac{p}{2\gamma+p}})$.

(B.2.3) $0 < c < P(D_i = k \mid \boldsymbol{X}_i) < \infty$ for some c.

(B.2.4) $n_o^{-\gamma/(2\gamma+p)}\|\boldsymbol{r}_{k,i,s}(\boldsymbol{\theta}_{0,k}) - \tilde{\boldsymbol{r}}_{k,i,s}(\boldsymbol{\theta}_{0,k})\|_2 = o(n_o^{-1/2})$ *for* $s = 1, \cdots, s_k$.

Lemma B.3. *For $k \in \mathcal{S}_m$, under Assumptions B.1.1, B.1.4 and $j_{n_k} = o(n_o)$, $n_o \to \infty$, we have,*

$$\|\hat{\boldsymbol{\eta}}_{k,i} - \boldsymbol{\eta}_{k,i}\|_\infty = o_p(1),$$

and for all $\boldsymbol{\theta}_k \in \mathcal{B}$,

$$\|\hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_k) - \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)\|_2 = O_p\left(\sqrt{\frac{j_{n_k}}{n_o}} + (j_{n_k})^{-\gamma/p}\right).$$

The Lemma can be proved by closely following the techniques given in the proof of Proposition B.1 in *Chen et al.* (2007). Thus, we omit the details.

Next for $k \in \mathcal{S}_m$ and $s = 1, \ldots, s_k$, we let $\Gamma_{k,s}(\boldsymbol{\theta}_k, \boldsymbol{\eta}_k)[\boldsymbol{\eta}'_k - \boldsymbol{\eta}_k]$ denote the functional derivative of $E_k\big[\nabla\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)^T\boldsymbol{V}_{k,i,s}\{\boldsymbol{Y}_i - \boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)\}\big]$ at $\boldsymbol{\eta}_{k,i}$ in the direction of $[\boldsymbol{\eta}'_{k,i} - \boldsymbol{\eta}_{k,i}]$, namely

$$\begin{aligned}\Gamma_{k,s}(\boldsymbol{\theta}_k, \boldsymbol{\eta}_k)[\boldsymbol{\eta}'_k - \boldsymbol{\eta}_k] = &- E_k\big[\nabla\boldsymbol{\eta}^T_{k,i}(\boldsymbol{\theta}_k)\boldsymbol{V}_{k,i,s}\{\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k) - \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)\}\big] \\ &+ E_k\big[\{\nabla\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k) - \nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)\}^T\boldsymbol{V}_{k,i,s}\{\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k}) - \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)\}\big].\end{aligned}$$

Lemma B.4. *Let $\delta_n > 0$ and $s = 1, \ldots, s_k$. Under Assumptions B.1.1-B.1.5, as $\delta_n \to 0$, for all $\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{0,k}\| < \delta_n$ and $\|\boldsymbol{\eta}'_{k,i} - \boldsymbol{\eta}_{k,i}\|_\infty < \delta_n$, we have*

(i) L_2 continuity of $\nabla\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)^T\boldsymbol{V}_{k,i,s}\{\boldsymbol{Y}_i - \boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)\}$ at $(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i})$, i.e.,

$$\begin{aligned}E_k\Big[\sup_{\|\boldsymbol{\theta}_{0,k}-\boldsymbol{\theta}_k\|<\delta_n, \|\boldsymbol{\eta}'_{k,i}-\boldsymbol{\eta}_{k,i}\|_\infty<\delta_n} &\|\nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})^T\boldsymbol{V}_{k,i,s}\{\boldsymbol{Y}_i - \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})\} \\ &- \nabla\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)^T\boldsymbol{V}_{k,i,s}\{\boldsymbol{Y}_i - \boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)\}\|^2\Big] \le const\,\delta_n^2.\end{aligned}$$

(ii)

$$\|\Gamma_{k,s}(\boldsymbol{\theta}_k,\boldsymbol{\eta}_k)(\boldsymbol{\eta}'_k-\boldsymbol{\eta}_k)-\Gamma_{k,s}(\boldsymbol{\theta}_{0,k},\boldsymbol{\eta}_k)(\boldsymbol{\eta}'_k-\boldsymbol{\eta}_k)\| \le a_n\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{0,k}\|, \text{ with } a_n = o(1).$$

(iii)

$$\begin{aligned}&\Big\|E_k\big[\nabla\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)\boldsymbol{V}_{k,i,s}\{\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})-\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)\}\\ &\quad-\nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)\boldsymbol{V}_{k,i,s}\{\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})-\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)\}\big]-\Gamma_s(\boldsymbol{\theta}_k,\boldsymbol{\eta}_k)(\boldsymbol{\eta}'_{k,i}-\boldsymbol{\eta}_{k,i})\Big\|\\ &\quad\le const\|\boldsymbol{\eta}'_{k,i}-\boldsymbol{\eta}_{k,i}\|_2\|\nabla\boldsymbol{\eta}'_{k,i}-\nabla\boldsymbol{\eta}_{k,i}\|_2 = o(n_o^{-1/2}).\end{aligned}$$

Proof of Lemma B.4. (i) To establish (i), we consider some neighborhood around $(\boldsymbol{\theta}_{0,k},\boldsymbol{\eta}_{k,i})$. Then for all $(\boldsymbol{\theta}_k,\boldsymbol{\eta}'_{k,i})$ with $\|\boldsymbol{\theta}_k-\boldsymbol{\theta}_{0,k}\|<\delta_n$ and $\|\boldsymbol{\eta}'_{k,i}-\boldsymbol{\eta}_{k,i}\|_\infty<\delta_n$.

$$\begin{aligned}&\nabla\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)^T\boldsymbol{V}_{k,i,s}\{\boldsymbol{Y}_i-\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)\}-\nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})^T\boldsymbol{V}_{k,i,s}\{\boldsymbol{Y}_i-\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})\}\\ =&\{\nabla\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)-\nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})\}^T\boldsymbol{V}_{k,i,s}\{\boldsymbol{Y}_i-\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})\}\\ &+\{\nabla\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)-\nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})\}^T\boldsymbol{V}_{k,i,s}\{\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})-\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)\}\\ &+\nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})^T\boldsymbol{V}_{k,i,s}\{\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})-\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)\}\\ =&I_1+I_2+I_3.\end{aligned}$$

For any $\boldsymbol{a}$ with $\boldsymbol{a}^T\boldsymbol{a} = 1$, by Assumptions B.1.2, B.1.3, and B.1.5

$$\begin{aligned}
&E_k(\boldsymbol{a}^T I_1 I_1^T \boldsymbol{a})\\
\le& C\|V_{k,i,s}\|^2 E_k\{\boldsymbol{a}^T\{\nabla\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k) - \nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})\}^T\{\nabla\boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k) - \nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})\}\boldsymbol{a}\}\\
\le& C\|V_{k,i,s}\|^2 E_k\{\|\nabla\boldsymbol{\eta}'_{k,i} - \nabla\boldsymbol{\eta}_{k,i}\|_\infty + \|\nabla^2\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_*)\|\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{0,k}\|\}^2\\
\le& C\|V_{k,i,s}\|^2 E_k\{\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{0,k}\|^2\|\nabla^2\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_*)\|^2 + \|\nabla\boldsymbol{\eta}'_{k,i} - \nabla\boldsymbol{\eta}_{k,i}\|_\infty^2 +\\
&2\|\nabla^2\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_*)\|\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{0,k}\|\|\nabla\boldsymbol{\eta}'_{k,i} - \nabla\boldsymbol{\eta}_{k,i}\|_\infty\}\\
\le& const\delta_n^2,
\end{aligned}$$

where $\boldsymbol{\theta}_*$ is between $\boldsymbol{\theta}_{0,k}$ and $\boldsymbol{\theta}_k$, $\|\nabla^2\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_*)\|^2 \le \infty$ by Assumption B.1.5, and $\|\boldsymbol{V}_{k,i,s}\| \le \infty$ by Assumption B.1.3 and special structure of basis matrix $\boldsymbol{M}_{k,s}$. For I_3, we can show that

$$\begin{aligned}
E\|I_3\|^2 \le& \|\boldsymbol{V}_{k,i,s}\|^2 E_k\{\|\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k}) - \boldsymbol{\eta}'_{k,i}(\boldsymbol{\theta}_k)\|^2\|\nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})\|^2\}\\
\le& \|\boldsymbol{V}_{k,i,s}\|^2 E_k\left[\{\|\nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{**})\|\|\boldsymbol{\theta}_{0,k} - \boldsymbol{\theta}_k\| + \|\boldsymbol{\eta}_{k,i} - \boldsymbol{\eta}'_{k,i}\|_\infty\}^2\|\nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})\|^2\right]\\
\le& \|\boldsymbol{V}_{k,i,s}\|^2 E_k(A^4\|\boldsymbol{\theta}_{0,k} - \boldsymbol{\theta}_k\|^2 + 2A^3\|\boldsymbol{\eta}_{k,i} - \boldsymbol{\eta}'_{k,i}\|_\infty\|\boldsymbol{\theta}_{0,k} - \boldsymbol{\theta}_k\|\\
&+ A^2\|\boldsymbol{\eta}_{k,i} - \boldsymbol{\eta}'_{k,i}\|_\infty^2\}\\
\le& const\delta_n^2
\end{aligned}$$

where $A = \sup_{\|\boldsymbol{\theta}_{0,k}-\boldsymbol{\theta}_k\|<\delta_n} \|\nabla\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_k)\|$. Similarly we can show $E_k\|I_2\|^2 \le const\delta_n^2$. Combing the results for I_1, I_2 and I_3 we can show (i) hold.

Second, we can prove the result (ii) in a similar way to that used to establish (i).

The third result (iii) can be shown by applying Lemma B.3 and Assumpton B.2.2.

□

Lemma B.5. *Under Assumptions B.1 and B.2, for $s = 1, \ldots, s_k$*

$$\Gamma_{k,s}(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_k)[\hat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k] = \frac{1}{n_o} \sum_{D_i \in \mathcal{S}_o} \boldsymbol{r}_{k,i,s}(\boldsymbol{\theta}_k)\{\boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k}) - \boldsymbol{h}_{k,i}(\boldsymbol{\theta}_{0,k})\} + o_p(n_o^{-1/2}).$$

Proof of Lemma B.5. First note that

$$\Gamma_{k,s}(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_k)[\hat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k] = -E_k\big[\nabla \boldsymbol{\eta}_{k,i}^T(\boldsymbol{\theta}_{0,k}) \boldsymbol{V}_{k,i,s}\{\hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_{0,k}) - \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})\}\big],$$

suggesting that the estimation of $\nabla \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})$ affects $\hat{\boldsymbol{\theta}}_k$ not by bringing an additional component into the asymptotic variance of $\hat{\boldsymbol{\theta}}_k$ but also by the accuracy in the estimation of $\nabla \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})$.

The dimension of $\Gamma_{k,s}(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_k)[\hat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k]$ is $p + q$. Thus we show the lemma holds for each component of $\Gamma_{k,s}(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_k)[\hat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k]$. Let $\boldsymbol{\zeta}_l$ be the l-th column of $\nabla \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})$ for $l = 1, \ldots, p+q$. According to Riesz representation theorem, the Riesz representer of functional $E_k\big[\boldsymbol{\zeta}_l^T \boldsymbol{V}_{k,i,s}\{\hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_{0,k}) - \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})\}\big]$ is $\boldsymbol{r}_{k,i,s}^l(\boldsymbol{\theta}_k)$, which is the l-th row of $\boldsymbol{r}_{k,i,s}(\boldsymbol{\theta}_k)$. Let $\tilde{\boldsymbol{r}}_{k,i,s}^l(\boldsymbol{\theta}_{0,k})$ denote the orthogonal projection of $\boldsymbol{r}_{k,i,s}^l(\boldsymbol{\theta}_{0,k})$ onto the linear span of j_{n_k} basis functions in $\|\cdot\|_2$. For simplicity, we suppress the index s of $\boldsymbol{r}_{k,i,s}^l(\boldsymbol{\theta}_{0,k})$ and $\tilde{\boldsymbol{r}}_{k,i,s}^l(\boldsymbol{\theta}_{0,k})$ in the following derivations. Suppose t_n is a sequence of positive numbers satisfying $t_n = o(n_o^{-1/2})$. By the definition of least square sieve estimator, $\hat{\boldsymbol{\eta}}_{k,i}$ minimizes $\mathbb{P}_{n_o}\{\|\boldsymbol{\eta}_k - \boldsymbol{h}_k\|^2\}$ in the sieve space using data from studies in $\mathcal{S}_o$, which leads to

$$\begin{aligned}
0 &\leq \sqrt{n_o}\Big[\mathbb{P}_{n_o}\{\|\hat{\boldsymbol{\eta}}_k + t_n\tilde{\boldsymbol{r}}_k - \boldsymbol{h}_k\|^2\} - \mathbb{P}_{n_o}\{\|\hat{\boldsymbol{\eta}}_k - \boldsymbol{h}_k\|^2\}\Big] \\
&\leq \sqrt{n_o}\Big[2\mathbb{P}_{n_o}\{(\hat{\boldsymbol{\eta}}_k - \boldsymbol{h}_k)^T\tilde{\boldsymbol{r}}_k\} - 2E_o\{(\hat{\boldsymbol{\eta}}_{k,i} - \boldsymbol{h}_{k,i})^T\tilde{\boldsymbol{r}}_{k,i}\} + 2E_o\{(\hat{\boldsymbol{\eta}}_{k,i} - \boldsymbol{h}_{k,i})^T\tilde{\boldsymbol{r}}_{k,i}\} \\
&\quad + t_n\mathbb{P}_{n_o}\{\|\tilde{\boldsymbol{r}}_k\|^2\}\Big] \\
&= 2\mathbb{G}_{n_o}\{(\hat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k)^T\tilde{\boldsymbol{r}}_k\} + 2\mathbb{G}_{n_o}\{(\boldsymbol{\eta}_k - \boldsymbol{h}_k)^T\tilde{\boldsymbol{r}}_k\} + 2\sqrt{n_o}E_o\{(\hat{\boldsymbol{\eta}}_{k,i} - \boldsymbol{\eta}_{k,i})^T\tilde{\boldsymbol{r}}_{k,i}\} \\
&\quad + t_n\sqrt{n_o}\mathbb{P}_{n_o}\{\|\tilde{\boldsymbol{r}}_k\|^2\} \\
&= 2\mathbb{G}_{n_o}\{(\hat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k)^T\tilde{\boldsymbol{r}}_k\} + 2\mathbb{G}_{n_o}\{(\boldsymbol{\eta}_k - \boldsymbol{h}_k)^T(\tilde{\boldsymbol{r}}_k - \boldsymbol{r}_k)\} + 2\mathbb{G}_{n_o}\{(\boldsymbol{\eta}_k - \boldsymbol{h}_k)^T\boldsymbol{r}_k\} \\
&\quad + 2\sqrt{n_o}E_o\{(\hat{\boldsymbol{\eta}}_{k,i} - \boldsymbol{\eta}_{k,i})^T(\tilde{\boldsymbol{r}}_{k,i} - \boldsymbol{r}_{k,i})\} + 2\sqrt{n_o}E_o\{(\hat{\boldsymbol{\eta}}_{k,i} - \boldsymbol{\eta}_{k,i})^T\boldsymbol{r}_{k,i}\} \\
&\quad + \sqrt{n_o}t_n\mathbb{P}_{n_o}\{\|\tilde{\boldsymbol{r}}_k\|^2\}.
\end{aligned}$$

Applying similar arguments in *Chen et al.* (2007), we establish the following results:

$$\mathbb{G}_{n_o}\{(\boldsymbol{\eta}_k - \boldsymbol{h})^T(\tilde{\boldsymbol{r}}_k - \boldsymbol{r}_k)\} = o_p(1), \tag{B.1}$$

$$\sqrt{n_o}E_o\{(\hat{\boldsymbol{\eta}}_{k,i} - \boldsymbol{\eta}_{k,i})^T(\tilde{\boldsymbol{r}}_{k,i} - \boldsymbol{r}_{k,i})\} = o_p(1), \tag{B.2}$$

$$\mathbb{G}_{n_o}\{(\hat{\boldsymbol{\eta}}_k - \boldsymbol{\eta}_k)^T\tilde{\boldsymbol{r}}_k\} = o_p(1), \tag{B.3}$$

$$t_n\sqrt{n_o}\mathbb{P}_{n_o}\tilde{\boldsymbol{r}}_k^2 = o(1)O_p(1) = o_p(1). \tag{B.4}$$

In the following, we briefly outline the key steps needed to prove (B.1)-(B.4). To show (B.1), by assumptions B.2.4 and B.2.2, we obtain

$$\begin{aligned}
P(\mathbb{P}_{n_o}(\tilde{\boldsymbol{r}}_k - \boldsymbol{r}_k)^T(\boldsymbol{\eta}_k - \boldsymbol{h}) > \frac{\epsilon}{n_o^{1/2}}) &\leq \frac{E\{(\tilde{\boldsymbol{r}}_k - \boldsymbol{r}_k)^T(\boldsymbol{\eta}_{k,i} - \boldsymbol{h}_i)\}n_o^{1/2}}{\epsilon} \\
&\leq \frac{\|\boldsymbol{\eta}_{k,i} - \boldsymbol{h}_i\|_2\|\tilde{\boldsymbol{r}}_k - \boldsymbol{r}_k\|_2 n_o^{1/2}}{\epsilon} = o(1),
\end{aligned}$$

which implies (B.1). (B.2) can be verified by the Cauchy-Schwartz inequality and assumption B.2.4. (B.4) can be proved similarly under assumption B.2.4.

To show (B.3), we first define $\mathcal{F}_n = \{\tilde{\boldsymbol{r}}_k^T\boldsymbol{\eta}_k(\boldsymbol{\theta}_{0,k}) : \boldsymbol{\eta}_k(\boldsymbol{\theta}_{0,k}) \in \Lambda^\gamma(\mathcal{X})\}$, and then check $\log N_{[]}\{\delta, \mathcal{F}_n, \|\cdot\|_2\} \leq const.(\frac{c}{\delta})^{d/\gamma}$ for any $\delta > 0$. Using results of (B.1)-(B.4), we obtain

$$0 \leq -\mathbb{G}_{n_o}\{\boldsymbol{r}_k^T(\boldsymbol{h}_k - \boldsymbol{\eta}_k)\} + \sqrt{n_o}E_o\{\boldsymbol{r}_{k,i}^T(\hat{\boldsymbol{\eta}}_{k,i} - \boldsymbol{\eta}_{k,i})\} + o_p(1),$$

where $E_o(\cdot) = E(\cdot \mid D_i \in \mathcal{S}_o)$ denotes the expectation for all studies with fully observed data, and similarly with t_n replaced by $-t_n$,

$$0 \leq \mathbb{G}_{n_o}\{\boldsymbol{r}_k^T(\boldsymbol{h}_k - \boldsymbol{\eta}_k)\} - \sqrt{n_o}E_o\{\boldsymbol{r}_{k,i}^T(\hat{\boldsymbol{\eta}}_{k,i} - \boldsymbol{\eta}_{k,i})\} + o_p(1).$$

Therefore

$$\begin{aligned}\sqrt{n_o}E_k\big[\boldsymbol{\zeta}_l^T\boldsymbol{V}_{k,i,s}\{\hat{\boldsymbol{\eta}}_{k,i}(\boldsymbol{\theta}_{0,k}) - \boldsymbol{\eta}_{k,i}(\boldsymbol{\theta}_{0,k})\}\big] &= \sqrt{n_o}E_o\{\boldsymbol{r}_{k,i}^T(\hat{\boldsymbol{\eta}}_{k,i} - \boldsymbol{\eta}_{k,i})\} \\ &= \mathbb{G}_{n_o}\{\boldsymbol{r}_k^T(\boldsymbol{\eta}_k - \boldsymbol{h}_k)\} + o_p(1).\end{aligned}$$

Combining results for $l = 1, \ldots, p+q$, we prove the lemma. □

Proof of Theorem III.2. We prove the theorem by verifying conditions of Theorem III.2 of *Chen et al.* (2003). Conditions 2.1 and 2.2 (*Chen et al.*, 2003) are satisfied by the construction of $\hat{\boldsymbol{\theta}}_k$ and Assumption $B.2.1$. Conditions 2.3 and 2.4 of *Chen et al.* (2003) are proved in Lemmas B.3 and B.4. According to Lemma B.5, conditions 2.5 of *Chen et al.* (2003) holds as well. Thus

$$n_k^{1/2}\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_k) = n_k^{1/2}\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_k) + \tau_k^{1/2}n_o^{1/2}\bar{\boldsymbol{q}}_k(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_k) + o_p(1),$$

and $n_k^{1/2}\bar{\boldsymbol{g}}_k(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_k)$ converges to a normal distribution with variance

$$\boldsymbol{\Sigma}_k = Var\{\boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{\eta}_{k,i}) + \tau_k^{1/2}\boldsymbol{q}_{k,i}(\boldsymbol{\theta}_{0,k}, \boldsymbol{h}_{k,i})\} = \boldsymbol{\Sigma}_{k,1} + \tau_k\boldsymbol{\Sigma}_{k,2},$$

where $\boldsymbol{\Sigma}_{k,1}$ and $\boldsymbol{\Sigma}_{k,2}$ are given in Theorem III.2. Next we can establish the following approximation:

$$\sup_{\|\hat{\boldsymbol{\eta}}_{k,i}-\boldsymbol{\eta}_{k,i}\|_\infty=o(1)} \left\| \frac{1}{n_k} \sum_{D_i=k} \nabla \boldsymbol{g}_{k,i}(\boldsymbol{\theta}_{0,k}, \hat{\boldsymbol{\eta}}_{k,i}) - \boldsymbol{G}_k \right\| = o_p(1).$$

Finally standard arguments for GMM estimator will lead to

$$\sqrt{n_k}(\hat{\boldsymbol{\theta}}_k - \boldsymbol{\theta}_{0,k}) \xrightarrow{d} N\{\mathbf{0}, (\boldsymbol{G}_k \boldsymbol{\Sigma}_k^{-1} \boldsymbol{G}_k)^{-1}\}.$$

□

BIBLIOGRAPHY

BIBLIOGRAPHY

Afeiche, M., et al. (2011), Prenatal lead exposure and weight of 0- to 5-year-old children in mexico city, *Environmental Health Perspectives*, *119*(10), 1436–1441.

Anderson, T. (2003), *An Introduction to Multivariate Statistical Analysis, Third Edition*, Wiley-Interscience, New Jersey.

Andrews, D. W. K. (1994), Asymptotics for semiparametric econometric models via stochastic equicontinuity, *Econometrica*, *62*(1), 43–72.

Aschard, H., W. Qiu, B. Pasaniuc, N. Zaitlen, M. Cho, and V. Carey (2011), Combining effects from rare and common genetic variants in an exome-wide association study of sequence data, *BMC Proceedings*, *5*(Suppl 9), S44.

Becker, B. J. (2007), Multivariate meta-analysis: Contributions of ingram olkin, *Statistical Science*, *22*(3), 401–406.

Bondell, H., and B. Reich (2008a), Simultaneous factor selection and collapsing levels in anova., *Biometrics*, *65*(1), 169–177.

Bondell, H. D., and B. J. Reich (2008b), Simultaneous Regression Shrinkage, Variable Selection, and Supervised Clustering of Predictors with OSCAR, *Biometrics*, *64*(1), 115–123.

Breslow, N. E., and N. E. Day (1980), *Statistical Methods in Cancer Research: The Analysis of Case-control Studies Vol 1*, International Agency for Research on Cancer, lyon.

Cabana, M. D., K. K. Slish, D. Evans, R. B. Mellins, R. W. Brown, X. Lin, N. Kaciroti, and N. M. Clark (2006), Impact of physician asthma care education on patient outcomes, *Pediatrics*, *117*(6), 2149–57.

Chen, X., O. Linton, and I. V. Keilegom (2003), Estimation of semiparametric models when the criterion function is not smooth, *Econometrica*, *71*(5), 1591–1608.

Chen, X., H. Hong, and E. Tamer (2005), Measurement error models with auxiliary data, *Review of Economic Studies*, *72*(2), 343–366.

Chen, X., H. Hong, and A. Tarozzi (2007), Semiparametric efficiency in gmm models with auxiliary data, *Annals of Statistics*, *36*(2), 808–843.

Crainiceanu, C. M., and D. Ruppert (2004), Likelihood ratio tests in linear mixed models with one variance component, *Journal of the Royal Statistical Society: Series B*, *66*(1), 165–185.

Craven, P., and G. Wahba (1979), Smoothing noisy data with spline functions, *Numerische Mathematik*, *31*(4), 377–403.

Crowder, M. (1995), On the use of a working correlation matrix in using generalised linear models for repeated measures, *Biometrika*, *82*(2), 407–410.

Davidian, M., et al. (2005), Semiparametric estimation of treatment effect in a pretest-posttest study with missing data [with comments and rejoinder], *Statistical Science*, *20*(3), 261–301.

Davidson, J. (2001), *Stochastic Limit Theory*, Oxford University Press, New York.

Davis, W. W., V. L. Parsons, D. Xie, N. Schenker, T. E. Raghunathan, and E. J. Feuer (2010), State-based estimates of mammography screening rates based on information from two health surveys, *Public Health Rep*, *125*, 567–78.

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

Fan, J., and R. Li (2001), Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, *96*(456), 1348–1360.

Fitzmaurice, G. M., S. R. Lipsitz, and J. G. Ibrahim (2007), A note on permutation tests for variance components in multilevel generalized linear mixed models, *Biometrics*, *63*(3), 942–946.

Follmann, D. A., and M. A. Proschan (1999), Valid inference in random effects meta-analysis, *Biometrics*, *55*(3), 732–727.

Friedman, J. H., T. Hastie, and R. Tibshirani (2010), Regularization paths for generalized linear models via coordinate descent, *Journal of Statistical Software*, *33*(1), 1–22.

Hall, P., and J. L. Horowitz (1996), Bootstrap critical values for tests based on generalized method of moments estimators, *Econometrica*, *64*(4), 891–916.

Hartung, J., G. Knapp, and B. K. Sinha (2008), *Statistical Meta-Analysis with Applications*, John Wiley and Sons, New Jersey.

Harville, D. A. (2008), *Matrix Algebra from A Statistician's Perspective,corrected edition*, Springer, New York.

He, X., Z.-Y. Zhu, and W.-K. Fung (2002), Estimation in a semiparametric model for longitudinal data with unspecified dependence structure, *Biometrika*, *89*(3), 579–590.

Hedges, L. V., and I. Olkin (1985), *Statistical Methods for Meta-analysis*, Academic Press, New York.

Horn, R. A., and C. R. Johnson (1990), *Matrix Analysis*, Cambridge University Press, New York.

Hu, F., and J. V. Zidek (2002), The weighted likelihood, *The Canadian Journal of Statistics*, *30*(3), 347–371.

Hu, Y., and P. X.-K. Song (2012), Sample size determination for quadratic inference functions in longitudinal design with dichotomous outcomes, *Statistics in Medicine*, *31*(8), 787–800.

Ichimura, H., and S. Lee (2010), Characterization of the asymptotic distribution of semiparametric m-estimators, *Journal of Econometrics*, *159*(2), 252–266.

Inoue, L. Y. T., R. Etzioni, E. H. Slate, C. Morrell, and D. F. Penson (2004), Combining longitudinal studies of psa, *Biostatistics*, *5*(3), 483–500.

Ishak, K. J., R. W. Platt, L. Joseph, J. A. Hanley, and J. J. Caro (2007), Meta-analysis of longitudinal studies, *Clinical Trials*, *4*, 525–539.

Ishak, K. J., R. W. Platt, L. Joseph, and J. A. Hanley (2008), Impact of approximating or ignoring within-study covariances in multivariate meta-analyses, *Statistics in Medicine*, *27*(5), 670–686.

Kim, J. K., and W. Fuller (2004), Fractional hot deck imputation, *Biometrika*, *91*(3), 559–578.

Kim, S.-J., K. Koh, S. P. Boyd, and D. M. Gorinevsky (2009), l_1 trend filtering, *SIAM Review*, *51*(2), 339–360.

Laird, N. M., and J. H. Ware (1982), Random-effects models for longitudinal data, *Biometrics*, *38*(4), 963–974.

Leng, C., W. Zhang, and J. Pan (2010), Semiparametric mean covariance regression analysis for longitudinal data, *Journal of the American Statistical Association*, *105*(489), 181–193.

Lipsitz, S. R., L. P. Zhao, and G. Molenberghs (1998), A semiparametric method of multiple imputation, *Journal of the Royal Statistical Society. Series B*, *60*(1), 127–144.

Little, R. J. A., and D. B. Rubin (1987), *Statistical Analysis with Missing Data*, Wiley Series in Probability and Statistics, 1st ed., Wiley, New York.

Localio, A. R., J. A. Berlin, T. R. T. Have, and S. E. Kimmel (2001), Adjustments for center in multicenter studies: An overview, *Annals of Internal Medicine*, *135*(2), 112–123.

Lopes, H. F., P. Müller, and G. L. Rosner (2003), Bayesian meta-analysis for longitudinal data models using multivariate mixture priors, *Biometrics*, *59*(1), 66–75.

Mancl, L. A., and T. A. DeRouen (2001), A covariance estimator for gee with improved small-sample properties, *Biometrics*, *57*(1), 126–134.

Mendenhall, W., R. J. Beaver, and B. M. Beaver (2008), *Introduction to Probability and Statistics*, Duxbury Press.

Molenberghs, G., and M. G. Kenward (2007), *Missing Data in Clinical Studies*, Wiley.

Müller, P., F. Quintana, and G. Rosner (2004), A method for combining inference across related nonparametric bayesian models, *Journal of the Royal Statistical Society: Series B*, *66*(3), 735–749.

Newey, W. K. (1994), The asymptotic variance of semiparametric estimators, *Econometrica*, *62*(6), 1349–1382.

Newey, W. K. (1997), Convergence rates and asymptotic normality for series estimators, *Journal of Econometrics*, *79*(1), 147–168.

Pan, W. (2001), On the robust variance estimator in generalised estimating equations, *Biometrika*, *88*(3), 901–906.

Pourahmadi, M. (1999), Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation, *Biometrika*, *86*(3), 677–690.

Powell, J. L. (1986), Estimation of semiparametric models, in *Handbook of Econometrics*, *Handbook of Econometrics*, vol. 4, edited by R. F. Engle and D. McFadden, chap. 41, pp. 2443–2521.

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

Qu, A., B. G. Lindsay, and L. Lu (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.

Reilly, M., and M. S. Pepe (1995), A mean score method for missing and auxiliary covariate data in regression models, *Biometrika*, *82*(2), 299–314.

Robins, J. M., and A. Rotnitzky (1995), Semiparametric efficiency in multivariate regression models with missing data, *Journal of the American Statistical Association*, *90*(429), 122–129.

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

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

Rubin, D. B. (1987), *Multiple Imputation for Nonresponse in Surveys*, Wiley, New Jersey.

Rubin, D. B. (1996), Multiple imputation after 18+ years, *Journal of the American Statistical Association*, *91*(434), 473–489.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999), Adjusting for nonignorable drop-out using semiparametric nonresponse models, *Journal of the American Statistical Association*, *94*(448), 1096–1120.

Schumaker, L. L. (1981), *Spline Functions: Basic Theory*, Pure and Applied Mathematics.

Schwarz, G. (1978), Estimating the dimension of a model, *The Annals of Statistics*, *6*(2), 461–464.

Shen, X. (1997), On methods of sieves and penalization, *The Annals of Statistics*, *25*(6), 2555–2591.

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

Sinha, S. K. (2009), Bootstrap tests for variance components in generalized linear mixed models, *Canadian Journal of Statistics*, *37*(2), 219–234.

Song, P. X.-K. (2007), *Correlated Data Analysis: Modeling, analytics, and Applications*, Springer, New York.

Stone, C. J. (1982), Optimal global rates of convergence for nonparametric regression, *The Annals of Statistics*, *10*(4), 1040–1053.

Stram, D. O., and J. W. Lee (1994), Variance components testing in the longitudinal mixed effects model, *Biometrics*, *50*(4), 1171–1177.

Sun, X., N. Wang, D. Li, X. Zheng, S. Qu, L. Wang, F. Lu, K. Poundstone, and L. Wang (2007), The development of HIV/AIDS surveillance in china, *AIDS*, *21*, 33–38.

Thase, M. E., S. G. Kornstein, J.-M. Germain, Q. Jiang, C. Guico-Pabia, and P. T. Ninan (2009), An integrated analysis of the efficacy of desvenlafaxine compared with placebo in patients with major depressive disorder, *CNS Spectrums*, *14*(3), 144–154.

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

Ueki, M. (2009), A note on automatic variable selection using smooth-threshold estimating equations, *Biometrika*, *96*(4), 1005–1011.

Ueki, M., and Y. Kawasaki (2011), Automatic grouping using smooth-threshold estimating equations, *Electronic Journal of Statistics*, *5*, 309–328.

Wang, F., L. Wang, and P. X. K. Song (2012), Quadratic inference function approach to merging longitudinal studies: validation and joint estimation, *Biometrika*, *99*(3), 755–762.

Wang, L., and A. Qu (2009), Consistent model selection and data-driven smooth tests for longitudinal data in the estimating equations approach, *Journal Of The Royal Statistical Society: Series B*, *71*(1), 177–190.

Wang, M., and Q. Long (2011), Modified robust variance estimator for generalized estimating equations with improved small-sample performance, *Statistics in Medicine*, *30*(11), 1278–1291.

Wang, X., and J. V. Zidek (2005), Selecting likelihood weights by cross-validation, *The Annals of Statistics*, *33*(2), 463–500.

Xie, Y., and C. Ahn (2010), Statistical methods for integrating multiple types of high-throughput data, *Methods in Molecular Biology*, *620*, 511–529.

Yuan, M., M. Yuan, Y. Lin, and Y. Lin (2006), Model selection and estimation in regression with grouped variables, *Journal of the Royal Statistical Society, Series B*, *68*, 49–67.

Zeger, S. L., K.-Y. Liang, and P. S. Albert (1988), Models for longitudinal data: A generalized estimating equation approach, *Biometrics*, *44*(4), 1049–1060.

Zhang, D., and X. Lin (2008), Variance component testing in generalized linear mixed models for longitudinal/clustered data and other related topics.

Zhang, W., J. Fan, and Y. Sun (2009), A semiparametric model for cluster data, *The Annal of Statistics*, *37*(5A), 2377–2408.

Zhang, Z., D. Chen, and D. A. Fenstermacher (2007), Integrated analysis of independent gene expression microarray datasets improves the predictability of breast cancer outcome, *BMC Genomics*, *8*(1), 331.

Zou, H. (2006), The adaptive lasso and its oracle properties, *Journal of the American Statistical Association*, *101*, 1418–1429.

Zou, H., and M. Yuan (2008), The f_∞-norm support vector machine, *Statistica Sinica*, *18*(1), 379–398.