Spring 2019

Regression for Pooled Testing Data with Biomedical Applications

Juexin Lin

Regression for pooled testing data with biomedical applications

by

Juexin Lin

Bachelor of Science
Jilin University, 2012

Master of Science
George Washington University, 2014

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2019

Accepted by:

Dewei Wang, Major Professor

Joshua M. Tebbs, Committee Member

Karl B. Gregory, Committee Member

Zhu Wang, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Dedication

This dissertation is dedicated to my loving family

My Husband

Yawei Liang

My Parents

Dequan Lin & Yahe Ye

My Sister

Jueying Lin

Acknowledgments

I would first like to express my deepest and sincere gratitude to my advisor Dr. Dewei Wang, who brought me to this field of study and guided me through last four years. His innovative thinking and conscientious attitude of tackling issues affect me all the time. I will never forget his continuous guidance in shaping me become more capable of things not only in research but also in life. Without his persistent encouragement and guidance, I could not be possible to complete my doctoral dissertation.

I am thankful to Drs. Joshua Tebbs, Karl Gregory, and Zhu Wang for spending time serving on my committee and giving constructive suggestions on my dissertation. I would especially like to thank Dr. Joshua Tebbs for reading and sharing the insightful comments and advice on my academic papers.

I also like to thank all the professors and faculty members in the department of statistics for teaching me the statistical knowledge and offering me various resources and activities during the five-year doctoral life. Thanks also go to my friends for making this journey colorful.

Last but not least, I want to thank my parents and my sister for always supporting me and always being there for me. I would not be the person I am today without their love and support. I owe a special appreciation to my loving husband, Yawei Liang, who has accompanied me since I came to the United States. Thanks for his patience, toleration and love making our past seven years precious.

Abstract

Since first introduced by Dorfman in 1943, pooled testing has been widely used as a cost and time effective testing protocol in the variety of applications. This dissertation consists of three projects that reveal the use of pooling techniques in the disease prevention from the perspective of regression. For disease monitoring and control, individual covariates information are often of practical interest and yield meaningful interpretations. It is natural to model the outcome of interest, which can be either a disease status (binary) or a biomarker concentration index (continuous), with individual-specific covariates through a regression analysis. Chapter 2 focuses on the pooled biomarker assessment, where a pooling procedure is implemented to measure a continuous outcome of interest. A semi-parametric single-index model is developed to model the mean trend of biomarker concentration. In spite of pooled biomarker assessment, this dissertation also focuses on the group testing problems in infectious disease studies. In Chapter 3, we propose a multivariate logistic regression model for the multiple-infection group testing data. To facilitate the variable selection and model interpretation, we further develop a regularized approach which selects the active risk factors for each infection. Other than significant cost savings, pooling strategy provides more precise biomarker mean curve estimations (in Chapter 2), and more accurate variable selections (in Chapter 3). With these cheerful benefits from pooling strategy, for the purpose of promoting group testing to laboratories, in Chapter 4, we further discuss how to simplify the pooled testing routine realistically without significant impairments on regression estimation.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Over the past decade, pooling techniques have been widely used for the purpose of cost or time reduction when the financial resource is limited or population of screening is large. The pooled testing (or group testing) was first introduced by Dorfman (1943) for screening syphilis among War World II US soldiers. The idea is to mix multiple individual specimens (e.g. blood, urine, swabs, etc.) to form a single pool to perform the test. The greatest benefit of pooling is cost saving. Beneficial from this economic efficiency, pooling has been applied to various areas of biometrical studies. The applications of pooling often result in two types of response interests: continuous and binary outcomes. Biological marker (biomarker) investigation and infectious disease screening are two typical applications of pooling methods on continuous and binary outcomes respectively.

We introduce the development of pooling for both cases in the following sections.

1.1 Pooled testing for biomarker investigation

Biomarkers are valuable biological indicators often used for case identification, early diagnosis, monitoring or predicting the progress of clinical response to an intervention during the treatment course. For example, the biomarker of susceptibility is used to indicate the risk of developing a disease in a patient without any clinically apparent symptoms; the predictive biomarker is used to identify patients who are more likely to experience a favorable outcome after medical treatment; the prognostic biomarker is used to identify the recurrence of a disease (FDA-NIH, 2016). In addition to the

disease-related biomarkers, there are some other drug-related biomarkers which are used to indicate the drug effectiveness in a drug development process. With those positive roles in disease control and prevention as well as medical product development, biomarkers have been widely used in the investigation of many diseases. For instance, cytokines and chemokines are predictive markers of miscarriage (Tong et al., 2004), MMPs is a diagnostic and prognostic marker for human cancer (Loukopoulos et al., 2003), and interleukin-6 of inflammation is a diagnostic biomarker for myocardial infarction in coronary heart disease (Faraggi et al., 2003). The benefits of biomarker have been acknowledged. However, to identify a selective and accurate testing method of biomarker diagnosis might require large samples, high cost and long time in laboratory.

As a remedy, the pooling technique (physically combining specimens into pools prior to performing laboratory assays) has been proposed as a cost and time effective analytical testing mechanism for biomarker evaluation in disease and medical research. Apart from the significant cost reduction, pooling has various practical and analytical advantages, such as preserving irreplaceable specimens, tackling the hindrance of detection limits or reducing the impact of potential outliers (Schisterman et al., 2011).

Along with the development of pooling strategy in practice, methods of analyzing data from such experiments have been investigated to explore further potential benefits. The statistical literature in biomarker pooling mainly focus on two topics: diagnostic efficacy evaluation and regression analysis. In the former, methodologies are proposed to evaluate the diagnostic efficacy of a biomarker by estimating the receiver-operating characteristic (ROC) curve and the corresponding area under the curve (AUC) from pooled biomarker measurements. The latter topic is to investigate the biomarkers measured in pools by the use of regression techniques on subject-specific characteristics, which is one of particular interest in this dissertation.

1.1.1 Diagnostic efficacy

In biomarker assessment, one critical qualifying attribute is the discriminatory ability by evaluating the associated ROC curve. Faraggi et al. (2003) investigated the effect of pooling on ROC curve analysis without the consideration of a detection limit effect for normally distributed markers. On account of the instruments' limit of detection, Mumford et al. (2006) concluded that pooling could increase the efficiency of estimation of the AUC through analyzing the pooled biomarker data. A generalization of ROC curve analysis on pooled biomarker was introduced in Vexler et al. (2008) to allow for the existence of additive measurement errors and a wide family of biomarker distribution assumptions.

1.1.2 Regression analysis

In many clinical studies, some patient information would be collected before assays, such as demographics, clinical symptoms, or historical risk factors. It is natural to incorporate such information into biomarker pooling study. With the collection of individual covariates, researchers are interested in studying the regression inferences such as estimating the biomarker distribution, modeling the biomarker trend, performing the hypothesis test, or predicting the biomarker concentration for a future subject. Ma et al. (2011) initially proposed the use of a linear regression model for the pooled biomarker data. Under the assumption that the observed pooled biomarker concentration conditional on individual-specific covariates follows the Gaussian distribution, the authors proposed a simple linear regression method and then discovered an optimal pooling design via the D-efficiency criteria computed from variance estimations. Malinovsky et al. (2012) proposed a random effects model for the longitudinal data with repeated measurements, but also assuming the outcomes are conditionally Gaussian. In practice, many biomarkers tend to be right-skew distributed. To generalize the techniques to allow for right-skewed outcomes, Mitchell et al. (2014) explored the

regression analysis involving a log transformation and then proposed a Monto Carlo expectation maximization algorithm to estimate regression coefficients. Although many regression methods have been developed, the aforementioned methodologies are all parametric with strict biomarker distribution assumptions. If the restrictions are violated, estimations under parametric models may be misleading. My dissertation is to relax those restrictions.

1.2 Pooled testing for infectious disease screening

Infectious diseases are the illness that can be spread from person to person generated by the transmissible pathogenetic microorganisms, such as bacteria, germs, nematodes, or prions. The infectious diseases are recognized as one of the top causes of death globally, particularly for young children in low-income countries. As World Health Organization (Last accessed 2018) reported, three of the top ten global causes of death in 2016 belonged to infectious disease, where lower respiratory infection ranked 4, diarrheal diseases ranked 9, and tuberculosis ranked 10. The situation is more serious in low-income countries where about 70% of deaths were due to infectious/transmissible diseases, particularly in the young. Infectious diseases have plagued humans throughout the past decades, such as SARS in 2003, H1N1 in 2009 or HIV/AIDS from 1990 to present. As such severity on human health, many efforts have been made on infectious disease research. Nevertheless, it is still a challenge to prevent and control these diseases. For instance, screening the infections among a large number of individuals one-by-one could be unfeasible when facing the laboratory's constraints. National Institutes of Health (NIH) and Centers for Disease Control and Prevention (CDC) have been investing a large amount of funding every year on infectious disease research, for example, the previous Infertility Prevention Project and the current ongoing Sexually Transmitted Disease Surveillance Network. Both of them are working on screening sexually transmitted diseases (STDs) like

chlamydia, gonorrhea and syphilis across the United States. To address this kind of large-scale screening problem, pooled testing has been widely used as a rapid and low-cost testing method.

The pooled testing first applied to infectious disease screening is proposed by Dorfman (1943) as a solution to test for the presence of syphilitic antigen in the blood samples of US soldiers in World War II. Rather than performing a test on each inductee's blood sample one by one, Dorfman suggested pooling multiple samples across individuals may be advantageous. Dorfman justified that the pooling blood specimen by mixing the negative serum would not contain the antigen as well. Thus, it is rational to diagnose all involved individual specimens as negative for syphilis when the pool tests negative. In contrast, if at least one of pool segment has the antigen, then the resultant pool would contain this antigen as well. In this case, Dorfman proposed that all individual specimens belonging to the positive pool should be retested separately for final diagnosis. Since Dorfman's seminal work, pooled testing has been applied to screen for many other infectious diseases in practice, such as chlamydia, gonorrhea (Tebbs et al., 2013), HIV, HBV, HCV (Stramer et al., 2013), West Nile virus (Busch et al., 2005), malaria (Wang et al., 2014a), influenza (Edouard et al., 2015), and herpes (Hill et al., 2016). Besides disease screening, many other areas, including genetics (Gastwirth, 2000), veterinary science (Munoz-Zanzi et al., 2000), medical entomology (Venette et al., 2002), blood safety (Dodd et al., 2002), and drug discovery (Remlinger et al., 2006), have also used the method of pooling.

Generally speaking, the literature in pooled testing can be divided into two categories: classification and estimation. For classification problems, many pooling algorithms have been studied, in order to improve case identification accuracy but meanwhile reduce testing cost. On the other hand, the estimation problems target at estimating the population-level or individual-level characteristics. This dissertation falls in the latter category.

1.2.1 Case identification

Numerous of pooled testing decoding algorithms have been developed for the purpose of identifying individual-level status as either presence or absence of the characteristic of interest. The simplest one is Dorfman decoding with two-stage classification. The procedure is to construct pools of non-overlapping groups of specimens and test the resultant pools in the first stage. The individual specimens that belonging to positive pools would be retested in the second stage. Sterrett (1957) generalized Dorfman's idea to a multiple-stage decoding strategy whereby repeatedly performing pooled testing randomly on the individuals from positive pools until all individuals are classified. Sterrett proved his procedure would increase the testing efficiency by about 6% compared to Dorfman decoding does. Litvak et al. (1994) proposed a halving decoding procedure, which splits the positive pool to two equally (or closest equally) sized sub-pools to test. The process begins from testing on the entire collection of specimens and proceeds until all specimens are declared. Phatarfod and Sudbury (1994) introduced array testing as a screening method for infectious diseases. Before performing the assay, all individual specimens are assigned to a grid, and the pools are formed by pooling the ones that belong to the same row or same column. By having the pools, the array testing is employed on those pools, then retesting is conducted for the specimens located at the intersection of a positive row and a positive column separately. Later on, researchers start to explore how much perception can the individual-specific covariate information provide on the disease identification. Bilder et al. (2010) generalized the Sterrett's (1957) multiple-stage algorithm to involve the individual covariate information. McMahan et al. (2012a) extended Dorfman decoding to a "greedy" heterogeneous algorithm which finds out the optimal group sizes by minimizing the expected number of tests. Black et al. (2012) proposed a modified halving decoding in which a positive pool is evenly divided in the order of individual risk probabilities. McMahan et al. (2012c) extended the array testing procedures to

informative array designs by the use of population heterogeneity. Generally, the goal of this research path is utmostly improving test efficiency or maximizing diagnostic accuracy.

1.2.2 Estimation

Another type of problems that are of particular interest is estimating disease risk from pooled testing data. Thompson (1962) was the first to consider modeling the observed pooled data to estimate the population prevalence of plant aster-yellow virus transmission by insects. Suppose k insects are caged per test plant and the probability that a randomly chosen insect being a vector is p. Thompson justified that the test plant would be non-infected with a probability of $(1-p)^k$. Based on this rational thought, Thompson (1962) provided a maximum likelihood estimator (MLE) of p and derived its asymptotic properties under the assumptions of a common p, independence of infection statuses among insects, and perfect testing. To generalize Thompson's approach which simply used the initial pool results, Sobel and Elashoff (1975) proposed the models that are applicable when the individual retesting information is available. Swallow (1985) showed that the MLE of p derived by Thompson was positively biased and the bias was monotone increased as the group size k. Burrows (1987) derived an alternative estimator which is much superior to the MLE provided in Thompson (1962) in terms of bias and mean square error. Hughes-Oliver and Swallow (1994) proposed an adaptive method to estimate the population prevalence, of which a priori prevalence value was used at the first stage to establish the pools. This type of adaptive estimator resulted in an improvement over the non-adaptive estimators in terms of the asymptotic performance. Hung and Swallow (1999) studied the robustness of group testing in two dilution-effect models and a serial correlation model by considering that both the absence of testing errors and independent individuals were violated.

The prevalence estimation in group testing from the Bayesian perspective has also been studied. Chaubey and Li (1995) proposed two Bayesian estimators of population prevalence with the consideration of two pre-specified prior distributions, and both of which were shown to be superior to the MLEs. Bilder and Tebbs (2005) proposed a new estimator from an empirical Bayesian approach which got rid of specifying hyperparameters distribution prior, and further formulated newly credible intervals for population prevalence. Hanson et al. (2006) proposed a Bayesian approach to estimate the population prevalence on a two-stage screening protocol, and meanwhile, it allowed for the estimation of the sensitivity and specificity of the testing assay.

More recently, research has shifted focus to explore the regression methodologies that incorporate the individual covariate information collected in the laboratories. Farrington (1992) first took the covariate information into account while modeling with group testing data. A generalized linear model with a complementary log-log link function was proposed in his work under the restrictive assumptions that individuals within each pool shared the identical covariates and the assay was perfect. Vansteelandt et al. (2000) extended Farrington (1992) to a more general framework which allowed for heterogeneous covariates inside pools, arbitrary link functions and testing errors. Xie (2001) introduced an expectation-maximization algorithm for the MLE, which adapted for a more flexible class of regression models and the additional decoding information. Bilder and Tebbs (2009) assessed pooling size and pooling strategy on the parameter estimation by comparing both non-homogeneous and homogeneous pooling with the traditional individual testing. Chen et al. (2009) proposed mixed-effects models for the group testing data with random-effect terms and described the estimation via a maximum likelihood approach. Huang and Tebbs (2009) examined the model misspecification for structural measurement error models with group testing responses. The aforementioned methodologies were specifically developed under parametric regression models, but more recently, Delaigle and Meister (2011) and

Delaigle and Hall (2012) proposed nonparametric methods for non-homogeneous and homogeneous pooling data respectively with the initial pool results only. Wang et al. (2014b) and Delaigle et al. (2014) developed the semi-parametric modeling approaches based on the single-index model allowing for multiple individual-level covariates. To take the dilution effect into account, McMahan et al. (2012b) and Delaigle and Hall (2015) suggested parametric and nonparametric estimators respectively. McMahan et al. (2017) presented a general Bayesian approach to analyze data arising from any group testing strategy and can be able to estimate accuracy probabilities along with the covariate effects.

1.2.3 Multiple infections group testing

Nowadays, multiplex assay, as an assay for multiple target analytes, becomes more and more popular because of its ability to extract more information than singleplex assay. For example, Aptima Combo 2 Assay (Gen-Prob, San Diego) is able to test chlamydia and gonorrhea at once in swab or urine biospecimens, and Procleix Ultrio Assay (Gen-Prob, San Diego) is an efficient method to detect HIV-1 RNA, HCV RNA, and HBV RNA simultaneously in individual blood samples. Initially, group testing is designed for one single infection. However, the use of multiplex assays makes pooled testing data with multiple infections widely available. Several statistical works have been proposed for case identification problem with multiple infections under various group testing strategies. Bilder et al. (2010) provided an informative retesting algorithm with the use of covariate information to structure how retesting was performed within positive groups. Tebbs et al. (2013) introduced a two-stage Dorfman decoding algorithm for multiple infections group testing data. Hou et al. (2017) generalized the two-stage hierarchical algorithm described in Tebbs et al. (2013) to a general S-stage setting with $S \geq 2$. Nevertheless, the research focusing on estimation with multiple-infection group testing data is scarce. A few works have studied the estimation of

disease prevalence. Hughes-Oliver and Rosenberger (2000) proposed an approach to estimate the prevalence of multiple rare traits under the assumption of no classification error. Concerned with misclassification of testing assay, Tebbs et al. (2013) and Warasi et al. (2016) provided a frequentist EM algorithm and a Bayesian approach to simultaneously estimate population prevalence under the assumption of the homogeneous population. Li et al. (2017) found the optimal group size in estimating the prevalence of two correlated diseases using the D-optimal criterion. Regarding regression analysis for multiple group testing data, this area remains mostly untapped. To the best of our knowledge, the only work is Zhang et al. (2013), of which approach was based on generalized estimating equations to estimate individual-level risk probability of each infection. Their approach only considered the initial pooled results.

1.3 Outlines

In this dissertation, several models for pooled testing data are investigated and presented from the perspective of regression estimation. In Chapter 2, a generalized semi-parametric framework is proposed for the regression analysis of pooled biomarker assessments. We introduce a dimension reduction model called single-index model where the continuous pooled biomarker measurement is modeled as the non-linear function of a linear combination of subject-specific covariates. A sequential estimation procedure is proposed to estimate the covariates coefficients and nonparametric curve. The proposed methodology not only overcomes the "curse of dimensionality" issue in nonparametric estimation but also maintains some nonparametric flexibility allowing us to capture the mean trend curve of the biomarker of interest. In addition, another desired feature of our model is the accessibility of meaningful interpretation for each covariate.

In Chapter 3, aiming at the multiple infections group testing data, we develop a joint multivariate logistic regression model which utilizes both subject-specific covari-

ates information and the potential individually decoding information. Although the data is modeled jointly, our proposed methodology can still draw the interpretable regression inference for each infection simultaneously. In the meantime, we allow the assay sensitivity and specificity to be unknown and estimate them simultaneously. Besides, we present a variable selection algorithm to identify relevant risk factors for each infection. We illustrate the performance of our methodology under two considered testing protocols through the simulation and real data application.

In Chapter 4, to encourage more laboratories to use pooling strategy, we investigate if the laboratory could simplify the data collection routine, for example, records only individual final diagnoses (and group memberships), how much regression inference would compromise. To answer this question, we conduct simulation studies under three considered data collection procedures. 1) Only collect the final diagnosis of each individual. 2) Collect both the group memberships and final individual diagnoses, but the pooled testing outcomes are not recorded. 3) Collect the whole data, including group constructing information, the pooled testing outcomes, and the individual diagnoses. We provide the estimation approaches for regression coefficients and compare the estimation performances across different scenarios under the two-stage Dorfman decoding algorithm of a single infection.

Chapter 2

Single-index regression for pooled biomarker data

Summary: Laboratory assays used to evaluate biomarkers (biological markers) are often prohibitively expensive. As an efficient data collection mechanism to save on testing costs, pooling has become more commonly used in epidemiological research. Useful statistical methods have been proposed to relate pooled biomarker measurements to individual covariate information. However, most of these regression techniques have proceeded under parametric linear assumptions. To relax such assumptions, we propose a semiparametric approach that originates from the context of the single-index model. Unlike with traditional single-index methodologies, we face a challenge in that the observed data are biomarker measurements on pools rather than individual specimens. In this chapter, we propose a method that addresses this challenge. The asymptotic properties of our estimators are derived. We illustrate the finite sample performance of our estimators through simulation and by applying it to a diabetes data set and a chemokine data set.

2.1 Literature review of single-index model

In this chapter, we propose a semiparametric method to model pooled data with continuous responses and individual covariate information, which overcomes the curse of dimensionality and maintains the important advantage of nonparametric flexibility. The new methodology is proposed in the context of the single-index model.

Instead of assuming a linear model, the single-index model assumes the mean of an individual response is related to a linear combination of the covariates through an unknown smooth function. It is a popular semiparametric model to accommodate multi-dimensional covariates while retaining the interpretability of the regression coefficients, see, for example, Ichimura (1993), Hardle et al. (1993), Klein and Spady (1993), Xia et al. (2002), Xia (2006), Zhu and Xue (2006) and Cui et al. (2011), who consider responses available on the individual level. In pooled testing, of course, the data are measured on pools. Existing single-index methods in pooled testing were proposed by Wang et al. (2014b) and Delaigle et al. (2014) for binary responses. This chapter presents a new single-index technique to analyze continuous pooled outcomes. We illustrate that when the population size is fixed, pooling could significantly reduce testing costs with only a minor loss in accuracy. On the other hand, when the number of assays is limited, testing pooled specimens does not compromise the estimation when compared to testing individual specimens.

The rest of this chapter is organized as follows. In Section 2.2, we present our semiparametric regression model to analyze biomarkers measured on pooled specimens, and in Section 2.3, we establish the asymptotic properties of the proposed estimators. We assess the performance of our methods using simulation in Section 2.4 and apply them to a diabetes data set from the National Health and Nutrition Examination Survey (NHANES) and a chemokine data set obtained from the Collaborative Perinatal Project (CPP) in Section 2.5. A discussion is given in Section 2.6. Proofs, and additional simulation results are provided in the Appendix A.

2.2 Model and methodology

2.2.1 Assumptions

We consider the situation in which J laboratory assays are taken on pools to measure a continuous biomarker of interest. The jth pool is formed by mixing c_j specimens, each

of which is obtained from an individual. The total number of individuals is denoted by $N = \sum_{j=1}^{J} c_j$. We let Y_{ij} and $\boldsymbol{X}_{ij} = (X_{ij1}, \ldots, X_{ijp})^{\mathrm{T}}$ denote the continuous biomarker level and the p-dimensional covariate of the ith individual in the jth pool, respectively, where $i = 1, \ldots, c_j$ and $j = 1, \ldots, J$. Assume throughout that the $(Y_{ij}, \boldsymbol{X}_{ij})$'s are independent and identical distributed (iid) versions of $(Y, \boldsymbol{X})$, where the mean and variance of Y given $\boldsymbol{X} = \boldsymbol{x}$ are

$$E(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \eta(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}) \text{ and } V(Y \mid \boldsymbol{X} = \boldsymbol{x}) = \sigma^2,$$

respectively, where $\eta(\cdot)$ is an unknown smooth curve, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is an unknown p-dimensional regression parameter, and $\sigma^2 > 0$. Note that we do not assume the type of the distribution of $Y \mid \boldsymbol{X} = \boldsymbol{x}$ to be known in advance. To ensure identifiability of a single-index model Lin and Kulasekera (2007), we assume that the support of the covariates, denoted by $\mathbb{X}$, is a bounded convex set that contains at least one interior point and the parameter space of $\boldsymbol{\beta}$ is $\boldsymbol{\mathcal{B}} = \{\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} : \beta_1 > 0, \|\boldsymbol{\beta}\| = 1\}$ where $\|\boldsymbol{\beta}\| = (\sum_{j=1}^{p} \beta_j^2)^{1/2}$. If the $(Y_{ij}, \boldsymbol{X}_{ij})$'s are observed, then traditional single-index modeling techniques can be applied to estimate $\eta(\cdot)$ and $\boldsymbol{\beta}$; e.g., see Ichimura (1993), Xia (2006), Wang et al. (2010), and Cui et al. (2011). However, in pooled testing, because assays are not taken on each individual, the Y_{ij}'s are all latent and the responses available to us are on the pooled level.

Denote the biomarker level of the jth pooled specimen by Z_j. In this chapter, we assume that $Z_j = c_j^{-1} \sum_{i=1}^{c_j} Y_{ij}$; i.e., the observed biomarker response is the arithmetic average of the individual biomarker levels. This is a common assumption in the statistical literature on biomarker pooling (Weinberg and Umbach, 1999; Faraggi et al., 2003; Vexler et al., 2008; Malinovsky et al., 2012; Lyles et al., 2015; McMahan et al., 2016). We view this to be reasonable as long as each individual contributes the same amount to the pool and there is no neutralization while pooling. The observed

data are $\{(Z_j, \boldsymbol{X}_{1j}, \ldots, \boldsymbol{X}_{c_j j}) : j = 1, \ldots, J\}$, where

$$E(Z_j \mid \boldsymbol{X}_{1j} = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_{c_j j} = \boldsymbol{x}_{c_j}) = \frac{1}{c_j} \sum_{i=1}^{c_j} \eta(\boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{\beta})$$

and $V(Z_j \mid \boldsymbol{X}_{1j} = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_{c_j j} = \boldsymbol{x}_{c_j}) = c_j^{-1} \sigma^2$. The goal of this work is to estimate $\eta(\cdot)$ and $\boldsymbol{\beta}$ based on the observed data $\{(Z_j, \boldsymbol{X}_{1j}, \ldots, \boldsymbol{X}_{c_j j}) : j = 1, \ldots, J\}$.

2.2.2 Estimation

In what follows, we propose a method to consistently estimate $\eta(\cdot)$ and $\boldsymbol{\beta}$. If $\eta(\cdot)$ was known, then one could immediately obtain an estimate of $\boldsymbol{\beta}$ by minimizing the weighted least squares objective function,

$$S\{\boldsymbol{\beta}, \eta(\cdot)\} = \sum_{j=1}^{J} c_j \left\{ Z_j - \frac{1}{c_j} \sum_{i=1}^{c_j} \eta(\boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta}) \right\}^2,$$

with respect to $\boldsymbol{\beta}$. A primary challenge herein is to account for the dependence between the infinite-dimensional parameter $\eta(\cdot)$ and the finite-dimensional parameter $\boldsymbol{\beta}$. To acknowledge this dependence in our notation, we write $\eta(\cdot)$ as $\eta_{\boldsymbol{\beta}}(\cdot)$; i.e., $\eta_{\boldsymbol{\beta}}(t) = E(Y_{ij} \mid \boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta} = t)$ for a given $\boldsymbol{\beta}$. If one can find a consistent estimator $\hat{\eta}_{\boldsymbol{\beta}}(\cdot)$ of $\eta_{\boldsymbol{\beta}}(\cdot)$, then our estimator of $\boldsymbol{\beta}$ can be obtained as $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}}{\operatorname{argmin}}\, S\{\boldsymbol{\beta}, \hat{\eta}_{\boldsymbol{\beta}}(\cdot)\}$. When each $\boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta}$ has its own response Y_{ij} available, $\hat{\eta}_{\boldsymbol{\beta}}(\cdot)$ could be obtained as the Nadaraya-Watson or the local-polynomial estimator between the Y_{ij}'s and the $\boldsymbol{X}_{ij}\boldsymbol{\beta}$'s (see, Ichimura, 1993; Cui et al., 2011). However, in our context, all the Y_{ij}'s are latent and $\{\boldsymbol{X}_{ij}\}_{i=1}^{c_j}$ share the same pooled response Z_j for each j. Constructing $\hat{\eta}_{\boldsymbol{\beta}}(\cdot)$ is not straightforward.

To circumvent this, we simply treat Z_j to be the response for each $\boldsymbol{X}_{ij}$ and find out what is $E(Z_j \mid \boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta} = t)$. Noting that $Z_j = c_j^{-1} \sum_{i=1}^{c_j} Y_{ij}$.

$$E(c_j Z_j \mid \boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta} = t) = E\left(\sum_{i=1}^{c_j} Y_{ij} \mid \boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta} = t \right) = \sum_{l \neq i} E(Y_{lj}) + E(Y_{ij} \mid \boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta} = t).$$

Because Y_{ij}'s are iid, we denote by μ the marginal expectation of Y_{ij}; i.e., $\mu = E(Y_{ij})$. Consequently, we have

$$E(c_j Z_j \mid \boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta} = t) = (c_j - 1)\mu + \eta_{\boldsymbol{\beta}}(t). \tag{2.1}$$

Comparing to the case where Y_{ij}'s are available, viewing Z_j as the response of $\boldsymbol{X}_{ij}$ adds one extra intercept term in the form of $(c_j - 1)\mu$. Equation (2.1) inspires the construction of $\hat{\eta}_{\boldsymbol{\beta}}(\cdot)$.

We first estimate μ. Marginally, one could easily recognize that the Z_j's are independent variables with mean μ, so a natural estimator of μ is $\hat{\mu} = N^{-1}\sum_{j=1}^{J} c_j Z_j$. Then, for a given $\boldsymbol{\beta}$ and t, we obtain the local linear kernel estimator $\hat{\eta}_{\boldsymbol{\beta}}(t)$ through minimizing the local least squares objective function,

$$\sum_{j=1}^{J}\sum_{i=1}^{c_j}\{c_j Z_j - (c_j-1)\hat{\mu} - \eta_{\boldsymbol{\beta}}(t) - \eta'_{\boldsymbol{\beta}}(t)(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - t)\}^2 K_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - t), \tag{2.2}$$

with respect to $\eta_{\boldsymbol{\beta}}(t)$ and $\eta'_{\boldsymbol{\beta}}(t)$, where $\eta'_{\boldsymbol{\beta}}(t)$ denotes the derivative of $\eta_{\boldsymbol{\beta}}(t)$, $K(\cdot)$ is a kernel function, h is a user-selected bandwidth, and $K_h(\cdot) = h^{-1}K(\cdot/h)$. The objective function in (2.2) utilizes a local linear approximation that approximates $\eta(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})$ by $\eta(t) + (\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - t)\eta'(t)$ at a given t. Because the accuracy of such an approximation depends on the distance between $\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}$ and t, the kernel term $K_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - t)$ weights $\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}$ more (less) if $\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}$ is close to (far away from) t. The local linear approximation became a well-accepted nonparametric regression technique due to the seminal work Fan (1993) where the optimality of local linear smoothers was demonstrated for the nonparametric regression. One could easily extend our method to incorporate a local polynomial (with a higher order) approximation (Fan and Gijbels, 1996) if $\eta_{\boldsymbol{\beta}}(\cdot)$ is smooth enough.

It is worthwhile to point out that the minimizer $\hat{\eta}_{\boldsymbol{\beta}}(t)$ can be expressed explicitly as

$$\hat{\eta}_{\boldsymbol{\beta}}(t) = \frac{S_{N2}(t,\boldsymbol{\beta})\hat{T}_{N0}(t,\boldsymbol{\beta}) - S_{N1}(t,\boldsymbol{\beta})\hat{T}_{N1}(t,\boldsymbol{\beta})}{S_{N0}(t,\boldsymbol{\beta})S_{N2}(t,\boldsymbol{\beta}) - S_{N1}^2(t,\boldsymbol{\beta})}, \tag{2.3}$$

where

$$S_{Nl}(t,\boldsymbol{\beta}) = N^{-1}h^{-l}\sum_{j=1}^{J}\sum_{i=1}^{c_j} K_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - t)\left(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - t\right)^l,$$

and

$$\hat{T}_{Nl}(t,\boldsymbol{\beta}) = N^{-1}h^{-l}\sum_{j=1}^{J}\sum_{i=1}^{c_j}\{c_j Z_j - (c_j-1)\hat{\mu}\}K_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - t)\left(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - t\right)^l,$$

for $l \in \{0, 1, 2\}$. Our final estimators are

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}}{\operatorname{argmin}} \; S\{\boldsymbol{\beta}, \hat{\eta}_{\boldsymbol{\beta}}(\cdot)\} \;\; \text{and} \;\; \hat{\eta}(\cdot) = \hat{\eta}_{\hat{\boldsymbol{\beta}}}(\cdot). \tag{2.4}$$

Directly minimizing $S\{\boldsymbol{\beta}, \hat{\eta}_{\boldsymbol{\beta}}(\cdot)\}$ in $\boldsymbol{\mathcal{B}} = \{\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} : \beta_1 > 0, \|\boldsymbol{\beta}\| = 1\}$ might be numerically difficult, because $\boldsymbol{\mathcal{B}}$ is a part of the surface of the unit ball. To reduce such a computational burden, we rewrite $\boldsymbol{\beta}$ to be $\boldsymbol{\beta} = (\sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2}, \boldsymbol{\beta}^{(1)\mathrm{T}})^{\mathrm{T}}$ where $\boldsymbol{\beta}^{(1)} = (\beta_2, \ldots, \beta_p)^{\mathrm{T}}$. Consequently, the parameter space is transformed from $\boldsymbol{\mathcal{B}}$ to be $\boldsymbol{\mathcal{B}}^{(1)} = \{\boldsymbol{\beta}^{(1)} = (\beta_2, \ldots, \beta_p)^{\mathrm{T}} : \|\boldsymbol{\beta}^{(1)}\| < 1\}$; i.e., the interior of the unit ball in $\mathbb{R}^{(p-1)}$. A numerical search inside a ball of a lower dimension is easier than on the surface of a ball of a higher dimension, even though theoretically they are the same.

2.3 Asymptotic properties

In this section, we present the asymptotic properties of our proposed estimators. To derive these properties, we assume that the group sizes remain finite as $N \to \infty$. We view this assumption to be reasonable because, in practice, the characteristics of the assay used often bound the pool size; i.e., larger pool sizes, at a point, could adversely affect an assay's accuracy and therefore would not be employed. For example, in a study on the relationship between chemokine levels and miscarriage, the levels of monocyte chemotactic protein-1 (MCP1) were measured using pools of size 2 (Whitcomb et al., 2007). In a BioCycle study, the F2-isoprostane level (a biomarker that measures oxidative stress) was measured in pools of size 3 (Malinovsky et al., 2012). To examine whether the pro-inflammatory cytokine interleukin-6 is a good predictor of myocardial infarction, pools of sizes 2 and 4 were used (McMahan et al., 2016). Besides practical concerns, diverging group sizes could also lead to theoretical issues. For instance, if $c_j \to \infty$ as $N \to \infty$, we have $E(Z_j \mid \boldsymbol{X}_{1j}, \ldots, \boldsymbol{X}_{c_j j}) = c_j^{-1} \sum_{i=1}^{c_j} \eta(\boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta})$ converges in probability to μ and $V(Z_j \mid \boldsymbol{X}_{1j}, \ldots, \boldsymbol{X}_{c_j j}) = c_j^{-1} \sigma^2$ converges to zero. In other words, when c_j's are large, all the Z_j's become nearly the same which makes

the estimation of $\eta(\cdot)$ and $\boldsymbol{\beta}$ very challenging. Hence, in this chapter, we focus on the scenario where c_j's are all finite.

Because Y_{ij}'s are latent as long as $c_j > 1$, using the method of pooling increases the theoretical complexity when comparing to traditional single-index models. Equation (2.1) provides an idea to consistently estimate the dependence between $\eta(\cdot)$ and $\boldsymbol{\beta}$ using pooled responses. It treats Z_j as the response for each covariate $\boldsymbol{X}_{ij}$ in the jth group. As a result, $(Z_j, \boldsymbol{X}_{ij})$'s are not iid observations as $(Y_{ij}, \boldsymbol{X}_{ij})$'s. Further, one needs to estimate the extra intercept term $(c_j - 1)\mu$ in advance. Despite these theoretical complications caused by pooling, we obtained the asymptotic properties of our estimators $\hat{\eta}(\cdot)$ and $\hat{\boldsymbol{\beta}}$. Before presenting the results, we need to introduce some notation. Because the group sizes (all positive integers) does not change with N, we denote the collection of the values of c_j by $\{c^{(m)} : m = 1, \ldots, M\}$, where M is also a fixed value. More explicitly, for each j, there exists an m such that $c_j = c^{(m)}$. For each m, we let J_m denote the number of pools having size $c^{(m)}$. The ratio $J_m c^{(m)}/N$ represents the proportion of individuals that were involved in pools of size $c^{(m)}$. When $N \to \infty$, we assume that this proportion converges to $\gamma_m \in [0, 1]$ where $\sum_{m=1}^{M} \gamma_m = 1$. Further, we denote the true parameters by $\eta_0(\cdot)$ and $\boldsymbol{\beta}_0 = (\beta_{01}, \boldsymbol{\beta}_0^{(1)\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{\beta}_0^{(1)} = (\beta_{02}, \ldots, \beta_{0p})^{\mathrm{T}}$. Let $\boldsymbol{\mathcal{J}}_0$ be the value of $\partial B(\boldsymbol{\beta}^{(1)})/\partial \boldsymbol{\beta}^{(1)}$ evaluated at $\boldsymbol{\beta}^{(1)} = \boldsymbol{\beta}_0^{(1)}$ where $B(\boldsymbol{\beta}^{(1)}) = (\sqrt{1 - \|\boldsymbol{\beta}^{(1)}\|^2}, \boldsymbol{\beta}^{(1)\mathrm{T}})^{\mathrm{T}}$. Moreover, denote by $\boldsymbol{\Omega}_0(\boldsymbol{X}) = E[\boldsymbol{X}\boldsymbol{X}^{\mathrm{T}}|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0] - E[\boldsymbol{X}|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0]E[\boldsymbol{X}^{\mathrm{T}}|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0]$ and $\boldsymbol{\Omega} = E[\eta_0'^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)\boldsymbol{\Omega}_0(\boldsymbol{X})]$.

Before we present the asymptotic properties, we first provide some regularity conditions. These conditions are common in the single-index literature.

C1 : The curves $\boldsymbol{d}_\beta(t) = E(\boldsymbol{X}|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} = t)$ and $\eta_\beta(t)$ have bounded and continuous second order derivatives.

C2 : The probability density function of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$ is bounded away from zero and satisfies the Lipschitz condition of order 1 on $\{t = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{x} \in \mathbb{X}\}$.

C3 : As $N \to \infty$, $h \to 0$, $Nh^4 \to \infty$, and $Nh/\log N \to \infty$.

C4 : $K(\cdot)$ is a bounded and symmetric kernel function with bounded first derivative.

C5 : Conditional on $\boldsymbol{X}$, Y has a finite fourth moment.

C6 : The equation $\boldsymbol{u}^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{u} = 0$ has the unique root $\boldsymbol{u} = \boldsymbol{\beta}_0$ in $\boldsymbol{\mathcal{B}}$.

Conditions C1-C4 are common smoothness assumptions (see Xia, 2006; Wang et al., 2010; Cui et al., 2011). The Lipschitz condition in C2 allows for the discrete components in the covariates. Condition C5 is similar to the one used in Wang et al. (2010). Condition C6 assures that the matrix $\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}_a\boldsymbol{\mathcal{J}}_0$ is positive definite.

Under the mild regularity conditions provided above, we present the asymptotic properties of $\hat{\boldsymbol{\beta}}$ and $\hat{\eta}(\cdot)$ in Theorem 2.1. The proofs are provided in the Appendix A.1.

Theorem 2.1. Under conditions C1-C6 stated above, as $N \to \infty$,

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{d} \mathrm{N}(\mathbf{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{\Sigma} = \sigma^2(\sum_{m=1}^{M} \gamma_m / c^{(m)})^{-1}\boldsymbol{\mathcal{J}}_0(\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\mathcal{J}}_0)^{-1}\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}$ and $\xrightarrow{d}$ means convergence in distribution. Furthermore,

$$\sup_{\boldsymbol{x}\in\mathbb{X}} |\hat{\eta}(\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) - \eta_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)| = O_p\{(Nh/\log N)^{-1/2}\}.$$

Theorem 2.1 reveals the asymptotic normality of $\hat{\boldsymbol{\beta}}$ and the consistency of $\hat{\eta}(\cdot)$. We would like to point out that, When c_j's are all 1, our pooled biomarker data Z_j's are exactly the individual-level responses Y_{ij}'s. Thus, the proposed estimator is the same as the classical single-index estimator based on all individual-level data; i.e., the asymptotic normality includes $c_j = 1$ as a special case. From the asymptotic variance, we could see some patterns that might help us understand the theoretical impact of pooling. For simplicity, let us consider all the pools to be of the same size; i.e., $c^{(m)} = c$, $\gamma_m = 1$ and $N = cJ$. We see that the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is $c\sigma^2\boldsymbol{\mathcal{J}}_0(\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\mathcal{J}}_0)^{-1}\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}/N$. Consequently, if the number of individuals (N)

is fixed in applications, pooling more individuals in a group would lead to a loss of information and yield a larger variability in the resulting estimates of $\boldsymbol{\beta}$. If the number of individuals is not limited but the budge is limited up to J assays, we could rewrite the asymptotic variance to be $\sigma^2 \boldsymbol{\mathcal{J}}_0 (\boldsymbol{\mathcal{J}}_0^{\mathrm{T}} \boldsymbol{\Omega} \boldsymbol{\mathcal{J}}_0)^{-1} \boldsymbol{\mathcal{J}}_0^{\mathrm{T}} / J$ which does not depend on the pool sizes. Thus, pooling does not compromise the asymptotic efficiency of $\hat{\boldsymbol{\beta}}$.

One must note that Theorem 2.1 holds when Condition C4 is satisfied; i.e., as $N \to \infty$, $h \to 0$, $Nh^4 \to \infty$, and $Nh / \log N \to \infty$. Thus, it is important to select a suitable value for the bandwidth h. One could use the traditional cross-validation method. That is leaving one group of data out and fitting the model using the remaining data to predict the response that was left out. After predicting all responses, the bandwidth is chosen to be the one that minimizes the sum of the squares of all the prediction errors. In other words, this traditional approach has to numerically search for an estimator of $\boldsymbol{\beta}$ when leaving each group out. When the number of groups J is large, the traditional cross-validation could cause a huge computational burden. In order to make our method more applicable in real applications, we suggest using a revised version of the traditional cross-validation method. This method was originally proposed by Hardle et al. (1993). Denote by $\hat{\eta}_{\boldsymbol{\beta}}^{(-j)}(u)$ the leave-one-out estimator of $\eta_{\boldsymbol{\beta}}(u)$ obtained via the explicit formula (3.5) without using the data pertaining to the jth pool. Our proposed bandwidth $\widetilde{h}$ is chosen so that $(\widetilde{\boldsymbol{\beta}}, \widetilde{h})$ minimizes

$$S_{cv}(\boldsymbol{\beta}, h) = \sum_{j=1}^{J} c_j \left\{ Z_j - \frac{1}{c_j} \sum_{i=1}^{c_j} \hat{\eta}_{\boldsymbol{\beta}}^{(-j)}(X_{ij}^{\mathrm{T}} \boldsymbol{\beta}) \right\}^2 .$$

Further, we use the value of $\widetilde{\boldsymbol{\beta}}$ as a sensible starting point to compute $\hat{\boldsymbol{\beta}}$.

2.4 Simulation studies

In this section, we illustrate the finite sample performance of our proposed method through simulation. Before presenting our results, we note that, to the best of our knowledge, the literature does not contain any competing methods for simultaneously

estimating both $\boldsymbol{\beta}$ and $\eta(\cdot)$ based on continuous pooled assessments. McMahan et al. (2016) proposed a parametric approach to estimate $\boldsymbol{\beta}$ by assuming $\eta(\cdot)$ is linear. Therefore, besides the main goal of illustrating of the performance of our proposed procedures under a variety of different settings, we have also compare our method with the one proposed by McMahan et al. (2016).

To illustrate that our estimation procedure does not rely on the distribution of biomarker levels; i.e., the distribution of $Y_{ij} \mid \boldsymbol{X}_{ij}$, we consider the following examples:

(D1) : $Y \mid \boldsymbol{X} = \boldsymbol{x} \sim \mathrm{N}\{\eta_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0), \sigma^2\}$

(D2) : $Y \mid \boldsymbol{X} = \boldsymbol{x} \sim \mathrm{Gamma}\{\text{shape} = \eta_0^2(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)/\sigma^2, \text{rate} = \eta_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)/\sigma^2\}$

(D3) : $Y \mid \boldsymbol{X} = \boldsymbol{x} \sim \text{Log-Normal}\{\mu_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0), g_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}$, where
$\mu_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0) = \log\{\eta_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\} - g_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)/2$, and $g_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0) = \log\{\sigma^2/\eta_0^2(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0) + 1\}$.

The normal distribution is used to simulate symmetric biomarker data, while the other two cases are used to emulate right skewed distributions. These distributions are used in simulating biomarker levels, but are not used in the part of estimation. Parameters in these distributions are chosen to satisfy our model assumption that $E(Y|\boldsymbol{X} = \boldsymbol{x}) = \eta_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)$, and $V(Y|\boldsymbol{X} = \boldsymbol{x}) = \sigma^2$. We set $\boldsymbol{\beta}_0 = (0.5, 0.5, \sqrt{2}/2)^{\mathrm{T}}$ and $\sigma = 0.5$. For $\eta_0(\cdot)$, we consider the following four models:

(M1) : $\eta_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta} + 2$

(M2) : $\eta_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)^2$

(M3) : $\eta_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0) = (\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)^2 \exp(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)$

(M4) : $\eta_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0) = \sin(a\pi\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0) + 1$, where $a = 1$ or 2.

One of the most attractive features of a single-index approach is that it does not force any shapes on the regression curve while being able to consistently estimate the regression coefficients. Model (M1) is chosen to be linear purposely. Through this

setting, we would like to see the consequences of using our method if ignoring the truly linearity. The regression curves in (M2)–(M4) are nonlinear. They are similar to those discussed in Cui et al. (2011). Models (M2) and (M3) offer smooth curves; in contrast, Model (M4) results in an oscillating curve with the frequency being controlled by a; i.e., larger values of a result in a larger degree of oscillatory behavior of the curve over the range of $\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0$. These nonlinear curves could illustrate the benefits of using our method if the regression curve is not linear. The vector of covariates $\boldsymbol{X} = (X_1, X_2, X_3)^{\mathrm{T}}$ contains continuous X_1 and X_2 following Uniform$(-1, 1)$ and $\mathrm{N}(0, 0.3^2)$ distributions, respectively, and discrete X_3 with $P(X_3 = \pm 0.5) = 0.5$.

To generate pooled observations, we consider two scenarios. In the first, the number of individual specimens to be tested is fixed. Testing the specimens individually is ideal providing full information; however, this may be impractical due to the financial limitations and thus pooling is used. We choose $N \in \{2500, 5000\}$ and specify a common pool size $c_j = c$ for all $j = 1, \ldots, J$, where $c \in \{1, 2, 5, 10\}$. Different values of c indicate different levels of savings. For example, $(N, c) = (2500, 5)$ means an 80% reduction in testing cost when compared to $(N, c) = (2500, 1)$ where each individual is tested separately. In this scenario, we are able to evaluate how the reduction of the number of tests would affect the accuracy of estimating $\boldsymbol{\beta}_0$ and $\eta_0(\cdot)$. For each combination of (D1)–(D3), (M1)–(M4), and $N \in \{2500, 5000\}$, we randomly generate N samples of $(Y, \boldsymbol{X})$ according to the covariate setting, the selected conditional distribution of $Y \mid \boldsymbol{X}$, and $\eta_0(\cdot)$. Then for each $c \in \{1, 2, 5, 10\}$, we randomly assign the N samples into $J = N/c$ pools, and label them by $(Y_{ij}, \boldsymbol{X}_{ij})$ where $i = 1, \ldots, c$, $j = 1, \ldots, N/c$. The testing response of the jth pooled specimen is determined by $Z_j = c^{-1} \sum_{i=1}^{c} Y_{ij}$.

In the second scenario, the investigator may have only J assays available due to the limitation of financial resources. The choice is between testing J specimens one-by-one or testing cJ specimens using pools of size c. We consider $J \in \{250, 500\}$ and

$c_j = c \in \{1, 2, 5, 10\}$. For example, $(J, c) = (250, 10)$ implies that even though there are only 250 assays, pooling could involve 10 times the number of specimens than testing individual specimens; i.e., $(J, c) = (250, 1)$. Through these settings, we are able to see whether the extra number of individuals could provide more information and thus improve the estimation accuracy. For each combination of (D1)–(D3), (M1)–(M4), and $J \in \{250, 500\}$, we randomly generate $10 \times J$ copies of $(Y, \boldsymbol{X})$ to form the specimen bank. Then for each $c \in \{1, 2, 5, 10\}$, we randomly sample $N = cJ$ individuals from the specimen bank and assign them to J pools. After labeling them by $(Y_{ij}, \boldsymbol{X}_{ij})$ where $i = 1, \ldots, c$, $j = 1, \ldots, J$, we generate the testing response of the jth pool by $Z_j = c^{-1} \sum_{i=1}^{c} Y_{ij}$.

Within each configuration in both scenarios, we repeat the data generating process 500 times for all considered pool sizes and apply our methodology to estimate $\boldsymbol{\beta}_0$ and $\eta_0(\cdot)$. We specify the kernel function $K(\cdot)$ in (2.2) to be the probability density function of the standard normal distribution. The bandwidth h is selected via the leave-one-out cross-validation method described at the end of Section 2.3. In order to reveal the robustness of our method to the shape of a regression curve, we also fit each data under the parametric linear assumption. The applied parametric method is from McMahan et al. (2016).

Tables 2.1 and 2.2 summarize the results for Model (M1) under all the considered distributions (D1)–(D3) when $N \in \{2500, 5000\}$ and when $J \in \{250, 500\}$, respectively. These summary statistics include the sample mean and the standard deviation of the 500 estimates of $\boldsymbol{\beta}_0$. In order to illustrate what role the pool size c plays, we use the average estimation error (AEE), defined by $\text{AEE} = \sum_{k=1}^{p} |\hat{\beta}_{0k} - \beta_{0k}|$, as an overall measure of the estimation accuracy for $\boldsymbol{\beta}_0$ and the empirical mean squared error (MSE), calculated by $\text{MSE} = N^{-1} \sum_{j=1}^{J} \sum_{i=1}^{c} \{\hat{\eta}(\boldsymbol{X}_{ij}^{\mathrm{T}} \hat{\boldsymbol{\beta}}) - \eta_0(\boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0)\}^2$, to evaluate the accuracy of estimating the entire regression curve $\eta_0(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_0)$. The sample mean of the 500 AEE's and MSE$\times$10's are also included in the tables.

Table 2.1: Simulation results of the estimators for (M1) using our proposed method and the parametric method proposed in McMahan et al. (2016). Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE$\times$10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $N \in \{2500, 5000\}$.

			Proposed Method				Parametric Method			
	N		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	2500	β_{01}	0.498(0.028)	0.497(0.039)	0.497(0.044)	0.497(0.065)	0.500(0.018)	0.500(0.026)	0.500(0.042)	0.500(0.056)
		β_{02}	0.498(0.033)	0.494(0.050)	0.498(0.061)	0.494(0.088)	0.499(0.035)	0.498(0.047)	0.500(0.072)	0.499(0.096)
		β_{03}	0.708(0.025)	0.710(0.035)	0.705(0.039)	0.702(0.058)	0.706(0.021)	0.707(0.029)	0.708(0.045)	0.710(0.064)
		AEE(MSE$\times$10)	0.308(0.011)	0.324(0.019)	0.340(0.009)	0.372(0.017)	0.058(0.004)	0.080(0.007)	0.128(0.017)	0.172(0.030)
	5000	β_{01}	0.501(0.012)	0.498(0.031)	0.502(0.029)	0.499(0.050)	0.500(0.012)	0.500(0.018)	0.499(0.027)	0.498(0.040)
		β_{02}	0.501(0.019)	0.499(0.036)	0.498(0.042)	0.497(0.064)	0.500(0.025)	0.501(0.033)	0.503(0.052)	0.503(0.077)
		β_{03}	0.706(0.013)	0.707(0.027)	0.705(0.026)	0.704(0.043)	0.707(0.014)	0.708(0.020)	0.708(0.031)	0.707(0.043)
		AEE(MSE$\times$10)	0.294(0.002)	0.305(0.007)	0.317(0.003)	0.336(0.008)	0.041(0.002)	0.056(0.004)	0.089(0.008)	0.128(0.016)
(D2)	2500	β_{01}	0.495(0.040)	0.500(0.025)	0.496(0.046)	0.492(0.069)	0.499(0.018)	0.497(0.024)	0.497(0.039)	0.498(0.056)
		β_{02}	0.498(0.044)	0.500(0.038)	0.496(0.065)	0.485(0.099)	0.501(0.034)	0.502(0.049)	0.498(0.074)	0.500(0.111)
		β_{03}	0.709(0.033)	0.706(0.022)	0.707(0.043)	0.710(0.054)	0.708(0.021)	0.708(0.028)	0.709(0.043)	0.711(0.062)
		AEE(MSE$\times$10)	0.309(0.022)	0.315(0.005)	0.347(0.008)	0.386(0.020)	0.058(0.004)	0.080(0.007)	0.125(0.016)	0.182(0.032)
	5000	β_{01}	0.500(0.013)	0.497(0.032)	0.500(0.029)	0.497(0.046)	0.501(0.013)	0.500(0.017)	0.499(0.027)	0.498(0.037)
		β_{02}	0.500(0.020)	0.498(0.038)	0.499(0.046)	0.501(0.066)	0.499(0.023)	0.499(0.035)	0.497(0.052)	0.499(0.074)
		β_{03}	0.707(0.012)	0.708(0.027)	0.705(0.029)	0.703(0.042)	0.707(0.015)	0.708(0.020)	0.705(0.032)	0.703(0.044)
		AEE(MSE$\times$10)	0.296(0.002)	0.308(0.012)	0.320(0.004)	0.339(0.006)	0.041(0.002)	0.057(0.004)	0.090(0.008)	0.123(0.015)
(D3)	2500	β_{01}	0.498(0.029)	0.500(0.033)	0.497(0.048)	0.496(0.065)	0.499(0.016)	0.499(0.024)	0.499(0.039)	0.501(0.053)
		β_{02}	0.498(0.034)	0.498(0.042)	0.496(0.069)	0.492(0.099)	0.499(0.032)	0.499(0.046)	0.504(0.074)	0.509(0.110)
		β_{03}	0.708(0.024)	0.707(0.027)	0.706(0.047)	0.702(0.064)	0.706(0.020)	0.706(0.029)	0.706(0.046)	0.702(0.065)
		AEE(MSE$\times$10)	0.306(0.009)	0.316(0.005)	0.344(0.015)	0.372(0.019)	0.056(0.004)	0.079(0.007)	0.127(0.016)	0.181(0.032)
	5000	β_{01}	0.500(0.012)	0.499(0.023)	0.500(0.028)	0.496(0.053)	0.500(0.012)	0.500(0.018)	0.500(0.028)	0.500(0.040)
		β_{02}	0.499(0.018)	0.500(0.030)	0.498(0.045)	0.495(0.071)	0.499(0.023)	0.501(0.034)	0.502(0.053)	0.502(0.075)
		β_{03}	0.707(0.012)	0.707(0.021)	0.706(0.029)	0.706(0.047)	0.707(0.014)	0.706(0.019)	0.707(0.031)	0.707(0.046)
		AEE(MSE$\times$10)	0.293(0.002)	0.303(0.007)	0.323(0.004)	0.341(0.006)	0.039(0.002)	0.056(0.004)	0.090(0.008)	0.129(0.016)

Table 2.2: Simulation results of the estimators for (M1) using our proposed method and the parametric method proposed in McMahan et al. (2016). Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE$\times$10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $J \in \{250, 500\}$.

			Proposed Method				Parametric Method			
	J		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	250	β_{01}	0.497(0.064)	0.497(0.064)	0.497(0.062)	0.499(0.069)	0.499(0.054)	0.501(0.056)	0.501(0.057)	0.501(0.054)
		β_{02}	0.490(0.091)	0.497(0.087)	0.498(0.090)	0.490(0.099)	0.501(0.111)	0.504(0.105)	0.501(0.110)	0.503(0.110)
		β_{03}	0.705(0.059)	0.700(0.058)	0.699(0.059)	0.701(0.063)	0.709(0.065)	0.710(0.063)	0.711(0.062)	0.712(0.063)
		AEE(MSE$\times$10)	0.372(0.085)	0.362(0.033)	0.366(0.018)	0.365(0.017)	0.184(0.041)	0.179(0.070)	0.181(0.166)	0.182(0.316)
	500	β_{01}	0.495(0.055)	0.495(0.050)	0.494(0.049)	0.496(0.045)	0.501(0.037)	0.501(0.037)	0.501(0.039)	0.501(0.038)
		β_{02}	0.497(0.069)	0.496(0.067)	0.497(0.069)	0.498(0.068)	0.500(0.076)	0.498(0.079)	0.504(0.073)	0.498(0.077)
		β_{03}	0.706(0.052)	0.707(0.046)	0.706(0.046)	0.705(0.042)	0.707(0.044)	0.705(0.044)	0.705(0.043)	0.710(0.044)
		AEE(MSE$\times$10)	0.349(0.039)	0.347(0.029)	0.345(0.011)	0.340(0.005)	0.126(0.019)	0.128(0.035)	0.124(0.077)	0.127(0.156)
(D2)	250	β_{01}	0.496(0.066)	0.498(0.063)	0.496(0.063)	0.497(0.065)	0.495(0.054)	0.504(0.056)	0.501(0.056)	0.502(0.054)
		β_{02}	0.492(0.096)	0.498(0.089)	0.497(0.095)	0.491(0.092)	0.495(0.106)	0.493(0.110)	0.492(0.111)	0.505(0.113)
		β_{03}	0.703(0.060)	0.699(0.059)	0.700(0.060)	0.704(0.061)	0.710(0.066)	0.710(0.064)	0.704(0.063)	0.708(0.064)
		AEE(MSE$\times$10)	0.367(0.075)	0.365(0.028)	0.376(0.019)	0.370(0.015)	0.181(0.040)	0.182(0.072)	0.183(0.166)	0.185(0.325)
	500	β_{01}	0.499(0.048)	0.495(0.053)	0.497(0.047)	0.498(0.048)	0.499(0.040)	0.502(0.038)	0.496(0.039)	0.502(0.038)
		β_{02}	0.491(0.067)	0.492(0.067)	0.496(0.071)	0.494(0.069)	0.500(0.075)	0.504(0.077)	0.503(0.076)	0.501(0.075)
		β_{03}	0.708(0.047)	0.710(0.048)	0.706(0.046)	0.707(0.043)	0.706(0.044)	0.707(0.047)	0.706(0.046)	0.706(0.043)
		AEE(MSE$\times$10)	0.347(0.042)	0.351(0.041)	0.350(0.041)	0.347(0.033)	0.127(0.020)	0.129(0.036)	0.127(0.084)	0.124(0.150)
(D3)	250	β_{01}	0.502(0.062)	0.500(0.061)	0.493(0.065)	0.492(0.064)	0.500(0.054)	0.499(0.057)	0.500(0.054)	0.505(0.054)
		β_{02}	0.494(0.092)	0.491(0.095)	0.488(0.095)	0.490(0.095)	0.500(0.102)	0.502(0.111)	0.507(0.106)	0.498(0.105)
		β_{03}	0.699(0.056)	0.702(0.059)	0.709(0.061)	0.708(0.061)	0.706(0.064)	0.706(0.062)	0.713(0.059)	0.707(0.065)
		AEE(MSE$\times$10)	0.367(0.072)	0.374(0.029)	0.377(0.020)	0.376(0.015)	0.174(0.040)	0.183(0.074)	0.175(0.152)	0.181(0.313)
	500	β_{01}	0.497(0.052)	0.499(0.042)	0.500(0.041)	0.498(0.049)	0.499(0.040)	0.501(0.038)	0.498(0.040)	0.500(0.038)
		β_{02}	0.491(0.066)	0.501(0.062)	0.503(0.062)	0.495(0.068)	0.499(0.077)	0.495(0.076)	0.500(0.071)	0.501(0.073)
		β_{03}	0.709(0.047)	0.702(0.041)	0.699(0.039)	0.706(0.044)	0.706(0.046)	0.707(0.046)	0.708(0.045)	0.706(0.044)
		AEE(MSE$\times$10)	0.348(0.041)	0.335(0.012)	0.330(0.005)	0.342(0.008)	0.131(0.021)	0.128(0.035)	0.124(0.079)	0.123(0.148)

Table 2.3: Simulation results of the estimators for (M2) using our proposed method and the parametric method proposed in McMahan et al. (2016). Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE$\times$10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $N \in \{2500, 5000\}$.

			Proposed Method				Parametric Method			
	N		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	2500	β_{01}	0.500(0.041)	0.502(0.056)	0.523(0.080)	0.535(0.107)	0.000(0.022)	0.000(0.029)	−0.001(0.046)	0.000(0.062)
		β_{02}	0.496(0.045)	0.496(0.063)	0.504(0.085)	0.502(0.119)	−0.001(0.041)	−0.002(0.057)	0.001(0.085)	0.000(0.115)
		β_{03}	0.706(0.049)	0.701(0.063)	0.673(0.079)	0.653(0.096)	0.001(0.022)	0.002(0.030)	0.005(0.052)	0.007(0.071)
		AEE(MSE×10)	0.331(0.017)	0.356(0.021)	0.363(0.026)	0.386(0.045)	1.707(0.672)	1.707(0.676)	1.702(0.688)	1.700(0.705)
	5000	β_{01}	0.501(0.026)	0.503(0.039)	0.506(0.059)	0.525(0.077)	0.001(0.016)	0.000(0.020)	−0.001(0.032)	−0.002(0.043)
		β_{02}	0.501(0.030)	0.501(0.043)	0.507(0.065)	0.506(0.088)	0.000(0.029)	0.001(0.039)	0.001(0.059)	−0.003(0.085)
		β_{03}	0.704(0.032)	0.700(0.045)	0.689(0.066)	0.669(0.080)	0.000(0.017)	−0.001(0.023)	−0.001(0.036)	−0.001(0.050)
		AEE(MSE×10)	0.310(0.008)	0.325(0.010)	0.343(0.015)	0.366(0.020)	1.706(0.671)	1.707(0.673)	1.709(0.679)	1.713(0.688)
(D2)	2500	β_{01}	0.503(0.031)	0.504(0.046)	0.512(0.074)	0.524(0.097)	−0.001(0.021)	−0.001(0.028)	−0.002(0.044)	−0.002(0.063)
		β_{02}	0.503(0.036)	0.506(0.054)	0.508(0.083)	0.517(0.107)	−0.003(0.037)	0.001(0.050)	0.000(0.079)	0.003(0.114)
		β_{03}	0.700(0.035)	0.695(0.053)	0.679(0.080)	0.654(0.097)	0.001(0.024)	0.000(0.032)	−0.001(0.046)	−0.002(0.071)
		AEE(MSE×10)	0.314(0.014)	0.332(0.017)	0.359(0.023)	0.382(0.032)	1.703(0.671)	1.708(0.674)	1.710(0.684)	1.709(0.704)
	5000	β_{01}	0.502(0.025)	0.503(0.036)	0.511(0.054)	0.521(0.075)	0.001(0.019)	0.002(0.033)	0.000(0.052)	0.000(0.054)
		β_{02}	0.503(0.028)	0.502(0.041)	0.505(0.060)	0.512(0.084)	0.001(0.028)	0.002(0.041)	0.001(0.062)	0.004(0.085)
		β_{03}	0.702(0.029)	0.700(0.044)	0.689(0.057)	0.669(0.076)	−0.001(0.018)	−0.002(0.022)	0.000(0.048)	0.000(0.065)
		AEE(MSE×10)	0.308(0.007)	0.322(0.009)	0.336(0.013)	0.347(0.019)	1.706(0.672)	1.705(0.676)	1.708(0.687)	1.705(0.696)
(D3)	2500	β_{01}	0.501(0.040)	0.504(0.054)	0.518(0.077)	0.542(0.103)	0.001(0.020)	0.000(0.028)	0.001(0.045)	0.000(0.062)
		β_{02}	0.498(0.046)	0.499(0.061)	0.506(0.087)	0.507(0.121)	−0.003(0.041)	−0.007(0.055)	−0.005(0.085)	−0.004(0.115)
		β_{03}	0.703(0.047)	0.698(0.061)	0.675(0.079)	0.645(0.093)	0.000(0.022)	0.000(0.032)	0.000(0.050)	0.000(0.070)
		AEE(MSE×10)	0.331(0.016)	0.352(0.019)	0.360(0.026)	0.372(0.036)	1.708(0.673)	1.713(0.677)	1.711(0.689)	1.711(0.707)
	5000	β_{01}	0.501(0.028)	0.503(0.041)	0.507(0.060)	0.522(0.083)	0.001(0.015)	0.001(0.021)	0.000(0.031)	0.001(0.044)
		β_{02}	0.500(0.032)	0.500(0.043)	0.498(0.064)	0.503(0.089)	0.001(0.029)	0.001(0.039)	0.001(0.059)	−0.001(0.082)
		β_{03}	0.704(0.034)	0.701(0.047)	0.695(0.066)	0.673(0.079)	0.001(0.016)	0.000(0.022)	0.002(0.035)	0.003(0.051)
		AEE(MSE×10)	0.314(0.008)	0.326(0.010)	0.347(0.015)	0.365(0.024)	1.705(0.671)	1.705(0.673)	1.704(0.678)	1.705(0.688)

Table 2.4: Simulation results of the estimators for (M2) using our proposed method and the parametric method proposed in McMahan et al. (2016). Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE$\times$10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $J \in \{250, 500\}$.

			Proposed Method				Parametric Method			
	J		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	250	β_{01}	0.509(0.069)	0.508(0.080)	0.517(0.079)	0.512(0.077)	−0.001(0.067)	−0.001(0.066)	−0.001(0.065)	−0.002(0.062)
		β_{02}	0.500(0.081)	0.503(0.083)	0.502(0.085)	0.510(0.084)	−0.006(0.127)	−0.003(0.125)	0.007(0.122)	−0.004(0.121)
		β_{03}	0.689(0.073)	0.685(0.080)	0.679(0.081)	0.678(0.078)	−0.002(0.069)	−0.004(0.070)	0.001(0.075)	−0.001(0.071)
		AEE(MSE$\times$10)	0.456(0.067)	0.468(0.045)	0.442(0.027)	0.441(0.021)	1.716(0.695)	1.716(1.399)	1.701(3.535)	1.714(7.057)
	500	β_{01}	0.513(0.058)	0.511(0.072)	0.511(0.076)	0.512(0.069)	−0.002(0.047)	0.000(0.045)	−0.003(0.043)	−0.001(0.041)
		β_{02}	0.505(0.074)	0.497(0.077)	0.500(0.078)	0.509(0.081)	−0.005(0.087)	0.000(0.086)	−0.008(0.081)	−0.005(0.088)
		β_{03}	0.685(0.056)	0.690(0.067)	0.687(0.074)	0.680(0.071)	0.002(0.051)	0.000(0.052)	0.004(0.050)	−0.002(0.051)
		AEE(MSE$\times$10)	0.401(0.059)	0.422(0.037)	0.404(0.024)	0.402(0.017)	1.712(0.688)	1.707(1.376)	1.713(3.436)	1.716(6.887)
(D2)	250	β_{01}	0.517(0.112)	0.510(0.122)	0.524(0.126)	0.518(0.122)	0.004(0.058)	0.004(0.061)	0.000(0.061)	0.000(0.060)
		β_{02}	0.489(0.127)	0.482(0.128)	0.498(0.128)	0.507(0.129)	−0.002(0.129)	−0.007(0.114)	0.006(0.132)	−0.015(0.113)
		β_{03}	0.673(0.109)	0.680(0.115)	0.657(0.117)	0.657(0.111)	−0.004(0.069)	0.003(0.072)	−0.003(0.071)	0.001(0.068)
		AEE(MSE$\times$10)	0.600(0.187)	0.490(0.136)	0.448(0.103)	0.422(0.098)	1.719(0.698)	1.709(1.414)	1.712(3.542)	1.704(7.097)
	500	β_{01}	0.505(0.076)	0.507(0.084)	0.515(0.082)	0.528(0.080)	0.003(0.041)	−0.002(0.044)	0.000(0.043)	−0.001(0.041)
		β_{02}	0.497(0.093)	0.499(0.087)	0.500(0.092)	0.501(0.089)	0.000(0.084)	−0.006(0.093)	−0.001(0.077)	0.000(0.084)
		β_{03}	0.690(0.085)	0.687(0.082)	0.680(0.084)	0.671(0.079)	−0.001(0.046)	0.001(0.050)	0.003(0.052)	0.001(0.048)
		AEE(MSE$\times$10)	0.546(0.088)	0.548(0.048)	0.409(0.040)	0.400(0.022)	1.705(0.679)	1.714(1.372)	1.705(3.432)	1.707(6.861)
(D3)	250	β_{01}	0.524(0.114)	0.512(0.119)	0.515(0.118)	0.526(0.121)	0.000(0.065)	0.001(0.063)	0.000(0.061)	0.004(0.063)
		β_{02}	0.500(0.120)	0.497(0.116)	0.488(0.124)	0.500(0.130)	−0.008(0.129)	−0.005(0.127)	−0.010(0.122)	0.000(0.124)
		β_{03}	0.660(0.108)	0.672(0.106)	0.675(0.109)	0.656(0.107)	−0.004(0.070)	0.003(0.075)	0.005(0.070)	−0.001(0.072)
		AEE(MSE$\times$10)	0.436(0.160)	0.447(0.120)	0.455(0.097)	0.436(0.095)	1.719(0.698)	1.709(1.414)	1.712(3.542)	1.704(7.097)
	500	β_{01}	0.529(0.084)	0.508(0.094)	0.515(0.101)	0.520(0.103)	0.000(0.047)	−0.002(0.046)	−0.002(0.045)	0.000(0.043)
		β_{02}	0.501(0.099)	0.504(0.103)	0.494(0.114)	0.507(0.114)	0.003(0.092)	0.005(0.087)	−0.003(0.091)	0.002(0.086)
		β_{03}	0.669(0.061)	0.679(0.084)	0.678(0.091)	0.664(0.095)	0.003(0.051)	−0.001(0.054)	−0.002(0.051)	0.002(0.051)
		AEE(MSE$\times$10)	0.405(0.134)	0.414(0.074)	0.406(0.068)	0.389(0.065)	1.702(0.681)	1.705(1.380)	1.714(3.448)	1.704(6.891)

Tables 2.3 and 2.4 summarize the same results for estimating Model (M2) under all considered distributions (D1)–(D3) when $N \in \{2500, 5000\}$ and when $J \in \{250, 500\}$, respectively. Results for Models (M3) and (M4) are similar. Hence, we include them in the Appendix A.2.

From all the tables, one could see that our estimates of $\boldsymbol{\beta}_0$ are generally on target across all models and exhibit little bias. As N or J increases, both the bias and the sample standard deviation of the estimates of $\boldsymbol{\beta}_0$ decrease, so do the AEE and MSE. These patterns reinforce the consistency of $\hat{\boldsymbol{\beta}}$ and $\hat{\eta}(\cdot)$ shown in Theorem 2.1. Further, the overall measures (AEE and MSE) are seldom affected by (D1)–(D3). By comparing our results with the ones of McMahan et al. (2016), one could see that from Tables 2.1 and 2.2 when the curve is truly linear, both methods yield reasonable estimates of $\boldsymbol{\beta}_0$. The variability of their estimates is smaller than the one of ours. This is expected because our procedure has to estimate the curve η_0 which is given as known to their method. However, when the curve is not linear (Tables 2.3 and 2.4), their estimates suffer from a huge bias while ours are still on target. For example, in Tables 2.3 and 2.4, almost all estimates of $\boldsymbol{\beta}$ from the parametric method are centered around 0. If inferences are made based on these estimates, one would incorrectly conclude that all the covariates are insignificant; i.e., a misspecified curve would greatly compromise statistical inferences. However, such concerns do not exist if using our method. To sum up, all the above observations demonstrate that our estimators are robust to the biomarker distribution and the shape of regression curve.

Now we look at the impact of pool sizes. When the number of individual N is fixed (Tables 2.1 and 2.3), as one might expect, all the standard deviations increase with the pool size c. The loss in estimation efficiency is the price paid for the significant cost reduction realized by pooling. In terms of estimating the entire mean curve $\eta_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)$, one could see that the MSE's only increase a little when c increases. For example, in Table 2.3 for (D1) and $N = 2500$, the MSE changes from 0.0017 to

0.0021 when c_j increases from 1 to 2. Note that, $c = 2$ represents a 50% saving in cost when comparing to $c = 1$. These results suggest that pooling could provide estimates similar to or not much worse than those obtained from individual testing while conferring a significant cost reduction.

Tables 2.2 and 2.4 correspond to the second scenario where the number of assays J is fixed. The results reinforce our findings from Theorem 2.1 which are that the pool size c does not affect the efficiency of estimating $\boldsymbol{\beta}_0$ across different pool sizes when J is fixed. For example, in Table 2.4 for (D2) and $J = 500$, when $c = 1$, the standard deviations of estimates of $(\beta_{01}, \beta_{02}, \beta_{03})$ are $(0.076, 0.094, 0.085)$ which change to $(0.080, 0.089, 0.079)$ when $c = 10$, respectively. As an overall measure, AEE actually decreases from 0.600 (when $c = 1$) to 0.422 (when $c = 10$). Further, the MSE strictly decreases with c. These patterns indicate that measuring biomarkers on pools will provide nearly the same or even more precise estimates on both $\boldsymbol{\beta}_0$ and $\eta_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)$ when compared to individually testing.

Lastly, we consider the case where $V(Y_{ij} \mid \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})$ depends on covariates. We set $V(Y_{ij} \mid \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) = (0.5\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})^2$ and repeated the whole simulation study described above. Because the patterns of these results are similar, we present them in the Appendix A.2. One conclusion is that our method also performs well if $V(Y_{ij} \mid \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})$ changes with covariates.

2.5 Real data analysis

2.5.1 NHANES diabetes data

We first illustrate our proposed methodology by applying it to a diabetes data set obtained from NHANES available at `https://wwwn.cdc.gov/nchs/nhanes/search/nhanes09_10.aspx`. The data consists of a continuous response variable for each individual, Y, which denotes a patient's two-hour plasma glucose level concentration (mg/dL), which has been identified as a viable biomarker for detecting diabetes mel-

litus. In addition, a set of explanatory variables are considered; namely, X_1 gender, X_2 age in month, X_3 the log body mass index (kg/m^2), X_4 systolic blood pressure (mm Hg), X_5 diastolic blood pressure (mm Hg), X_6 the log fasting plasma glucose level (mg/dL), X_7 the log triglycerides level (mg/dL), and X_8 the log HDL-cholesterol level (mg/dL), so that the covariate vector $\boldsymbol{X} = (X_1, \ldots, X_8)^{\mathrm{T}}$ for each individual. This data set contains $N = 2574$ individual observations with 2318 of them having all of the explanatory variables listed above. In this section, we analyze the diabetes data set of $N = 2318$ individuals with full covariates information. It is important to notice that instead of analyzing actual pooled testing data, it is more advantageous to artificially construct pooled responses using individual level data, because it allows us to investigate the effect that pool size and composition (in terms of the covariates) has on parameter estimation.

The first focus of our analysis is to compare our pool response model to the analogous model in which the individual level data is fully observed. To accomplish this, we randomly assigned individuals to pools of size c, where $c \in \{2, \ldots, 10\}$. Note that the sample size $N = 2318$ cannot be divided by some values of c; in such cases, we pool the remainders as the last group (e.g., when $c = 10$, the pool response data consists of 231 pools of size 10 and 1 of size 8). Pooling responses for the pools were then determined according to $Z_j = c^{-1} \sum_{i=1}^{c} Y_{ij}$. We repeated the above procedure 500 times and applied our proposed model to each of the resulting pooled data sets. We standardized the continuous covariates so that they had mean 0 and variance 1, while the discrete binary covariates were recoded to be -0.5 or 0.5, respectively.

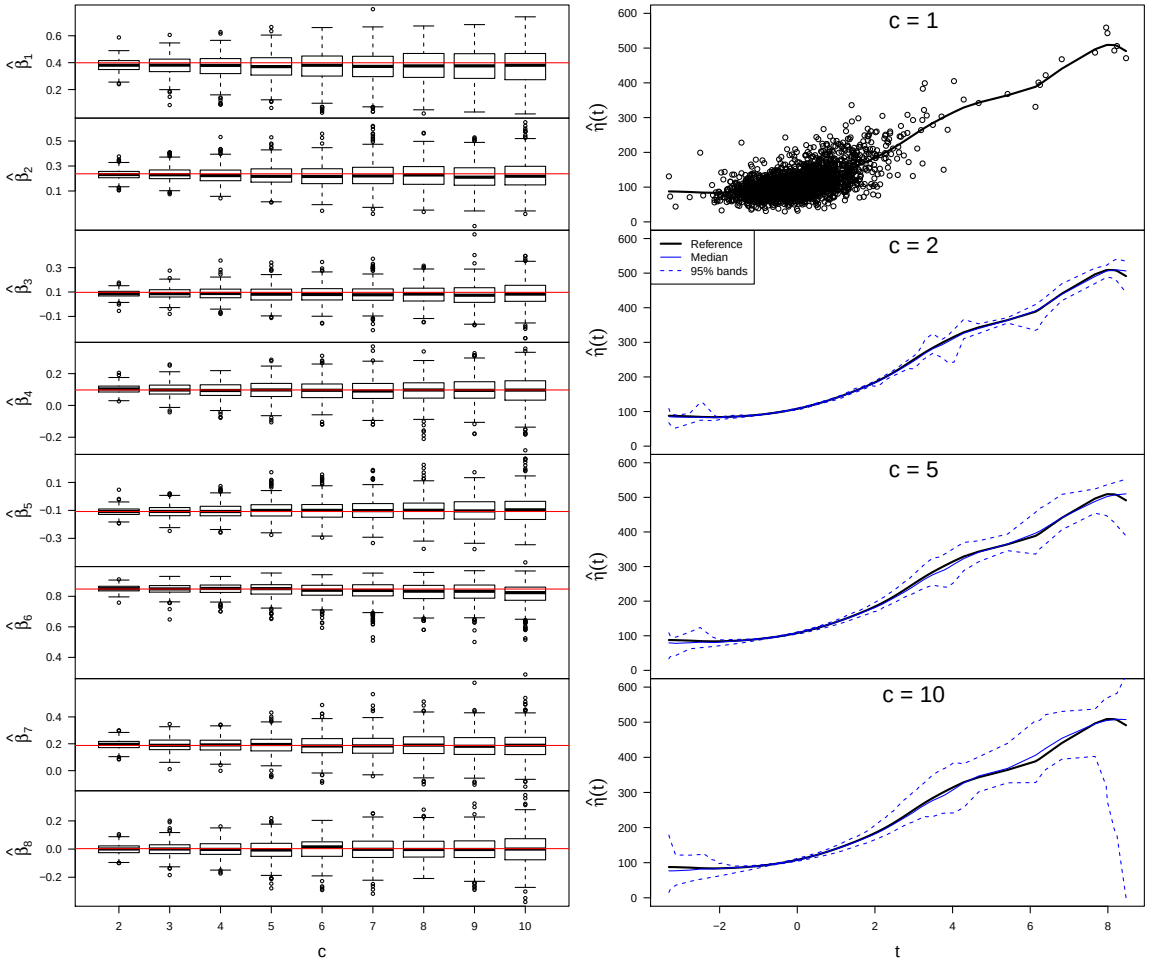

Figure 2.1: Left: Box-plots of the 500 estimates of $\boldsymbol{\beta} = (\beta_{01}, \ldots, \beta_{08})^{\mathrm{T}}$ across $c \in \{2, \ldots, 10\}$. Right: the points in the top figure denote the patient's two-hour plasma glucose level. The remaining three figures depict the estimate curve of $\eta(t)$ and the quantile plots of the estimates of $\eta(t)$ for $c \in \{2, 5, 10\}$. Specifically, at every value of t we plot the 2.5th, 50th, and 97.5th percentiles of the 500 estimates of $\eta(t)$. The solid lines in the figures denote the estimates of $\eta(t)$, when $N = 2318$ and $c = 1$.

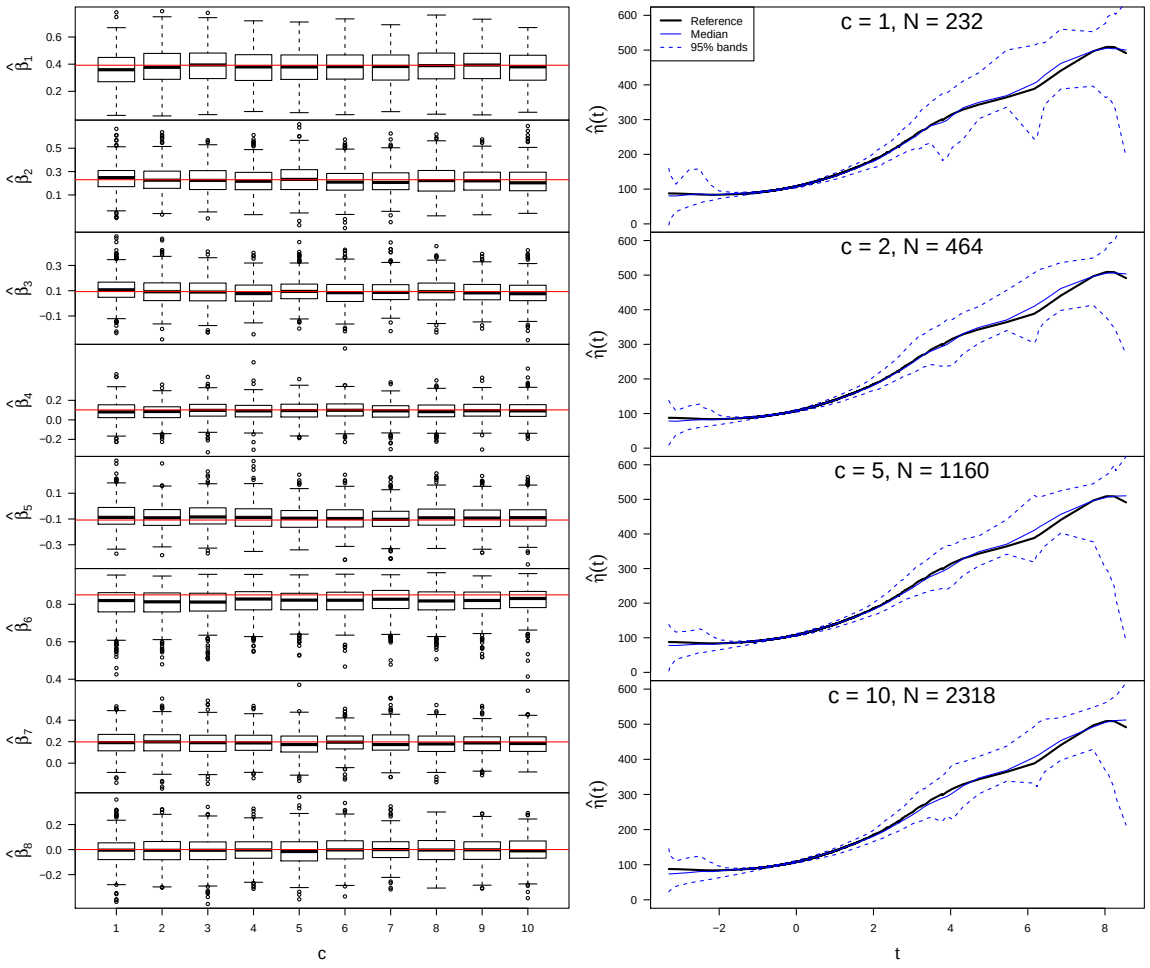

Figure 2.2: Left: Box-plots of the 500 estimates of $\boldsymbol{\beta} = (\beta_{01}, \ldots, \beta_{08})^{\mathrm{T}}$ across $c \in \{1, 2, \ldots, 10\}$ when J is fixed to be 232. Right: the fourth figures depict the estimate curve of $\eta(t)$ and the quantile plots of the estimates of $\eta(t)$ for $c \in \{1, 2, 5, 10\}$ when J is fixed to be 232. Specifically, at every value of t we plot the 2.5th, 50th, and 97.5th percentiles of the 500 estimates of $\eta(t)$. The solid lines in the figures denote the estimates of $\eta(t)$, when $N = 2318$ and $c = 1$.

Figure 2.1 presents box-plots of the 500 estimates of $\boldsymbol{\beta}$ obtained from our method across $c \in \{2, 3, \ldots, 10\}$. Also included in the figure are quantile plots of the estimates of $\eta_0(t)$ for pool sizes of $c = 1, 2, 5, 10$. For purpose of comparison we use the $c = 1$ case as a reference by which our estimates can be compared. Note that the reference estimates suggest a nonlinear shape of $\eta_0(\cdot)$ which supports the use of our single-index model. From these results, it is apparent that the estimates of $\boldsymbol{\beta}_0$ are largely in agreement with the estimates based on the individual-level data. This can also be said for our estimates of $\eta_0(t)$ across all considered pool sizes. We again observe that the variability in our parameter estimates tends to increase with the pool size, which is expected due to the significant cost reduction. Additionally, one will note that our estimates of $\eta_0(t)$ exhibit evidence of instability toward the upper bound of $\boldsymbol{X}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ for larger pool size (e.g., $c = 10$). Again this is an expected phenomenon, since the number of observations that occur in that region is relatively small. The second primary focus is to assess the effect of pooling when the number of assays J is fixed. For this purpose, we set $J = 232$ and consider $c \in \{1, 2, \ldots, 10\}$. The pool response data for each c was constructed by randomly sampling cJ specimens from the 2318 individuals and assigning them to pools of size c. Once the pools have been established, we determine the testing response for the jth pool by $Z_j = c^{-1} \sum_{i=1}^{c} Y_{ij}$. Again, we repeated the procedure 500 times and applied our proposed method to those data sets. Figure 2.2 presents box-plots of the 500 estimates of $\boldsymbol{\beta}_0$'s across $c \in \{1, 2, \ldots, 10\}$ and quantile plots of the estimates of $\eta_0(t)$ for $c = 1, 2, 5, 10$. It can be seen that the estimates of $\boldsymbol{\beta}_0$ and $\eta_0(\cdot)$ generally agree with the reference estimates (obtained when $N = 2318$ and $c = 1$). Further, the box-plots are nearly the same across all pool sizes. The variability of estimates of $\eta_0(t)$ when $c > 2$ is smaller than the one when $c = 1$; e.g., the width of the 95% quantile bands when $c = 10$ is apparently smaller than the one when $c = 1$. These results reinforce the main findings of the second scenario in Section 2.4.

2.5.2 CPP pooled chemokine data

We now illustrate the proposed methodology using a pooled data. This data was collected from the CPP, a study conducted from 1957 to 1974 to assess various aspects of maternal and child health (e.g., see Whitcomb et al., 2007). In 2007, stored serum samples from CPP participants were measured for levels of many chemokines to study whether these biomarkers are related to miscarriage risk. In this chapter, we focus on the biomarker macrophage inhibitory protein (MIP)-1α which was measured in pools of size $c = 2$. We consider only the pools with participants whose full covariate information were available. Considered covariates include age (standardized; x_1), race (1=Afriance-American/0=otherwise; x_2), and miscarriage status (1=yes/0=no; x_3). After removing missing values, the number of pools is $J = 330$. Our goal is to apply our single-index technique to the pooled measurements so that one can estimate the individual-level mean trend of the MIP-1α given the covariate information.

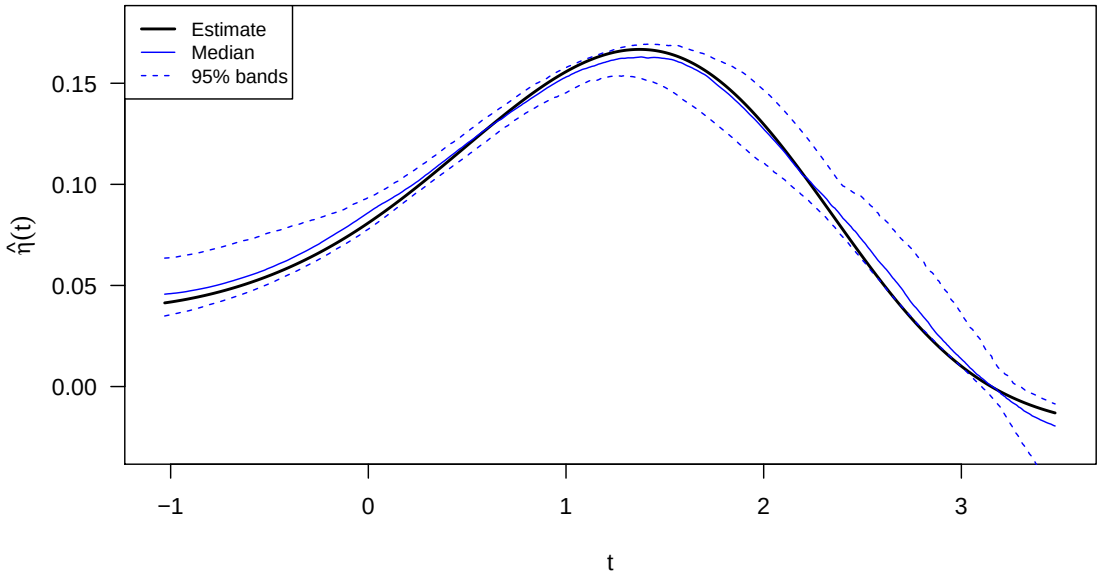

Figure 2.3: CPP pooled MIP-1α data: this figure includes the estimate curve of $\eta(t)$ and the quantile plots of the 500 bootstrap estimates of $\eta(t)$ based on bootstrap samples. Specifically, at every value of t we plot the 2.5th, 50th, and 97.5th percentiles of the 500 bootstrap estimates of $\eta(t)$.

Applying our methodology yields a bandwidth $h = 0.692$ and estimates of the regression coefficients $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)^{\mathrm{T}} = (0.703, 0.700, 0.127)^{\mathrm{T}}$. The estimated mean curve $\hat{\eta}(t)$ is plotted (the black line) in Figure 2.3. In order to obtain valid inference, we adopted a bootstrapping method. A general description of this bootstrapping method is presented in the Appendix A.3, where a simulation studied is also included to illustrate its performance. We bootstrapped the pooled data for 500 times. On each bootstrap sample, we applied our methodology and obtained 500 bootstrap estimates of $(\boldsymbol{\beta}, \eta(\cdot))$. The standard deviation of these bootstrap estimates of $\boldsymbol{\beta}$ can be used to estimate the standard error of our point estimates. The resulting estimated standard errors are $\mathrm{SE}(\hat{\beta}_1) = 0.178$, $\mathrm{SE}(\hat{\beta}_2) = 0.480$, and $\mathrm{SE}(\hat{\beta}_1) = 0.368$, which suggest that at least age has a significant impact on the individual's MIP-1α mean level. Pointwise quantile plots (the 2.5th, 50th, and 97.5th percentiles) of the 500 bootstrap estimates of $\eta(\cdot)$ are also included in Figure 2.3. Clearly, one can see a nonlinear mean relationship between the linear predictor ($t = \boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}$) and the MIP-1$\alpha$ level. This nonlinear relationship further demonstrates the contribution and the flexibility of our proposed single-index methodology.

2.6 Discussion

In spite of the wide and lasting interest in pooling strategy under restrictive parametric assumptions, nonparametric or semiparametric estimation based on continuous pooled biomarker data received relatively less attention. In this chapter, we proposed a general semiparametric framework for modeling pooled biomarker data allowing for the incorporation of individual covariates. Compared to existing works (Ma et al., 2011; Malinovsky et al., 2012; McMahan et al., 2016), our approach does not force the regression function to be linear nor the type of biomarker distribution to be known. We have shown that our estimates enjoy nice asymptotic properties. To illustrate the performance of our methodology, we have considered two scenarios. In the first,

the population size is fixed. Our numerical studies suggest that pooling could reduce the cost substantially with only a minor loss of estimation accuracy. In the second, the number of assays is fixed. We found out that the pooling strategy could be superior providing more information than testing specimens separately. Our estimates performed well under either symmetric or right-skewed biomarker distribution settings.

Because pooling biomarker is now more common in practical applications (see Lyles et al., 2015; Mitchell et al., 2015; Perrier et al., 2016), we believe it would be very beneficial to develop more statistical methods that are flexible to model such data. In this work, we assumed that pools are constructed by randomly mixing individual specimens. One interesting future work is to consider the situation where pooling is performed within stratification of population on the basis of some demographic variables, such as age or gender, which might potentially improve the estimation performance. Caudill (2010) and Mitchell et al. (2014) adopted such grouping criteria to characterize population and analyze biomarker data, respectively. Another interesting extension of our work is to incorporate pooled exposures as a part of the covariate information, which is a more complex problem and received many attention recently (Linton and Whang, 2002; Whitcomb et al., 2012; Delaigle and Zhou, 2015).

Chapter 3

Regression analysis and variable selection for two-stage multiple-infection group testing data

Summary: Group testing, as a cost-effective strategy, has been widely used to perform large-scale screening for rare infections. Recently, the use of multiplex assays has transformed the goal of group testing from detecting a single disease to diagnosing multiple infections simultaneously. Existing research on multiple-infection group testing data either exclude individual covariate information or ignore possible retests on suspicious individuals. To incorporate both, we propose a new regression model. This new model allows us to perform regression analysis for each infection using multiple-infection group testing data. Furthermore, we introduce an efficient variable selection method to reveal truly relevant risk factors for each disease. Our methodology also allows for the estimation of the assay sensitivity and specificity when they are unknown. We examine the finite sample performance of our method through extensive simulation studies and apply it to a chlamydia and gonorrhea screening data set to illustrate its practical usefulness.

3.1 Motivation

This article is motivated by the annual chlamydia trachomatis (CT) and neisseria gonorrhoeae (NG) screening practice conducted by the State Hygienic Laboratory (SHL) in Iowa. The CT and NG are two of the most common notifiable STDs in the United States. Over two million cases were reported to the CDC in 2016 (Dis-

ease Control and Prevention, Last accessed 2018(a)). Both infections are commonly asymptomatic in women. If left untreated, they could cause pelvic inflammatory disease and further lead to tubal infertility, ectopic pregnancy, or chronic pelvic pain (Lewis et al., 2012). In addition, both diseases could facilitate the transmission of HIV and human papillomavirus infection (Samoff et al., 2005). Concerned by these severe sequelae, CDC continually supports nationwide CT/NG screening and recommends annual CT/NG screening for all sexually active women under 25 years old (Disease Control and Prevention, Last accessed 2018(b)).

In this nationwide screening practice, specimens (swab or urine) are collected across each state and shipped to major state laboratories to be tested. Due to different budgets, laboratories conduct the screening differently. For example, the Nebraska Public Health Laboratory (NPHL) uses a traditional individual testing protocol which tests individual specimens one-by-one. The SHL tests male specimens and female urine specimens individually, but tests female swab specimens according to a two-stage pooling protocol:

The SHL Pooling Protocol

- Individual swab specimens are randomly assigned to non-overlapping groups of size four. A pool is constructed by mixing individual specimens in the same group.
- Stage 1: Each pool is tested for CT and NG simultaneously using a multiplex assay. If a pool tests negative for both infections, all the involved individuals are diagnosed as negative for each infection with no additional tests; otherwise, the protocol proceeds to the next stage.
- Stage 2: Swabs of individuals in pools that test positive for either infection are retested separately using the same multiplex assay for final diagnosis.

The most practical reason for using pooling is cost reduction. When a pool tests negative for both infections, four individuals are diagnosed at the expense of one

assay. Since switching from individual testing to pooling in 1999, Iowa has saved over $2.2 million in the CT/NG screening (Jirsa, 2008).

As per the screening guidelines, many risk factors are collected as well, such as age, number of partners, any symptoms of the infections, etc. A motivating question is how to incorporate these covariate information so that one can identify truly relevant risk factors for each infection and understand their effects. Challenges to this question arise from the use of the multiplex Aptima Combo 2 Assay (Gen-Prob, San Diego), an imperfect discriminatory test that produces diagnoses for both diseases simultaneously. Due to the imperfectness of the assay, it is possible to observe some discrepancies between testing outcomes of the two stages, as shown in Figure 3.1. Whenever a discrepancy occurs, the SHL ignores pooled-level results from Stage 1 and makes the diagnosis solely based on individual testing from Stage 2. However, when the objective is probing the impact of risk factors rather than case identification, disregarding testing outcomes from any stage could impair the estimation. It is important to seamlessly incorporate outcomes from both stages. Towards this goal, we need to account for how likely the retests were triggered by either infection.

Most of existing literature in modeling multiple-infection pooled testing data (see Hughes-Oliver and Rosenberger, 2000, Tebbs et al., 2013, Zhang et al., 2013, Warasi et al., 2016, Li et al., 2017) assumed that there were some preliminary studies to provide those misclassification parameters. However, this assumption could be impractical because the preliminary study might have used unrepresentative samples (Huang et al., 2017). If inaccurate values of assay sensitivity and specificity were used for estimation, it could compromise inference. In this chapter, we keep the testing error rates as unknown and estimate them from the data along with the regression.

Existing literature has not considered the combination of incorporating retesting results into regression and estimating misclassification parameters in the context of multiple-infection group testing. Only one Bayesian work has provided inference for

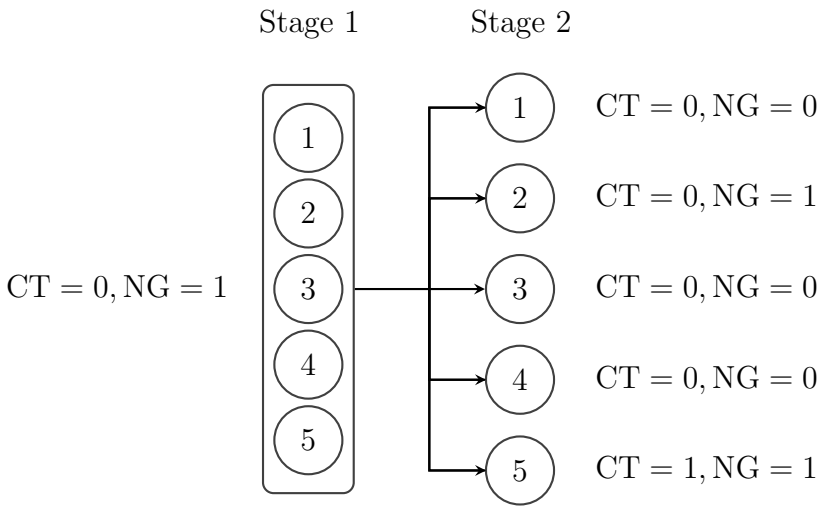

Figure 3.1: An example of the SHL pooling data when the group size is 5: the rectangle with rounded corners represents the pooled specimen that is constructed by mixing 5 individual specimens (in circles) together. Though the pool tests negative for CT (i.e., CT= 0), positivity is shown for NG (i.e., NG= 1). As per the SHL pooling protocol, this NG positivity triggers the second stage of screening for both CT and NG.

disease prevalence and estimates of assay sensitivity and specificity without consideration of individual covariates (Warasi et al., 2016). In this chapter, we propose a copula-based multivariate binary regression model to incorporate the covariates. We introduce a generalized expectation-maximization (GEM) algorithm to facilitate the numerical computation of the maximum likelihood estimates (MLEs) of the regression coefficients and misclassification parameters. When compared to the traditional EM algorithm, the GEM only requires the maximization step to search for an increase in the objective function rather than achieving the maximum (Wu, 1983, Neal and Hinton, 1998). This feature greatly accelerates the computation of the MLE.

In addition, we provide a variable selection technique that can identify truly relevant risk factors for each infection. A recent work has introduced a regularized regression technique for group testing (Gregory et al., in print). But it is for a single infection. Our work is designed to allow for multiple infections. We believe a package of regression, estimation of misclassification parameters, and variable selection can provide a useful toolbox for the epidemiology study of CT and NG based on group

testing data.

The rest of the chapter is organized as follows. In Section 3.2, we propose a new copula-based regression model for multiple-infection group testing data. In Section 3.3, we introduce the GEM algorithm that accelerates the computation of the MLE. Section 3.4 presents a variable selection method that can identify important risk factors for each infection. In Section 3.5, we use simulation to illustrate that, with the use of a fewer number of tests, the SHL pooling protocol can lead to more efficient regression estimates, better prediction of infection probabilities, and more accurate variable selection than traditional individual testing. These advantages are further demonstrated by analyzing a CT/NG screening data set in Section 3.6. We conclude the chapter with a discussion in Section 3.7. All technical details and additional numerical results are relegated to Appendix B.

3.2 Model

Suppose N individuals are to be tested. We randomly assign each individual to one of J groups, each of size c_j; i.e., $N = \sum_{j=1}^{J} c_j$. For generality, we allow group size c_j to vary across groups. Motivated by the CT/NG screening practice, we mainly consider two infections. Section 3.7 discusses an extension of more than two diseases. The true infection statuses of the ith individual in the jth group are denoted by a binary vector $\widetilde{\boldsymbol{Y}}_{ij} = (\widetilde{Y}_{ij1}, \widetilde{Y}_{ij2})^{\mathrm{T}}$, where $\widetilde{Y}_{ijk} = 1$ if the individual is positive for the kth infection, $\widetilde{Y}_{ijk} = 0$ otherwise, for $i = 1, \ldots, c_j$, $j = 1, \ldots, J$, and $k = 1, 2$. Denote the covariates (risk factors and an intercept term) of the ith individual in the jth group by a $(p+1)$-dimensional vector $\mathbf{x}_{ij} = (1, x_{ij1}, \ldots, x_{ijp})^{\mathrm{T}}$. We assume that $\widetilde{\boldsymbol{Y}}_{ij}|\mathbf{x}_{ij}$'s are independent across ij and $\widetilde{Y}_{ijk}$ is related to a linear predictor $\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_k$ via

$$\mathrm{pr}(\widetilde{Y}_{ijk} = 1|\mathbf{x}_{ij}) = g_k(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_k), \text{ for } k = 1, 2, \tag{3.1}$$

where $\boldsymbol{\beta}_k = (\beta_{k0}, \beta_{k1}, \ldots, \beta_{kp})^{\mathrm{T}}$ is a vector of $(p+1)$ regression coefficients that will be estimated and g_k is a user-chosen known link function (e.g., the inverse of the logit or probit link). One could use different links for different infections. Equation (3.1) builds marginal probability models of the random vector $\widetilde{\boldsymbol{Y}}_{ij}|\mathbf{x}_{ij}$.

In pooled testing, the true infection statuses are often latent due to pooling and potential misclassification. In each group, individual specimens are mixed together to form a pool. We denote the true status of the jth pool by $\tilde{\mathbf{Z}}_j = (\widetilde{Z}_{j1}, \widetilde{Z}_{j2})^{\mathrm{T}}$ where $\widetilde{Z}_{jk} = \max\{\widetilde{Y}_{ijk} : i = 1, \ldots, c_j\}$; i.e., $\widetilde{Z}_{jk} = 1$ if the pool involves at least one individual who is positive for the kth infection, $\widetilde{Z}_{jk} = 0$ otherwise. With the use of an imperfect assay, both $\widetilde{\boldsymbol{Y}}_{ij}$'s and $\tilde{\mathbf{Z}}_j$'s are latent. Observed data are the testing outcomes from the imperfect multiplex assay. Pools are tested in Stage 1. We denote the testing outcomes of the jth pool by $\boldsymbol{Z}_j = (Z_{j1}, Z_{j2})^{\mathrm{T}}$, where $Z_{jk} = 1(0)$ if the pool tests positive (negative) for the kth infection. If $\boldsymbol{Z}_j = (0,0)^{\mathrm{T}}$, then $\boldsymbol{Z}_j$ is the only observed test response for the jth group of individuals. Otherwise, those individuals are tested separately in Stage 2. We denote by $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2})^{\mathrm{T}}$ the retesting outcome of the ith individual in the jth group; i.e., $Y_{ijk} = 1(0)$ if the individual is retested as positive (negative) for the kth infection. Note that, the $\boldsymbol{Y}_{ij}$'s can only be observed if $\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}}$. In summary, observed testing outcomes from the jth group, denoted by $\boldsymbol{\mathcal{P}}_j$, take one of the two forms, either $\boldsymbol{Z}_j = (0,0)^{\mathrm{T}}$, or $\boldsymbol{Z}_j \in \{(1,0)^{\mathrm{T}}, (0,1)^{\mathrm{T}}, (1,1)^{\mathrm{T}}\}$ and $\boldsymbol{Y}_{1j}, \ldots, \boldsymbol{Y}_{c_j j}$.

The discrepancy between true statuses and testing outcomes is often measured by assay sensitivity and specificity. Denote by $S_{e:k}$ and $S_{p:k}$ the assay sensitivity and specificity, respectively, for the kth infection. In practice, an assay used for large-scale screening is often imperfect. We let $S_{e:k}$'s and $S_{p:k}$'s be in $(0,1)$. Our methodology posits three assumptions on these misclassification parameters. Assumption 1 is that $S_{e:k}$'s and $S_{p:k}$'s do not depend on the group size; e.g., $S_{e:k} = \mathrm{pr}(Z_{jk} = 1|\widetilde{Z}_{jk} = 1) = \mathrm{pr}(Y_{ijk} = 1|\widetilde{Y}_{ijk} = 1)$ and $S_{p:k} = \mathrm{pr}(Z_{jk} = 0|\widetilde{Z}_{jk} = 0) = \mathrm{pr}(Y_{ijk} = 0|\widetilde{Y}_{ijk} = 0)$ hold for

all i, j, and k. Assumption 2 assumes that conditioning on the true statuses of the specimens being tested, testing responses are independent across each other and also across infections. Assumption 3 further assumes that given the true statuses, testing responses are independent of the covariates; e.g., $\text{pr}(Z_{j1}=0, Z_{j2}=1, Y_{ij1}=1, Y_{ij2}=0|\widetilde{Z}_{j1}=0, \widetilde{Z}_{j1}=0, \widetilde{Y}_{ij1}=1, \widetilde{Y}_{ij2}=1, \boldsymbol{x}_{ij}) = \text{pr}(Z_{j1}=0|\widetilde{Z}_{j1}=0)\text{pr}(Z_{j2}=1|\widetilde{Z}_{j2}=0)\text{pr}(Y_{ij1}=1|\widetilde{Y}_{ij1}=1)\text{pr}(Y_{ij2}=0|\widetilde{Y}_{ij2}=1) = S_{p:1}(1-S_{p:2})S_{e:1}(1-S_{e:2})$. All these assumptions are standard in group testing literature (see most references in Section 1.2). In practice, one may need to conduct proper assay calibration to ensure the applicability of these assumptions.

Our primary goal is to estimate $\boldsymbol{\beta}_k$'s, $S_{e:k}$'s and $S_{p:k}$'s. Towards this goal, we want to incorporate the retesting outcomes for two main reasons: 1) ignoring the retesting outcomes could severely inflate the variance of the estimators of β_k's (see Appendix B.1 for a numerical illustration). 2) Including the retesting outcomes gives us repeated measurements (i.e., many specimens are tested in pools and also individually) which provide valuable information to estimate misclassification parameters. To seamlessly incorporate all retesting outcomes, we propose a copula-based multivariate binary regression model. We assume that there exists a vector of standard uniform random variables, $\boldsymbol{U}_{ij} = (U_{ij1}, U_{ij2})^{\text{T}}$, such that the event $\{\widetilde{Y}_{ijk}=1|\mathbf{x}_{ij}\}$ is equivalent to the event $\{U_{ijk} \leq g_k(\mathbf{x}_{ij}^{\text{T}}\boldsymbol{\beta}_k)\}$, where $\boldsymbol{U}_{ij}$'s are independent and follow a bivariate copula (Nelsen, 2007). Denote the chosen copula by $\mathcal{C}\{u_1, u_2|\delta\}$, where $u_1, u_2 \in (0,1)$ and $\mathcal{C}$ is known up to a parameter δ (which could be a vector). Then the marginal regression models in (3.1) naturally hold, and the co-infection probability is

$$\text{pr}(\widetilde{Y}_{ij1}=1, \widetilde{Y}_{ij2}=1|\mathbf{x}_{ij}) = \mathcal{C}\{g_1(\mathbf{x}_{ij}^{\text{T}}\boldsymbol{\beta}_1), g_2(\mathbf{x}_{ij}^{\text{T}}\boldsymbol{\beta}_2)|\delta\}. \tag{3.2}$$

Combining (3.1) and (3.2) together defines our joint probability model of $\widetilde{\boldsymbol{Y}}_{ij}|\mathbf{x}_{ij}$.

3.3 Estimation

We maximize the likelihood function to obtain our estimators of $\boldsymbol{\beta}_k$'s, $S_{e:k}$'s, $S_{p:k}$'s, and δ. For notation simplicity, we write $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1^{\mathrm{T}}, \boldsymbol{\beta}_2^{\mathrm{T}}, \delta)^{\mathrm{T}}$, $\boldsymbol{\theta}_2 = (S_{e:1}, S_{e:2}, S_{p:1}, S_{p:2})^{\mathrm{T}}$, and $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\mathrm{T}}, \boldsymbol{\theta}_2^{\mathrm{T}})^{\mathrm{T}}$. Furthermore, we denote by $p_{ijy_1y_2}(\boldsymbol{\theta}_1)$ the cell probability $\mathrm{pr}(\widetilde{Y}_{ij1} = y_1, \widetilde{Y}_{ij2} = y_2|\mathbf{x}_{ij})$ defined by (3.1) and (3.2) under $\boldsymbol{\theta}_1$ for $y_1, y_2 \in \{0, 1\}$, $i = 1, \ldots, c_j$, and $j = 1, \ldots, J$. Then $p_{ij11}(\boldsymbol{\theta}_1) = \mathcal{C}\{g_1(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1), g_2(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_2)|\delta\}$, $p_{ij10}(\boldsymbol{\theta}_1) = g_1(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1) - p_{ij11}(\boldsymbol{\theta}_1)$, $p_{ij01}(\boldsymbol{\theta}_1) = g_2(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_2) - p_{ij11}(\boldsymbol{\theta}_1)$, and $p_{ij00}(\boldsymbol{\theta}_1) = 1 - p_{ij11}(\boldsymbol{\theta}_1) - p_{ij10}(\boldsymbol{\theta}_1) - p_{ij01}(\boldsymbol{\theta}_1)$. In Appendix B.2, we derive an expression of the log-likelihood function $\ell(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \mathbf{X})$ where $\boldsymbol{\mathcal{P}}$ and $\mathbf{X}$ denote the collections of $\boldsymbol{\mathcal{P}}_j$'s and $\mathbf{x}_{ij}$'s, respectively. However, due to the complexity of $\ell(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \mathbf{X})$, a direct maximization could be time-consuming. Appendix B.3 includes a numerical illustration of this disadvantage.

We propose a GEM algorithm to accelerate the computation. The algorithm incorporates $\widetilde{\boldsymbol{Y}} = \{\widetilde{\boldsymbol{Y}}_{11}, \ldots, \widetilde{\boldsymbol{Y}}_{c_JJ}\}$ as latent variables. The complete log-likelihood function of $\boldsymbol{\theta}$, derived from the conditional distribution of $\boldsymbol{\mathcal{P}}$ and $\widetilde{\boldsymbol{Y}}$ given $\mathbf{X}$, can be written by $\ell_c(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}}, \mathbf{X}) = \ell_{c1}(\boldsymbol{\theta}_1|\widetilde{\boldsymbol{Y}}, \mathbf{X}) + \ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}})$, where

$$\ell_{c1}(\boldsymbol{\theta}_1|\widetilde{\boldsymbol{Y}}, \mathbf{X}) = \sum_{j=1}^{J}\sum_{i=1}^{c_j}\Big[(1-\widetilde{Y}_{ij1})(1-\widetilde{Y}_{ij2})\log p_{ij00}(\boldsymbol{\theta}_1) + \widetilde{Y}_{ij1}(1-\widetilde{Y}_{ij2})\log p_{ij10}(\boldsymbol{\theta}_1) + (1-\widetilde{Y}_{ij1})\widetilde{Y}_{ij2}\log p_{ij01}(\boldsymbol{\theta}_1) + \widetilde{Y}_{ij1}\widetilde{Y}_{ij2}\log p_{ij11}(\boldsymbol{\theta}_1)\Big] \tag{3.3}$$

and

$$\begin{aligned}\ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}}) = \sum_{j=1}^{J}\sum_{k=1}^{2}\Bigg[&\left\{\widetilde{Z}_{jk}Z_{jk} + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}})\sum_{i=1}^{c_j}\widetilde{Y}_{ijk}Y_{ijk}\right\}\log S_{e:k} \\ &+ \left\{\widetilde{Z}_{jk}(1-Z_{jk}) + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}})\sum_{i=1}^{c_j}\widetilde{Y}_{ijk}(1-Y_{ijk})\right\}\log(1-S_{e:k}) \\ &+ \left\{(1-\widetilde{Z}_{jk})(1-Z_{jk}) + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}})\sum_{i=1}^{c_j}(1-\widetilde{Y}_{ijk})(1-Y_{ijk})\right\}\log S_{p:k} \\ &+ \left\{(1-\widetilde{Z}_{jk})Z_{jk} + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}})\sum_{i=1}^{c_j}(1-\widetilde{Y}_{ijk})Y_{ijk}\right\}\log(1-S_{p:k})\Bigg],\end{aligned} \tag{3.4}$$

in which $\widetilde{Z}_{jk} = \max\{\widetilde{Y}_{ijk} : i = 1, \ldots, c_j\}$ and $I(\cdot)$ is the indicator function.

Our GEM algorithm starts at an initial value, and then iterates between an E-step and an M-step to update the value until reaching a numerical convergence. At a current value $\boldsymbol{\theta}^{(d)}$, the E-step calculates $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(d)}) = \mathcal{Q}_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}^{(d)}) + \mathcal{Q}_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}^{(d)})$, where

$$\mathcal{Q}_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}^{(d)}) = E\{\ell_{c1}(\boldsymbol{\theta}_1|\widetilde{\boldsymbol{Y}}, \mathbf{X})|\boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta}^{(d)}\}$$

and

$$\mathcal{Q}_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}^{(d)}) = E\{\ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}})|\boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta}^{(d)}\}.$$

After an inspection of (3.3) and (3.4), it suffices to calculate $\eta^{(d)}_{ij00}$, $\eta^{(d)}_{ij10}$, $\eta^{(d)}_{ij01}$, $\eta^{(d)}_{ij11}$ (for $\mathcal{Q}_1$) and $\eta^{(d)}_{\mathcal{P},jk}$ (for $\mathcal{Q}_2$), where

$$\eta^{(d)}_{ijy_1y_2} = \text{pr}(\widetilde{Y}_{ij1} = y_1, \widetilde{Y}_{ij2} = y_2|\boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta}^{(d)}) \text{ and } \eta^{(d)}_{\mathcal{P},jk} = \text{pr}(\widetilde{Z}_{jk} = 1|\boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta}^{(d)}), \tag{3.5}$$

for $i = 1, \ldots, c_j$, $j = 1, \ldots, J$, $y_1, y_2 \in \{0, 1\}$, and $k = 1, 2$. Though $\eta^{(d)}_{ijy_1y_2}$'s have been studied without the consideration of $\mathbf{X}$ (Tebbs et al., 2013), they were not updated in closed forms and thus a Gibbs sampler was employed to approximate these quantities. However, in the regression context, using such approximations requires enlarging the tolerance of the numerical convergence and hence might induce bias. To improve the computational accuracy, we calculate all the probabilities in (3.5) exactly (see Appendix B.4 for details).

With the probabilities in (3.5) calculated, we rewrite $\mathcal{Q}_1(\boldsymbol{\theta}_1|\boldsymbol{\theta}^{(d)})$ by

$$\mathcal{Q}_1(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \delta|\boldsymbol{\theta}^{(d)}) = \sum_{j=1}^{J}\sum_{i=1}^{c_j}\sum_{y_1=0}^{1}\sum_{y_2=0}^{1}\eta^{(d)}_{ijy_1y_2}\log p_{ijy_1y_2}(\boldsymbol{\theta}_1),$$

and $\mathcal{Q}_2(\boldsymbol{\theta}_2|\boldsymbol{\theta}^{(d)})$ by

$$\sum_{k=1}^{2}\Big\{W^{(d)}_{1k}\log S_{e:k} + W^{(d)}_{2k}\log(1 - S_{e:k}) + W^{(d)}_{3k}\log S_{p:k} + W^{(d)}_{4k}\log(1 - S_{p:k})\Big\}, \tag{3.6}$$

where

$$
\begin{aligned}
W_{1k}^{(d)} &= \sum_{j=1}^{J} \left\{ \eta_{\mathcal{P},jk}^{(d)} Z_{jk} + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}}) \sum_{i=1}^{c_j} \eta_{ij,k}^{(d)} Y_{ijk} \right\}, \\
W_{2k}^{(d)} &= \sum_{j=1}^{J} \left\{ \eta_{\mathcal{P},jk}^{(d)} (1 - Z_{jk}) + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}}) \sum_{i=1}^{c_j} \eta_{ij,k}^{(d)} (1 - Y_{ijk}) \right\}, \\
W_{3k}^{(d)} &= \sum_{j=1}^{J} \left\{ (1 - \eta_{\mathcal{P},jk}^{(d)})(1 - Z_{jk}) + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}}) \sum_{i=1}^{c_j} (1 - \eta_{ij,k}^{(d)})(1 - Y_{ijk}) \right\}, \\
W_{4k}^{(d)} &= \sum_{j=1}^{J} \left\{ (1 - \eta_{\mathcal{P},jk}^{(d)}) Z_{jk} + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}}) \sum_{i=1}^{c_j} (1 - \eta_{ij,k}^{(d)}) Y_{ijk} \right\},
\end{aligned}
$$

in which, $\eta_{ij,1}^{(d)} = \eta_{ij11}^{(d)} + \eta_{ij10}^{(d)}$ and $\eta_{ij,2}^{(d)} = \eta_{ij11}^{(d)} + \eta_{ij01}^{(d)}$. The M-step in our GEM algorithm updates $\boldsymbol{\theta}_1^{(d)}$ by $\boldsymbol{\theta}_1^{(d+1)} = (\boldsymbol{\beta}_1^{(d+1)^{\mathrm{T}}}, \boldsymbol{\beta}_2^{(d+1)^{\mathrm{T}}}, \delta^{(d+1)})^{\mathrm{T}}$ where

$$
\begin{aligned}
\boldsymbol{\beta}_1^{(d+1)} &= \mathrm{argmax}_{\boldsymbol{\beta}_1} \ \mathcal{Q}_1(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2^{(d)}, \delta^{(d)} | \boldsymbol{\theta}^{(d)}), \\
\boldsymbol{\beta}_2^{(d+1)} &= \mathrm{argmax}_{\boldsymbol{\beta}_2} \ \mathcal{Q}_1(\boldsymbol{\beta}_1^{(d+1)}, \boldsymbol{\beta}_2, \delta^{(d)} | \boldsymbol{\theta}^{(d)}), \\
\delta^{(d+1)} &= \mathrm{argmax}_{\delta} \ \mathcal{Q}_1(\boldsymbol{\beta}_1^{(d+1)}, \boldsymbol{\beta}_2^{(d+1)}, \delta | \boldsymbol{\theta}^{(d)}).
\end{aligned}
$$

The value of $\boldsymbol{\theta}_2^{(d+1)}$ is obtained by maximizing (3.6) and can be written as

$$
\boldsymbol{\theta}_2^{(d+1)} = (S_{e:1}^{(d+1)}, S_{e:2}^{(d+1)}, S_{p:1}^{(d+1)}, S_{p:2}^{(d+1)})^{\mathrm{T}}
$$

where $S_{e:k}^{(d+1)} = W_{1k}^{(d)} / (W_{1k}^{(d)} + W_{2k}^{(d)})$ and $S_{p:k}^{(d+1)} = W_{3k}^{(d)} / (W_{3k}^{(d)} + W_{4k}^{(d)})$, for $k = 1, 2$. Combining $\boldsymbol{\theta}_1^{(d+1)}$ and $\boldsymbol{\theta}_2^{(d+1)}$ provides $\boldsymbol{\theta}^{(d+1)}$. Because $\mathcal{Q}(\boldsymbol{\theta}^{(d+1)} | \boldsymbol{\theta}^{(d)}) \geq \mathcal{Q}(\boldsymbol{\theta}^{(d)} | \boldsymbol{\theta}^{(d)})$, the convergence of $\{\boldsymbol{\theta}^{(d)}\}_{d=1}^{\infty}$ is guaranteed (Wu, 1983). We denote by $\hat{\boldsymbol{\theta}}$ the limit of $\boldsymbol{\theta}^{(d)}$'s.

Denote by $\mathcal{I}(\boldsymbol{\theta})$ the observed data information matrix. Following the standard arguments of the MLE (Lehmann, 1983), we have $\mathcal{I}(\hat{\boldsymbol{\theta}})^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ converges in distribution to $\mathcal{N}(0, \boldsymbol{I}_{2p+7})$ as $N \to \infty$, where $\boldsymbol{I}_m$ denotes the m-dimensional identity matrix. Applying Louis' method (Louis, 1982) provides

$$
\mathcal{I}(\boldsymbol{\theta}) = -E \left\{ \frac{\partial^2 \ell_c(\boldsymbol{\theta} | \boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}}, \mathbf{X})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\mathrm{T}}} \Big| \boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta} \right\} - \mathrm{cov} \left\{ \frac{\partial \ell_c(\boldsymbol{\theta} | \boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}}, \mathbf{X})}{\partial \boldsymbol{\theta}} \Big| \boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta} \right\}.
$$

Again, instead of approximating $\mathcal{I}(\boldsymbol{\theta})$ via the Gibbs sampling approach (Tebbs et al., 2013), we are able to calculate it exactly. The calculations are included in Appendix B.5. With $\mathcal{I}(\hat{\boldsymbol{\theta}})$, one can make large sample Wald-type inferences. For example, let θ_l, $\hat{\theta}_l$ and $\hat{\sigma}^2_{ll}$ be the lth component of $\boldsymbol{\theta}$, the lth component of $\hat{\boldsymbol{\theta}}$ and the lth diagonal entry of $\mathcal{I}(\hat{\boldsymbol{\theta}})^{-1}$, respectively, for $l = 1, \ldots, 2p+7$. The estimated standard error (SE) of $\hat{\theta}_l$ is $\hat{\sigma}_{ll}$ and an approximated $100(1-\alpha)\%$ confidence interval of θ_l is $\hat{\theta}_l \pm z_{\alpha/2}\hat{\sigma}_{ll}$, where z_α is the αth upper quantile of $\mathcal{N}(0,1)$.

3.4 Variable selection for each infection

With $\hat{\boldsymbol{\theta}}$ and $\mathcal{I}(\hat{\boldsymbol{\theta}})$ computed, we further identify which risk factors are truly relevant for each infection. Denote by $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ the values of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ that generate the true individual statuses $\widetilde{\boldsymbol{Y}}$, respectively, where $\boldsymbol{\beta}_k^* = (\beta_{k0}^*, \beta_{k1}^*, \ldots, \beta_{kp}^*)^{\mathrm{T}}$ for $k = 1, 2$. One can index the significant risk factors to the kth infection by $\mathcal{M}_k = \{j \in \mathcal{M} : \beta_{kj}^* \neq 0\}$, where we take $\mathcal{M} = \{1, 2, \ldots, p\}$ by defaulting that an intercept term is always included in the model. One must note that $\mathcal{M}_1$ and $\mathcal{M}_2$ might be different.

We apply a shrinkage method to simultaneously select $\mathcal{M}_k$'s and estimate nonzero β_{kj}^*'s. To unify notation, we write $\boldsymbol{\theta}_{\mathcal{T}}$ and $\hat{\boldsymbol{\Sigma}}_{\mathcal{T}\mathcal{T}}$ as the sub-vector and the sub-matrix of $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\Sigma}}$ according to an index set $\mathcal{T} \subset \{1, \ldots, 2p+7\}$, respectively. Let $\mathcal{A} = \{2, \ldots, p+1, p+3, \ldots, 2p+2\}$. Our shrinkage estimator of $\boldsymbol{\theta}_{\mathcal{A}}$ is defined by

$$\widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda} = \underset{\boldsymbol{\theta}_{\mathcal{A}}}{\operatorname{argmin}} \left\{ \frac{1}{2}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{\mathcal{A}})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}} (\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \boldsymbol{\theta}_{\mathcal{A}}) + \sum_{k=1}^{2} \lambda_k \sum_{j=1}^{p} \omega_{kj} |\beta_{kj}| \right\}, \tag{3.7}$$

where $\lambda_k \sum_{j=1}^{p} \omega_{kj}|\beta_{kj}|$ is an adaptive LASSO penalty (Zou, 2006), $\lambda_k \geq 0$ is a tuning parameter that controls the shrinkage level, and $\omega_{kj} = |\hat{\beta}_{kj}|^{-1}$ is an adaptive weight. When λ_k's are 0, $\widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda} = \hat{\boldsymbol{\theta}}_{\mathcal{A}}$. When λ_k's increase, due to the singularity of the absolute value function at the origin, components of $\widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda}$ are penalized to zero one-by-one. Writing $\widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda} = (\widetilde{\beta}_{11,\lambda}, \ldots, \widetilde{\beta}_{1p,\lambda}, \widetilde{\beta}_{21,\lambda}, \ldots, \widetilde{\beta}_{2p,\lambda})^{\mathrm{T}}$, we estimate $\mathcal{M}_1$ and $\mathcal{M}_2$ by $\widetilde{\mathcal{M}}_{1,\lambda} = \{j \in \mathcal{M} : \widetilde{\beta}_{1j,\lambda} \neq 0\}$ and $\widetilde{\mathcal{M}}_{2,\lambda} = \{j \in \mathcal{M} : \widetilde{\beta}_{2j,\lambda} \neq 0\}$, respectively.

Computing $\widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda}$ is fast. The objective function in (3.7) is simply a summation of a quadratic function and a weighted l_1-norm of $\boldsymbol{\theta}_{\mathcal{A}}$ and therefore can be quickly minimized by slightly modifying the seminal least angle regression (Efron et al., 2004). Let $\mathcal{A}^c = \{1, 2, \ldots, 2p+7\} \setminus \mathcal{A}$ and $\ell(\boldsymbol{\theta}_{\mathcal{A}}|\boldsymbol{\mathcal{P}}, \mathbf{X}, \hat{\boldsymbol{\theta}}_{\mathcal{A}^c})$ be the log-likelihood function $\ell(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \mathbf{X})$ with $\boldsymbol{\theta}_{\mathcal{A}^c}$ fixed to be $\hat{\boldsymbol{\theta}}_{\mathcal{A}^c}$. One could also construct a shrinkage estimator by the traditional penalized MLE (Fan and Li, 2001) which minimizes $-\ell(\boldsymbol{\theta}_{\mathcal{A}}|\boldsymbol{\mathcal{P}}, \mathbf{X}, \hat{\boldsymbol{\theta}}_{\mathcal{A}^c}) + \sum_{k=1}^{2} \lambda_k \sum_{j=1}^{p} \omega_{kj}|\beta_{kj}|$. As the quadratic term in (3.7) is the leading component of the Taylor's expansion of $-\ell(\boldsymbol{\theta}_{\mathcal{A}}|\boldsymbol{\mathcal{P}}, \mathbf{X}, \hat{\boldsymbol{\theta}}_{\mathcal{A}^c})$ at $\boldsymbol{\theta}_{\mathcal{A}} = \hat{\boldsymbol{\theta}}_{\mathcal{A}}$, it can be easily shown that $\widetilde{\boldsymbol{\theta}}_{\mathcal{A}}$ and the penalized MLE are asymptotically equivalent. However, the computation cost of obtaining penalized MLE will be a lot higher due to the complexity of the log-likelihood function.

The use of adaptive weights ω_{kj}'s is critical to achieve the *oracle* properties (Zou, 2006). It assigns sufficiently large penalties to insignificant covariates so that they would be excluded from the model; on the other hand, it imposes mild penalties to significant ones in order that they would be retained in the model. The oracle properties are stated as follows. As $N \to \infty$, if $\max(\lambda_1, \lambda_2)/\sqrt{N} \to 0$ and $\min(\lambda_1, \lambda_2) \to \infty$, we have both the selection consistency, $\mathrm{pr}(\widetilde{\mathcal{M}}_{1,\lambda} = \mathcal{M}_1, \widetilde{\mathcal{M}}_{2,\lambda} = \mathcal{M}_2) \to 1$, and the estimation consistency, $\sup_{k,j} \ \|\widetilde{\beta}_{kj,\lambda} - \beta^*_{kj}\| = O_p(N^{-1/2})$. The proof follows similar arguments in the proofs of Theorems 1 and 2 in Wang and Leng (2007) and thus is omitted.

To select λ_1 and λ_2, we propose to minimize a BIC-type criterion (Schwarz, 1978),

$$\mathrm{BIC}(\lambda_1, \lambda_2) = (\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}} (\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda}) + \{df_{1,\lambda} + df_{2,\lambda}\} \log N, \tag{3.8}$$

where $df_{k,\lambda} = |\widetilde{\mathcal{M}}_{k,\lambda}|$ for $k = 1, 2$. Following the proof of Theorem 3 in Wang et al. (2009), one can show that with the optimal (λ_1, λ_2) from (3.8), $\mathrm{pr}(\widetilde{\mathcal{M}}_{1,\lambda} = \mathcal{M}_1, \widetilde{\mathcal{M}}_{2,\lambda} = \mathcal{M}_2) \to 1$ as $N \to \infty$. In other words, any (λ_1, λ_2) that does not lead to the correct variable selection cannot be selected by (3.8) when the number of individuals is large.

The purpose of this subsection is to provide a shrinkage estimator of the regression coefficients, of which the sparsity pattern can help us identify the truly relevant risk factor for each infection. Inference procedures, such as constructing a confidence interval or conducting hypothesis testing, based on this shrinkage estimator are beyond the scope of this work. There are numerous studies demonstrating that even in classical linear regression, finite-sample inference procedures based on asymptotic properties of the adaptive LASSO estimator perform poorly (Minnier et al., 2011). Developing valid inferential methods for shrinkage estimators in group testing, even with a single infection, could be an interesting but challenging future research topic. In this article, it is the variable selection of primary interest.

3.5 Numerical studies

We consider three different settings for the joint distribution of $\widetilde{\boldsymbol{Y}}_{ij}|\mathbf{x}_{ij}$. In all of them, we keep both g_1 and g_2 in the marginal regression model (3.1) being the inverse of the logit link function, and use a Gumbel copula (Gumbel, 1960), $\mathcal{C}(u_1, u_2|\delta) = \exp\{-[(-\log u_1)^{1/\delta} + (-\log u_2)^{1/\delta}]^{\delta}\}$ with $\delta = 0.3$, to generate the co-infection probability (3.2). The difference across the three settings comes from the choices of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \mathbf{x})$, where $\mathbf{x}$ is a generic notation of $\mathbf{x}_{ij}$'s:

- (S1) $\boldsymbol{\beta}_1 = (-5, -3, 2, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (-5, -3, 0, 3, 0, 0)^{\mathrm{T}}$ and $\mathbf{x} = (1, x_1, \cdots, x_5)^{\mathrm{T}}$, where we independently simulate x_1 from $\mathcal{N}(0, 1)$, x_2 and x_3 from Bernoulli(0.4), x_4 from Uniform$(-0.5, 0.5)$, and x_5 from $\mathcal{N}(0, 0.75^2)$

- (S2) $\boldsymbol{\beta}_1 = (-4, -2, 2, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (-5, -2, 0, -2, 0, 0)^{\mathrm{T}}$ and $\mathbf{x} = (1, x_1, \cdots, x_5)^{\mathrm{T}}$, where $\mathbf{x}$ is simulated from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ with $[\boldsymbol{\Omega}]_{st} = 1$ if $s = t$ and $[\boldsymbol{\Omega}]_{st} = 0.5$ if $s \neq t$.

- (S3) $\boldsymbol{\beta}_1 = (-5, (-2, -2, -2, 2, 2) \otimes (1, 0))^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (-6, (-3, -3, 2, 3, 0) \otimes (1, 0))^{\mathrm{T}}$ and $\mathbf{x} = (1, x_1, \cdots, x_{10})^{\mathrm{T}}$, where $\otimes$ is the Kronecker product, $\mathbf{x}$ is simulated

from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ with $[\boldsymbol{\Omega}]_{st} = 1$ if $s = t$ and $[\boldsymbol{\Omega}]_{st} = 0.5$ if $s \neq t$.

Note that $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ have different sparsity patterns (e.g., in S1, x_2 is significant to the first infection but not to the second infection). This is used to emulate the situation where two infections have different sets of significant risk factors. The values of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are chosen in a way such that the prevalence of each infection is about 7%–10%.

Under each setting, we simulate two types of data: individual testing data and the SHL pooled testing data. To do so, we first generate $N = 3000$ individual covariates. Given a set of covariates, we calculate the individual's cell probabilities ($p_{ijy_1y_2}$'s) using the specified copula-based multivariate binary regression model, and then generate the true infection statuses for both infections from a multinomial distribution with those cell probabilities. We denote the covariates and the true infection statuses of the nth individual by $\mathbf{x}_n$ and $\widetilde{\boldsymbol{Y}}_n = (\widetilde{Y}_{n1}, \widetilde{Y}_{n2})^{\mathrm{T}}$, respectively, for $n = 1, \ldots, 3000$. Herein, because groups have not been created yet, we use the subscript n instead of the ij (in $\widetilde{\boldsymbol{Y}}_{ij}$ and $\mathbf{x}_{ij}$). Given $(\widetilde{\boldsymbol{Y}}_n, \mathbf{x}_n)$'s, we simulate individual testing data and the SHL pooled testing data. We let $S_{e:k} = S_{p:k} = 0.95$ for $k = 1, 2$. Values other than 0.95 are considered in Appendix B.6.

Based on $\widetilde{\boldsymbol{Y}}_n$'s, we generate individual testing outcomes of the nth specimen by $\boldsymbol{T}_n = (T_{n1}, T_{n2})^{\mathrm{T}}$ where $T_{nk} \sim \text{Bernoulli}\{S_{e:k}\widetilde{Y}_{nk} + (1 - S_{p:k})(1 - \widetilde{Y}_{nk})\}$. Then we estimate $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \delta)^{\mathrm{T}}$ from $(\boldsymbol{T}_n, \mathbf{x}_n)$'s. This estimation procedure is similar to the one outlined in Section 3.3. We also use a GEM-algorithm to compute the MLEs and Louis' method to calculate the observed data information matrix for making large sample Wald-type inferences. Furthermore, we slightly modify our variable selection method (in Section 4) to accommodate individual testing data. All the details are provided in Appendix B.7. It is worthwhile to note that $S_{e:k}$'s and $S_{p:k}$'s are not estimable in individual testing data. Hence, with individual testing data $(\boldsymbol{T}_n, \mathbf{x}_n)$'s, we have to assume the true values of $S_{e:k}$'s and $S_{p:k}$'s as known to estimate $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \delta)$.

We generate the SHL pooled testing data from $\widetilde{\boldsymbol{Y}}_n$'s. A common group size is used in our simulations; i.e., $c_j = c$, and $c \in \{2, 5, 10\}$. For a fixed c, we randomly assign the 3000 individuals to one of $J = 3000/c$ groups. With the group membership identified, we relabel $(\widetilde{\boldsymbol{Y}}_n, \mathbf{x}_n)$'s by $(\widetilde{\boldsymbol{Y}}_{ij}, \mathbf{x}_{ij})$ where $i = 1, \ldots, c$ and $j = 1, \ldots, J$. The true statuses of the jth pool are calculated as $\widetilde{Z}_{jk} = \max_i \widetilde{Y}_{ijk}$ where $k = 1, 2$. Then we generate the pooled testing outcomes by $\boldsymbol{Z}_j = (Z_{j1}, Z_{j2})^{\mathrm{T}}$, where $Z_{jk} \sim \text{Bernoulli}\{S_{e:k}\widetilde{Z}_{jk} + (1 - S_{p:k})(1 - \widetilde{Z}_{jk})\}$. As per the SHL pooling protocol, only if $\max(Z_{j1}, Z_{j2}) = 1$, we generate retesting outcomes of the ith individual in this group by $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2})^{\mathrm{T}}$, where $Y_{ijk} \sim \text{Bernoulli}\{S_{e:k}\widetilde{Y}_{ijk}+(1-S_{p:k})(1-\widetilde{Y}_{ijk})\}$. Collecting all $\boldsymbol{Z}_j$'s and $\boldsymbol{Y}_{ij}$'s yields the SHL pooled testing data $\boldsymbol{\mathcal{P}}$. Note that the number of tests that were used to obtain $\boldsymbol{\mathcal{P}}$ is the summation of J and the number of $\boldsymbol{Y}_{ij}$'s. From $\boldsymbol{\mathcal{P}}$ and $\mathbf{x}_{ij}$'s, we estimate $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \delta, S_{e:1}, S_{e:2}, S_{p:1}, S_{p:2})$.

We repeat 500 times the process of generating $\boldsymbol{T}_n$'s and $\boldsymbol{\mathcal{P}}$ for each $c \in \{2, 5, 10\}$. For each set of individual testing data or the SHL pooled testing data, we first treat the diagnosis results for each infection as the true statues and fit them using our copula-based multivariate binary regression model. The resulting MLE of $\boldsymbol{\theta}_1$ is used as the initial value of $\boldsymbol{\theta}_1$. The initial values of the assay sensitivity and specificity are chosen to be 0.9. Then we run our GEM algorithm to compute the MLE and use Louis' method to construct a 95% confidence interval for each unknown parameter (see the last paragraph of Section 3.3). In addition to the BIC-type shrinkage estimator, we also compute an AIC-type (Akaike, 1974) and an ERIC-type (Hui et al., 2015) estimator using the tuning parameters selected by minimizing

$$\text{AIC}(\lambda_1, \lambda_2) = (\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda}) + 2\{df_{1,\lambda} + df_{2,\lambda}\}$$

and

$$\text{ERIC}(\lambda_1, \lambda_2) = (\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda})^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A}}(\hat{\boldsymbol{\theta}}_{\mathcal{A}} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda}) + df_{1,\lambda}\log(N/\lambda_1) + df_{2,\lambda}\log(N/\lambda_2),$$

respectively. For individual testing data, slightly modified versions are available in Appendix B.7.3.

To compare the overall performance of the MLE and three shrinkage estimators, we consider the prediction error,

$$\text{PE} = N^{-1} \sum_{j=1}^{J} \sum_{i=1}^{c_j} \{ \sum_{y_1=0}^{1} \sum_{y_2=0}^{1} (\hat{p}_{ijy_1y_2} - p^*_{ijy_1y_2})^2 \}^{1/2},$$

where $p^*_{ijy_1y_2}$'s are the true cell probabilities and $\hat{p}_{ijy_1y_2}$'s are the predicted cell probabilities using an estimator of $(\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \delta)$. To evaluate the variable selection performance of shrinkage estimators, we define by the selection rate (SR) the proportion of the true model being exactly selected by a shrinkage estimator. Results from the 500 replications under S1–S3 are summarized in Tables 3.1–3.4.

Tables 1–3 provide summary statistics of the MLEs for S1–S3, respectively. Under both individual testing and the SHL pooling protocol, the MLEs of the unknown parameters obtained by our GEM algorithm exhibit little, if any, evidence of bias, across all considered settings. Regarding the use of Louis' method, we notice that the average standard errors are in agreement with the sample standard deviations of the estimates. In addition, the empirical coverage probabilities for 95% confidence intervals are predominantly at the nominal level. These results indicate that the observed data information matrix is estimated correctly via Louis' method.

To examine the performance of the variable selection, Table 4 provides the SR (in parenthesis) of each shrinkage estimator across all considered settings. One can see that our BIC-type estimator performs the best in identifying the true model in each scenario. For example, in S3 when $c = 2$, the SR using the BIC criterion is 0.820 which is significantly larger than the ones using the AIC (0.294) and the ERIC (0.448) criterion. These results demonstrate the advantage of using the BIC criterion in identifying risk factors that are truly relevant for each infection.

Table 3.1: Summary statistics of the 500 MLEs obtained under S1, including the sample mean (Mean), the sample standard deviation (SD), the average of the estimated standard errors (SE), and the empirical coverage (EC) of 95% confidence intervals under either individual testing (IT) or the SHL pooling with $c = 2, 5, 10$. The average number of tests (# of tests) under each protocol is also provided. The prevalence (averaged over 500 repetitions) of the first and second infections are 7.64% and 8.22%, respectively.

		IT		$c = 2$		$c = 5$		$c = 10$	
# of tests		3000		2351		2078		2445	
	Truth	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)
β_{10}	-5	-5.08(0.36)	0.94(0.37)	-5.06(0.29)	0.94(0.29)	-5.06(0.31)	0.94(0.29)	-5.07(0.34)	0.95(0.32)
β_{11}	-3	-3.05(0.25)	0.96(0.26)	-3.03(0.21)	0.94(0.21)	-3.04(0.22)	0.94(0.21)	-3.04(0.24)	0.95(0.23)
β_{12}	2	2.03(0.27)	0.94(0.27)	2.02(0.24)	0.94(0.24)	2.02(0.25)	0.94(0.24)	2.03(0.26)	0.95(0.25)
β_{13}	0	-0.01(0.24)	0.95(0.23)	-0.01(0.22)	0.95(0.21)	-0.01(0.22)	0.95(0.21)	-0.01(0.22)	0.94(0.21)
β_{14}	0	0.01(0.38)	0.95(0.39)	0.01(0.34)	0.96(0.35)	-0.01(0.35)	0.96(0.35)	0.00(0.37)	0.96(0.36)
β_{15}	0	0.00(0.19)	0.96(0.20)	0.00(0.17)	0.97(0.18)	0.00(0.17)	0.97(0.18)	0.00(0.19)	0.94(0.19)
β_{20}	-5	-5.08(0.37)	0.95(0.37)	-5.05(0.28)	0.94(0.30)	-5.05(0.30)	0.94(0.30)	-5.04(0.33)	0.96(0.32)
β_{21}	-3	-3.04(0.26)	0.95(0.26)	-3.03(0.21)	0.94(0.22)	-3.03(0.22)	0.94(0.21)	-3.02(0.24)	0.95(0.23)
β_{22}	0	-0.01(0.24)	0.94(0.23)	-0.01(0.21)	0.93(0.21)	0.00(0.22)	0.93(0.21)	-0.01(0.23)	0.94(0.21)
β_{23}	3	3.04(0.33)	0.94(0.32)	3.03(0.27)	0.94(0.27)	3.03(0.29)	0.94(0.27)	3.03(0.30)	0.94(0.29)
β_{24}	0	0.00(0.40)	0.95(0.38)	0.02(0.34)	0.95(0.35)	0.01(0.35)	0.95(0.35)	0.00(0.36)	0.96(0.36)
β_{25}	0	0.01(0.20)	0.95(0.20)	0.00(0.18)	0.94(0.18)	0.00(0.18)	0.94(0.18)	0.00(0.19)	0.95(0.19)
δ	0.3	0.28(0.09)	0.97(0.10)	0.29(0.06)	0.95(0.06)	0.29(0.06)	0.95(0.06)	0.29(0.07)	0.95(0.07)
$S_{e:1}$	0.95	–	–	0.95(0.02)	0.93(0.02)	0.95(0.02)	0.93(0.02)	0.95(0.02)	0.90(0.02)
$S_{e:2}$	0.95	–	–	0.95(0.01)	0.95(0.01)	0.95(0.02)	0.91(0.01)	0.95(0.02)	0.92(0.02)
$S_{p:1}$	0.95	–	–	0.95(0.01)	0.94(0.01)	0.95(0.01)	0.94(0.01)	0.95(0.01)	0.93(0.01)
$S_{p:2}$	0.95	–	–	0.95(0.01)	0.94(0.01)	0.95(0.01)	0.93(0.01)	0.95(0.01)	0.93(0.01)

Table 3.2: Summary statistics of the 500 MLEs obtained under S2, including the sample mean (Mean), the sample standard deviation (SD), the average of the estimated standard errors (SE), and the empirical coverage (EC) of 95% confidence intervals under either individual testing (IT) or the SHL pooling with $c = 2, 5, 10$. The average number of tests (# of tests) under each protocol is also provided. The prevalence (averaged over 500 repetitions) of the first and the second infections are 6.77% and 9.98%, respectively.

		IT		$c = 2$		$c = 5$		$c = 10$	
# of tests		3000		2493		2312		2678	
	Truth	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)
β_{10}	-4	-4.05(0.25)	0.95(0.26)	-4.02(0.20)	0.95(0.19)	-4.03(0.19)	0.97(0.20)	-4.03(0.21)	0.96(0.21)
β_{11}	-2	-2.03(0.19)	0.96(0.20)	-2.02(0.15)	0.96(0.15)	-2.03(0.16)	0.96(0.16)	-2.02(0.17)	0.95(0.17)
β_{12}	2	2.03(0.19)	0.94(0.20)	2.02(0.16)	0.95(0.16)	2.02(0.16)	0.96(0.16)	2.02(0.17)	0.95(0.17)
β_{13}	0	0.00(0.13)	0.95(0.14)	0.00(0.12)	0.95(0.12)	0.00(0.12)	0.96(0.12)	0.00(0.12)	0.95(0.13)
β_{14}	0	0.00(0.13)	0.96(0.14)	0.00(0.12)	0.95(0.12)	0.00(0.12)	0.96(0.12)	-0.01(0.13)	0.95(0.13)
β_{15}	0	0.00(0.14)	0.95(0.14)	0.00(0.11)	0.96(0.12)	0.00(0.12)	0.95(0.12)	0.00(0.13)	0.95(0.13)
β_{20}	-5	-5.06(0.36)	0.94(0.35)	-5.04(0.26)	0.96(0.27)	-5.03(0.28)	0.97(0.29)	-5.04(0.32)	0.94(0.33)
β_{21}	-2	-2.04(0.20)	0.95(0.20)	-2.03(0.16)	0.97(0.17)	-2.02(0.17)	0.97(0.17)	-2.03(0.19)	0.95(0.19)
β_{22}	0	0.01(0.13)	0.97(0.13)	0.00(0.12)	0.95(0.12)	0.01(0.12)	0.96(0.12)	0.01(0.13)	0.95(0.13)
β_{23}	-2	-2.04(0.20)	0.95(0.20)	-2.03(0.17)	0.96(0.17)	-2.02(0.17)	0.95(0.17)	-2.02(0.19)	0.94(0.18)
β_{24}	0	0.01(0.14)	0.93(0.13)	0.01(0.12)	0.93(0.12)	0.00(0.12)	0.94(0.12)	0.00(0.13)	0.95(0.13)
β_{25}	0	0.01(0.13)	0.95(0.13)	0.01(0.12)	0.95(0.12)	0.00(0.12)	0.95(0.12)	0.01(0.12)	0.96(0.13)
δ	0.3	0.30(0.08)	0.99(0.11)	0.30(0.06)	0.97(0.07)	0.30(0.07)	0.97(0.07)	0.30(0.08)	0.97(0.08)
$S_{e:1}$	0.95	–	–	0.95(0.02)	0.93(0.02)	0.95(0.02)	0.93(0.02)	0.95(0.02)	0.92(0.02)
$S_{e:2}$	0.95	–	–	0.95(0.02)	0.95(0.01)	0.95(0.02)	0.92(0.01)	0.95(0.02)	0.91(0.02)
$S_{p:1}$	0.95	–	–	0.95(0.01)	0.97(0.01)	0.95(0.01)	0.95(0.01)	0.95(0.01)	0.92(0.01)
$S_{p:2}$	0.95	–	–	0.95(0.01)	0.92(0.01)	0.95(0.01)	0.93(0.01)	0.95(0.01)	0.92(0.01)

Table 3.3: Summary statistics of the 500 MLEs obtained under S3, including the sample mean (Mean), the sample standard deviation (SD), the average of the estimated standard errors (SE), and the empirical coverage (EC) of 95% confidence intervals under either individual testing (IT) or the SHL pooling with $c = 2, 5, 10$. The average number of tests (# of tests) under each protocol is also provided. The prevalence (averaged over 500 repetitions) of the first and the second infections are 9.97% and 8.54%, respectively.

		IT		$c = 2$		$c = 5$		$c = 10$	
# of tests		3000		2508		2337		2701	
	Truth	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)
β_{10}	-5	-5.10(0.39)	0.94(0.36)	-5.07(0.30)	0.93(0.27)	-5.10(0.33)	0.92(0.30)	-5.09(0.36)	0.95(0.34)
β_{11}	-2	-2.04(0.22)	0.94(0.21)	-2.03(0.18)	0.95(0.17)	-2.05(0.20)	0.93(0.18)	-2.04(0.20)	0.94(0.20)
β_{12}	0	0.00(0.13)	0.97(0.14)	0.00(0.12)	0.98(0.13)	0.00(0.13)	0.96(0.13)	0.00(0.13)	0.97(0.14)
β_{13}	-2	-2.04(0.22)	0.93(0.21)	-2.03(0.18)	0.94(0.17)	-2.04(0.19)	0.94(0.18)	-2.04(0.21)	0.93(0.20)
β_{14}	0	0.01(0.14)	0.96(0.14)	0.00(0.12)	0.95(0.13)	0.00(0.13)	0.94(0.13)	0.00(0.13)	0.97(0.14)
β_{15}	-2	-2.05(0.22)	0.94(0.21)	-2.04(0.18)	0.94(0.17)	-2.05(0.19)	0.93(0.18)	-2.05(0.21)	0.92(0.19)
β_{16}	0	0.01(0.14)	0.96(0.14)	0.00(0.13)	0.95(0.13)	0.01(0.13)	0.96(0.13)	0.01(0.14)	0.95(0.14)
β_{17}	2	2.04(0.22)	0.95(0.21)	2.03(0.18)	0.93(0.17)	2.05(0.19)	0.93(0.18)	2.04(0.21)	0.94(0.20)
β_{18}	0	0.00(0.14)	0.96(0.14)	0.00(0.12)	0.96(0.13)	0.00(0.13)	0.95(0.13)	0.01(0.14)	0.96(0.14)
β_{19}	2	2.04(0.21)	0.94(0.21)	2.03(0.18)	0.93(0.17)	2.04(0.19)	0.94(0.18)	2.04(0.20)	0.96(0.20)
β_{110}	0	0.00(0.15)	0.94(0.14)	0.00(0.13)	0.96(0.13)	0.01(0.13)	0.95(0.13)	0.00(0.14)	0.95(0.14)
β_{20}	-6	-6.13(0.49)	0.95(0.48)	-6.10(0.36)	0.96(0.36)	-6.10(0.41)	0.95(0.39)	-6.12(0.43)	0.95(0.43)
β_{21}	-3	-3.07(0.30)	0.95(0.29)	-3.05(0.24)	0.94(0.24)	-3.05(0.26)	0.94(0.25)	-3.06(0.27)	0.94(0.27)
β_{22}	0	0.00(0.16)	0.95(0.16)	0.01(0.14)	0.95(0.14)	0.00(0.15)	0.95(0.15)	0.00(0.15)	0.95(0.15)
β_{23}	-3	-3.07(0.30)	0.96(0.29)	-3.05(0.24)	0.94(0.24)	-3.05(0.26)	0.94(0.25)	-3.06(0.27)	0.94(0.27)
β_{24}	0	0.00(0.16)	0.96(0.16)	0.00(0.13)	0.94(0.14)	0.00(0.14)	0.95(0.15)	0.01(0.15)	0.95(0.16)
β_{25}	2	2.04(0.21)	0.97(0.23)	2.04(0.19)	0.96(0.19)	2.04(0.20)	0.95(0.20)	2.04(0.20)	0.95(0.20)
β_{26}	0	0.01(0.17)	0.94(0.16)	0.01(0.15)	0.93(0.14)	0.01(0.15)	0.95(0.15)	0.00(0.16)	0.94(0.16)
β_{27}	3	3.06(0.29)	0.95(0.29)	3.04(0.24)	0.95(0.24)	3.04(0.26)	0.94(0.25)	3.05(0.27)	0.95(0.27)
β_{28}	0	0.01(0.16)	0.96(0.16)	0.00(0.14)	0.96(0.14)	0.00(0.15)	0.95(0.15)	0.01(0.15)	0.96(0.16)
β_{29}	0	0.00(0.16)	0.95(0.16)	-0.01(0.14)	0.94(0.14)	-0.01(0.15)	0.95(0.15)	-0.01(0.16)	0.96(0.15)
β_{210}	0	0.01(0.16)	0.95(0.16)	0.01(0.14)	0.94(0.14)	0.01(0.15)	0.95(0.15)	0.01(0.15)	0.94(0.16)
δ	0.3	0.29(0.09)	0.98(0.13)	0.28(0.07)	0.99(0.08)	0.29(0.07)	0.98(0.09)	0.29(0.07)	0.99(0.11)
$S_{e:1}$	0.95	–	–	0.95(0.01)	0.93(0.01)	0.95(0.01)	0.94(0.01)	0.95(0.02)	0.94(0.01)
$S_{e:2}$	0.95	–	–	0.95(0.01)	0.95(0.01)	0.95(0.02)	0.93(0.01)	0.95(0.02)	0.91(0.02)
$S_{p:1}$	0.95	–	–	0.95(0.01)	0.96(0.01)	0.95(0.01)	0.93(0.01)	0.95(0.01)	0.92(0.01)
$S_{p:2}$	0.95	–	–	0.95(0.01)	0.96(0.01)	0.95(0.01)	0.94(0.01)	0.95(0.01)	0.93(0.01)

Table 3.4: The average prediction error PE $\times$ 100 and the SR value (provided in parenthesis) of the MLE and the shrinkage estimates under the AIC, BIC, and ERIC tuning parameter criterion over 500 replications under S1 – S3 across individual testing (IT) and the SHL pooling with $c = 2, 5$ and 10. Recall that the SR (selection rate) is defined to be the proportion of the true model being exactly selected by a shrinkage estimator. The highest SR value under each setting is underlined.

		IT	$c = 2$	$c = 5$	$c = 10$
Setting	Estimate	PE×100(SR)	PE×100(SR)	PE×100(SR)	PE×100(SR)
S1	MLE	0.148(0.000)	0.126(0.000)	0.130(0.000)	0.142(0.000)
	AIC	0.106(0.414)	0.092(0.430)	0.092(0.442)	0.102(0.462)
	BIC	0.079(0.910)	0.071(0.908)	0.073(<u>0.926</u>)	0.083(0.898)
	ERIC	0.085(0.724)	0.075(0.736)	0.076(0.744)	0.085(0.734)
S2	MLE	0.133(0.000)	0.106(0.000)	0.117(0.000)	0.121(0.000)
	AIC	0.095(0.414)	0.074(0.414)	0.084(0.436)	0.087(0.418)
	BIC	0.074(0.908)	0.059(<u>0.910</u>)	0.067(0.892)	0.069(0.876)
	ERIC	0.084(0.702)	0.064(0.696)	0.074(0.702)	0.074(0.702)
S3	MLE	0.284(0.000)	0.231(0.000)	0.250(0.000)	0.266(0.000)
	AIC	0.193(0.266)	0.160(0.294)	0.175(0.274)	0.184(0.298)
	BIC	0.158(0.818)	0.130(<u>0.820</u>)	0.145(0.786)	0.153(0.808)
	ERIC	0.183(0.428)	0.150(0.448)	0.163(0.420)	0.170(0.448)

Table 4 also provides the average PE×100 values of the MLE and the three shrinkage estimators across all settings. It is clear that all the shrinkage estimators produce smaller prediction errors than the MLE. For example, the BIC-type estimator can reduce almost 50% of the prediction error of the MLE. This is because that the adaptive LASSO penalty in (3.7) could eliminate unnecessary risk factors. Furthermore, because our BIC-type estimator outperforms the other two in term of variable selection, its prediction errors are the smallest under all settings. In conclusion, using the BIC-type shrinkage estimator not only provides a large chance of identifying truly relevant covariates, but also yields a high prediction accuracy.

Finally, we want to see whether the SHL pooling protocol causes a loss of information and thus compromises regression inference, when compared to individual testing. To find the answer, we revisit Tables 1–4. This time we focus on the comparison between individual testing and the SHL pooling. Tables 1–3 provide the average number of tests under each setting. Obviously, the SHL pooling protocol uses fewer

tests than individual testing (saves about 16% costs). This is an expected appealing feature of the SHL pooling (Tebbs et al., 2013). And we observe more: (i) In Tables 1–3, the standard deviations obtained using pooling data are uniformly less than the ones obtained using individual testing, suggesting that the SHL pooling could provide a less variational MLE; (ii) All the averaged standard errors under the SHL pooling are smaller than the ones under individual testing, meaning that one could use the SHL pooled testing data to construct narrower confidence intervals while maintaining the same nominal level; (iii) The advantage of pooling also holds when comparing the average PE$\times$100 values in Table 4, indicating that the SHL pooling enables one to make a better prediction of an individual's infection probabilities; (iv) In terms of variable selection, the highest SR value (in Table 4) always occurs at $c > 1$ under each setting; that is, using the SHL pooled testing data has a larger chance to identify the true model. Hence, instead of compromising regression inference, the SHL pooling produces more precise inference. In addition, one must note that these advantages are achieved with a less amount of costs and a larger number of parameters to be estimated. This finding could be very encouraging to laboratories that are not using pooling (such as the NHPL).

3.6 A CT/NG screening data set

To further encourage the use of pooling, we analyze a data set collected from the NPHL which currently uses individual testing for the CT/NG screening. We will illustrate, if switching from individual testing to the two-stage hierarchical pooling used by the SHL, what benefits could be achieved for regression. To do so, we first reiterate how the SHL is using the pooling protocol (Tebbs et al., 2013). Only female swab specimens are screened using the SHL pooling protocol. The testing is carried out by the TECAN DTS platform with the Aptima Combo 2 assay. The platform is calibrated for a group size $c = 4$. The sensitivity and specificity of the assay are

$S_{e:1} = 0.942$ ($S_{e:2} = 0.992$) and $S_{p:1} = 0.976$ ($S_{p:2} = 0.987$) for CT (NG), respectively (Gen-Probe, San Diego).

In 2009, 14530 female swab specimens were tested individually in the NPHL. The employed assay was also the Aptima Combo 2 Assay. We are provided with the diagnosed results of each specimen for CT and NG. Based on these diagnoses, the approximated prevalence of CT and NG are 0.069 and 0.013, respectively. To reveal the benefits of pooling, we mimic the SHL screening practice in the most realistic way. We use a group size $c = 4$ which is used by the SHL. Then we construct pools by assigning specimens according to their arrival time at the NPHL. Because the arrival time of specimens at the NPHL are random, our way of pooling is also random. We treat the diagnoses as "true" statuses and simulate a two-stage group testing data set using the above testing error rates. For comparison, we also simulate an individual testing data set using the same testing error rates. The considered covariates include age, prenatal, symptoms, cervical friability, pelvic inflammatory disease, cervicitis, multiple partners, new partner in the last 90 days, and contact with someone who has an STD. All covariates, except age, are binary. With these covariates on each individual, we first fit the individual diagnoses results by viewing them as the truth. The resulting estimates are used as the "reference" estimates. We then fit the individual testing data and the two-stage group testing data using the regression and variable selection methods previously described. In our analysis, we standardize age and code dichotomous covariates as either -0.5 or 0.5.

Table 3.5 summarizes the parameter estimates and variable selection results. The estimates from both testing protocols are close to the "reference" estimates, but the SEs under $c = 4$ are uniformly less than the ones under individual testing. The testing error rates are estimated accurately from the group testing data.

Table 3.5: The NPHL screening data analysis: parameter estimates (MLE), estimated standard errors (SE) and variable selection results (using the AIC, BIC, and ERIC criterion) from the reference estimates (Reference), individual testing estimates (IT) and the SHL pooling estimates with a group size 4 ($c = 4$). The number of tests under each is provided as well.

		Reference				IT				$c = 4$			
number of tests		–				14530				7737			
		MLE(SE)	AIC	BIC	ERIC	MLE(SE)	AIC	BIC	ERIC	MLE(SE)	AIC	BIC	ERIC
	Intercept	-1.382(0.241)	–	–	–	-1.528(0.286)	–	–	–	-1.269(0.262)	–	–	–
	Age	-0.559(0.045)	✓	✓	✓	-0.535(0.057)	✓	✓	✓	-0.561(0.051)	✓	✓	✓
	Prenatal	0.390(0.220)	✓	✓	✓	0.141(0.291)	×	×	×	0.480(0.229)	✓	✓	✓
	Symptoms	0.356(0.079)	✓	✓	✓	0.324(0.095)	✓	✓	✓	0.356(0.088)	✓	✓	✓
CT	Cervical F	0.065(0.163)				-0.058(0.202)				0.003(0.182)			
	PID	0.443(0.392)	✓	✓	✓	0.443(0.448)	✓	✓	✓	0.492(0.427)	✓	✓	✓
	Cervicitis	0.611(0.106)	✓	✓	✓	0.746(0.118)	✓	✓	✓	0.645(0.116)	✓	✓	✓
	Multi Partner	0.476(0.099)	✓	✓	✓	0.522(0.116)	✓	✓	✓	0.532(0.109)	✓	✓	✓
	New Partner	-0.069(0.091)				-0.205(0.116)	✓			-0.067(0.102)			
	Contact STD	1.006(0.098)	✓	✓	✓	1.023(0.111)	✓	✓	✓	1.048(0.108)	✓	✓	✓
	Intercept	-2.426(0.416)	–	–	–	-2.727(0.595)	–	–	–	-2.683(0.507)	–	–	–
	Age	-0.251(0.083)	✓	✓	✓	-0.278(0.112)	✓	✓	×	-0.258(0.087)	✓	✓	✓
	Prenatal	0.283(0.591)				0.003(0.929)				-0.073(0.750)			
	Symptoms	1.202(0.164)	✓	✓	✓	1.176(0.219)	✓	✓	✓	1.234(0.174)	✓	✓	✓
NG	Cervical F	0.277(0.288)				0.290(0.327)	✓	✓	✓	0.270(0.301)			
	PID	1.032(0.496)	✓	✓	✓	0.719(0.635)	✓	✓	✓	0.879(0.554)	✓	✓	✓
	Cervicitis	0.625(0.199)	✓	✓	✓	0.746(0.225)	✓	✓	✓	0.712(0.201)	✓	✓	✓
	Multi Partner	1.070(0.177)	✓	✓	✓	0.894(0.216)	✓	✓	✓	1.106(0.185)	✓	✓	✓
	New Partner	-0.130(0.189)				-0.060(0.229)				-0.127(0.198)			
	Contact STD	1.405(0.173)	✓	✓	✓	1.208(0.216)	✓	✓	✓	1.402(0.180)	✓	✓	✓
	δ	0.573(0.030)	–	–	–	0.604(0.042)	–	–	–	0.563(0.033)	–	–	–
	$S_{e:1} = 0.942$	–	–	–	–	–	–	–	–	0.922(0.016)	–	–	–
	$S_{e:2} = 0.992$	–	–	–	–	–	–	–	–	0.989(0.029)	–	–	–
	$S_{p:1} = 0.976$	–	–	–	–	–	–	–	–	0.974(0.004)	–	–	–
	$S_{p:2} = 0.987$	–	–	–	–	–	–	–	–	0.985(0.002)	–	–	–

In terms of variable selection, the reference shrinkage estimates identified different sets of significant risk factors for the two infections, where prenatal is significant to CT but not to NG. The same results are identified by the three shrinkage estimates based on the group testing data. However, based on the individual testing data, none of the three shrinkage estimates can select prenatal for CT. These comparisons reinforce our conclusion that, in addition to a significant cost reduction (i.e., it saves $14530 - 7737 = 6793$ tests), the two-stage pooling protocol leads to more precise inference than individual testing while estimating the testing error rates simultaneously. In addition, we have considered randomly assigning individuals into groups as in Section 3.5 and used group sizes varying from 2 to 10. Appendix B.8 includes these results which reinforce the aforementioned conclusion on the advantages of the two-stage pooling protocol when compared to individual testing. We believe these numerical findings could encourage more laboratories to consider the two-stage pooling protocol.

3.7 Discussion

Motivated by the SHL CT/NG screening practice, we have developed a regression method for the two-stage hierarchical pooling data. Our proposed technique jointly models the unobserved individual disease statuses and produces interpretable marginal inference for each infection. The assay sensitivity and specificity for each infection can be estimated as well. In addition, we further developed a shrinkage estimator to consistently select truly relevant risk factors for each infection.

From the simulation studies and the CT/NG screening data analysis, it is exciting to observe that, as compared to individual testing, the SHL pooling protocol can significantly reduce cost and yet produce more efficient regression estimators. An interesting future project would be to theoretically investigate how to construct groups to obtain the most efficient regression estimators for each infection within a budget limit. Intuitively, individuals with high probabilities of being infected should

be tested individually and those with low probabilities could be tested in pools. But what is the criterion to differentiate between high and low probabilities? How to know these probabilities before the screening? For those tested in pools, what is the optimal pool size that should be used for inference? These are interesting but challenging questions to be answered in future works. Possible guidance could be found in McMahan et al. (2012a) and Huang et al. (2017).

In our simulation studies, we used a Gumbel copula. We chose it for two reasons. 1) When compared to Gaussian copulas, it has an analytic expression which facilitates the computation. 2) It is able to deliver robust estimates of the regression coefficients and misclassification parameters even when the true copula is not Gumbel. To reveal this robustness, we have included a simulation study in Appendix B.9. In practice, users are welcome to choose other copulas, such as Gaussian, Clayton, or Frank (Nelsen, 2007). Besides, the logistic function for g_k's could also be changed to the inverse of the link in probit or complementary log-log models. Our GEM algorithm has the generality to incorporate those choices.

Though this work mainly focuses on two infections, the model can be extended to incorporate more infections. For example, suppose there are three infections. We have $\widetilde{\boldsymbol{Y}}_{ij} = (\widetilde{Y}_{ij1}, \widetilde{Y}_{ij2}, \widetilde{Y}_{ij3})^{\mathrm{T}}$. A joint model for $\widetilde{\boldsymbol{Y}}_{ij}|\boldsymbol{x}_{ij}$ is built by assuming that there exists a random vector $\boldsymbol{U}_{ij} = (U_{ij1}, U_{ij2}, U_{ij3})^{\mathrm{T}}$, of which the distribution function is a three-dimensional copula $\mathcal{C}(u_1, u_2, u_3|\delta)$, such that the event $\{\widetilde{Y}_{ijk} = 1|\boldsymbol{x}_{ij}\}$ is equivalent to $\{U_{ijk} \leq g_k(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_k)\}$ for $k = 1, 2, 3$. Consequently, the marginal regression model (3.1) naturally holds for each disease, and the cell probabilities of $\widetilde{\boldsymbol{Y}}_{ij}|\boldsymbol{x}_{ij}$ can be calculated in terms of $\mathcal{C}$; e.g., $\mathrm{pr}(\widetilde{Y}_{ij1} = 1, \widetilde{Y}_{ij2} = 1, \widetilde{Y}_{ij3} = 0|\boldsymbol{x}_{ij}) = \mathcal{C}\{g_1(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1), g_2(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_2), 1|\delta\} - \mathcal{C}\{g_1(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1), g_2(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_2), g_3(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_3)|\delta\}$. Our GEM algorithm can be generalized to incorporate more than two infections as well. We omit details but include some simulation results in Appendix B.10 to demonstrate this generalizability.

Lastly, we discuss the three assumptions (Assumptions 1–3) on the assay sensitivity and specificity and possible ways to relax them. For Assumption 1, when the assay utilizes the concentration level of a specific biological marker (biomarker) to make a diagnosis, mixing a positive specimen with negative ones could dilute the concentration level and affect the assay sensitivity and specificity significantly when group size changes. This "dilution effect" can be taken into consideration if the distribution of the biomarker concentration is provided in advance (Wang et al., 2015; Wang et al., 2018). To relax Assumption 2, one could use a multinomial distribution to account for the cross-disease dependency of the testing outcomes when the true statuses are given. Then the number of misclassification parameters increases from 4 to 12 when the number of diseases is two. One could modify the GEM algorithm to estimate the twelve parameters along with the regression. However, some of these parameters may require an impractical large sample size to be accurately estimated. The last assumption can be relaxed by assuming a covariate-adjusted model for misclassification parameters (Janes and Pepe, 2008). But caution must be taken for model identifiability when the covariate-adjusted misclassification parameters are to be estimated along with the regression.

Chapter 4

Estimation in group testing: what can be thrown away?

Summary: For large-scale screening problems, pooled testing has been demonstrated to be more efficient than traditional individual testing (see most references in Section 1). However, at present, most laboratories still conduct individual testing protocol to collect data for estimating infection probabilities of rare diseases. The practical reason for not using pooled testing might be the complexity of a group testing algorithm. Many group testing algorithms contain multiple stages and require to perform retesting tests on suspicious individuals (Kim et al., 2007), which complicates the laboratorians' work particularly on recording the testing outcome, the structure of the data, and all the group members. Even when these information are accurately recorded, the complexity increases the difficulty for the lab technicians to analyze these highly structured data. To make the analysis of group testing data easier for the lab, we discuss potential ways to simplify data collection routine and examine how the regression estimation accuracy would be affected by the use of fewer data. The simulation studies are conducted under the two-stage Dorfman decoding for a single infection. It is natural to expect the use of less information would lead to less accurate estimators. However, when the pool responses are blind, i.e. only the final diagnoses and group memberships are collected, the regression estimation accuracy does not sacrifice greatly from using the entire structured group testing data. In other words, for implementation efficiency, in practice, the laboratory is suggested to

record only individual diagnoses and group members while performing group testing.

4.1 Model and assumptions

Suppose a total of N individuals need to be examined for one binary characteristic, and each individual is randomly assigned to one of J non-overlapping pools. For simplicity, we consider a fixed group size c herein; i.e. $cJ = N$. With the group assignments, we let ij denote the ith individual specimen in the jth pool. Then, we denote the true infection status and the covariates (risk factors and an intercept term) of the ith individual in the jth pool by a binary variable $\widetilde{Y}_{ij}$ and a $(m+1)$-dimensional vector $\mathbf{x}_{ij} = (1, x_{ij1}, \ldots, x_{ijm})^{\mathrm{T}}$, respectively. We assume the $\widetilde{Y}_{ij} \mid \mathbf{x}_{ij}$ are independent across ij and have the following relationship

$$\mathrm{pr}(\widetilde{Y}_{ij} = 1 \mid \mathbf{x}_{ij}) = g(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}), \tag{4.1}$$

where $g()$ is a user pre-specified function; e.g. the inverse of logit or probit link.

Throughout the study, we consider one infection Dorfman group testing algorithm. In the first stage, we denote the testing response of the jth group and specimens belonging to the jth group by Z_j and $\mathcal{G}_j$ respectively, for $j = 1, \cdots, J$. When the jth pool tests positive in the first stage i.e. $Z_j = 1$, we retest the corresponding subjects through the same testing assay in the second stage, of which testing outcomes are denoted by Y_{ij}'s for $i \in \mathcal{G}_j$ and $j = 1, \ldots, J$. Classified with errors, the observed testing outcome might be different from the true infection status. We denote by $\widetilde{Z}_j$ the true status for the jth pool, where $\widetilde{Z}_j = 0$ indicates the infection is truly negative for pool $\mathcal{G}_j$ and $\widetilde{Z}_j = 1$ otherwise. When the assay is perfect, a pool tests negative only if it is a mix of negative subjects, the relationship of true infection statuses before and after pooling is $\widetilde{Z}_j = \max_{i \in \mathcal{G}_j} \widetilde{Y}_{ij}$. The diagnostic ability of testing assay is described by assay sensitivity and specificity, which are denoted by S_e and S_p respectively. The sensitivity (specificity) is defined as the probability that a specimen

tests positive (negative) given that it is truly positive (negative). Note that the specimen could be either in a pooled unit from Stage 1 or an individual subject from Stage 2. Hence, we have $S_e = \text{pr}(Z_j = 1 \mid \widetilde{Z}_j = 1) = \text{pr}(Y_{ij} = 1 \mid \widetilde{Y}_{ij} = 1)$ and $S_p = \text{pr}(Z_j = 0 \mid \widetilde{Z}_j = 0) = \text{pr}(Y_{ij} = 0 \mid \widetilde{Y}_{ij} = 0)$.

To deduce the methodology in next section, we have assumptions enumerated below. They are commonly used assumptions in group testing literature; see Xie (2001), Kim et al. (2007), and Wang et al. (2014b).

A1. We assume only one discrimination assay is applied throughout the protocol and the sensitivity and specificity are constant across stages. In addition, these misclassification errors do not depend on the pool size nor the proportion of affected subjects within the pool.

A2. Given the true status, the testing outcome is independent with individual risk factors; e.g. $\text{pr}(Y_{ij} = 1 \mid \widetilde{Y}_{ij} = 1, \mathbf{x}_{ij}, \boldsymbol{\beta}) = \text{pr}(Y_{ij} = 1 \mid \widetilde{Y}_{ij} = 1)$.

A3. Conditioning on the true statuses, testing outcomes are independent; e.g. $\text{pr}(Y_{ij} = 1, Z_j = 1 \mid \widetilde{Y}_{ij} = 1, \widetilde{Z}_j = 1) = \text{pr}(Y_{ij} = 1 \mid \widetilde{Y}_{ij} = 1)\text{pr}(Z_j = 1 \mid \widetilde{Z}_j = 1)$.

4.2 Estimation

Recall that the aim of this study is to discover an efficient data collection strategy for group testing that can simplify laboratory routine and yet maintain good regression estimation performance. Because individual final diagnoses are always the final goal when laboratories are conducting disease screening, we assume that the least information that we have to collect under group testing is the disease diagnosis for every individual. Particularly, under two-stage Dorfman group testing, it comes from Stage 1 if the pool tests negative, otherwise is testing outcome from Stage 2. We denote by D_{ij} the diagnosis of the ith individual from the jth pool. It is easy to observe that

$D_{ij} = Z_j Y_{ij}$. Hence, we have

$$\text{pr}(D_{ij} = 0) = \text{pr}(Z_j = 0) + \text{pr}(Z_j = 1, Y_{ij} = 0),$$

and

$$\text{pr}(D_{ij} = 1) = \text{pr}(Z_j = 1, Y_{ij} = 1).$$

Other than individual diagnoses, some laboratories could also record group assignment information intending to illustrate how pooled testing is performed. In addition to those two types of information, current data collection strategy of group testing would also collect pooled testing outcomes in order to describe the protocol structure. With those accessible information, we consider the following three scenarios $\mathcal{S}1 - \mathcal{S}3$ from the use of the least required data to the entire data.

$\mathcal{S}1$. Unknown group memberships and pooled outcomes

$\mathcal{S}2$. Unknown pooled outcomes

$\mathcal{S}3$. Known all information

In the next, we provide the estimators of regression coefficients under the above scenarios.

4.2.1 $\mathcal{S}1$. Unknown group memberships and pooled outcomes

In this subsection, we estimate $\boldsymbol{\beta}$ from the observed individual diagnoses. Due to the unknown grouping structure, instead of using ij, we denote individual diagnosis, true infection status, and individual covariate vector of the nth subject by D_n, $\widetilde{Y}_n$ and $\mathbf{x}_n$ respectively. We further assume the $\widetilde{Y}_n$'s are independent and identically distributed with a homogeneous prevalence of infection; i.e., $\widetilde{Y}_n \sim$ Bernoulli(p), where $p = \text{pr}(\widetilde{Y}_n = 1)$, and $n = 1, \cdots, N$. With a population prevalence, we provide the estimation steps for the unknown parameters (the population prevalence p and the regression coefficients $\boldsymbol{\beta}$) from the observed data $\{D_n : n = 1, \cdots, N\}$.

Prevalence estimation

To estimate of the population prevalence p, we consider the method of moment. For the nth subject, according to the Law of Total Probability (LTP), the first moment of D_n is

$$\begin{aligned} E(D_n) &= \text{pr}(D_n = 1) \\ &= \text{pr}(D_n = 1 \mid \widetilde{Y}_n = 1)\text{pr}(\widetilde{Y}_n = 1) + \text{pr}(D_n = 1 \mid \widetilde{Y}_n = 0)\text{pr}(\widetilde{Y}_n = 0). \end{aligned} \tag{4.2}$$

Under the assumption A3,

$$\text{pr}(D_n = 1 \mid \widetilde{Y}_n = 1) = \text{pr}(Z_j = 1, Y_n = 1 \mid \widetilde{Y}_n = 1) = S_e^2. \tag{4.3}$$

Applying the LTP again, we have

$$\begin{aligned} \text{pr}(D_n = 1 \mid \widetilde{Y}_n = 0) &= \text{pr}(Z_j = 1, Y_n = 1 \mid \widetilde{Y}_n = 0) \\ &= \text{pr}(Z_j = 1, Y_n = 1 \mid \widetilde{Z}_j = 0, \widetilde{Y}_n = 0)\text{pr}(\widetilde{Z}_j = 0 \mid \widetilde{Y}_n = 0) \\ &\quad + \text{pr}(Z_j = 1, Y_n = 1 \mid \widetilde{Z}_j = 1, \widetilde{Y}_n = 0)\text{pr}(\widetilde{Z}_j = 1 \mid \widetilde{Y}_n = 0) \\ &= (1 - S_p)^2 q^{c-1} + S_e(1 - S_p)(1 - q^{c-1}), \end{aligned} \tag{4.4}$$

where $q = 1 - p$. Combining (4.3) and (4.4), the first moment of D_n in (4.2) has the explicit formula

$$E(D_n) = S_e^2 p + (1 - S_p)q\{(1 - S_p)q^{c-1} + S_e(1 - q^{c-1})\}. \tag{4.5}$$

With known S_e and S_p, an estimator of p could be obtained by minimizing the least squares objective function $\sum_{j=1}^{J}\sum_{i=1}^{c}\{D_n - E(D_n)\}^2$. Namely, the $\hat{p}$ is obtained via

$$\hat{p} = \underset{p}{\text{argmin}} \sum_{j=1}^{J}\sum_{i=1}^{c}\{D_n - E(D_n)\}^2, \tag{4.6}$$

where $E(D_n)$ is provided in (4.5).

Coefficients estimation

Next, we demonstrate the method to estimate regression coefficients $\boldsymbol{\beta}$ in (4.1). Incorporating covariates information, the first moment of $D_n \mid \mathbf{x}_n$ equals

$$\begin{aligned} E(D_n \mid \mathbf{x}_n) &= \mathrm{pr}(D_n = 1 \mid \mathbf{x}_n) \\ &= \mathrm{pr}(D_n = 1 \mid \widetilde{Y}_n = 1, \mathbf{x}_n)\mathrm{pr}(\widetilde{Y}_n = 1 \mid \mathbf{x}_n) \\ &\quad + \mathrm{pr}(D_n = 1 \mid \widetilde{Y}_n = 0, \mathbf{x}_n)\mathrm{pr}(\widetilde{Y}_n = 0 \mid \mathbf{x}_n). \end{aligned}$$

Under the assumption A2, combining the (4.1), (4.3) and (4.4), we have

$$E(D_n \mid \mathbf{x}_n, p) = S_e^2 g(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}) + (1 - S_p)\{(1 - S_p)q^{c-1} + S_e(1 - q^{c-1})\}\{1 - g(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta})\}.$$

Having the $\hat{p}$ provided in (4.6), we simply replace the unknown parameter q in above equation by $\hat{q} = 1 - \hat{p}$. Consequently, the estimator of $\boldsymbol{\beta}$ could be achieved through minimizing the least squares objective function $\sum_{j=1}^{J}\sum_{i=1}^{c}\{D_n - E(D_n \mid \mathbf{x}_n, \hat{p})\}^2$ with respect to $\boldsymbol{\beta}$. Hence, the final estimator of $\boldsymbol{\beta}$ is obtained in the form of

$$\begin{aligned} \hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\mathrm{argmin}} \sum_{j=1}^{J}\sum_{i=1}^{c}\{D_n - S_e^2 g(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}) - (1 - S_p)\{(1 - S_p)\hat{q}^{c-1} \\ + S_e(1 - \hat{q}^{c-1})\}\{1 - g(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta})\}\}^2. \end{aligned} \tag{4.7}$$

4.2.2 $\mathcal{S}2$. Unknown pooled outcomes

In this scenario, in spite of final individual diagnoses, we consider that the information of group membership are also available to use, however, pooled testing outcomes from the first stage are still assumed as unknown. Hence, we estimate $\boldsymbol{\beta}$ from observed information of $\{\mathcal{G}_j : j = 1, \cdots, J\}$ and $\boldsymbol{D} = \{D_{ij} : i \in \mathcal{G}_j, j = 1, \cdots, J\}$. The relationship of observed and latent variables can be described by the conditional

distributions below.

$$\widetilde{Y}_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\beta} \sim \text{Bernoulli}\{p_{ij} = g(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\},$$
$$Z_j \mid \widetilde{Z}_j \sim \text{Bernoulli}\{S_e^{\widetilde{Z}_j}(1-S_p)^{1-\widetilde{Z}_j}\},$$
$$\text{if } Z_j = 1, D_{ij} \mid \widetilde{Y}_{ij} \sim \text{Bernoulli}\{S_e^{\widetilde{Y}_j}(1-S_p)^{1-\widetilde{Y}_j}\}, \text{ else } D_{ij} = 0, \text{ for } i \in \mathcal{G}_j.$$

Let $\widetilde{\boldsymbol{Y}}$ and $\mathbf{x}$ be the collection of $\widetilde{Y}_{ij}$'s and $\mathbf{x}_{ij}$'s respectively. Under the assumptions in Section 4.1, the joint distribution of the observed data D_{ij} and latent data $\widetilde{Y}_{ij}$ can be written as

$$\begin{aligned}
&f(\boldsymbol{D}, \widetilde{\boldsymbol{Y}} \mid \mathbf{x}, \boldsymbol{\beta}) \\
&= \prod_{j=1}^{J} \prod_{i \in \mathcal{G}_j} f(\widetilde{Y}_{ij} \mid \mathbf{x}_{ij}, \boldsymbol{\beta}) \\
&\quad \times \prod_{j=1}^{J} I\Big(\max_{i \in \mathcal{G}_j} D_{ij} = 0\Big)\Big\{\text{pr}(Z_j = 0 \mid \widetilde{Z}_j) + \text{pr}(Z_j = 1 \mid \widetilde{Z}_j) \prod_{i \in \mathcal{G}_j} f(D_{ij} \mid \widetilde{Y}_{ij})\Big\} \\
&\quad \times \prod_{j=1}^{J} I\Big(\max_{i \in \mathcal{G}_j} D_{ij} = 1\Big)\Big\{\text{pr}(Z_j = 1 \mid \widetilde{Z}_j) \prod_{i \in \mathcal{G}_j} f(D_{ij} \mid \widetilde{Y}_{ij})\Big\} \\
&\propto \prod_{j=1}^{J} \prod_{i \in \mathcal{G}_j} g(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})^{\widetilde{Y}_{ij}}\{1 - g(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\}^{1-\widetilde{Y}_{ij}}.
\end{aligned}$$

Since S_e and S_p are assumed as known, the complete log-likelihood reduces to be

$$l_c(\boldsymbol{\beta} \mid \widetilde{\boldsymbol{Y}}, \mathbf{x}) = \sum_{j=1}^{J}\sum_{i=1}^{c}\Big[(1-\widetilde{Y}_{ij})\log\{1 - g(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\} + \widetilde{Y}_{ij}\log g(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\Big].$$

Due to the latency of $\widetilde{Y}_{ij}$'s, EM approach is applied to estimate $\boldsymbol{\beta}$. The algorithm starts from an initial value of $\boldsymbol{\beta}$ then finds the maximum likelihood estimator by applying E-step and M-step iteratively until a convergence. At a current estimator of the parameters $\boldsymbol{\beta}^{(d)}$, the E-step calculates

$$\begin{aligned}
\mathcal{Q}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(d)}) &= E\{l_c(\boldsymbol{\beta} \mid \widetilde{\boldsymbol{Y}}, \mathbf{x}) \mid \boldsymbol{D}, \mathbf{x}, \boldsymbol{\beta}^{(d)}\} \\
&= \sum_{j=1}^{J}\sum_{i=1}^{c}\Bigg[\text{pr}(\widetilde{Y}_{ij} = 0 \mid \boldsymbol{D}, \mathbf{x}, \boldsymbol{\beta}^{(d)})\log\{1 - g(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\} \\
&\qquad + \text{pr}(\widetilde{Y}_{ij} = 1 \mid \boldsymbol{D}, \mathbf{x}, \boldsymbol{\beta}^{(d)})\log g(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\Bigg].
\end{aligned}$$

Because of the independence assumption, the expected true disease status of ith individual is only related to the information from the jth pool, where $i \in \mathcal{G}_j$. Thus, we have $\text{pr}(\widetilde{Y}_{ij} = y \mid \boldsymbol{D}, \mathbf{x}, \boldsymbol{\beta}^{(d)}) = \text{pr}(\widetilde{Y}_{ij} = y \mid \boldsymbol{D}_j, \mathbf{x}_j, \boldsymbol{\beta}^{(d)})$, where $\boldsymbol{D}_j$ and $\mathbf{x}_j$ are the collection of D_{ij}'s and $\mathbf{x}_{ij}$'s from the jth pool, respectively. Then, it suffices to compute

$$\text{pr}(\widetilde{Y}_{ij} = y \mid \boldsymbol{D}_j, \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) = \frac{\text{pr}(\widetilde{Y}_{ij} = y, \boldsymbol{D}_j \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)})}{\sum_{y' \in \{0,1\}} \text{pr}(\widetilde{Y}_{ij} = y', \boldsymbol{D}_j \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)})}, \tag{4.8}$$

for $i \in \mathcal{G}_j$. Note that

$$\begin{aligned}\text{pr}(\widetilde{Y}_{ij} = y, \boldsymbol{D}_j \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) = I(\boldsymbol{D}_j = \mathbf{0})\text{pr}(Z_j = 0, \boldsymbol{D}_j, \widetilde{Y}_{ij} = y \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) \\ + \text{pr}(Z_j = 1, \boldsymbol{D}_j, \widetilde{Y}_{ij} = y \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)}).\end{aligned} \tag{4.9}$$

Let $p_{ij}(\boldsymbol{\beta}^{(d)}) = g(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}^{(d)})$ and $q_{ij}(\boldsymbol{\beta}^{(d)}) = 1 - p_{ij}(\boldsymbol{\beta}^{(d)})$. When $Z_j = 0$, then

$$\begin{aligned}\text{pr}(Z_j = 0, \boldsymbol{D}_j, \widetilde{Y}_{ij} = 0 \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) &= \text{pr}(Z_j = 0, \widetilde{Y}_{ij} = 0 \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) \\ &= \text{pr}(Z_j = 0 \mid \widetilde{Z}_j = 0)\text{pr}(\widetilde{Z}_j = 0, \widetilde{Y}_{ij} = 0 \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) \\ &\quad + \text{pr}(Z_j = 0 \mid \widetilde{Z}_j = 1)\text{pr}(\widetilde{Z}_j = 1, \widetilde{Y}_{ij} = 0 \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) \\ &= \left\{1 - S_e + (S_e + S_p - 1) \prod_{l \in \mathcal{G}_j \setminus \{i\}} q_{lj}(\boldsymbol{\beta}^{(d)})\right\} q_{ij}(\boldsymbol{\beta}^{(d)}),\end{aligned} \tag{4.10}$$

and

$$\begin{aligned}\text{pr}(Z_j = 0, \boldsymbol{D}_j, \widetilde{Y}_{ij} = 1 \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) &= \text{pr}(Z_j = 0, \widetilde{Y}_{ij} = 1 \mid \mathbf{x}_{ij}, \boldsymbol{\beta}^{(d)}) \\ &= \text{pr}(Z_j = 0 \mid \widetilde{Z}_j = 1)\text{pr}(\widetilde{Y}_{ij} = 1 \mid \mathbf{x}_{ij}, \boldsymbol{\beta}^{(d)}) \\ &= (1 - S_e)p_{ij}(\boldsymbol{\beta}^{(d)}).\end{aligned} \tag{4.11}$$

When Z_j =1,

$$
\begin{aligned}
&\text{pr}(Z_j = 1, \boldsymbol{D}_j, \widetilde{Y}_{ij} = 0 \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) \\
&= \text{pr}(Z_j = 1 \mid \widetilde{Z}_j = 0)\text{pr}(\widetilde{Z}_j = 0, \widetilde{Y}_{ij} = 0, \boldsymbol{D}_j \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) \\
&\quad + \text{pr}(Z_j = 1 \mid \widetilde{Z}_j = 1)\text{pr}(\widetilde{Z}_j = 1, \widetilde{Y}_{ij} = 0, \boldsymbol{D}_j \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) \\
&= S_p^{1-D_{ij}}(1-S_p)^{D_{ij}} q_{ij}(\boldsymbol{\beta}^{(d)}) \Bigg\{ (1-S_p) \prod_{l \in \mathcal{G}_j \setminus \{i\}} S_p^{1-D_{lj}}(1-S_p)^{D_{lj}} q_{lj}(\boldsymbol{\beta}^{(d)}) \\
&\quad + S_e \Bigg[\prod_{l \in \mathcal{G}_j \setminus \{i\}} \Big\{ S_p^{1-D_{lj}}(1-S_p)^{D_{lj}} q_{lj}(\boldsymbol{\beta}^{(d)}) + (1-S_e)^{1-D_{lj}} S_e^{D_{lj}} p_{lj}(\boldsymbol{\beta}^{(d)}) \Big\} \\
&\qquad - \prod_{l \in \mathcal{G}_j \setminus \{i\}} S_p^{1-D_{lj}}(1-S_p)^{D_{lj}} q_{lj}(\boldsymbol{\beta}^{(d)}) \Bigg] \Bigg\}, \qquad (4.12)
\end{aligned}
$$

and

$$
\begin{aligned}
&\text{pr}(Z_j = 1, \boldsymbol{D}_j, \widetilde{Y}_{ij} = 1 \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) \\
&= \text{pr}(Z_j = 1 \mid \widetilde{Z}_j = 1)\text{pr}(\widetilde{Z}_j = 1, \widetilde{Y}_{ij} = 1, \boldsymbol{D}_j \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) \\
&= S_e(1-S_e)^{1-D_{ij}} S_e^{D_{ij}} p_{ij}(\boldsymbol{\beta}^{(d)}) \\
&\quad \prod_{l \in \mathcal{G}_j \setminus \{i\}} \Big\{ S_p^{1-D_{lj}}(1-S_p)^{D_{lj}} q_{lj}(\boldsymbol{\beta}^{(d)}) + (1-S_e)^{1-D_{lj}} S_e^{D_{lj}} p_{lj}(\boldsymbol{\beta}^{(d)}) \Big\}. \qquad (4.13)
\end{aligned}
$$

Therefore, combining (4.10) – (4.13) and (4.9) completes the computation of (4.8), and further finishes the E-step. The M-step updates the $\boldsymbol{\beta}^{(d)}$ by maximizing $\mathcal{Q}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(d)})$ with respect to $\boldsymbol{\beta}$; i.e. $\boldsymbol{\beta}^{(d+1)} = \underset{\beta}{\text{argmax}} \mathcal{Q}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(d)})$. The final estimator of $\boldsymbol{\beta}$ is achieved by repeating the E-step and M-step until reaching a certain criterion.

4.2.3 $\mathcal{S}3$. Unknown all information

We now assume the group memberships, the pooled level results, and individual-level final diagnoses are all recorded by laboratorians and available to use for estimation. The method herein is simply the one-dimensional version of the one introduced in Chapter 3. Again, the EM algorithm is applied to estimate the regression coefficients. Since we now observe the $\{\mathcal{G}_j : j = 1, \cdots, J\}$, $\boldsymbol{Z} = \{Z_j : j = 1, , J\}$, and $\boldsymbol{D} = \{D_{ij} :$

$i \in \mathcal{G}_j, j = 1, \cdots, J\}$, the E-step becomes to compute

$$\begin{aligned}\mathcal{Q}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(d)}) &= E\{l_c(\boldsymbol{\beta} \mid \widetilde{\boldsymbol{Y}}, \mathbf{x}) \mid \boldsymbol{Z}, \boldsymbol{D}, \mathbf{x}, \boldsymbol{\beta}^{(d)}\} \\ &= \sum_{j=1}^{J} \sum_{i=1}^{c} \Big[\mathrm{pr}(\widetilde{Y}_{ij} = 0 \mid \boldsymbol{Z}, \boldsymbol{D}, \mathbf{x}, \boldsymbol{\beta}^{(d)}) \log\{1 - g(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\} \\ &\qquad + \mathrm{pr}(\widetilde{Y}_{ij} = 1 \mid \boldsymbol{Z}, \boldsymbol{D}, \mathbf{x}, \boldsymbol{\beta}^{(d)}) \log g(\mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) \Big],\end{aligned}$$

at a current estimate of $\boldsymbol{\beta}^{(d)}$. By the assumption of independence across pools, it suffices to compute

$$\begin{aligned}\mathrm{pr}(\widetilde{Y}_{ij} = y \mid \boldsymbol{Z}, \boldsymbol{D}, \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) &= \mathrm{pr}(\widetilde{Y}_{ij} = y \mid Z_j = z, \boldsymbol{D}_j, \mathbf{x}_j, \boldsymbol{\beta}^{(d)}) \\ &= \frac{\mathrm{pr}(\widetilde{Y}_{ij} = y, Z_j = z, \boldsymbol{D}_j \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)})}{\sum_{y' \in \{0,1\}} \mathrm{pr}(\widetilde{Y}_{ij} = y', Z_j = z, \boldsymbol{D}_j \mid \mathbf{x}_j, \boldsymbol{\beta}^{(d)})},\end{aligned}$$

for $y \in \{0, 1\}$ and $z \in \{0, 1\}$. Note that all the required terms have been provided in (4.10) – (4.13). By combining these probabilities, we accomplish the calculation in the E-step. In the M-step, the $\boldsymbol{\beta}^{(d)}$ is updated through maximizing $\mathcal{Q}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(d)})$ with updated formula $\boldsymbol{\beta}^{(d+1)} = \underset{\boldsymbol{\beta}}{\mathrm{argmax}} \mathcal{Q}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(d)})$. Finally, the maximum likelihood estimator of $\boldsymbol{\beta}$ is obtained by iteratively repeating the two updating steps until a numerical convergence.

4.3 Simulation

In this section, we conduct simulation studies to illustrate and compare the performance of the proposed estimators over all considered scenarios $\mathcal{S}1$ – $\mathcal{S}3$.

The group testing data are generated with sample size $N = 3000$. Since the groups have not constructed yet, we denote the index of subject by n for the nth individual, for $n = 1, \cdots, N$. We first generate the true individual variables. For each single subject, we generate a $m + 1$-dimensional covariate vector $\mathbf{x}_n = (1, x_{n1}, \cdots, x_{nm})^{\mathrm{T}}$, where $(x_{n1}, \cdots, x_{nm})^{\mathrm{T}}$ is simulated from a multivariate normal distribution with a correlated variance-covariance matrix $\boldsymbol{\Sigma}_{l_1,l_2} = 0.9^{|l_1 - l_2|}$ and $0 \leq l_1, l_2 \leq m$. We have

experimented with model dimensions $m \in (5, 10)$ with the corresponding parameter settings presented below.

- $\mathcal{M}5$: $\boldsymbol{\beta} = (-6, -3, -1, -3, 2, 3)^{\mathrm{T}}$
- $\mathcal{M}10$: $\boldsymbol{\beta} = (-5, 3, 1, -2, 1, 1, -4, 1, -1, -1, 1)^{\mathrm{T}}$
- $\mathcal{SM}10$: $\boldsymbol{\beta} = (-5, 3, 0, -1, 0, 0, -4, 0, 0, 1, 0)^{\mathrm{T}}$

The first two parameter settings experience low-dimensional and high-dimensional parameters $\boldsymbol{\beta}$ with $m = 5, 10$ respectively. The third setting considers sparse parameters of $\boldsymbol{\beta}$. It is worthwhile to point out, the true regression coefficients $\boldsymbol{\beta}$ are set to ensure the infectious prevalence being around 8% under each parameter setting. The individual true infection status $\widetilde{Y}_n$ is simulated from a Bernoulli random variable with the probability of infection computed from (4.1) and an inverse of logit link; i.e. $\mathrm{pr}(\widetilde{Y}_n = 1 \mid \mathbf{x}_n) = \exp(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta})/\{1 + \exp(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta})\}$. In the next, we mimic the two-stage Dorfman testing protocol. The individuals are randomly assigned to J non-overlapping pools. Without loss of generality, we consider a common group size across pools and $c \in \{2, 5, 10\}$. Then we use ij to reindex the ith individual from the jth pool, for $i \in \mathcal{G}_j$ and $j = 1, \cdots, J$. With pre-specified testing errors (S_e, S_p), the first-stage pooled outcome of the jth pool is generated by $Z_j \sim$ Bernoulli $\{S_e\widetilde{Z}_j + (1-S_p)(1-\widetilde{Z}_j)\}$, where $\widetilde{Z}_j = \max_{i \in \mathcal{G}_j} \widetilde{Y}_{ij}$. Only if $Z_j = 1$, the protocol proceeds to the second stage, and we generate the retesting outcome of the ith individual from positive pools by $Y_{ij} \sim$ Bernoulli $\{S_e\widetilde{Y}_{ij} + (1-S_p)(1-\widetilde{Y}_{ij})\}$. According to our definition in Section 4.2, the individual diagnoses are recorded as $D_{ij} = Z_j Y_{ij}$ for $i \in \mathcal{G}_j$ and $j = 1, \cdots, J$. Finally, the observed testing data are $\{D_n : n = 1, \cdots, N\}$ for $\mathcal{S}1$, $\{D_{ij} : i \in \mathcal{G}_j, j = 1, \cdots, J\}$ for $\mathcal{S}2$ and $\{(Z_j, D_{1j}, \cdots, D_{cj}) : j = 1, \cdots, J\}$ for $\mathcal{S}3$.

The simulation is repeated 500 times under each parameter, group size and data collection setting. To evaluate the overall estimation performance, we use the empir-

ical mean squared error (MSE), calculated by MSE $= E\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^{\mathrm{T}}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\}$, where $\boldsymbol{\beta}^*$ is the true $\boldsymbol{\beta}$ in the parameter setting which was used to generate data.

The results of parameter settings $\mathcal{M}5$, $\mathcal{M}10$ and $\mathcal{SM}10$ are reported in Tables 4.1 – 4.3, respectively. For the purpose of comparison, we also mimic the individual testing procedure and provide regression results for all parameter settings. We observe that the two-stage group testing could reduce up to about 45% of the testing cost compared to individual testing does. The details of estimation step under individual testing is provided in Appendix C. As for parameter estimation, in general, the estimates are close to the truth exhibiting small bias regardless of simulation settings. Although the average sample standard deviation increases as group size, it is an expected phenomenon due to the loss of individual status information while testing on large pools. Let's compare the model estimation over considered data collection settings $\mathcal{S}1$ – $\mathcal{S}3$. One can observe that the MSEs uniformly decrease from $\mathcal{S}1$ to $\mathcal{S}2$ and $\mathcal{S}2$ to $\mathcal{S}3$, which implies the use of more data would result in a better model estimation. However, for all simulation settings, $\mathcal{S}2$ beats individual testing in terms of producing estimates with lower MSEs. Regarding $\mathcal{S}1$ of using purely individual diagnoses, we found that the moment estimators do not perform well when the number of covariates is small (Table 1.1). In contrast, when regressing on a larger number of covariates (see Tables 4.2 and 4.3), the MSEs of estimator from the "method of moment" ($\mathcal{S}1$) are greatly improved no matter whether the sparsity of $\boldsymbol{\beta}$. In particular, for some cases, the estimation performance under $\mathcal{S}1$ is even better than that of $\mathcal{S}2$ and individual testing. Therefore, at a concern of estimation stability and practical efficiency, when performing group testing for screening, we strongly recommend laboratories recording only individual diagnoses (and the group memberships) instead of keeping track of every testing outcomes.

Table 4.1: Summary statistics of the estimates under parameter setting $\mathcal{M}5$, data collection scenarios $\mathcal{S}1$ – $\mathcal{S}3$ of two-stage group testing with $c \in \{2, 5, 10\}$ and individual testing (IT). Reported are the average values over 500 simulation runs, with the standard deviations in parentheses. The average numbers of test are 2053.40 ($c = 2$), 1652.48 ($c = 5$) and 1944.92 ($c = 10$). The average prevalence of infection is 7.79%.

			$c = 2$			$c = 5$			$c = 10$		
		IT	$\mathcal{S}1$	$\mathcal{S}2$	$\mathcal{S}3$	$\mathcal{S}1$	$\mathcal{S}2$	$\mathcal{S}3$	$\mathcal{S}1$	$\mathcal{S}2$	$\mathcal{S}3$
	True	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)
β_0	-6	-6.14(0.52)	-6.17(0.54)	-6.11(0.42)	-6.10(0.41)	-6.14(0.55)	-6.11(0.43)	-6.10(0.41)	-6.16(0.53)	-6.10(0.46)	-6.12(0.44)
β_1	-3	-3.08(0.38)	-3.08(0.41)	-3.06(0.35)	-3.05(0.34)	-3.07(0.39)	-3.07(0.36)	-3.06(0.35)	-3.08(0.42)	-3.04(0.39)	-3.06(0.37)
β_2	-1	-1.01(0.40)	-1.05(0.39)	-1.03(0.34)	-1.03(0.33)	-1.04(0.40)	-1.05(0.38)	-1.03(0.37)	-1.01(0.42)	-1.03(0.41)	-1.02(0.39)
β_3	-3	-3.09(0.48)	-3.07(0.46)	-3.06(0.42)	-3.06(0.42)	-3.05(0.50)	-3.05(0.45)	-3.04(0.44)	3.08(0.50)	3.06(0.47)	3.05(0.45)
β_4	2	2.05(0.42)	2.03(0.42)	2.03(0.38)	2.03(0.38)	2.05(0.47)	2.05(0.41)	2.05(0.40)	2.03(0.44)	2.03(0.41)	2.01(0.39)
β_5	3	3.08(0.39)	3.10(0.46)	3.08(0.40)	3.07(0.40)	3.06(0.49)	3.06(0.43)	3.04(0.42)	3.08(0.51)	3.06(0.46)	3.05(0.44)
	MSE	1.1809	1.2736	0.8500	0.8190	1.4266	0.9418	0.8899	1.3123	1.0627	0.9357

Table 4.2: Summary statistics of the estimates under parameter setting $\mathcal{M}10$, data collection scenarios $\mathcal{S}1$ – $\mathcal{S}3$ of two-stage group testing with $c \in \{2, 5, 10\}$ and individual testing (IT). Reported are the average values over 500 simulation runs, with the standard deviations in parentheses. The average numbers of test are 2060.25 ($c = 2$), 1664.83 ($c = 5$) and 1964.48 ($c = 10$). The average prevalence of infection is 7.91%.

			$c = 2$			$c = 5$			$c = 10$		
		IT	$\mathcal{S}1$	$\mathcal{S}2$	$\mathcal{S}3$	$\mathcal{S}1$	$\mathcal{S}2$	$\mathcal{S}3$	$\mathcal{S}1$	$\mathcal{S}2$	$\mathcal{S}3$
	True	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)
β_0	-5	-5.09(0.40)	-5.13(0.35)	-5.09(0.29)	-5.10(0.28)	-5.08(0.34)	-5.11(0.31)	-5.11(0.31)	-5.10(0.37)	-5.08(0.35)	-5.13(0.34)
β_1	3	3.08(0.37)	3.07(0.32)	3.05(0.31)	3.05(0.30)	3.08(0.32)	3.08(0.32)	3.08(0.31)	3.08(0.34)	3.12(0.38)	3.09(0.36)
β_2	1	1.02(0.35)	1.05(0.34)	1.03(0.33)	1.03(0.32)	1.00(0.33)	1.03(0.34)	1.03(0.33)	0.98(0.36)	1.01(0.39)	0.98(0.37)
β_3	-2	-2.06(0.40)	-2.05(0.38)	-2.03(0.37)	-2.03(0.36)	-2.02(0.33)	-2.06(0.35)	-2.05(0.34)	-2.02(0.36)	-2.05(0.38)	-2.03(0.37)
β_4	1	1.01(0.35)	1.01(0.37)	1.00(0.35)	1.00(0.34)	1.00(0.34)	1.03(0.34)	1.02(0.34)	1.01(0.36)	1.06(0.37)	1.05(0.36)
β_5	1	1.05(0.36)	1.02(0.36)	1.02(0.34)	1.02(0.34)	1.00(0.36)	1.01(0.35)	1.01(0.34)	1.02(0.36)	1.06(0.37)	1.05(0.36)
β_6	-4	-4.09(0.50)	-4.07(0.42)	-4.06(0.39)	-4.07(0.38)	-4.00(0.41)	-4.08(0.43)	-4.07(0.43)	-4.07(0.46)	-4.15(0.47)	-4.12(0.45)
β_7	1	1.01(0.36)	1.01(0.34)	1.01(0.33)	1.01(0.33)	0.99(0.33)	1.03(0.33)	1.02(0.33)	1.02(0.37)	1.05(0.38)	1.03(0.37)
β_8	-1	-1.02(0.36)	-1.00(0.33)	-1.00(0.33)	-0.99(0.33)	-1.02(0.33)	-1.04(0.35)	-1.04(0.34)	-1.01(0.33)	-1.05(0.35)	-1.03(0.34)
β_9	-1	-1.03(0.39)	-1.03(0.33)	-1.03(0.32)	-1.03(0.32)	-1.00(0.34)	-1.03(0.34)	-1.03(0.34)	-1.02(0.35)	-1.04(0.36)	-1.02(0.35)
β_{10}	1	1.02(0.29)	1.03(0.25)	1.02(0.24)	1.02(0.24)	1.02(0.26)	1.03(0.26)	1.03(0.25)	1.03(0.27)	1.04(0.28)	1.02(0.28)
	MSE	1.5978	1.3505	1.2059	1.1770	1.2485	1.3018	1.2669	1.4364	1.5517	1.4693

Table 4.3: Summary statistics of the estimates under parameter setting $\mathcal{SM}10$, data collection scenarios $\mathcal{S}1$ – $\mathcal{S}3$ of two-stage group testing with $c \in \{2, 5, 10\}$ and individual testing (IT). Reported are the average values over 500 simulation runs, with the standard deviations in parentheses. The average numbers of test are 2067.44 ($c = 2$), 1672.91 ($c = 5$) and 1983.30 ($c = 10$). The average prevalence of infection is 8.02%.

			$c = 2$			$c = 5$			$c = 10$		
		IT	$\mathcal{S}1$	$\mathcal{S}2$	$\mathcal{S}3$	$\mathcal{S}1$	$\mathcal{S}2$	$\mathcal{S}3$	$\mathcal{S}1$	$\mathcal{S}2$	$\mathcal{S}3$
	True	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)
β_0	-5	-5.17(0.41)	-5.02(0.31)	-5.09(0.28)	-5.08(0.28)	-5.06(0.34)	-5.10(0.33)	-5.10(0.33)	-5.03(0.32)	-5.04(0.34)	-5.09(0.33)
β_1	3	3.10(0.38)	3.02(0.30)	3.05(0.29)	3.05(0.29)	3.04(0.32)	3.06(0.32)	3.05(0.31)	3.08(0.32)	3.04(0.33)	3.08(0.33)
β_2	0	0.01(0.35)	-0.02(0.32)	-0.01(0.32)	-0.01(0.32)	0.01(0.33)	0.01(0.33)	0.01(0.33)	-0.02(0.36)	-0.02(0.36)	-0.02(0.35)
β_3	-1	-1.05(0.39)	-0.98(0.33)	-1.01(0.35)	-1.01(0.34)	-1.03(0.34)	-1.05(0.34)	-1.04(0.34)	-1.00(0.34)	-1.03(0.35)	-1.02(0.34)
β_4	0	0.01(0.36)	0.00(0.32)	0.02(0.33)	0.02(0.33)	0.02(0.31)	0.02(0.33)	0.02(0.33)	-0.02(0.32)	0.00(0.35)	0.00(0.34)
β_5	0	0.00(0.36)	-0.03(0.32)	-0.03(0.31)	-0.03(0.30)	-0.02(0.31)	0.00(0.33)	0.00(0.33)	0.01(0.33)	0.01(0.35)	0.01(0.34)
β_6	-4	-4.14(0.52)	-3.97(0.38)	-4.05(0.39)	-4.05(0.39)	-4.00(0.42)	-4.11(0.42)	-4.10(0.41)	-3.99(0.39)	-4.09(0.46)	-4.08(0.45)
β_7	0	0.00(0.36)	-0.03(0.31)	-0.02(0.32)	-0.02(0.32)	0.00(0.33)	0.02(0.32)	0.03(0.32)	-0.01(0.33)	0.01(0.34)	0.01(0.33)
β_8	0	0.02(0.36)	0.01(0.30)	0.01(0.30)	0.01(0.30)	-0.01(0.33)	-0.01(0.33)	-0.01(0.32)	-0.01(0.33)	-0.01(0.34)	-0.01(0.34)
β_9	1	1.03(0.37)	1.02(0.31)	1.03(0.31)	1.04(0.30)	0.98(0.33)	1.00(0.33)	1.00(0.33)	0.98(0.34)	1.03(0.36)	1.01(0.35)
β_{10}	0	-0.01(0.25)	0.00(0.25)	-0.01(0.25)	-0.01(0.24)	0.01(0.25)	0.01(0.24)	0.01(0.23)	0.03(0.25)	0.01(0.26)	0.01(0.25)
	MSE	1.6413	1.1024	1.1130	1.0876	1.2093	1.2262	1.1901	1.2067	1.3564	1.3182

4.4 Analysis of a chlamydia data

In this section, we apply the proposed method to analyze the chlamydia data collected from the Nebraska Public Health Laboratory (NPHL) in 2009 which implemented the traditional individual testing protocol for screening. As another participating laboratory of annual chlamydia and gonorrhea screening, the State Hygienic Laboratory (SHL) in Iowa City has already applied the two-stage group testing protocol for the purpose of cost reduction. In contrast, the NPHL suffers from the high expense of using individual testing for this large-scale screening practice. Intending to provide an alternative option to the NPHL, the analysis focuses on the regression estimation under two-stage Dorfman testing with only a fraction of cost when compared to current individual testing for understanding how risk factors influence on the status of infected chlamydia. We will illustrate the benefits by switching from individual testing to two-stage Dorfman testing, and further discuss the data collection strategy which is convenient to conduct if the lab decides to use this cost-effective protocol.

The available data consists of 14530 individual diagnoses of female swab specimens with an overall prevalence of chlamydia being about 6.90%. For the purpose of comparison, we emulate the individual testing and two-stage group testing protocol of a common group size of 5, then artificially construct individual testing data and group testing data by pretending the individual diagnoses as the "true" disease statuses and with the use of a testing assay sensitivity of 0.942 and a specificity of 0.976. Both of the misclassification values are obtained from the manufacturer's document of the Aptima Combo 2 assay, which is the current screening assay for chlamydia in both the NPHL and SHL (Gen-Probe, San Diego). We consider 9 covariates including age, whether in the prenatal period, whether presented with a symptom of infection, clinical indicators of cervical friability, pelvic inflammatory disease and cervicitis, three-month sexual history indicators of multiple partners and new partner, as well as whether the individual had contact with someone having a sexually transmitted

Table 4.4: Summary statistics of the estimates for the NPHL chlamydia data under individual testing (IT) and two-stage Dorfman testing of $c = 5$ and data collection scenarios $\mathcal{S}1$ – $\mathcal{S}3$ and. Reported are the average values over 500 simulation runs, with the standard deviations in parentheses. The "Ref" indicates the reference estimates. The average number of test for $c = 5$ is 7255.77.

			$c = 5$		
		IT	$\mathcal{S}1$	$\mathcal{S}2$	$\mathcal{S}3$
	Ref	Mean(SD)	Mean(SD)	Mean(SD)	Mean(SD)
Intercept	-1.42	-1.43(0.14)	-1.51(0.11)	-1.44(0.11)	-1.44(0.11)
Age	-0.56	-0.55(0.04)	-0.53(0.02)	-0.56(0.02)	-0.56(0.02)
Prenatal	0.40	0.39(0.14)	0.33(0.12)	0.39(0.10)	0.39(0.10)
Symptoms	0.36	0.36(0.05)	0.38(0.04)	0.36(0.04)	0.36(0.04)
Cervical F	0.02	0.02(0.10)	-0.13(0.10)	0.06(0.07)	0.07(0.07)
PID	0.38	0.38(0.21)	0.41(0.20)	0.34(0.19)	0.35(0.18)
Cervicitis	0.62	0.64(0.06)	0.72(0.06)	0.62(0.05)	0.62(0.05)
Multi Partner	0.47	0.48(0.06)	0.47(0.05)	0.47(0.05)	0.47(0.04)
New Partner	-0.07	-0.11(0.06)	-0.19(0.05)	-0.07(0.05)	-0.07(0.04)
Contact STD	1.00	1.00(0.05)	1.00(0.05)	1.00(0.04)	1.00(0.04)
	MSE	0.1144	0.1636	0.0762	0.0712

disease. With those covariates, we fit our proposed method to individual testing data and two-stage group testing data, then report a summary of estimates from 500 replications in Table 4.4. In addition, we also fit the diagnoses through a logistic regression and treat the coefficient estimates as a "reference" result seeing Table 4.4. Here, the MSE is calculated via MSE $= E\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\#})^{\intercal}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\#})\}$, where $\boldsymbol{\beta}^{\#}$ is the reference estimates described above.

In Table 4.4, the observation reinforces what we concluded in simulation studies. First of all, switching to the two-stage Dorfman group testing indeed helps the lab save nearly half of testing cost with about 7256 required tests compared to 14530 from individual testing. As for regression estimation accuracy, the $\mathcal{S}2$ and $\mathcal{S}3$ both provide better estimates than individual testing in terms of lower MSEs. Moreover, the performance of $\mathcal{S}2$ and $\mathcal{S}3$ are similar with only 0.005 difference in the MSE, which indicates that the compromise in regression estimation is negligibly small if laboratorians skip recording pooled outcomes.

4.5 Discussion

In this chapter, we discuss three scenarios of collecting data from group testing design in practice and provide regression analysis for the corresponding dataset. Our simulation studies reveal that collecting entire information (group memberships, pooled testing outcomes, and individual diagnoses) is too complicated and not recommended when the target is achieving precise regression estimations. Skipping collecting pooled outcomes can greatly optimize the laboratory process, and yet it does not significantly compromise the regression estimation. We hope this realistic simplification would make group testing more recognized in practice and more employed in the future.

Our study is specifically for two-stage Dorfman group testing algorithm of one infection. To further demonstrate the applicability of simplifying data collection in our suggested way, the work can be extended for other group testing algorithms like halve decoding or array-based group testing or to the cases of pooling with multiple infections.

Bibliography

Akaike, H. (1974). "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19, pp. 716–723.

Bilder, C. and Tebbs, J. (2005). "Empirical Bayesian estimation of the disease transmission probability in multiple-vector-transfer designs". In: *Biometrical Journal* 47, pp. 502–516.

Bilder, C. and Tebbs, J. (2009). "Bias, efficiency, and agreement for group-testing regression models". In: *Journal of Statistical Computation and Simulation* 79, pp. 67–80.

Bilder, C., Tebbs, J., and Chen, P. (2010). "Informative retesting". In: *Journal of the American Statistical Association* 105, pp. 942–955.

Black, M., Bilder, C., and Tebbs, J. (2012). "Group testing in heterogeneous populations by using halving algorithms". In: *Journal of the Royal Statistical Society, Series C* 61, pp. 277–290.

Burrows, P. (1987). "Improved estimation of pathogen transmission rates by group testing". In: *Phytopathology* 77, pp. 363–365.

Busch, M., Caglioti, S., Robertson, E., McAuley, J., Tobler, L., Kamel, H., Linnen, J., Shyamala, V., Tomasulo, P., and S., Kleinman (2005). "Screening the blood supply for West Nile virus RNA by nucleic acid amplification testing". In: *New England Journal of Medicine* 353, pp. 460–467.

Caudill, S. (2010). "Characterizing populations of individuals using pooled samples". In: *Journal of Exposure Science and Environmental Epidemiology* 20, pp. 29–37.

Chaubey, Y. and Li, W. (1995). "Comparison between maximum likelihood and Bayes methods for estimation of binomial probability with samples composition". In: *Journal of Official Statistics* 11, pp. 379–390.

Chen, P., Tebbs, J., and Bilder, C. (2009). "Group testing regression models with fixed and random effects". In: *Biometrics* 65, pp. 1270–1278.

Cui, X., Hardle, W., and Zhu, L. (2011). "The EFM approach for single-index models". In: *The Annals of Statistics* 39, pp. 1658–1688.

Delaigle, A. and Hall, P. (2012). "Nonparametric regression with homogeneous group testing data". In: *Annal of Statistics* 40, pp. 131–158.

Delaigle, A. and Hall, P. (2015). "Nonparametric methods for group testing data, taking dilution into account". In: *Biometrika* 102, pp. 871–887.

Delaigle, A. and Meister, A. (2011). "Nonparametric regression analysis for group testing data". In: *Journal of the American Statistical Association* 106, pp. 640–650.

Delaigle, A. and Zhou, W. (2015). "Nonparametric and parametric estimators of prevalence from group testing data with aggregated covariates". In: *Journal of the American Statistical Association* 110, pp. 1785–1796.

Delaigle, A., Hall, P., and Wishart, J. (2014). "New approaches to nonparametric and semiparametric regression for univariate and multivariate group testing data". In: *Biometrika* 101, pp. 567–585.

Disease Control, Centers for and Prevention (Last accessed 2018[a]). "2016 STD surveillance report". In: Available at `https://www.cdc.gov/std/stats16/default.htm`.

Disease Control, Centers for and Prevention (Last accessed 2018[b]). "STDs & Infertility". In: Available at `https://www.cdc.gov/std/infertility/default.htm`.

Dodd, R., Notari, E., and Stramer, S. (2002). "Current prevalence and incidence of infectious disease markers and estimated window-period risk in the American Red Cross blood donor population". In: *Transfusion* 42, pp. 975–979.

Dorfman, R. (1943). "The detection of defective members of large populations". In: *Annals of Mathematical Statistics* 14, pp. 436–440.

Edouard, S., Prudent, E., Gautret, P., Memish, Z., and Raoult, D. (2015). "Cost-effective pooling of DNA from nasopharyngeal swab samples for large-scale detection of bacteria by real-time PCR". In: *Journal of Clinical Microbiology* 52, pp. 1002–1004.

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

Fan, J (1993). "Local linear regression smoothers and their minimax efficiencies". In: *The Annals of Statistics* 21, pp. 196–216.

Fan, J. and Gijbels, I. (1996). "Local polynomial modelling and its applications". In: *Monographs on Statistics and Applied Probability. Chapman & Hall/CRC* 66.

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

Faraggi, D., Reiser, B., and Schisterman, E. (2003). "ROC curve analysis for biomarkers based on pooled assessments". In: *Statistics in Medicine* 22, pp. 2515–2527.

Farrington, C. (1992). "Estimating prevalence by group testing using generalized linear models". In: *Statistics in Medicine* 11, pp. 1591–1597.

FDA-NIH, Biomarker-Working-Group (2016). *BEST (Biomarkers, EndpointS, and other Tools) resource*. Food and Drug Administration (US).

Gastwirth, J. (2000). "The efficiency of pooling in the detection of rare mutations". In: *American Journal of Human Genetics* 67, pp. 1036–1039.

Gregory, K., Wang, D., and McMahan, C. (in print). "Adaptive elastic net for group testing". In: *Biometrics* 00, pp. 1–13.

Gumbel, E. (1960). "Bivariate exponential distributions". In: *Journal of the American Statistical Association* 55, pp. 698–707.

Hanson, T., Johnson, W., and Gastwirth, J. (2006). "Bayesian inference for prevalence and diagnostic test accuracy based on dual-pooled screening". In: *Biostatistics* 7, pp. 41–57.

Hardle, W., Hall, P., and Ichimura, H. (1993). "Optimal smoothing in single-index models". In: *The Annals of Statistics* 21, pp. 157–178.

Hill, J., HallSedlak, R., Magaret, A., Huang, M., Zerr, D., Jerome, K., and Boeckh, M. (2016). "Efficient identification of inherited chromosomally integrated human herpesvirus 6 using specimen pooling". In: *Journal of Clinical Virology* 77, pp. 71–76.

Hou, P., Tebbs, J., Bilder, C., and McMahan, C. (2017). "Hierarchical group testing for multiple infections". In: *Biometrics* 73, pp. 656–665.

Huang, S., Huang, M. Lo, Shedden, K., and Wong, W. (2017). "Optimal group testing designs for estimating prevalence with uncertain testing errors". In: *Journal of the Royal Statistical Society: Series B* 79, pp. 1547–1563.

Huang, X. and Tebbs, J. (2009). "On Latent-Variable Model Misspecification in Structural Measurement Error Models for Binary Response". In: *Biometrics* 65, pp. 710–718.

Hughes-Oliver, J. and Rosenberger, W. (2000). "Efficient estimation of the prevalence of multiple rare traits". In: *Biometrika* 87, pp. 315–327.

Hughes-Oliver, J. and Swallow, W. (1994). "A two-stage adaptive group-testing procedure for estimating small proportions". In: *Journal of the American Statistical Association* 89, pp. 982–993.

Hui, F., Warton, D., and Foster, S. (2015). "Tuning parameter selection for the adaptive lasso using ERIC". In: *Journal of the American Statistical Association* 110, pp. 262–269.

Hung, M. and Swallow, W. (1999). "Robustness of group testing in the estimation of proportions". In: *Biometrics* 55, pp. 231–237.

Ichimura, H. (1993). "Semiparametric least squares (SLS) and weighted SLS estimation of single-index models". In: *Journal of Econometrics* 58, pp. 71–120.

Janes, H. and Pepe, M. (2008). "Adjusting for covariates in studies of diagnostic, screening, or prognostic markers: an old concept in a new setting". In: *American Journal of Epidemiology* 168, pp. 89–97.

Jirsa, S. (2008). "Pooling specimens: A decade of successful cost savings." In: *National STD Prevention Conference, 2008. Chicago, IL.*

Kim, H., Hudgens, M., Dreyfuss, J., Westreich, D., and Pilcher, C. (2007). "Comparison of group testing algorithms for case identification in the presence of testing error". In: *Biometrics* 63, pp. 1152–1163.

Klein, R. and Spady, R. (1993). "An efficient semiparametric estimator for binary response models". In: *Econometrica: Journal of the Econometric Society* 61, pp. 387–421.

Lehmann, E. (1983). *Theory of point estimation*. Pacific Grove, CA: Wadsworth and Brooks/Cole.

Lewis, J., Lockary, V., and Kobic, S. (2012). "Cost savings and increased efficiency using a stratified specimen pooling strategy for Chlamydia trachomatis and Neisseria gonorrhoeae". In: *Sexually Transmitted Diseases* 39, pp. 46–48.

Li, Q., Liu, A., and Xiong, W. (2017). "D-optimality of group testing for joint estimation of correlated rate diseases with misclassification". In: *Statistica Sinica* 27, pp. 823–838.

Lin, W. and Kulasekera, K. (2007). "Identifiability of single-index models and additive-index models". In: *Biometrika* 94, pp. 496–501.

Linton, O. and Whang, Y. (2002). "Nonparametric estimation with aggregated data". In: *Econometric Theory* 18, pp. 420–468.

Litvak, E., Tu, X., and Pagano, M. (1994). "Screening for the presence of a disease by pooling sera samples". In: *Journal of the American Statistical Association* 89, pp. 424–434.

Louis, T. (1982). "Finding the observed information matrix when using the EM algorithm". In: *Journal of the Royal Statistical Society. Series B* 44, pp. 226–233.

Loukopoulos, P., Mungall, B., Straw, R., Thornton, J., and Robinson, W. (2003). "Matrix metalloproteinase-2 and- 9 involvement in canine tumors". In: *Veterinary Pathology* 40, pp. 382–394.

Lyles, R., Van Domelen, D., Mitchell, E., and Schisterman, E. (2015). "A discriminant function approach to adjust for processing and measurement error when a biomarker is assayed in pooled samples". In: *International Journal of Environmental Research and Public Health* 12, pp. 14723–14740.

Ma, CX, Vexler, A., Schisterman, E., and Tian, L. (2011). "Cost-efficient designs based on linearly associated biomarkers". In: *Journal of Applied Statistics* 38, pp. 2739–2750.

Malinovsky, Y., Albert, P., and Schisterman, E. (2012). "Pooling designs for outcomes under a Gaussian random effects model". In: *Biometrics* 68, pp. 45–52.

McMahan, C., Tebbs, J., and Bilder, C. (2012a). "Informative Dorfman screening". In: *Biometrics* 68, pp. 287–296.

McMahan, C., Tebbs, J., and Bilder, C. (2012b). "Regression models for group testing data with pool dilution effects". In: *Biostatistics* 14, pp. 284–298.

McMahan, C., Tebbs, J., and Bilder, C. (2012c). "Two-dimensional informative array testing". In: *Biometrics* 68, pp. 793–804.

McMahan, C., McLain, A., Gallagher, C., and Schisterman, E. (2016). "Estimating covariate-adjusted measures of diagnostic accuracy based on pooled biomarker assessments". In: *Biometrical Journal* 58, pp. 944–961.

McMahan, C., Tebbs, J., Hanson, T., and Bilder, C. (2017). "Bayesian regression for group testing data". In: *Biometrics* 73, pp. 1443–1452.

Minnier, J., Tian, L., and Cai, T. (2011). "A perturbation method for inference on regularized regression estimates". In: *Journal of the American Statistical Association* 106, pp. 1371–1382.

Mitchell, E., Lyles, R., Manatunga, A., Danaher, M., Perkins, N., and Schisterman, E. (2014). "Regression for skewed biomarker outcomes subject to pooling". In: *Biometrics* 70, pp. 202–211.

Mitchell, E., Lyles, R., Manatunga, A., and Schisterman, E. (2015). "Semiparametric regression models for a right-skewed outcome subject to pooling". In: *American Journal of Epidemiology* 181, pp. 541–548.

Mumford, S., Schisterman, E., Vexler, A., and Liu, A. (2006). "Pooling biospecimens and limits of detection: effects on ROC curve analysis". In: *Biostatistics* 7, pp. 585–598.

Munoz-Zanzi, C., Johnson, W., Thurmond, M., and Hietala, S. (2000). "Pooled-sample testing as a herd-screening tool for detection of bovine viral diarrhea virus persistently infected cattle". In: *Journal of Veterinary Diagnostic Investigation* 12, pp. 195–203.

Neal, R. and Hinton, G. (1998). "A view of the EM algorithm that justifies incremental, sparse, and other variants". In: *Learning in Graphical Models*, pp. 355–368.

Nelsen, R. (2007). *An introduction to copulas*. New York: Springer Science Business Media, Inc.

Ortega, J. and Rheinboldt, W. (1970). *Iterative solution of nonlinear equations in several variables*. Vol. 30. Academic Press, New York and London.

Perrier, F., Giorgis-Allemand, L., Slama, R., and Philippat, C. (2016). "Within-subject pooling of biological samples to reduce exposure misclassification in biomarker-based studies". In: *Epidemiology* 27, pp. 378–388.

Petrov, V. (1995). *Limit theorems of probability theory: sequences of independent random variables*. Tech. rep. Oxford, New York.

Phatarfod, R. and Sudbury, A. (1994). "The use of a square array scheme in blood testing". In: *Statistics in Medicine* 13, pp. 2337–2343.

Remlinger, K., Hughes-Oliver, J., Young, S., and Lam, R. (2006). "Statistical design of pools using optimal coverage and minimal collision". In: *Technometrics* 48, pp. 133–143.

Samoff, E., Koumans, E., Markowitz, L., Sternberg, M., Sawyer, M., Swan, D., Papp, J., Black, C., and Unger, E. (2005). "Association of Chlamydia trachomatis with persistence of high-risk types of human papillomavirus in a cohort of female adolescents". In: *American Journal of Epidemiology* 162, pp. 668–675.

Schisterman, E., Vexler, A., Ye, A., and Perkins, N. (2011). "A combined efficient design for biomarker data subject to a limit of detection due to measuring instrument sensitivity". In: *The Annals of Applied Statistics* 5, pp. 2651–2667.

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

Sobel, M. and Elashoff, R. (1975). "Group testing with a new goal, estimation". In: *Biometrika* 62, pp. 182–193.

Sterrett, A. (1957). "On the detection of defective members of large populations". In: *Annals of Mathematical Statistics* 28, pp. 1033–1036.

Stramer, S., Krysztof, D., Brodsky, J., Fickett, T., Reynolds, B., Dodd, R., and Kleinman, S. (2013). "Comparative analysis of triplex nucleic acid test assays in United States blood donors". In: *Transfusion* 53, pp. 2525–2537.

Swallow, W. (1985). "Group testing for estimating infection rates and probabilities of disease transmission". In: *Phytopathology* 75, pp. 882–889.

Tebbs, J., McMahan, C., and Bilder, C. (2013). "Two-stage hierarchical group testing for multiple infections with application to the Infertility Prevention Project". In: *Biometrics* 69, pp. 1064–1073.

Thompson, K. (1962). "Estimation of the proportion of vectors in a natural population of insects". In: *Biometrics* 18, pp. 568–578.

Tong, S., Marjono, B., Brown, D., Mulvey, S., Breit, S., Manuelpillai, U., and Wallace, E. (2004). "Serum concentrations of macrophage inhibitory cytokine 1 (MIC 1) as a predictor of miscarriage". In: *The Lancet* 363, pp. 129–130.

Vansteelandt, S., Goetghebeur, E., and Verstraeten, T. (2000). "Regression models for disease prevalence with diagnostic tests on pools of serum samples". In: *Biometrics* 56, pp. 1126–1133.

Venette, R., Moon, R., and Hutchison, W. (2002). "Strategies and statistics of sampling for rare individuals". In: *Annual Review of Entomology* 47, pp. 143–174.

Vexler, A., Schisterman, E., and Liu, A. (2008). "Estimation of ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures". In: *Statistics in Medicine* 27, pp. 280–296.

Wang, B., Han, S., Cho, C., Han, J., Cheng, Y., Lee, S., Galappaththy, G., Thimasarn, K., Soe, M., Oo, H., Kyaw, M., and Han, E. (2014a). "Comparison of microscopy, nested-PCR, and Real-Time-PCR assays using high-throughput screening of pooled samples for diagnosis of malaria in asymptomatic carriers from areas of endemicity in Myanmar". In: *Journal of Clinical Microbiology* 52, pp. 1838–1845.

Wang, D., McMahan, C., Gallagher, C., and Kulasekera, K. (2014b). "Semiparametric group testing regression models". In: *Biometrika* 101, pp. 587–598.

Wang, D., McMahan, C., and Gallagher, C. (2015). "A general regression framework for group testing data, which incorporates pool dilution effects". In: *Statistics in Medicine* 34, pp. 3606–3621.

Wang, D., McMahan, C., Tebbs, J., and Bilder, C. (2018). "Group testing case identification with biomarker information". In: *Computational Statistics & Data Analysis* 122, pp. 156–166.

Wang, H. and Leng, C. (2007). "Unified LASSO estimation by least squares approximation". In: *Journal of the American Statistical Association* 102, pp. 1039–1048.

Wang, H., Li, B., and Leng, C. (2009). "Shrinkage tuning parameter selection with a diverging number of parameters". In: *Journal of the Royal Statistical Society: Series B* 71, pp. 671–683.

Wang, J., Xue, L., Zhu, L., and Chong, Y. (2010). "Estimation for a partial-linear single-index model". In: *The Annals of Statistics* 38, pp. 246–274.

Warasi, M., Tebbs, J., McMahan, C., and Bilder, C. (2016). "Estimating the prevalence of multiple diseases from two-stage hierarchical pooling". In: *Statistics in Medicine* 35, pp. 3851–3864.

Weinberg, C. and Umbach, D. (1999). "Using pooled exposure assessment to improve efficiency in case-control studies". In: *Biometrics* 55, pp. 718–726.

Weisberg, S. and Welsh, A. (1994). "Adapting for the missing link". In: *The Annals of Statistics* 22, pp. 1674–1700.

Whitcomb, B., Schisterman, E., Klebanoff, M., Baumgarten, M., Rhoton-Vlasak, A., Luo, X., and Chegini, N. (2007). "Circulating chemokine levels and miscarriage". In: *American Journal of Epidemiology* 166, pp. 323–331.

Whitcomb, B., Perkins, N., Zhang, Z., Ye, A., and Lyles, R. (2012). "Assessment of skewed exposure in case-control studies with pooling". In: *Statistics in Medicine* 31, pp. 2461–2472.

World Health Organization, Geneva (Last accessed 2018). "Global Health Estimates 2016: Disease burden by Cause, Age, Sex, by Country and by Region, 2000-2016." In: Available at `https://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html`.

Wu, C. (1983). "On the convergence properties of the EM algorithm". In: *The Annals of Statistics* 11, pp. 95–103.

Xia, Y. (2006). "Asymptotic distributions for two estimators of the single-index model". In: *Econometric Theory* 22, pp. 1112–1137.

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

Xie, M. (2001). "Regression analysis of group testing samples". In: *Statistics in Medicine* 20, pp. 1957–1969.

Zhang, B., Bilder, R., and Tebbs, J. (2013). "Regression analysis for multiple-disease group testing data". In: *Statistics in Medicine* 32, pp. 4954–4966.

Zhu, L. and Xue, L. (2006). "Empirical likelihood confidence regions in a partially linear single-index model". In: *Journal of the Royal Statistical Society: Series B* 68, pp. 549–570.

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

Appendix A

Chapter 2 Supplementary Materials

A.1 Proofs from Section 2.3

A.1.1 A description of the proofs of Theorem 2.1

In the following, we denote $a_N = O_P(b_N)$ if a_N/b_N is bounded in probability, and $a_N = o_P(b_N)$ if a_N/b_N converges to zero in probability. We further denote the summation over all the pools with size $c^{(m)}$ by $\sum_{|j|=c^{(m)}}$. Then $\sum_{j=1}^{J}\sum_{i=1}^{c_j}$ can be written as $\sum_{m=1}^{M}\sum_{i=1}^{c^{(m)}}\sum_{|j|=c^{(m)}}$. A term of the form $\sum_{|j|=c^{(m)}} A_j$ indicates that A_js are from pools of size $c^{(m)}$. Since the function $B(\boldsymbol{\beta}^{(1)}) = \boldsymbol{\beta}$ is a one-to-one mapping from $\boldsymbol{\mathcal{B}}^{(1)} = \{\boldsymbol{\beta}^{(1)} \in \mathbb{R}^{p-1} : \|\boldsymbol{\beta}^{(1)}\| < 1\}$ to $\boldsymbol{\mathcal{B}}$, $\hat{\boldsymbol{\beta}}$ can be viewed as $\hat{\boldsymbol{\beta}} = B(\hat{\boldsymbol{\beta}}^{(1)})$ where $\hat{\boldsymbol{\beta}}^{(1)}$ is the minimizer of $S\{B(\boldsymbol{\beta}^{(1)}), \hat{\eta}_{B(\boldsymbol{\beta}^{(1)})}(\cdot)\}$ in $\boldsymbol{\mathcal{B}}^{(1)}$. Denote $\hat{G}(\boldsymbol{\beta}^{(1)})$ as the partial derivative of $S\{B(\boldsymbol{\beta}^{(1)}), \hat{\eta}_{B(\boldsymbol{\beta}^{(1)})}(\cdot)\}$ with respect to $\boldsymbol{\beta}^{(1)}$. It could be written as

$$\hat{G}(\boldsymbol{\beta}^{(1)}) = -2\boldsymbol{\mathcal{J}}_{\boldsymbol{\beta}}^{\mathrm{T}} \sum_{j=1}^{J} \left\{ Z_j - \frac{1}{c_j}\sum_{i=1}^{c_j} \hat{\eta}_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) \right\} \sum_{i=1}^{c_j} \hat{\eta}_{\boldsymbol{\beta}}^{(1)}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}),$$

where $\boldsymbol{\mathcal{J}}_{\boldsymbol{\beta}} = \partial B(\boldsymbol{\beta}^{(1)})/\partial\boldsymbol{\beta}^{(1)}$ and $\hat{\eta}_{\boldsymbol{\beta}}^{(1)}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) = \partial\hat{\eta}_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})/\partial\boldsymbol{\beta}$. An asymptotically equivalent version of $\hat{G}$ could be written as

$$G(\boldsymbol{\beta}^{(1)}) = -2\boldsymbol{\mathcal{J}}_{\boldsymbol{\beta}}^{\mathrm{T}} \sum_{j=1}^{J} \left\{ Z_j - \frac{1}{c_j}\sum_{i=1}^{c_j} \eta_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) \right\} \sum_{i=1}^{c_j} \eta'_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) \left\{ X_{ij} - d_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) \right\},$$

where $d_{\boldsymbol{\beta}}(t) = E(\boldsymbol{X} \mid \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta} = t)$. In the next subsection, we derive that (in Lemma A.3)

$$\sup_{X\in\mathbb{X},\boldsymbol{\beta}^{(1)}\in\boldsymbol{\mathcal{B}}_N^{(1)}} |\hat{\eta}_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) - \eta_0(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}_0)| = O_p\{(Nh/\log N)^{-1/2}\} \tag{A.1}$$

and (in Lemma A.4)

$$\sup_{\boldsymbol{\beta}^{(1)} \in \boldsymbol{\mathcal{B}}_N^{(1)}} \|\hat{G}(\boldsymbol{\beta}^{(1)}) - G(\boldsymbol{\beta}_0^{(1)}) + 2N\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}_1\boldsymbol{\mathcal{J}}_0(\boldsymbol{\beta} - \boldsymbol{\beta}_0)\| = o_p(N^{1/2}), \tag{A.2}$$

where $\boldsymbol{\mathcal{B}}_N^{(1)} = \{\boldsymbol{\beta}^{(1)} \in \boldsymbol{\mathcal{B}}^{(1)} : \|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}_0^{(1)}\| \leq CN^{-1/2}\}$ for some constant $C > 0$ and $\boldsymbol{\Omega}_1 = \sum_{m=1}^{M} \gamma_m / c^{(m)} \times \boldsymbol{\Omega}$. Further, Lemma A.5 shows that that

$$\left\|\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}\right\| = O_p(N^{-1/2}). \tag{A.3}$$

Consequently,

$$\hat{G}(\hat{\boldsymbol{\beta}}^{(1)}) = G(\boldsymbol{\beta}_0^{(1)}) - 2N\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\mathcal{J}}_0(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}) + o_p(N^{1/2}).$$

Since $\hat{G}(\hat{\boldsymbol{\beta}}^{(1)}) = 0$, we have

$$N^{1/2}(\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}) = (\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}_1\boldsymbol{\mathcal{J}}_0)^{-1}\{-N^{1/2}G(\boldsymbol{\beta}_0^{(1)})/2\} + o_p(1).$$

Further, we have

$$\begin{aligned}
-2^{-1}N^{-1/2}G(\boldsymbol{\beta}_0^{(1)}) &= \boldsymbol{\mathcal{J}}_0^{\mathrm{T}}N^{-1/2}\sum_{j=1}^{J} A_j(\boldsymbol{\beta}_0)B_j(\boldsymbol{\beta}_0) \\
&= \boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\sum_{m=1}^{M}(c^{(m)}J_m/N)^{1/2}\{c^{(m)}\}^{-1/2}J_m^{-1/2}\sum_{|j|=c^{(m)}} A_j(\boldsymbol{\beta}_0)B_j(\boldsymbol{\beta}_0) \\
&\xrightarrow{d} N(0, \sigma^2\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}_1\boldsymbol{\mathcal{J}}_0),
\end{aligned}$$

where $A_j(\boldsymbol{\beta}) = Z_j - c_j^{-1}\sum_{i=1}^{c_j}\eta_\beta(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})$ and $B_j(\boldsymbol{\beta}) = \sum_{i=1}^{c_j}\eta'_\beta(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\{X_{ij} - d_\beta(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\}$. Then the asymptotic normality of $\hat{\boldsymbol{\beta}}$ follows. Combining (A.1) with (A.3) gives

$$\sup_{\boldsymbol{x}\in\mathbb{X}}\left|\hat{\eta}(\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}) - \eta_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\right| = O_p\{(Nh/\log N)^{-1/2}\},$$

which completes the proof of Theorem 2.1. In the next section we prove equations (A.1), (A.2), and (A.3).

A.1.2 A.1.2 Detailed proofs

Before proceeding to the detailed proofs, we would like to introduce some notations. We write $a_N = O(b_N)$ if a_N/b_N is bounded; $a_N = o(b_N)$ if a_N/b_N converges to

zero; $a_N \simeq b_N$ if $a_N/b_N = O(1)$; $a_N \xrightarrow{a.s.} a$ if a_N converges almost surely to a; and $a_N = \mathcal{O}_r(b_N)$, if $E(|a_N|^r) = O(b_N^r)$. $E_T(\boldsymbol{X})$ denotes the conditional expectation of X given T. By Cauchy-Schwartz inequality, we have $\mathcal{O}_r(a_N)\mathcal{O}_r(b_N) = \mathcal{O}_{r/2}(a_N b_N)$. Let $K_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}, t; l) = h^{-1}K\{(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - t)/h\}\{(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - t)/h\}^l$ for $l = 0, 1, 2$.

We first introduce a useful equation which would help us find the bounds for the centralized rth moments of $\hat{\eta}_\beta(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})$ and $\hat{\eta}_\beta^{(1)}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})$. Let $X_1, \ldots, X_n$ be independent random variables, and $r \geq 2$. Then

$$E\left(\left|\sum_{i=1}^n X_i\right|^r\right) \simeq \sum_{i=1}^n E\left(|X_i|^r\right) + \left|\sum_{i=1}^n E(X_i)\right|^r + \left\{\sum_{i=1}^n E(X_i^2)\right\}^{r/2}. \tag{A.4}$$

For the proof of (A.4) we refer to Petrov (1995).

Proposition A.1. *Under Conditions C1–C5, we have, for any $\boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}$ and $r \geq 2$,*

$$E|\hat{\eta}_\beta(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) - \eta_\beta(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})|^2 = O(h^4 + \{Nh\}^{-1})$$

and

$$E|\hat{\eta}'_\beta(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) - \eta'_\beta(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})|^2 = O(h^2 + \{Nh^3\}^{-1})$$

over all (i, j)s.

Proof. We only show the result for $\hat{\eta}'_\beta$ as the first result can be proven similarly, but more easily. Let $\boldsymbol{X}$ be one of $\boldsymbol{X}_{ij}$s. After a little algebra, we obtain

$$h\{\hat{\eta}'_\beta(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) - \eta'_\beta(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\} = \frac{\hat{H}_{N1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\beta})S_{N0}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\beta}) - \hat{H}_{N0}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\beta})S_{N1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\beta})}{S_{N0}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\beta})S_{N2}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\beta}) - S_{N1}^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\beta})},$$

where

$$\hat{H}_{Nl}(t, \boldsymbol{\beta}) = N^{-1}\sum_{j=1}^J \sum_{i=1}^{c_j} \left\{c_j Z_j - (c_j - 1)\hat{\mu} - \eta_\beta(t) - \eta'_\beta(t)(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - t)\right\} \mathcal{K}_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}, t; l).$$

It is easy to see that, for $s \in \{2, 4\}$,

$$\hat{H}_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\beta}) = H_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{\beta}) + \mathcal{O}_s(N^{-1/2}),$$

where $H_{Nl}(u,\boldsymbol{\beta})$ is the version of $\hat{H}_{Nl}(u,\boldsymbol{\beta})$ by replacing $\hat{\mu}$ with μ. Thus, it leaves us to show that $H_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) = \mathcal{O}_s(h^2 + \{Nh\}^{-1/2})$ for $s \in \{2,4\}$. To this end, we rewrite it as

$$H_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) = \sum_{m=1}^{M} \frac{c^{(m)}J_m}{N} \cdot \frac{1}{c^{(m)}} \sum_{i=1}^{c^{(m)}} H_{Nlmi}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}),$$

where $H_{Nlmi}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) = \sum_{|j|=c^{(m)}} H_{Nlmij}$ with $H_{Nlmij} = J_m^{-1}\{c_j Z_j - (c_j-1)\mu - \eta_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) - \eta'_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\}\mathcal{K}_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}, \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}; l)$. By (A.4), for $s \in \{2,4\}$, we have

$$E_{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\left\{|H_{Nlmi}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})|^s\right\} \simeq \left|\sum_{|j|=c^{(m)}} E_{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\left\{H_{Nlmij}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})\right\}\right|^s \tag{A.5}$$

$$+\sum_{j=1}^{J_m} E_{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\left\{|H_{Nlmij}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})|^s\right\} + \left[\sum_{j=1}^{J_m} E_{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\left\{H_{Nlmij}^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})\right\}\right]^{s/2}. \tag{A.6}$$

Simple Taylor expansion provides that $\sum_{|j|=c^{(m)}} E_{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\left\{H_{Nlmij}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})\right\} = O(h^2)$ which implies that the term (A.5) is also of order $O(h^{2s})$. Further, note that

$$\begin{aligned} E_{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\left(|H_{Nlmij}|^s\right) =& \frac{h}{J_m^s h^s}\int [\{c_j Z_j - (c_j-1)\mu - \eta_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) - \eta'_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})(u - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})]^s \\ & \times h^{-1}K^s\left(\frac{u-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}}{h}\right)\left(\frac{u-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}}{h}\right)^{ls} f_{\boldsymbol{\beta}}(u)du \\ =& O(J_m^{-s}h^{1-s}), \end{aligned}$$

where $f_{\boldsymbol{\beta}}$ is the density of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}$. Therefore, the term (A.6) is of order $O(J_m^{-s/2}h^{-s/2})$. Consequently, $E_{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\left\{|H_{Nlmi}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})|^s\right\} = O(h^{2s}) + O(J_m^{-s/2}h^{-s/2})$. Moreover, the boundedness of $\mathbb{X}$ and $\boldsymbol{\mathcal{B}}$ yields that $E\{|H_{Nlmi}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})|^s\} = O(h^{2s} + \{J_m h\}^{-s/2})$. Then

$$\hat{H}_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) = O_s(h^2 + \{Nh\}^{-1/2}),$$

for $s \in \{2,4\}$. Similarly as in Wang et al. (2014b), we can show that for any $\boldsymbol{\beta} \in \boldsymbol{\mathcal{B}}$ and $s \geq 2$,

$$\begin{aligned} S_{N0}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) &= f_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\pi_0 + \mathcal{O}_s(h^2 + \{Nh\}^{-1/2}), \\ S_{N1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) &= h f'_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\pi_2 + \mathcal{O}_s(h^2 + \{Nh\}^{-1/2}), \end{aligned}$$

and

$$S_{N2}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) = \quad f_{\beta}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\pi_2 + \mathcal{O}_s(h^2 + \{Nh\}^{-1/2}),$$

where $\pi_l = \int K(u)u^l du$, and (see Zhu and Xue, 2006)

$$\inf_{\boldsymbol{x}\in\mathbb{X}} \left| S_{N0}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{N2}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) - S_{N1}^2(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) \right| \geq C > 0 \text{ almost surely}$$

for some constant C. Then, we have

$$\hat{H}_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) = \mathcal{O}_4(h^2 + \{Nh\}^{-1/2})\mathcal{O}_4(1) = \mathcal{O}_2(h^2 + \{Nh\}^{-1/2}).$$

Thus we have

$$E|h\{\hat{\eta}'_{\beta}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) - \eta'_{\beta}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\}|^2 = O(h^4 + \{Nh\}^{-1})$$

and

$$E|\hat{\eta}'_{\beta}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) - \eta'_{\beta}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})|^2 = O(h^2 + \{Nh^3\}^{-1}).$$

□

Proposition A.2. *Let* $\boldsymbol{\mathcal{B}}_N^{(1)} = \{\boldsymbol{\beta}^{(1)} \in \boldsymbol{\mathcal{B}}^{(1)} : \|\boldsymbol{\beta}^{(1)} - \boldsymbol{\beta}_0^{(1)}\| \leq CN^{-1/2}\}$ *for some constant* $C > 0$. *Under Conditions C1–C5, for any* $\boldsymbol{\beta}^{(1)} \in \boldsymbol{\mathcal{B}}_N^{(1)}$, *we have*

$$E\left\|\hat{\eta}_{\beta}^{(1)}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) - \eta'_{\beta}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\left(\boldsymbol{X}_{ij} - d_{\beta}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\right)\right\|^2 = O(h^2 + \{Nh^3\}^{-1})$$

over all (i,j)*s.*

Proof. Let X be one of the X_{ij}s. After some algebra, $\hat{\eta}_{\beta}^{(1)}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})$ can be written as

$$\begin{aligned}\hat{\eta}_{\beta}^{(1)}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) \;=\; & \frac{\hat{R}_{N0}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{N2}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) - \hat{R}_{N1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{N1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})}{S_{N2}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{N0}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) - S_{N1}^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})} \\ & + \hat{\eta}'_{\beta}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\{X - \hat{d}_{\beta}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\}\end{aligned}$$

where

$$\begin{aligned}\hat{R}_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) = & N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}\{c_j Z_j - (c_j-1)\hat{\mu} - \hat{\eta}_{\beta}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) \\ & -\hat{\eta}'_{\beta}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\Big\}\,\partial\{\mathcal{K}_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta};l)\}/\partial\boldsymbol{\beta},\end{aligned}$$

$$\hat{d}_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) = \frac{U_{N0}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{N2}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) - U_{N1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{N1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})}{S_{N2}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{N0}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) - S_{N1}^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})},$$

and

$$U_{Nl}(t,\boldsymbol{\beta}) = N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}\boldsymbol{X}_{ij}\mathcal{K}_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta};l).$$

Note that $\hat{d}_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})$ acts like a local linear estimator of $d_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})$. Similar to the proof of Proposition A.1, we have $\|\hat{d}_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) - d_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\| = \mathcal{O}_s(h^2 + \{Nh\}^{-1/2})$ for all (i,j)s and $s \geq 2$, and $\sup_{\boldsymbol{x}\in\mathbb{X},\boldsymbol{\beta}\in\boldsymbol{\mathcal{B}}}\|\hat{d}_{\boldsymbol{\beta}}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}) - d_{\boldsymbol{\beta}}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})\|$ converges to 0 almost surely. Consequently,

$$E\left\|\hat{\eta}'_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\left\{X_{ij} - \hat{d}_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\right\} - \eta'_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\left\{X_{ij} - d_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\right\}\right\|^2 = O(h^2 + \{Nh^3\}^{-1}).$$

Simple algebra provides that $\hat{R}_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})$ can be decomposed as following.

$$\begin{aligned}
&\hat{R}_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) \\
&= R_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) + (\mu - \hat{\mu})B_{N1l}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) \\
&\quad + \{\eta_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) - \hat{\eta}_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\}B_{N2l}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) \\
&\quad + h\{\eta'_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) - \hat{\eta}'_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\}B_{N3l}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) \\
&= R_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) + (\mu - \hat{\mu})B_{N1l}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) \\
&\quad + \frac{\hat{H}_{N0}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{N2}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) - \hat{H}_{N1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{N1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})}{S_{N0}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{N2}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) - S_{N1}^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})}B_{N2l}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) \\
&\quad + \frac{\hat{H}_{N1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{N0}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) - \hat{H}_{N0}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{N1}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})}{S_{N0}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{N2}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) - S_{N1}^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})}B_{N3l}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}),
\end{aligned}$$

where

$$\begin{aligned}
R_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) &= N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}\{c_jZ_j - (c_j-1)\mu - \eta_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) \\
&\qquad -\eta'_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\}\frac{\partial\mathcal{K}_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta};l)}{\partial\boldsymbol{\beta}}, \\
B_{N1l}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) &= N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}(c_j-1)\frac{\partial\mathcal{K}_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta};l)}{\partial\boldsymbol{\beta}}, \\
B_{N2l}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) &= N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}\frac{\partial\mathcal{K}_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta};l)}{\partial\boldsymbol{\beta}},
\end{aligned}$$

and

$$B_{N3l}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) = N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}\left(\frac{\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}}{h}\right)\frac{\partial\mathcal{K}_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta};l)}{\partial\boldsymbol{\beta}}.$$

It is easy to see that $B_{Nil}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) = \mathcal{O}_s(1)$ for $i = 1, 2, 3$, and $s \geq 2$. We also have shown that $\hat{\mu} = \mu + \mathcal{O}_s(N^{1/2})$, $\hat{H}_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) = \mathcal{O}_s(h^2 + \{Nh\}^{-1/2})$ for $s \in \{2, 4\}$, $S_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) = \mathcal{O}_s(1)$ for $s \geq 2$, and $\inf_{\boldsymbol{x}\in\mathbb{X}} |S_{N0}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})S_{N2}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) - S_{N1}^2(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})| \geq C > 0$ almost surely. It suffices to show that $R_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) = \mathcal{O}_s(h^2 + \{Nh^3\}^{-1/2})$ for $s \in \{2, 4\}$ and $\boldsymbol{\beta}^{(1)} \in \boldsymbol{\mathcal{B}}_N^{(1)}$ and thus completes the proof of Proposition A.2.

Rewrite $R_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})$ as

$$R_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) = \sum_{m=1}^{M}\frac{c^{(m)}J_m}{N}\cdot\frac{1}{c^{(m)}}\sum_{i=1}^{c^{(m)}}R_{Nlmi}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})$$

where $R_{Nlmi}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta}) = \sum_{|j|=c^{(m)}} R_{Nlmij}$ and $R_{Nlmij} = J_m^{-1}\{c_jZ_j - (c_j-1)\mu - \eta_\beta(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) - \eta'_\beta(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\}\{(\boldsymbol{X}_{ij}-\boldsymbol{X})/h\}\phi_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta};l)$, where $\phi_h(\cdot;l) = h^{-1}\phi(\cdot/h;l)$ and $\phi(u;l) = K'(u)u^l + K(u)I(l=1)$. By (A.4), for $s \in \{2, 4\}$, we have

$$\begin{aligned}E_{\mathbf{X}^{\mathrm{T}}\beta}\{|R_{Nlmi}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})|^s\} \simeq{}& \left|\sum_{|j|=c^{(m)}} E_{\mathbf{X}^{\mathrm{T}}\beta}\{R_{Nlmij}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})\}\right|^s \\ &+ \sum_{j=1}^{J_m} E_{\mathbf{X}^{\mathrm{T}}\beta}\{|R_{Nlmij}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})|^s\} + \left[\sum_{j=1}^{J_m} E_{\mathbf{X}^{\mathrm{T}}\beta}\left\{R_{Nlmij}^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})\right\}\right]^{s/2}.\end{aligned}$$

Using the smoothness of $\eta_\beta(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})$ and the condition $\|\boldsymbol{\beta} - \boldsymbol{\beta}_0\| = O(N^{-1/2})$, we can evaluate $E_{\mathbf{X}^{\mathrm{T}}\beta}\{R_{Nlmij}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})\} = O(J_m^{-1}h^2)$ since

$$\begin{aligned}&E_{\mathbf{X}^{\mathrm{T}}\beta}\Big[\{c_jZ_j - (c_j-1)\mu - \eta_\beta(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) - \eta'_\beta(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\} \\ &\qquad\times\frac{\boldsymbol{X}_{ij}-\boldsymbol{X}}{h}\phi_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta};l)\Big] \\ ={}& E_{\mathbf{X}^{\mathrm{T}}\beta}\Big[\{\eta_{\beta_0}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_0) - \eta_\beta(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta}) - \eta'_\beta(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\} \\ &\qquad\times\frac{\boldsymbol{X}_{ij}-\boldsymbol{X}}{h}\phi_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta};l)\Big]\end{aligned}$$

$$
\begin{aligned}
&= E_{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\left[\frac{\boldsymbol{X}_{ij}-\boldsymbol{X}}{h}\phi_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta};l)\right]\times O(N^{-1/2})\\
&\quad + E_{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\Bigg[\{\eta_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})-\eta_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})-\eta'_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\}\\
&\qquad\qquad \times\frac{\boldsymbol{X}_{ij}-\boldsymbol{X}}{h}\phi_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta};l)\Bigg]\\
&= E_{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\left[\frac{d_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})-d_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})}{h}\phi_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta};l)\right]\times O(N^{-1/2})\\
&\quad + E_{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\Bigg[\{\eta_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})-\eta_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})-\eta'_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})\}\\
&\qquad\qquad \times\frac{d_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})-d_{\boldsymbol{\beta}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta})}{h}\phi_h(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}-\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta};l)\Bigg]\\
&= O(N^{-1/2})-O(h^2)=O(h^2),
\end{aligned}
$$

Further $E_{\mathbf{X}^{\mathrm{T}}\boldsymbol{\beta}}\{|R_{Nlmij}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})|^s\}=O(J_m^{-s}h^{1-2s})$. Thus $R_{Nlmi}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})=\mathcal{O}_s(h^2+\{J_mh^3\}^{-1/2})$ which finishes the proof of $R_{Nl}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\beta},\boldsymbol{\beta})=\mathcal{O}_s(h^2+\{Nh^3\}^{-1/2})$ for $s\in\{2,4\}$ and $\boldsymbol{\beta}^{(1)}\in\boldsymbol{\mathcal{B}}_N^{(1)}$ and thus completes the proof of Proposition A.2. □

Lemma A.3. *Under Conditions C1–C5, we have*

$$
\sup_{\boldsymbol{x}\in\mathbb{X},\boldsymbol{\beta}^{(1)}\in\boldsymbol{\mathcal{B}}_N^{(1)}}|\hat{\eta}_{\boldsymbol{\beta}}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})-\eta_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)|=O_p\{(Nh/\log N)^{-1/2}\}
$$

and

$$
\sup_{\boldsymbol{x}\in\mathbb{X},\boldsymbol{\beta}^{(1)}\in\boldsymbol{\mathcal{B}}_N^{(1)}}\left\|\hat{\eta}_{\boldsymbol{\beta}}^{(1)}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta})-\eta'_0(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\{X-d_{\boldsymbol{\beta}_0}(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{\beta}_0)\}\right\|=O_p\{(Nh^3/\log N)^{-1/2}\}
$$

where $\boldsymbol{\mathcal{B}}_N^{(1)}=\{\boldsymbol{\beta}^{(1)}\in\boldsymbol{\mathcal{B}}^{(1)}:\|\boldsymbol{\beta}^{(1)}-\boldsymbol{\beta}_0^{(1)}\|\le CN^{-1/2}\}$ *for some constant* $C>0$.

Proof. Using Propositions A.1 and A.2, this proof directly follows Lemma A.1 in Wang et al. (2010). □

Lemma A.4. *Under Conditions C1–C5, we have*

$$
\sup_{\boldsymbol{\beta}^1\in\boldsymbol{\mathcal{B}}_N^{(1)}}\left\|\hat{G}(\boldsymbol{\beta}^{(1)})-G(\boldsymbol{\beta}_0^{(1)})+2N\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\mathcal{J}}_0(\boldsymbol{\beta}-\boldsymbol{\beta}_0)\right\|=o_p(N^{1/2}).
$$

Proof. We firstly denote $A_j(\boldsymbol{\beta}) = Z_j - c_j^{-1}\sum_{i=1}^{c_j}\eta_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})$, $\hat{A}_j(\boldsymbol{\beta}) = Z_j - c_j^{-1}\sum_{i=1}^{c_j}\hat{\eta}_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})$, $B_j(\boldsymbol{\beta}) = \sum_{i=1}^{c_j}\eta'_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\{X_{ij} - d_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})\}$, and $\hat{B}_j(\boldsymbol{\beta}) = \sum_{i=1}^{c_j}\hat{\eta}^{(1)}_{\boldsymbol{\beta}}(\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})$. Then we have $\hat{G}(\boldsymbol{\beta}^{(1)}) = -2\boldsymbol{\mathcal{J}}_{\boldsymbol{\beta}}^{\mathrm{T}}\sum_{j=1}^{J}\hat{A}_j(\boldsymbol{\beta})\hat{B}_j(\boldsymbol{\beta})$ and $G(\boldsymbol{\beta}) = -2\boldsymbol{\mathcal{J}}_{\boldsymbol{\beta}}^{\mathrm{T}}\sum_{j=1}^{J}A_j(\boldsymbol{\beta})B_j(\boldsymbol{\beta})$. Further we have the following decomposition,

$$\begin{aligned}
\frac{1}{2}\{G(\boldsymbol{\beta}_0^{(1)}) - \hat{G}(\boldsymbol{\beta}^{(1)})\} = &(\boldsymbol{\mathcal{J}}_{\boldsymbol{\beta}}^{\mathrm{T}} - \boldsymbol{\mathcal{J}}_0^{\mathrm{T}})\sum_{m=1}^{M}\sum_{|j|=c^{(m)}}A_j(\boldsymbol{\beta}_0)B_j(\boldsymbol{\beta}_0)\\
&+\boldsymbol{\mathcal{J}}_{\boldsymbol{\beta}}^{\mathrm{T}}\sum_{j=1}^{J}\left\{\hat{A}_j(\boldsymbol{\beta}) - \hat{A}_j(\boldsymbol{\beta}_0)\right\}B_j(\boldsymbol{\beta}_0)\\
&+\boldsymbol{\mathcal{J}}_{\boldsymbol{\beta}}^{\mathrm{T}}\sum_{j=1}^{J}\left\{\hat{A}_j(\boldsymbol{\beta}_0) - A_j(\boldsymbol{\beta}_0)\right\}B_j(\boldsymbol{\beta}_0)\\
&+\boldsymbol{\mathcal{J}}_{\boldsymbol{\beta}}^{\mathrm{T}}\sum_{j=1}^{J}A_j(\boldsymbol{\beta}_0)\left\{\hat{B}_j(\boldsymbol{\beta}) - B_j(\boldsymbol{\beta}_0)\right\}\\
&+\boldsymbol{\mathcal{J}}_{\boldsymbol{\beta}}^{\mathrm{T}}\sum_{j=1}^{J}\left\{\hat{A}_j(\boldsymbol{\beta}) - A_j(\boldsymbol{\beta}_0)\right\}\times\left\{\hat{B}_j(\boldsymbol{\beta}) - B_j(\boldsymbol{\beta}_0)\right\}\\
=&I_1(\boldsymbol{\beta}^{(1)}) + I_2(\boldsymbol{\beta}^{(1)}) + I_3(\boldsymbol{\beta}^{(1)}) + I_4(\boldsymbol{\beta}^{(1)}) + I_5(\boldsymbol{\beta}^{(1)}). \quad \text{(A.7)}
\end{aligned}$$

Since $\boldsymbol{\mathcal{J}}_{\boldsymbol{\beta}} - \boldsymbol{\mathcal{J}}_0 = O(N^{-1/2})$ for all $\boldsymbol{\beta}^{(1)} \in \boldsymbol{\mathcal{B}}_N^{(1)}$, and $\sum_{|j|=c^{(m)}}A_j(\boldsymbol{\beta}_0)B_j(\boldsymbol{\beta}_0)$ is a sum of identical and independent random variables with mean zero and bounded covariance matrix,

$$\sup_{\boldsymbol{\beta}^{(1)}\in\boldsymbol{\mathcal{B}}_N^{(1)}}\left\|I_1(\boldsymbol{\beta}^{(1)})\right\| = o_p(N^{1/2}). \quad \text{(A.8)}$$

Considering $I_2(\boldsymbol{\beta}^{(1)})$, for a suitable $\bar{\boldsymbol{\beta}}^{(1)} \in \boldsymbol{\mathcal{B}}_N^{(1)}$, a Taylor expansion gives

$$I_2(\boldsymbol{\beta}^{(1)}) = -\boldsymbol{\mathcal{J}}_{\boldsymbol{\beta}}^{\mathrm{T}}\left\{\sum_{j=1}^{J}c_j^{-1}B_j(\boldsymbol{\beta}_0)\hat{B}_j(\bar{\boldsymbol{\beta}})^{\mathrm{T}}\right\}\boldsymbol{\mathcal{J}}_{\bar{\boldsymbol{\beta}}}(\boldsymbol{\beta} - \boldsymbol{\beta}_0).$$

By $\bar{\boldsymbol{\beta}}^{(1)} \in \boldsymbol{\mathcal{B}}_N^{(1)}$ and Lemma A.3, we have that $\sup_j\sup_{\boldsymbol{\beta}^{(1)}\in\boldsymbol{\mathcal{B}}_N^{(1)}}\|\hat{B}_j(\boldsymbol{\beta}) - B_j(\boldsymbol{\beta}_0)\| = o_p(1)$. Then

$$\begin{aligned}
\frac{1}{N}\sum_{j=1}^{J}B_j(\boldsymbol{\beta}_0)\hat{B}_j(\bar{\boldsymbol{\beta}})^{\mathrm{T}} &= \sum_{m=1}^{M}\frac{c^{(m)}J_m}{N}\times\frac{1}{c^{(m)2}}\times\frac{1}{J_m}\sum_{|j|=c^{(m)}}B_j(\boldsymbol{\beta}_0)B_j(\boldsymbol{\beta}_0)^{\mathrm{T}} + o_p(1)\\
&= \sum_{m=1}^{M}\frac{\gamma_m}{c^{(m)2}}E\{B_j(\boldsymbol{\beta}_0)B_j(\boldsymbol{\beta}_0)^{\mathrm{T}}\} + o_p(1)\\
&= \boldsymbol{\Omega}_1 + o_p(1).
\end{aligned}$$

Noticing that $\boldsymbol{\mathcal{J}}_{\boldsymbol{\beta}} = \boldsymbol{\mathcal{J}}_0 + O(N^{-1/2})$, $\boldsymbol{\mathcal{J}}_{\bar{\boldsymbol{\beta}}} = \boldsymbol{\mathcal{J}}_0 + O(N^{-1/2})$, and $\boldsymbol{\beta} - \boldsymbol{\beta}_0 = O(N^{-1/2})$, we obtain

$$\sup_{\boldsymbol{\beta}^{(1)} \in \boldsymbol{\mathcal{B}}_N^{(1)}} \left\| I_2(\boldsymbol{\beta}^{(1)}) + N \boldsymbol{\mathcal{J}}_0^{\mathrm{T}} \boldsymbol{\Omega}_1 \boldsymbol{\mathcal{J}}_0 (\boldsymbol{\beta} - \boldsymbol{\beta}_0) \right\| = o_p(N^{1/2}). \tag{A.9}$$

For $I_3(\boldsymbol{\beta}^{(1)})$, by Lemma A.3 we have

$$\begin{aligned}
&\left\| \sum_{j=1}^{J} \{\hat{A}_j(\boldsymbol{\beta}_0) - A_j(\boldsymbol{\beta}_0)\} B_j(\boldsymbol{\beta}_0) \right\| \\
&= \left\| \sum_{j=1}^{J} \frac{1}{c_j} \sum_{i=1}^{c_j} \left\{ \eta_{\boldsymbol{\beta}_0}(\boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) - \hat{\eta}_{\boldsymbol{\beta}_0}(\boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \right\} \sum_{i=1}^{c_j} \eta_0'(\boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \{ \boldsymbol{X}_{ij} - \boldsymbol{d}_{\boldsymbol{\beta}_0}(\boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \} \right\| \\
&\leq \sup_{\boldsymbol{x} \in \mathbb{X}} |\eta_{\boldsymbol{\beta}_0}(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_0) - \hat{\eta}_{\boldsymbol{\beta}_0}(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_0)| \times \left\| \sum_{j=1}^{J} \sum_{i=1}^{c_j} \eta_0'(\boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \{ \boldsymbol{X}_{ij} - \boldsymbol{d}_{\boldsymbol{\beta}_0}(\boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \} \right\| \\
&= O_p\{(Nh/\log N)^{-1/2}\} \times O_p(N^{1/2}) = o_p(N^{1/2}).
\end{aligned} \tag{A.10}$$

Similarly, for $I_4(\boldsymbol{\beta}^{(1)})$ we have

$$\begin{aligned}
&\left\| \sum_{j=1}^{J} A_j(\boldsymbol{\beta}_0) \left\{ \hat{B}_j(\boldsymbol{\beta}) - B_j(\boldsymbol{\beta}_0) \right\} \right\| \\
&= \left\| \sum_{j=1}^{J} A_j(\boldsymbol{\beta}_0) \sum_{i=1}^{c_j} \left\{ \hat{\eta}_{\boldsymbol{\beta}}^{(1)}(\boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta}) - \eta_{\boldsymbol{\beta}_0}'(\boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta}) (\boldsymbol{X}_{ij} - d_{\boldsymbol{\beta}_0}(\boldsymbol{X}_{ij}^{\mathrm{T}} \boldsymbol{\beta}_0) \right\} \right\| \qquad \text{(A.11)} \\
&\leq \left| \sum_{j=1}^{J} A_j(\boldsymbol{\beta}_0) \right| \times \max_m c^{(m)} \times \sup_{\boldsymbol{x} \in \mathbb{X}, \boldsymbol{\beta}^{(1)} \in \boldsymbol{\mathcal{B}}_N^{(1)}} \left\| \hat{\eta}_{\boldsymbol{\beta}}^{(1)}(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}) - \eta_{\boldsymbol{\beta}_0}'(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}) (\boldsymbol{x} - d_{\boldsymbol{\beta}_0}(\boldsymbol{x}^{\mathrm{T}} \boldsymbol{\beta}_0) \right\| \qquad \text{(A.12)} \\
&= O_p(N^{1/2}) \times O_p\{(Nh^3/\log N)^{-1/2}\} = o_p(N^{1/2}). \qquad \text{(A.13)}
\end{aligned}$$

The bound for $I_5(\boldsymbol{\beta}^{(1)})$ follows Lemma A.3 as

$$\begin{aligned}
\sup_{\boldsymbol{\beta}^{(1)} \in \boldsymbol{\mathcal{B}}_N^{(1)}} \left\| I_5(\boldsymbol{\beta}^{(1)}) \right\| &\leq J \sup_j \sup_{\boldsymbol{\beta}^{(1)} \in \boldsymbol{\mathcal{B}}_N^{(1)}} \left| \hat{A}_j(\boldsymbol{\beta}) - A_j(\boldsymbol{\beta}_0) \right| \\
&\quad \times p \times \sup_j \sup_{\boldsymbol{\beta}^{(1)} \in \boldsymbol{\mathcal{B}}_N^{(1)}} \left\| \hat{\boldsymbol{\mathcal{J}}}_{\boldsymbol{\beta}} \{ B_j(\boldsymbol{\beta}) - B_j(\boldsymbol{\beta}_0) \} \right\| \\
&= J \times O_p[\{\log N/(Nh)\}^{1/2}] \times O_p[\{\log N/(Nh^3)\}^{1/2}] \\
&= o_p(N^{1/2}).
\end{aligned} \tag{A.14}$$

Combining (A.7)-(A.14) completes the proof of Lemma A.4. □

Lemma A.5. *Under Condition C6, $\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\mathcal{J}}_0$ is a positve definite matrix. Further if Conditions C1–C5 are satisfied, we have*

$$\left\|\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}\right\| = O_p(N^{-1/2}).$$

Proof. By the definition of $\boldsymbol{\Omega}$, it can be seen that $\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\mathcal{J}}_0$ is a positive semidefinite matrix. It suffices to show that 0 is not one of its eigenvalues. By Condition $C6$, $(\boldsymbol{\mathcal{J}}_0 u)^{\mathrm{T}}\boldsymbol{\Omega}(\boldsymbol{\mathcal{J}}_0 u) = 0$ if and only if $\boldsymbol{\mathcal{J}}_0 u = r\boldsymbol{\beta}_0$ for some constant $r > 0$ where

$$\boldsymbol{\mathcal{J}}_0 = \begin{pmatrix} -\frac{\beta_2}{\sqrt{1-\|\boldsymbol{\beta}_0^{(1)}\|^2}} & \cdots & -\frac{\beta_p}{\sqrt{1-\|\boldsymbol{\beta}_0^{(1)}\|^2}} \\ 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}.$$

Solving $\boldsymbol{\mathcal{J}}_0 u = r\boldsymbol{\beta}_0$ results in $u = 0$ and thus $r = 0$. It is a contradiction to $r > 0$. This indicates that $\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}\boldsymbol{\mathcal{J}}_0$ is a positive definite matrix and so is $\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}_1\boldsymbol{\mathcal{J}}_0$.

To show $\|\hat{\boldsymbol{\beta}}^{(1)} - \boldsymbol{\beta}_0^{(1)}\| = O_p(N^{-1/2})$, by (6.3.4) on page 163 of Ortega and Rheinboldt (1970), which is also used by Weisberg and Welsh (1994) and Wang et al. (2010), it suffices to show that for any small probability τ, there always exists a constant $C > 0$, such that

$$\liminf_N P\left(\sup_{u \in U_N} u^{\mathrm{T}}\hat{G}(\boldsymbol{\beta}^{(1)}) < 0\right) = 1 - \tau, \tag{A.15}$$

where $U_N = \{u \in \mathbb{R}^{p-1} : (\boldsymbol{\beta}_0^{(1)} + u) \in \boldsymbol{\mathcal{B}}^{(1)}, N^{1/2}\|u\| = C\}$. Let $\lambda_{\min}$ be the smallest eigenvalue of $\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}_1\boldsymbol{\mathcal{J}}_0$. Then

$$\begin{aligned} u^{\mathrm{T}}G(\boldsymbol{\beta}_0^{(1)}) - 2Nu^{\mathrm{T}}\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}_1\boldsymbol{\mathcal{J}}_0 u \times \; &\leq \; \|N^{1/2}u\| \times \|N^{-1/2}G(\boldsymbol{\beta}_0^{(1)})\| - \lambda_{\min}\|N^{1/2}u\|^2 \\ &= \; C \times \|N^{-1/2}G(\boldsymbol{\beta}_0^{(1)})\| - \lambda_{\min} \times C^2. \end{aligned} \tag{A.16}$$

Noting that (A.16) is a quadratic function in C with $\lambda_{\min} > 0$ and $\|N^{-1/2}G(\boldsymbol{\beta}_0^{(1)})\| = O_p(1)$, for any $\tau > 0$, if C is chosen large enough, we have (A.16) being negative with

probability at least $1-\tau$. Further by Lemma A.4, we have

$$\sup_{u\in U_n}\left|u^{\mathrm{T}}\hat{G}(\boldsymbol{\beta}^{(1)})-\left\{u^{\mathrm{T}}G(\boldsymbol{\beta}_0^{(1)})-2Nu^{\mathrm{T}}\boldsymbol{\mathcal{J}}_0^{\mathrm{T}}\boldsymbol{\Omega}_1\boldsymbol{\mathcal{J}}_0u\right\}\right|=o_p(1).$$

This proves (A.15) and hence completes the proof. □

A.2 Additional results from Section 2.4.

This appendix provides a comprehensive summary of the simulation results obtained from the study described in Section 2.4 of Chapter 2.

A.1: Summary results, (M3)–(M4) when $V(Y_{ij} \mid \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) = 0.5^2$

Table A.1: (M3), fixed N

Table A.2: (M3), fixed J

Table A.3: (M4) when $a = 1$, fixed N

Table A.4: (M4) when $a = 1$, fixed J

Table A.5: (M4) when $a = 2$, fixed N

Table A.6: (M4) when $a = 2$, fixed J

A.2: Summary results, (M1)–(M4) when $V(Y_{ij} \mid \boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta}) = (0.5\boldsymbol{X}_{ij}^{\mathrm{T}}\boldsymbol{\beta})^2$

Table A.7: (M1), fixed N

Table A.8: (M1), fixed J

Table A.9: (M2), fixed N

Table A.10: (M2), fixed J

Table A.11: (M3), fixed N

Table A.12: (M3), fixed J

Table A.13: (M4) when $a = 1$, fixed N

Table A.14: (M4) when $a = 1$, fixed J

Table A.15: (M4) when $a = 2$, fixed N

Table A.16: (M4) when $a = 2$, fixed J

Table A.1: Simulation results of the estimators for (M3) using our proposed method and the parametric method proposed in McMahan et al. (2016). Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE×10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $N \in \{2500, 5000\}$.

			Proposed Method				Parametric Method			
	N		$c=1$	$c=2$	$c=5$	$c=10$	$c=1$	$c=2$	$c=5$	$c=10$
(D1)	2500	β_{01}	0.502(0.031)	0.505(0.035)	0.510(0.058)	0.522(0.075)	0.331(0.027)	0.331(0.035)	0.331(0.051)	0.329(0.073)
		β_{02}	0.503(0.031)	0.508(0.036)	0.508(0.057)	0.518(0.073)	0.399(0.053)	0.398(0.067)	0.393(0.107)	0.394(0.143)
		β_{03}	0.701(0.038)	0.694(0.043)	0.686(0.067)	0.666(0.080)	0.346(0.026)	0.347(0.038)	0.346(0.063)	0.344(0.084)
		AEE(MSE×10)	0.313(0.030)	0.312(0.031)	0.337(0.057)	0.339(0.093)	0.633(1.679)	0.635(1.684)	0.653(1.700)	0.677(1.726)
	5000	β_{01}	0.499(0.027)	0.500(0.032)	0.508(0.043)	0.513(0.057)	0.332(0.018)	0.332(0.024)	0.333(0.036)	0.334(0.051)
		β_{02}	0.498(0.027)	0.498(0.032)	0.506(0.045)	0.508(0.058)	0.397(0.036)	0.397(0.044)	0.404(0.071)	0.407(0.099)
		β_{03}	0.707(0.032)	0.706(0.040)	0.693(0.054)	0.684(0.069)	0.346(0.019)	0.345(0.026)	0.342(0.039)	0.341(0.054)
		AEE(MSE×10)	0.309(0.024)	0.319(0.025)	0.329(0.033)	0.338(0.093)	0.631(1.681)	0.633(1.684)	0.633(1.691)	0.645(1.703)
(D2)	2500	β_{01}	0.503(0.024)	0.504(0.035)	0.508(0.055)	0.518(0.078)	0.330(0.026)	0.331(0.034)	0.332(0.052)	0.330(0.074)
		β_{02}	0.504(0.026)	0.503(0.039)	0.504(0.055)	0.504(0.082)	0.396(0.053)	0.396(0.063)	0.400(0.095)	0.399(0.127)
		β_{03}	0.701(0.031)	0.698(0.045)	0.691(0.063)	0.676(0.088)	0.346(0.026)	0.346(0.036)	0.344(0.056)	0.345(0.082)
		AEE(MSE×10)	0.305(0.022)	0.316(0.029)	0.331(0.056)	0.355(0.098)	0.636(1.678)	0.637(1.682)	0.647(1.697)	0.667(1.720)
	5000	β_{01}	0.502(0.021)	0.503(0.029)	0.505(0.039)	0.514(0.054)	0.332(0.018)	0.333(0.023)	0.332(0.037)	0.331(0.049)
		β_{02}	0.503(0.021)	0.503(0.030)	0.506(0.041)	0.508(0.055)	0.397(0.037)	0.397(0.046)	0.395(0.068)	0.392(0.094)
		β_{03}	0.703(0.027)	0.700(0.037)	0.695(0.049)	0.684(0.064)	0.346(0.017)	0.346(0.025)	0.346(0.038)	0.343(0.059)
		AEE(MSE×10)	0.300(0.014)	0.308(0.020)	0.320(0.030)	0.324(0.045)	0.632(1.682)	0.632(1.685)	0.638(1.692)	0.651(1.704)
(D3)	2500	β_{01}	0.503(0.030)	0.507(0.040)	0.513(0.061)	0.523(0.077)	0.331(0.024)	0.331(0.032)	0.332(0.049)	0.331(0.070)
		β_{02}	0.503(0.030)	0.506(0.042)	0.511(0.059)	0.510(0.081)	0.393(0.051)	0.396(0.067)	0.394(0.098)	0.395(0.138)
		β_{03}	0.700(0.038)	0.693(0.051)	0.681(0.070)	0.669(0.084)	0.345(0.026)	0.345(0.036)	0.342(0.062)	0.338(0.085)
		AEE(MSE×10)	0.310(0.027)	0.320(0.037)	0.327(0.073)	0.344(0.095)	0.638(1.676)	0.638(1.680)	0.654(1.696)	0.679(1.722)
	5000	β_{01}	0.501(0.022)	0.502(0.029)	0.507(0.042)	0.513(0.057)	0.331(0.018)	0.331(0.025)	0.331(0.036)	0.330(0.051)
		β_{02}	0.502(0.023)	0.502(0.031)	0.506(0.043)	0.510(0.057)	0.401(0.037)	0.400(0.048)	0.400(0.071)	0.398(0.096)
		β_{03}	0.704(0.028)	0.702(0.038)	0.693(0.053)	0.682(0.069)	0.345(0.018)	0.347(0.025)	0.349(0.040)	0.349(0.058)
		AEE(MSE×10)	0.302(0.016)	0.315(0.021)	0.326(0.032)	0.334(0.049)	0.629(1.683)	0.630(1.686)	0.632(1.693)	0.643(1.705)

Table A.2: Simulation results of the estimators for (M3) using our proposed method and the parametric method proposed in McMahan et al. (2016). Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE$\times$10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $J \in \{250, 500\}$.

			Proposed Method				Parametric Method			
	J		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	250	β_{01}	0.518(0.074)	0.518(0.073)	0.518(0.075)	0.524(0.073)	0.327(0.083)	0.337(0.073)	0.331(0.073)	0.330(0.078)
		β_{02}	0.513(0.075)	0.512(0.073)	0.515(0.080)	0.518(0.079)	0.406(0.167)	0.398(0.149)	0.395(0.139)	0.390(0.140)
		β_{03}	0.671(0.085)	0.672(0.083)	0.669(0.087)	0.662(0.085)	0.344(0.086)	0.349(0.081)	0.346(0.083)	0.337(0.081)
		AEE(MSE$\times$10)	0.373(0.226)	0.373(0.125)	0.376(0.109)	0.362(0.083)	0.696(1.700)	0.669(3.399)	0.672(8.633)	0.687(17.36)
	500	β_{01}	0.510(0.054)	0.507(0.055)	0.511(0.055)	0.517(0.060)	0.334(0.058)	0.334(0.055)	0.331(0.052)	0.329(0.052)
		β_{02}	0.515(0.057)	0.507(0.056)	0.506(0.055)	0.512(0.060)	0.398(0.124)	0.395(0.105)	0.393(0.100)	0.403(0.097)
		β_{03}	0.682(0.065)	0.689(0.067)	0.687(0.066)	0.677(0.072)	0.347(0.059)	0.346(0.060)	0.344(0.059)	0.346(0.056)
		AEE(MSE$\times$10)	0.343(0.109)	0.361(0.075)	0.367(0.052)	0.368(0.054)	0.658(1.680)	0.651(3.382)	0.652(8.484)	0.646(16.99)
(D2)	250	β_{01}	0.517(0.054)	0.514(0.058)	0.512(0.067)	0.523(0.070)	0.337(0.077)	0.328(0.071)	0.326(0.070)	0.333(0.072)
		β_{02}	0.522(0.061)	0.515(0.065)	0.506(0.068)	0.516(0.069)	0.403(0.170)	0.406(0.153)	0.410(0.142)	0.406(0.141)
		β_{03}	0.672(0.055)	0.677(0.069)	0.683(0.077)	0.666(0.079)	0.353(0.072)	0.340(0.078)	0.347(0.078)	0.347(0.087)
		AEE(MSE$\times$10)	0.471(0.170)	0.388(0.098)	0.388(0.087)	0.367(0.067)	0.673(1.703)	0.684(3.387)	0.669(8.552)	0.666(17.27)
	500	β_{01}	0.516(0.042)	0.506(0.052)	0.512(0.054)	0.513(0.056)	0.333(0.056)	0.328(0.050)	0.332(0.051)	0.331(0.054)
		β_{02}	0.511(0.046)	0.505(0.054)	0.504(0.054)	0.509(0.055)	0.393(0.119)	0.380(0.095)	0.405(0.100)	0.394(0.095)
		β_{03}	0.683(0.046)	0.692(0.064)	0.689(0.063)	0.683(0.065)	0.348(0.056)	0.342(0.056)	0.346(0.057)	0.343(0.054)
		AEE(MSE$\times$10)	0.434(0.080)	0.371(0.064)	0.358(0.049)	0.357(0.045)	0.661(1.677)	0.666(3.331)	0.643(8.503)	0.653(16.99)
(D3)	250	β_{01}	0.526(0.072)	0.525(0.072)	0.522(0.074)	0.528(0.071)	0.333(0.085)	0.325(0.074)	0.332(0.072)	0.333(0.072)
		β_{02}	0.520(0.069)	0.514(0.074)	0.515(0.075)	0.516(0.073)	0.396(0.164)	0.398(0.151)	0.401(0.140)	0.407(0.144)
		β_{03}	0.660(0.083)	0.665(0.081)	0.666(0.087)	0.662(0.076)	0.346(0.083)	0.345(0.083)	0.349(0.084)	0.345(0.082)
		AEE(MSE$\times$10)	0.357(0.193)	0.358(0.121)	0.369(0.087)	0.371(0.073)	0.688(1.675)	0.686(3.393)	0.668(8.607)	0.669(17.28)
	500	β_{01}	0.509(0.058)	0.510(0.058)	0.511(0.057)	0.516(0.058)	0.332(0.055)	0.333(0.053)	0.335(0.049)	0.337(0.049)
		β_{02}	0.508(0.057)	0.507(0.057)	0.511(0.055)	0.512(0.056)	0.406(0.124)	0.402(0.104)	0.398(0.097)	0.402(0.098)
		β_{03}	0.686(0.069)	0.687(0.069)	0.684(0.065)	0.678(0.067)	0.348(0.058)	0.344(0.058)	0.347(0.059)	0.351(0.057)
		AEE(MSE$\times$10)	0.358(0.121)	0.361(0.080)	0.357(0.059)	0.350(0.048)	0.655(1.696)	0.647(3.387)	0.641(8.511)	0.633(17.04)

Table A.3: Simulation results of the estimators for (M4) when $a = 1$ using our proposed method and the parametric method proposed in McMahan et al. (2016). Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE×10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $N \in \{2500, 5000\}$.

			Proposed Method				Parametric Method			
	N		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	2500	β_{01}	0.500(0.007)	0.501(0.011)	0.501(0.018)	0.501(0.027)	0.480(0.024)	0.480(0.035)	0.479(0.052)	0.479(0.072)
		β_{02}	0.500(0.012)	0.500(0.018)	0.497(0.031)	0.496(0.047)	0.397(0.047)	0.400(0.065)	0.402(0.099)	0.402(0.135)
		β_{03}	0.707(0.007)	0.706(0.010)	0.707(0.017)	0.707(0.025)	1.020(0.024)	1.020(0.036)	1.021(0.052)	1.018(0.075)
		AEE(MSE×10)	0.288(0.016)	0.293(0.021)	0.307(0.035)	0.323(0.050)	0.441(1.475)	0.448(1.480)	0.475(1.494)	0.506(1.518)
	5000	β_{01}	0.500(0.005)	0.500(0.008)	0.500(0.013)	0.500(0.018)	0.485(0.016)	0.485(0.022)	0.486(0.036)	0.487(0.050)
		β_{02}	0.500(0.009)	0.499(0.012)	0.497(0.023)	0.496(0.033)	0.397(0.034)	0.398(0.044)	0.403(0.068)	0.404(0.096)
		β_{03}	0.707(0.005)	0.708(0.006)	0.708(0.012)	0.709(0.017)	1.022(0.017)	1.022(0.026)	1.022(0.039)	1.023(0.059)
		AEE(MSE×10)	0.285(0.009)	0.289(0.011)	0.300(0.020)	0.311(0.030)	0.435(1.474)	0.438(1.477)	0.448(1.484)	0.468(1.497)
(D2)	2500	β_{01}	0.500(0.008)	0.500(0.012)	0.501(0.019)	0.500(0.029)	0.481(0.022)	0.482(0.031)	0.482(0.049)	0.485(0.067)
		β_{02}	0.499(0.013)	0.499(0.020)	0.496(0.033)	0.494(0.050)	0.395(0.049)	0.396(0.063)	0.398(0.104)	0.400(0.141)
		β_{03}	0.708(0.007)	0.707(0.011)	0.708(0.017)	0.708(0.026)	1.020(0.025)	1.020(0.037)	1.020(0.057)	1.020(0.076)
		AEE(MSE×10)	0.290(0.017)	0.298(0.022)	0.309(0.037)	0.325(0.052)	0.441(1.477)	0.448(1.482)	0.477(1.497)	0.507(1.519)
	5000	β_{01}	0.500(0.005)	0.500(0.007)	0.501(0.012)	0.500(0.019)	0.484(0.016)	0.483(0.023)	0.485(0.035)	0.487(0.050)
		β_{02}	0.500(0.010)	0.499(0.013)	0.498(0.022)	0.498(0.033)	0.398(0.033)	0.397(0.045)	0.401(0.065)	0.397(0.098)
		β_{03}	0.707(0.005)	0.708(0.007)	0.707(0.012)	0.708(0.018)	1.021(0.018)	1.022(0.025)	1.019(0.042)	1.018(0.057)
		AEE(MSE×10)	0.285(0.009)	0.289(0.012)	0.299(0.020)	0.308(0.030)	0.435(1.472)	0.441(1.475)	0.446(1.482)	0.469(1.494)
(D3)	2500	β_{01}	0.500(0.007)	0.501(0.010)	0.501(0.018)	0.500(0.028)	0.483(0.024)	0.483(0.033)	0.484(0.048)	0.488(0.071)
		β_{02}	0.499(0.013)	0.499(0.019)	0.497(0.032)	0.496(0.048)	0.400(0.048)	0.400(0.065)	0.397(0.101)	0.396(0.145)
		β_{03}	0.708(0.007)	0.707(0.010)	0.708(0.017)	0.707(0.024)	1.018(0.025)	1.019(0.034)	1.020(0.055)	1.020(0.080)
		AEE(MSE×10)	0.289(0.017)	0.294(0.022)	0.308(0.036)	0.323(0.051)	0.436(1.477)	0.446(1.482)	0.472(1.496)	0.514(1.523)
	5000	β_{01}	0.500(0.005)	0.500(0.007)	0.500(0.012)	0.499(0.016)	0.484(0.016)	0.484(0.023)	0.488(0.036)	0.488(0.051)
		β_{02}	0.500(0.010)	0.499(0.013)	0.501(0.022)	0.504(0.027)	0.396(0.034)	0.398(0.046)	0.401(0.070)	0.397(0.096)
		β_{03}	0.707(0.005)	0.707(0.007)	0.706(0.012)	0.704(0.016)	1.021(0.018)	1.021(0.025)	1.021(0.043)	1.020(0.060)
		AEE(MSE×10)	0.285(0.009)	0.289(0.012)	0.295(0.020)	0.299(0.030)	0.437(1.477)	0.439(1.479)	0.448(1.487)	0.472(1.500)

Table A.4: Simulation results of the estimators for (M4) when $a = 1$ using our proposed method and the parametric method proposed in McMahan et al. (2016). Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE×10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $J \in \{250, 500\}$.

			Proposed Method				Parametric Method			
	J		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	250	β_{01}	0.502(0.026)	0.500(0.026)	0.497(0.027)	0.500(0.028)	0.491(0.073)	0.477(0.070)	0.485(0.070)	0.487(0.072)
		β_{02}	0.496(0.045)	0.497(0.046)	0.500(0.049)	0.496(0.050)	0.403(0.154)	0.399(0.152)	0.403(0.138)	0.397(0.139)
		β_{03}	0.706(0.024)	0.707(0.025)	0.707(0.026)	0.707(0.025)	1.028(0.083)	1.024(0.079)	1.023(0.082)	1.020(0.076)
		AEE(MSE×10)	0.327(0.147)	0.325(0.090)	0.330(0.060)	0.327(0.051)	0.524(1.469)	0.525(3.020)	0.509(7.566)	0.512(15.17)
	500	β_{01}	0.502(0.018)	0.501(0.018)	0.500(0.018)	0.501(0.018)	0.482(0.051)	0.483(0.052)	0.486(0.048)	0.483(0.051)
		β_{02}	0.495(0.032)	0.497(0.032)	0.497(0.031)	0.497(0.033)	0.394(0.113)	0.392(0.098)	0.390(0.100)	0.398(0.093)
		β_{03}	0.708(0.016)	0.707(0.017)	0.708(0.017)	0.707(0.018)	1.017(0.061)	1.018(0.054)	1.018(0.056)	1.020(0.057)
		AEE(MSE×10)	0.312(0.070)	0.311(0.047)	0.312(0.036)	0.310(0.031)	0.480(1.479)	0.477(2.980)	0.475(7.473)	0.470(14.92)
(D2)	250	β_{01}	0.501(0.025)	0.501(0.025)	0.500(0.026)	0.502(0.027)	0.479(0.074)	0.483(0.070)	0.481(0.070)	0.481(0.070)
		β_{02}	0.498(0.044)	0.497(0.044)	0.499(0.045)	0.498(0.047)	0.395(0.144)	0.407(0.140)	0.400(0.130)	0.381(0.135)
		β_{03}	0.705(0.024)	0.706(0.025)	0.705(0.024)	0.705(0.026)	1.021(0.078)	1.015(0.081)	1.021(0.080)	1.024(0.080)
		AEE(MSE×10)	0.321(0.148)	0.324(0.082)	0.324(0.060)	0.328(0.051)	0.519(1.484)	0.503(3.016)	0.502(7.600)	0.520(15.24)
	500	β_{01}	0.499(0.018)	0.501(0.018)	0.500(0.018)	0.500(0.018)	0.485(0.058)	0.487(0.049)	0.486(0.050)	0.481(0.049)
		β_{02}	0.499(0.031)	0.501(0.030)	0.499(0.031)	0.498(0.030)	0.394(0.106)	0.401(0.101)	0.404(0.092)	0.399(0.093)
		β_{03}	0.707(0.017)	0.705(0.016)	0.707(0.016)	0.707(0.016)	1.023(0.057)	1.022(0.054)	1.019(0.056)	1.023(0.054)
		AEE(MSE×10)	0.316(0.077)	0.306(0.045)	0.311(0.036)	0.310(0.031)	0.485(1.479)	0.473(2.958)	0.462(7.459)	0.471(14.93)
(D3)	250	β_{01}	0.499(0.026)	0.500(0.026)	0.500(0.027)	0.500(0.027)	0.488(0.075)	0.483(0.072)	0.484(0.073)	0.483(0.071)
		β_{02}	0.502(0.046)	0.496(0.045)	0.497(0.049)	0.496(0.048)	0.397(0.158)	0.401(0.142)	0.400(0.138)	0.400(0.143)
		β_{03}	0.704(0.024)	0.708(0.024)	0.706(0.025)	0.707(0.026)	1.026(0.084)	1.024(0.082)	1.021(0.080)	1.018(0.081)
		AEE(MSE×10)	0.324(0.148)	0.328(0.088)	0.330(0.060)	0.334(0.051)	0.531(1.485)	0.516(2.997)	0.507(7.570)	0.512(15.20)
	500	β_{01}	0.500(0.019)	0.501(0.017)	0.500(0.018)	0.500(0.018)	0.483(0.054)	0.485(0.051)	0.484(0.051)	0.483(0.050)
		β_{02}	0.498(0.032)	0.499(0.030)	0.499(0.031)	0.497(0.032)	0.405(0.109)	0.408(0.099)	0.399(0.102)	0.406(0.094)
		β_{03}	0.707(0.017)	0.706(0.017)	0.706(0.016)	0.708(0.017)	1.025(0.056)	1.025(0.054)	1.020(0.057)	1.020(0.055)
		AEE(MSE×10)	0.318(0.063)	0.309(0.045)	0.312(0.035)	0.310(0.030)	0.484(1.465)	0.470(2.960)	0.476(7.469)	0.464(14.93)

Table A.5: Simulation results of the estimators for (M4) when $a = 2$ using our proposed method and the parametric method proposed in McMahan et al. (2016). Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE×10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $N \in \{2500, 5000\}$.

			Proposed Method				Parametric Method			
	N		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	2500	β_{01}	0.500(0.004)	0.501(0.006)	0.501(0.010)	0.501(0.017)	−0.371(0.030)	−0.371(0.043)	−0.373(0.062)	−0.372(0.090)
		β_{02}	0.500(0.006)	0.501(0.009)	0.499(0.016)	0.498(0.025)	−0.002(0.059)	−0.001(0.082)	0.006(0.131)	−0.004(0.186)
		β_{03}	0.707(0.004)	0.706(0.006)	0.707(0.010)	0.707(0.018)	0.004(0.034)	0.001(0.046)	−0.002(0.072)	−0.002(0.100)
		AEE(MSE×10)	0.282(0.028)	0.284(0.036)	0.293(0.058)	0.301(0.101)	2.076(4.533)	2.079(4.541)	2.076(4.566)	2.085(4.607)
	5000	β_{01}	0.500(0.003)	0.500(0.004)	0.500(0.007)	0.501(0.010)	−0.369(0.021)	−0.371(0.029)	−0.372(0.043)	−0.370(0.062)
		β_{02}	0.500(0.005)	0.500(0.007)	0.499(0.011)	0.499(0.016)	0.000(0.042)	−0.002(0.060)	0.004(0.092)	0.010(0.129)
		β_{03}	0.707(0.003)	0.707(0.004)	0.708(0.007)	0.707(0.010)	0.000(0.025)	0.001(0.034)	0.000(0.054)	−0.006(0.071)
		AEE(MSE×10)	0.281(0.017)	0.283(0.022)	0.288(0.036)	0.291(0.052)	2.076(4.541)	2.079(4.546)	2.076(4.558)	2.073(4.578)
(D2)	2500	β_{01}	0.500(0.004)	0.501(0.006)	0.501(0.009)	0.500(0.015)	−0.371(0.028)	−0.371(0.039)	−0.372(0.067)	−0.367(0.087)
		β_{02}	0.500(0.007)	0.501(0.009)	0.499(0.015)	0.499(0.023)	0.006(0.061)	0.009(0.082)	0.012(0.133)	0.016(0.189)
		β_{03}	0.707(0.004)	0.706(0.006)	0.707(0.009)	0.707(0.016)	0.001(0.034)	0.002(0.047)	−0.003(0.074)	−0.005(0.106)
		AEE(MSE×10)	0.282(0.028)	0.284(0.035)	0.291(0.059)	0.301(0.088)	2.071(4.537)	2.068(4.546)	2.069(4.573)	2.063(4.615)
	5000	β_{01}	0.500(0.003)	0.500(0.004)	0.501(0.006)	0.501(0.009)	−0.370(0.022)	−0.370(0.030)	−0.370(0.048)	−0.375(0.067)
		β_{02}	0.500(0.004)	0.500(0.006)	0.499(0.011)	0.498(0.015)	0.003(0.042)	−0.001(0.057)	0.000(0.090)	0.002(0.123)
		β_{03}	0.707(0.003)	0.707(0.004)	0.707(0.006)	0.708(0.009)	−0.001(0.023)	0.000(0.033)	0.002(0.053)	0.003(0.081)
		AEE(MSE×10)	0.280(0.015)	0.282(0.021)	0.287(0.036)	0.292(0.054)	2.076(4.539)	2.078(4.543)	2.075(4.556)	2.077(4.579)
(D3)	2500	β_{01}	0.500(0.004)	0.500(0.006)	0.500(0.009)	0.499(0.015)	−0.371(0.030)	−0.371(0.040)	−0.371(0.067)	−0.369(0.098)
		β_{02}	0.500(0.006)	0.500(0.009)	0.498(0.016)	0.498(0.025)	−0.001(0.060)	−0.004(0.081)	−0.019(0.129)	−0.009(0.172)
		β_{03}	0.707(0.004)	0.707(0.006)	0.708(0.010)	0.708(0.017)	0.003(0.034)	0.001(0.047)	−0.001(0.076)	−0.003(0.105)
		AEE(MSE×10)	0.282(0.028)	0.285(0.037)	0.292(0.059)	0.302(0.094)	2.076(4.538)	2.081(4.546)	2.098(4.573)	2.088(4.616)
	5000	β_{01}	0.500(0.003)	0.500(0.004)	0.500(0.007)	0.499(0.010)	−0.371(0.020)	−0.373(0.029)	−0.372(0.047)	−0.372(0.068)
		β_{02}	0.500(0.005)	0.500(0.007)	0.500(0.011)	0.501(0.016)	−0.004(0.042)	−0.004(0.057)	−0.009(0.093)	−0.008(0.129)
		β_{03}	0.707(0.003)	0.707(0.004)	0.707(0.007)	0.707(0.010)	0.001(0.024)	0.001(0.034)	−0.001(0.054)	−0.003(0.077)
		AEE(MSE×10)	0.281(0.016)	0.283(0.021)	0.286(0.036)	0.293(0.054)	2.081(4.540)	2.082(4.544)	2.090(4.557)	2.090(4.580)

Table A.6: Simulation results of the estimators for (M4) when $a = 2$ using our proposed method and the parametric method proposed in McMahan et al. (2016). Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE×10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $J \in \{250, 500\}$.

			Proposed Method				Parametric Method			
	J		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	250	β_{01}	0.501(0.013)	0.501(0.014)	0.500(0.014)	0.501(0.014)	−0.364(0.093)	−0.370(0.090)	−0.373(0.092)	−0.365(0.092)
		β_{02}	0.498(0.023)	0.498(0.022)	0.498(0.025)	0.498(0.023)	−0.005(0.191)	−0.020(0.177)	−0.013(0.175)	−0.019(0.180)
		β_{03}	0.707(0.015)	0.707(0.014)	0.708(0.015)	0.707(0.015)	0.009(0.108)	0.009(0.110)	−0.004(0.108)	0.005(0.111)
		AEE(MSE×10)	0.354(0.230)	0.351(0.138)	0.356(0.104)	0.345(0.092)	2.068(4.517)	2.089(9.118)	2.098(22.99)	2.086(46.21)
	500	β_{01}	0.501(0.009)	0.500(0.009)	0.500(0.009)	0.501(0.010)	−0.372(0.068)	−0.374(0.064)	−0.368(0.063)	−0.371(0.062)
		β_{02}	0.498(0.015)	0.499(0.014)	0.498(0.016)	0.499(0.015)	−0.007(0.129)	0.004(0.133)	−0.009(0.123)	0.002(0.131)
		β_{03}	0.707(0.009)	0.708(0.009)	0.708(0.010)	0.707(0.010)	−0.003(0.072)	−0.003(0.073)	0.000(0.075)	0.002(0.074)
		AEE(MSE×10)	0.298(0.113)	0.294(0.076)	0.296(0.062)	0.295(0.051)	2.089(4.522)	2.080(9.101)	2.084(22.86)	2.075(45.82)
(D2)	250	β_{01}	0.502(0.015)	0.501(0.013)	0.501(0.014)	0.500(0.015)	−0.371(0.096)	−0.376(0.087)	−0.372(0.096)	−0.374(0.091)
		β_{02}	0.497(0.022)	0.497(0.022)	0.498(0.022)	0.500(0.021)	−0.003(0.190)	−0.011(0.178)	−0.002(0.177)	−0.001(0.172)
		β_{03}	0.707(0.013)	0.708(0.014)	0.707(0.014)	0.707(0.014)	0.001(0.103)	−0.007(0.111)	−0.004(0.105)	−0.007(0.100)
		AEE(MSE×10)	0.303(0.375)	0.302(0.130)	0.304(0.101)	0.303(0.090)	2.080(4.496)	2.101(9.107)	2.085(22.93)	2.090(46.02)
	500	β_{01}	0.502(0.009)	0.500(0.010)	0.500(0.009)	0.501(0.010)	−0.374(0.069)	−0.373(0.068)	−0.370(0.068)	−0.371(0.064)
		β_{02}	0.498(0.015)	0.499(0.015)	0.499(0.015)	0.498(0.015)	0.001(0.136)	−0.002(0.124)	0.007(0.120)	0.000(0.129)
		β_{03}	0.707(0.009)	0.707(0.009)	0.707(0.009)	0.708(0.010)	−0.008(0.077)	0.000(0.077)	−0.001(0.075)	−0.002(0.077)
		AEE(MSE×10)	0.293(0.113)	0.294(0.072)	0.294(0.060)	0.295(0.053)	2.088(4.517)	2.082(9.098)	2.072(22.84)	2.081(45.74)
(D3)	250	β_{01}	0.501(0.014)	0.499(0.013)	0.500(0.013)	0.500(0.015)	−0.366(0.095)	−0.379(0.096)	−0.370(0.090)	−0.376(0.090)
		β_{02}	0.498(0.021)	0.500(0.022)	0.499(0.022)	0.498(0.024)	−0.007(0.193)	−0.001(0.187)	0.011(0.175)	0.003(0.192)
		β_{03}	0.707(0.014)	0.707(0.013)	0.707(0.014)	0.707(0.015)	0.005(0.111)	−0.001(0.109)	0.002(0.108)	−0.004(0.112)
		AEE(MSE×10)	0.306(0.248)	0.303(0.135)	0.301(0.106)	0.306(0.088)	2.076(4.503)	2.088(9.132)	2.064(22.95)	2.085(46.15)
	500	β_{01}	0.500(0.009)	0.500(0.009)	0.501(0.009)	0.501(0.010)	−0.375(0.068)	−0.363(0.066)	−0.369(0.064)	−0.368(0.065)
		β_{02}	0.501(0.014)	0.499(0.015)	0.499(0.016)	0.497(0.016)	−0.008(0.136)	−0.007(0.134)	0.005(0.127)	−0.002(0.128)
		β_{03}	0.706(0.008)	0.708(0.009)	0.707(0.010)	0.708(0.010)	0.001(0.075)	0.000(0.078)	−0.001(0.074)	0.000(0.072)
		AEE(MSE×10)	0.294(0.116)	0.294(0.077)	0.295(0.060)	0.297(0.054)	2.089(4.514)	2.078(9.112)	2.072(22.86)	2.078(45.82)

Table A.7: Simulation results of the estimators for (M1) using our proposed method and the parametric method proposed in McMahan et al. (2016) when $\sigma^2(t) = (0.5t)^2$. Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE$\times$10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $N \in \{2500, 5000\}$.

			Proposed Method				Parametric Method			
	N		$c=1$	$c=2$	$c=5$	$c=10$	$c=1$	$c=2$	$c=5$	$c=10$
(D1)	2500	β_{01}	0.500(0.006)	0.500(0.010)	0.500(0.019)	0.499(0.029)	0.501(0.009)	0.502(0.013)	0.501(0.020)	0.501(0.028)
		β_{02}	0.501(0.012)	0.500(0.017)	0.500(0.029)	0.501(0.044)	0.500(0.016)	0.500(0.024)	0.501(0.037)	0.501(0.054)
		β_{03}	0.706(0.006)	0.706(0.010)	0.706(0.017)	0.705(0.028)	0.708(0.010)	0.708(0.014)	0.709(0.022)	0.708(0.030)
		AEE(MSE$\times$10)	0.287(0.003)	0.292(0.002)	0.306(0.002)	0.320(0.003)	0.029(0.001)	0.040(0.002)	0.062(0.004)	0.087(0.008)
	5000	β_{01}	0.500(0.004)	0.500(0.007)	0.499(0.012)	0.499(0.024)	0.500(0.007)	0.500(0.009)	0.500(0.013)	0.499(0.019)
		β_{02}	0.500(0.008)	0.500(0.012)	0.501(0.020)	0.499(0.033)	0.501(0.012)	0.501(0.016)	0.502(0.025)	0.503(0.036)
		β_{03}	0.707(0.004)	0.707(0.007)	0.707(0.012)	0.707(0.021)	0.707(0.007)	0.707(0.010)	0.707(0.016)	0.708(0.022)
		AEE(MSE$\times$10)	0.283(0.001)	0.288(0.001)	0.296(0.001)	0.310(0.005)	0.020(0.001)	0.027(0.001)	0.043(0.002)	0.061(0.004)
(D2)	2500	β_{01}	0.500(0.006)	0.500(0.011)	0.500(0.018)	0.500(0.028)	0.500(0.010)	0.500(0.013)	0.500(0.019)	0.500(0.027)
		β_{02}	0.502(0.012)	0.500(0.017)	0.500(0.029)	0.499(0.044)	0.502(0.017)	0.499(0.024)	0.499(0.037)	0.501(0.052)
		β_{03}	0.706(0.005)	0.707(0.010)	0.706(0.017)	0.705(0.027)	0.707(0.009)	0.707(0.013)	0.707(0.021)	0.706(0.031)
		AEE(MSE$\times$10)	0.286(0.003)	0.294(0.002)	0.304(0.002)	0.317(0.003)	0.029(0.001)	0.039(0.002)	0.061(0.004)	0.087(0.007)
	5000	β_{01}	0.500(0.004)	0.500(0.017)	0.500(0.013)	0.499(0.023)	0.500(0.007)	0.500(0.009)	0.500(0.013)	0.498(0.018)
		β_{02}	0.501(0.008)	0.499(0.020)	0.501(0.021)	0.500(0.033)	0.500(0.013)	0.500(0.017)	0.503(0.027)	0.501(0.035)
		β_{03}	0.707(0.004)	0.707(0.014)	0.706(0.012)	0.706(0.022)	0.707(0.007)	0.707(0.010)	0.707(0.016)	0.705(0.022)
		AEE(MSE$\times$10)	0.283(0.001)	0.291(0.006)	0.296(0.001)	0.308(0.004)	0.020(0.001)	0.029(0.001)	0.044(0.002)	0.060(0.004)
(D3)	2500	β_{01}	0.498(0.029)	0.500(0.014)	0.499(0.027)	0.502(0.029)	0.500(0.010)	0.501(0.013)	0.500(0.020)	0.499(0.029)
		β_{02}	0.498(0.031)	0.499(0.021)	0.497(0.035)	0.499(0.043)	0.499(0.017)	0.499(0.023)	0.500(0.036)	0.501(0.051)
		β_{03}	0.708(0.024)	0.707(0.014)	0.708(0.023)	0.704(0.027)	0.707(0.009)	0.707(0.013)	0.707(0.021)	0.706(0.031)
		AEE(MSE$\times$10)	0.293(0.016)	0.294(0.003)	0.310(0.009)	0.318(0.003)	0.028(0.001)	0.039(0.002)	0.063(0.004)	0.088(0.008)
	5000	β_{01}	0.500(0.004)	0.500(0.007)	0.500(0.012)	0.500(0.018)	0.500(0.006)	0.500(0.009)	0.500(0.014)	0.500(0.019)
		β_{02}	0.500(0.008)	0.500(0.013)	0.499(0.020)	0.499(0.030)	0.500(0.012)	0.500(0.017)	0.501(0.026)	0.502(0.036)
		β_{03}	0.707(0.004)	0.707(0.007)	0.707(0.012)	0.707(0.019)	0.707(0.007)	0.708(0.009)	0.709(0.016)	0.708(0.022)
		AEE(MSE$\times$10)	0.283(0.001)	0.289(0.001)	0.296(0.001)	0.308(0.001)	0.020(0.001)	0.028(0.001)	0.045(0.002)	0.061(0.004)

Table A.8: Simulation results of the estimators for (M1) using our proposed method and the parametric method proposed in McMahan et al. (2016) when $\sigma^2(t) = (0.5t)^2$. Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE×10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $J \in \{250, 500\}$.

			Proposed Method				Parametric Method			
	J		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	250	β_{01}	0.499(0.021)	0.500(0.026)	0.499(0.026)	0.500(0.029)	0.499(0.029)	0.499(0.028)	0.499(0.027)	0.501(0.027)
		β_{02}	0.503(0.040)	0.500(0.042)	0.500(0.042)	0.494(0.045)	0.494(0.056)	0.501(0.053)	0.500(0.050)	0.498(0.051)
		β_{03}	0.704(0.018)	0.705(0.024)	0.705(0.025)	0.708(0.028)	0.706(0.030)	0.708(0.029)	0.707(0.030)	0.706(0.030)
		AEE(MSE×10)	0.310(0.040)	0.316(0.013)	0.320(0.004)	0.323(0.003)	0.091(0.010)	0.088(0.017)	0.086(0.037)	0.086(0.073)
	500	β_{01}	0.500(0.019)	0.500(0.017)	0.499(0.023)	0.500(0.019)	0.499(0.010)	0.500(0.021)	0.499(0.020)	0.499(0.019)
		β_{02}	0.499(0.030)	0.502(0.028)	0.498(0.031)	0.500(0.029)	0.500(0.039)	0.501(0.035)	0.498(0.038)	0.499(0.036)
		β_{03}	0.707(0.016)	0.705(0.016)	0.707(0.021)	0.707(0.018)	0.707(0.021)	0.707(0.021)	0.707(0.022)	0.707(0.022)
		AEE(MSE×10)	0.301(0.019)	0.302(0.006)	0.308(0.006)	0.307(0.001)	0.064(0.005)	0.061(0.008)	0.064(0.020)	0.060(0.036)
(D2)	250	β_{01}	0.499(0.022)	0.499(0.024)	0.499(0.025)	0.499(0.028)	0.502(0.031)	0.500(0.028)	0.498(0.028)	0.501(0.027)
		β_{02}	0.505(0.040)	0.501(0.042)	0.498(0.041)	0.500(0.042)	0.503(0.055)	0.498(0.052)	0.502(0.053)	0.503(0.051)
		β_{03}	0.703(0.020)	0.705(0.023)	0.707(0.026)	0.705(0.026)	0.710(0.031)	0.709(0.032)	0.707(0.027)	0.709(0.031)
		AEE(MSE×10)	0.309(0.048)	0.313(0.011)	0.318(0.005)	0.316(0.003)	0.093(0.011)	0.089(0.017)	0.087(0.037)	0.087(0.074)
	500	β_{01}	0.499(0.024)	0.499(0.025)	0.501(0.018)	0.500(0.018)	0.498(0.022)	0.500(0.020)	0.500(0.019)	0.501(0.018)
		β_{02}	0.500(0.035)	0.501(0.034)	0.499(0.028)	0.500(0.029)	0.497(0.039)	0.499(0.036)	0.500(0.035)	0.501(0.037)
		β_{03}	0.707(0.020)	0.706(0.022)	0.706(0.018)	0.706(0.018)	0.706(0.021)	0.708(0.022)	0.706(0.020)	0.707(0.023)
		AEE(MSE×10)	0.305(0.023)	0.305(0.010)	0.304(0.002)	0.305(0.001)	0.066(0.005)	0.062(0.008)	0.060(0.018)	0.064(0.038)
(D3)	250	β_{01}	0.499(0.021)	0.500(0.023)	0.500(0.028)	0.499(0.027)	0.501(0.029)	0.498(0.028)	0.501(0.025)	0.500(0.027)
		β_{02}	0.505(0.037)	0.499(0.040)	0.498(0.043)	0.496(0.044)	0.500(0.056)	0.495(0.054)	0.498(0.053)	0.499(0.050)
		β_{03}	0.703(0.017)	0.706(0.023)	0.706(0.026)	0.708(0.028)	0.709(0.029)	0.705(0.031)	0.709(0.030)	0.705(0.032)
		AEE(MSE×10)	0.310(0.042)	0.317(0.011)	0.322(0.005)	0.323(0.003)	0.091(0.010)	0.091(0.017)	0.085(0.037)	0.087(0.074)
	500	β_{01}	0.499(0.021)	0.499(0.017)	0.500(0.018)	0.501(0.019)	0.500(0.021)	0.500(0.020)	0.499(0.020)	0.500(0.018)
		β_{02}	0.502(0.032)	0.501(0.029)	0.501(0.028)	0.499(0.029)	0.501(0.039)	0.502(0.037)	0.501(0.036)	0.501(0.036)
		β_{03}	0.706(0.018)	0.706(0.016)	0.705(0.018)	0.706(0.018)	0.706(0.021)	0.707(0.022)	0.707(0.022)	0.708(0.022)
		AEE(MSE×10)	0.302(0.024)	0.306(0.005)	0.304(0.002)	0.305(0.001)	0.065(0.005)	0.063(0.009)	0.063(0.020)	0.061(0.036)

Table A.9: Simulation results of the estimators for (M2) using our proposed method and the parametric method proposed in McMahan et al. (2016) when $\sigma^2(t) = (0.5t)^2$. Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE×10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $N \in \{2500, 5000\}$.

			Proposed Method				Parametric Method			
	N		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	2500	β_{01}	0.500(0.037)	0.504(0.025)	0.503(0.043)	0.507(0.066)	0.000(0.015)	0.001(0.020)	−0.001(0.028)	0.000(0.042)
		β_{02}	0.502(0.040)	0.506(0.033)	0.505(0.050)	0.502(0.070)	0.000(0.028)	0.000(0.038)	−0.005(0.058)	−0.005(0.077)
		β_{03}	0.703(0.036)	0.698(0.028)	0.696(0.049)	0.690(0.071)	0.001(0.014)	0.001(0.019)	0.001(0.031)	0.000(0.047)
		AEE(MSE×10)	0.303(0.023)	0.301(0.010)	0.323(0.015)	0.353(0.025)	1.706(0.671)	1.705(0.672)	1.713(0.677)	1.712(0.686)
	5000	β_{01}	0.502(0.018)	0.493(0.047)	0.503(0.028)	0.504(0.047)	0.000(0.011)	0.000(0.014)	0.001(0.021)	0.001(0.028)
		β_{02}	0.502(0.018)	0.493(0.047)	0.503(0.028)	0.504(0.047)	0.001(0.021)	0.000(0.027)	0.000(0.041)	−0.001(0.057)
		β_{03}	0.702(0.018)	0.710(0.048)	0.700(0.034)	0.701(0.053)	0.000(0.011)	0.001(0.015)	0.001(0.023)	0.002(0.032)
		AEE(MSE×10)	0.289(0.009)	0.324(0.031)	0.313(0.006)	0.337(0.010)	1.705(0.668)	1.707(0.669)	1.705(0.671)	1.706(0.675)
(D2)	2500	β_{01}	0.498(0.038)	0.501(0.024)	0.502(0.047)	0.506(0.062)	0.000(0.015)	0.001(0.019)	0.002(0.030)	0.003(0.040)
		β_{02}	0.500(0.041)	0.506(0.029)	0.503(0.053)	0.502(0.073)	0.000(0.027)	0.000(0.035)	0.001(0.052)	−0.003(0.076)
		β_{03}	0.705(0.038)	0.700(0.026)	0.698(0.052)	0.692(0.066)	0.000(0.014)	0.000(0.020)	0.000(0.033)	0.001(0.046)
		AEE(MSE×10)	0.308(0.025)	0.304(0.009)	0.331(0.017)	0.352(0.019)	1.707(0.668)	1.706(0.669)	1.705(0.674)	1.707(0.682)
	5000	β_{01}	0.501(0.022)	0.490(0.054)	0.505(0.031)	0.505(0.046)	0.000(0.010)	0.000(0.013)	0.001(0.020)	0.001(0.028)
		β_{02}	0.503(0.023)	0.490(0.056)	0.501(0.033)	0.497(0.049)	0.000(0.020)	0.000(0.027)	0.001(0.039)	−0.001(0.053)
		β_{03}	0.703(0.022)	0.715(0.055)	0.700(0.033)	0.701(0.051)	0.000(0.009)	0.000(0.014)	0.000(0.025)	−0.002(0.033)
		AEE(MSE×10)	0.291(0.011)	0.336(0.038)	0.311(0.006)	0.340(0.008)	1.708(0.670)	1.707(0.671)	1.705(0.673)	1.709(0.677)
(D3)	2500	β_{01}	0.500(0.034)	0.503(0.025)	0.499(0.048)	0.504(0.064)	−0.001(0.015)	0.000(0.020)	0.000(0.029)	−0.002(0.041)
		β_{02}	0.500(0.037)	0.504(0.029)	0.499(0.050)	0.499(0.070)	0.000(0.028)	0.002(0.036)	0.003(0.053)	0.004(0.074)
		β_{03}	0.704(0.034)	0.700(0.028)	0.702(0.053)	0.695(0.066)	0.000(0.014)	0.000(0.020)	−0.001(0.032)	0.000(0.044)
		AEE(MSE×10)	0.304(0.021)	0.303(0.009)	0.336(0.020)	0.357(0.022)	1.708(0.670)	1.705(0.671)	1.705(0.676)	1.706(0.683)
	5000	β_{01}	0.503(0.017)	0.493(0.045)	0.505(0.030)	0.510(0.043)	0.000(0.011)	0.001(0.014)	0.000(0.019)	0.000(0.027)
		β_{02}	0.503(0.018)	0.493(0.047)	0.504(0.034)	0.503(0.048)	−0.001(0.020)	−0.001(0.026)	−0.002(0.037)	0.000(0.051)
		β_{03}	0.703(0.016)	0.712(0.046)	0.699(0.035)	0.694(0.049)	0.000(0.010)	0.000(0.015)	0.001(0.023)	0.001(0.031)
		AEE(MSE×10)	0.288(0.009)	0.324(0.024)	0.312(0.006)	0.325(0.008)	1.707(0.671)	1.707(0.672)	1.708(0.674)	1.707(0.677)

Table A.10: Simulation results of the estimators for (M2) using our proposed method and the parametric method proposed in McMahan et al. (2016) when $\sigma^2(t) = (0.5t)^2$. Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE×10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $J \in \{250, 500\}$.

			Proposed Method				Parametric Method			
	J		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	250	β_{01}	0.522(0.054)	0.505(0.065)	0.508(0.067)	0.507(0.068)	−0.004(0.047)	0.001(0.044)	−0.002(0.043)	0.003(0.040)
		β_{02}	0.509(0.067)	0.504(0.069)	0.497(0.071)	0.499(0.072)	−0.002(0.089)	−0.001(0.084)	−0.005(0.081)	−0.006(0.078)
		β_{03}	0.677(0.041)	0.691(0.062)	0.693(0.067)	0.692(0.070)	−0.002(0.043)	0.001(0.045)	−0.002(0.045)	−0.002(0.045)
		AEE(MSE×10)	0.331(0.082)	0.348(0.057)	0.353(0.033)	0.356(0.026)	1.714(0.666)	1.707(1.355)	1.716(3.420)	1.712(6.843)
	500	β_{01}	0.511(0.036)	0.508(0.041)	0.502(0.046)	0.505(0.047)	0.001(0.033)	0.002(0.031)	−0.001(0.028)	0.002(0.027)
		β_{02}	0.512(0.048)	0.503(0.046)	0.500(0.049)	0.503(0.051)	0.000(0.061)	−0.003(0.057)	0.002(0.057)	−0.002(0.054)
		β_{03}	0.686(0.031)	0.695(0.043)	0.700(0.049)	0.697(0.052)	0.000(0.032)	0.002(0.032)	−0.002(0.032)	−0.001(0.033)
		AEE(MSE×10)	0.310(0.038)	0.320(0.023)	0.334(0.019)	0.329(0.011)	1.707(0.670)	1.706(1.338)	1.708(3.375)	1.708(6.754)
(D2)	250	β_{01}	0.518(0.049)	0.501(0.062)	0.507(0.066)	0.507(0.065)	−0.001(0.049)	0.004(0.042)	−0.004(0.040)	0.002(0.040)
		β_{02}	0.517(0.067)	0.507(0.069)	0.504(0.074)	0.497(0.074)	−0.008(0.090)	−0.003(0.082)	−0.006(0.077)	0.002(0.078)
		β_{03}	0.675(0.036)	0.689(0.093)	0.689(0.069)	0.694(0.069)	−0.002(0.044)	0.001(0.047)	0.002(0.046)	−0.001(0.044)
		AEE(MSE×10)	0.316(0.094)	0.348(0.054)	0.355(0.030)	0.354(0.021)	1.718(0.669)	1.705(1.356)	1.715(3.405)	1.704(6.827)
	500	β_{01}	0.513(0.038)	0.508(0.039)	0.502(0.048)	0.501(0.048)	0.000(0.035)	0.002(0.031)	0.000(0.029)	0.000(0.029)
		β_{02}	0.506(0.048)	0.508(0.047)	0.500(0.051)	0.497(0.050)	−0.001(0.065)	0.003(0.056)	0.002(0.054)	0.005(0.058)
		β_{03}	0.690(0.034)	0.692(0.041)	0.700(0.052)	0.703(0.052)	0.000(0.032)	−0.001(0.032)	−0.001(0.033)	−0.001(0.032)
		AEE(MSE×10)	0.313(0.043)	0.322(0.020)	0.332(0.017)	0.335(0.010)	1.709(0.669)	1.703(1.350)	1.705(3.375)	1.704(6.770)
(D3)	250	β_{01}	0.515(0.050)	0.505(0.067)	0.507(0.063)	0.509(0.066)	−0.001(0.049)	0.001(0.046)	−0.001(0.041)	0.000(0.040)
		β_{02}	0.500(0.120)	0.521(0.064)	0.494(0.068)	0.500(0.070)	−0.001(0.089)	0.008(0.082)	−0.001(0.081)	0.008(0.079)
		β_{03}	0.675(0.083)	0.693(0.065)	0.696(0.066)	0.691(0.070)	0.002(0.047)	0.004(0.045)	0.003(0.046)	0.000(0.047)
		AEE(MSE×10)	0.314(0.079)	0.352(0.048)	0.356(0.028)	0.353(0.021)	1.708(0.669)	1.694(1.348)	1.706(3.413)	1.698(6.846)
	500	β_{01}	0.511(0.037)	0.507(0.043)	0.504(0.045)	0.503(0.044)	−0.001(0.033)	−0.001(0.031)	−0.002(0.030)	0.000(0.029)
		β_{02}	0.511(0.045)	0.506(0.048)	0.502(0.050)	0.501(0.048)	−0.001(0.058)	−0.001(0.059)	−0.001(0.053)	−0.003(0.053)
		β_{03}	0.688(0.030)	0.694(0.044)	0.698(0.049)	0.699(0.049)	−0.001(0.032)	−0.003(0.031)	−0.002(0.031)	0.001(0.030)
		AEE(MSE×10)	0.311(0.045)	0.317(0.024)	0.330(0.013)	0.332(0.009)	1.710(0.663)	1.713(1.342)	1.712(3.380)	1.710(6.767)

Table A.11: Simulation results of the estimators for (M3) using our proposed method and the parametric method proposed in McMahan et al. (2016) when $\sigma^2(t) = (0.5t)^2$. Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE$\times$10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $N \in \{2500, 5000\}$.

			Proposed Method				Parametric Method			
	N		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	2500	β_{01}	0.505(0.014)	0.504(0.026)	0.499(0.045)	0.503(0.059)	0.333(0.022)	0.333(0.028)	0.334(0.041)	0.336(0.054)
		β_{02}	0.506(0.017)	0.504(0.028)	0.498(0.045)	0.496(0.058)	0.398(0.047)	0.396(0.055)	0.398(0.077)	0.398(0.111)
		β_{03}	0.698(0.015)	0.700(0.030)	0.704(0.055)	0.700(0.070)	0.346(0.020)	0.346(0.028)	0.343(0.043)	0.341(0.060)
		AEE(MSE$\times$10)	0.285(0.018)	0.300(0.029)	0.335(0.051)	0.355(0.074)	0.630(1.683)	0.634(1.686)	0.638(1.694)	0.652(1.709)
	5000	β_{01}	0.489(0.053)	0.504(0.017)	0.502(0.032)	0.500(0.047)	0.333(0.014)	0.333(0.019)	0.333(0.027)	0.334(0.037)
		β_{02}	0.489(0.052)	0.504(0.019)	0.501(0.033)	0.496(0.045)	0.398(0.031)	0.398(0.038)	0.400(0.056)	0.396(0.077)
		β_{03}	0.716(0.054)	0.700(0.021)	0.702(0.039)	0.704(0.057)	0.346(0.013)	0.346(0.019)	0.346(0.030)	0.350(0.042)
		AEE(MSE$\times$10)	0.328(0.188)	0.289(0.016)	0.309(0.025)	0.335(0.051)	0.629(1.683)	0.630(1.685)	0.630(1.689)	0.634(1.696)
(D2)	2500	β_{01}	0.506(0.013)	0.501(0.028)	0.498(0.043)	0.499(0.060)	0.332(0.022)	0.333(0.027)	0.330(0.040)	0.331(0.053)
		β_{02}	0.508(0.014)	0.504(0.029)	0.499(0.045)	0.495(0.061)	0.398(0.043)	0.398(0.049)	0.401(0.077)	0.397(0.112)
		β_{03}	0.697(0.013)	0.702(0.032)	0.704(0.053)	0.703(0.072)	0.346(0.019)	0.346(0.027)	0.346(0.043)	0.344(0.061)
		AEE(MSE$\times$10)	0.283(0.016)	0.303(0.031)	0.333(0.057)	0.357(0.084)	0.632(1.677)	0.631(1.680)	0.636(1.688)	0.658(1.703)
	5000	β_{01}	0.489(0.051)	0.504(0.016)	0.500(0.035)	0.499(0.043)	0.333(0.014)	0.333(0.019)	0.333(0.029)	0.334(0.040)
		β_{02}	0.490(0.052)	0.505(0.018)	0.500(0.037)	0.498(0.047)	0.398(0.030)	0.400(0.038)	0.400(0.054)	0.400(0.072)
		β_{03}	0.703(0.027)	0.700(0.037)	0.695(0.049)	0.684(0.064)	0.347(0.013)	0.347(0.019)	0.347(0.030)	0.346(0.043)
		AEE(MSE$\times$10)	0.326(0.165)	0.287(0.012)	0.318(0.042)	0.337(0.051)	0.630(1.683)	0.627(1.684)	0.628(1.688)	0.633(1.695)
(D3)	2500	β_{01}	0.507(0.011)	0.505(0.023)	0.502(0.041)	0.505(0.058)	0.333(0.022)	0.334(0.026)	0.332(0.040)	0.332(0.055)
		β_{02}	0.507(0.012)	0.505(0.024)	0.500(0.042)	0.501(0.061)	0.398(0.043)	0.399(0.052)	0.398(0.074)	0.397(0.102)
		β_{03}	0.697(0.011)	0.698(0.027)	0.701(0.050)	0.694(0.071)	0.347(0.020)	0.347(0.028)	0.346(0.043)	0.346(0.061)
		AEE(MSE$\times$10)	0.283(0.011)	0.298(0.020)	0.327(0.041)	0.346(0.074)	0.630(1.682)	0.628(1.685)	0.639(1.693)	0.649(1.707)
	5000	β_{01}	0.491(0.050)	0.505(0.014)	0.499(0.036)	0.503(0.044)	0.333(0.014)	0.333(0.018)	0.331(0.027)	0.332(0.037)
		β_{02}	0.491(0.050)	0.504(0.015)	0.497(0.038)	0.500(0.045)	0.398(0.030)	0.398(0.037)	0.398(0.053)	0.399(0.074)
		β_{03}	0.715(0.052)	0.700(0.017)	0.707(0.044)	0.700(0.054)	0.347(0.013)	0.346(0.018)	0.347(0.030)	0.348(0.044)
		AEE(MSE$\times$10)	0.323(0.170)	0.289(0.010)	0.320(0.041)	0.326(0.040)	0.630(1.677)	0.631(1.678)	0.632(1.682)	0.633(1.689)

Table A.12: Simulation results of the estimators for (M3) using our proposed method and the parametric method proposed in McMahan et al. (2016) when $\sigma^2(t) = (0.5t)^2$. Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE×10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $J \in \{250, 500\}$.

			Proposed Method				Parametric Method			
	J		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	250	β_{01}	0.519(0.033)	0.510(0.050)	0.505(0.055)	0.500(0.061)	0.327(0.068)	0.334(0.058)	0.335(0.055)	0.332(0.055)
		β_{02}	0.523(0.039)	0.509(0.050)	0.503(0.057)	0.497(0.062)	0.390(0.137)	0.400(0.124)	0.403(0.104)	0.405(0.105)
		β_{03}	0.673(0.031)	0.687(0.056)	0.694(0.067)	0.700(0.073)	0.344(0.062)	0.349(0.059)	0.351(0.060)	0.348(0.059)
		AEE(MSE×10)	0.297(0.102)	0.324(0.084)	0.342(0.069)	0.354(0.094)	0.681(1.630)	0.656(3.375)	0.640(8.550)	0.642(17.13)
	500	β_{01}	0.516(0.023)	0.505(0.035)	0.500(0.046)	0.499(0.045)	0.332(0.046)	0.331(0.043)	0.332(0.037)	0.330(0.040)
		β_{02}	0.517(0.028)	0.507(0.038)	0.497(0.047)	0.498(0.043)	0.393(0.092)	0.396(0.083)	0.401(0.079)	0.401(0.070)
		β_{03}	0.681(0.021)	0.696(0.042)	0.704(0.056)	0.705(0.054)	0.345(0.042)	0.346(0.044)	0.347(0.044)	0.347(0.043)
		AEE(MSE×10)	0.289(0.051)	0.313(0.046)	0.338(0.061)	0.330(0.048)	0.649(1.660)	0.643(3.349)	0.637(8.458)	0.633(16.90)
(D2)	250	β_{01}	0.522(0.031)	0.509(0.046)	0.503(0.057)	0.504(0.058)	0.335(0.070)	0.335(0.063)	0.339(0.056)	0.331(0.054)
		β_{02}	0.521(0.041)	0.508(0.048)	0.499(0.060)	0.502(0.059)	0.403(0.140)	0.397(0.114)	0.404(0.106)	0.401(0.107)
		β_{03}	0.672(0.032)	0.690(0.053)	0.697(0.069)	0.695(0.068)	0.348(0.062)	0.349(0.058)	0.346(0.061)	0.345(0.060)
		AEE(MSE×10)	0.296(0.103)	0.323(0.070)	0.346(0.091)	0.345(0.070)	0.664(1.682)	0.650(3.373)	0.640(8.481)	0.652(17.03)
	500	β_{01}	0.513(0.024)	0.506(0.037)	0.501(0.042)	0.500(0.045)	0.334(0.047)	0.333(0.042)	0.328(0.038)	0.332(0.038)
		β_{02}	0.517(0.028)	0.505(0.040)	0.501(0.045)	0.496(0.045)	0.395(0.095)	0.397(0.084)	0.393(0.076)	0.400(0.072)
		β_{03}	0.684(0.022)	0.696(0.044)	0.701(0.052)	0.704(0.055)	0.344(0.041)	0.345(0.042)	0.347(0.042)	0.347(0.045)
		AEE(MSE×10)	0.291(0.051)	0.315(0.049)	0.330(0.050)	0.333(0.050)	0.651(1.647)	0.641(3.361)	0.646(8.412)	0.632(16.91)
(D3)	250	β_{01}	0.521(0.034)	0.510(0.048)	0.505(0.054)	0.504(0.059)	0.331(0.069)	0.326(0.060)	0.333(0.055)	0.336(0.056)
		β_{02}	0.520(0.042)	0.508(0.047)	0.505(0.057)	0.501(0.059)	0.396(0.132)	0.389(0.121)	0.407(0.113)	0.395(0.105)
		β_{03}	0.674(0.031)	0.689(0.053)	0.692(0.066)	0.695(0.072)	0.347(0.060)	0.343(0.059)	0.341(0.062)	0.345(0.059)
		AEE(MSE×10)	0.298(0.103)	0.321(0.078)	0.343(0.078)	0.346(0.079)	0.670(1.640)	0.676(3.346)	0.652(8.494)	0.650(17.06)
	500	β_{01}	0.517(0.025)	0.503(0.038)	0.500(0.043)	0.500(0.044)	0.333(0.047)	0.334(0.042)	0.332(0.037)	0.335(0.040)
		β_{02}	0.514(0.031)	0.505(0.038)	0.500(0.044)	0.498(0.047)	0.399(0.100)	0.403(0.083)	0.401(0.077)	0.406(0.072)
		β_{03}	0.682(0.025)	0.698(0.043)	0.702(0.053)	0.704(0.056)	0.347(0.043)	0.349(0.044)	0.346(0.042)	0.347(0.043)
		AEE(MSE×10)	0.297(0.055)	0.319(0.053)	0.327(0.054)	0.338(0.047)	0.646(1.675)	0.631(3.389)	0.633(8.484)	0.627(16.96)

Table A.13: Simulation results of the estimators for (M4) when $a = 1$ using our proposed method and the parametric method proposed in McMahan et al. (2016) when $\sigma^2(t) = (0.5t)^2$. Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE×10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $N \in \{2500, 5000\}$.

			Proposed Method				Parametric Method			
	N		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	2500	β_{01}	0.500(0.003)	0.500(0.005)	0.500(0.009)	0.501(0.014)	0.483(0.018)	0.483(0.024)	0.482(0.035)	0.482(0.046)
		β_{02}	0.501(0.005)	0.500(0.009)	0.498(0.016)	0.495(0.027)	0.395(0.039)	0.395(0.049)	0.396(0.075)	0.397(0.099)
		β_{03}	0.707(0.002)	0.707(0.004)	0.708(0.008)	0.709(0.014)	1.020(0.018)	1.019(0.025)	1.020(0.040)	1.020(0.060)
		AEE(MSE×10)	0.280(0.011)	0.284(0.015)	0.293(0.026)	0.302(0.036)	0.439(1.475)	0.441(1.478)	0.453(1.485)	0.470(1.497)
	5000	β_{01}	0.500(0.002)	0.500(0.003)	0.501(0.006)	0.501(0.009)	0.483(0.013)	0.483(0.017)	0.483(0.025)	0.484(0.033)
		β_{02}	0.501(0.004)	0.500(0.006)	0.499(0.011)	0.497(0.018)	0.399(0.027)	0.398(0.033)	0.397(0.050)	0.396(0.069)
		β_{03}	0.707(0.001)	0.707(0.003)	0.707(0.006)	0.709(0.009)	1.021(0.013)	1.021(0.018)	1.021(0.029)	1.021(0.041)
		AEE(MSE×10)	0.279(0.006)	0.282(0.009)	0.286(0.016)	0.294(0.023)	0.433(1.471)	0.436(1.472)	0.442(1.476)	0.452(1.482)
(D2)	2500	β_{01}	0.500(0.003)	0.500(0.005)	0.501(0.009)	0.501(0.015)	0.482(0.019)	0.483(0.025)	0.481(0.035)	0.479(0.050)
		β_{02}	0.500(0.005)	0.499(0.008)	0.496(0.016)	0.495(0.027)	0.398(0.038)	0.397(0.051)	0.396(0.074)	0.393(0.102)
		β_{03}	0.707(0.002)	0.707(0.004)	0.709(0.008)	0.710(0.013)	1.021(0.018)	1.022(0.027)	1.022(0.040)	1.020(0.057)
		AEE(MSE×10)	0.280(0.011)	0.284(0.015)	0.294(0.026)	0.306(0.035)	0.438(1.469)	0.443(1.472)	0.456(1.479)	0.478(1.491)
	5000	β_{01}	0.500(0.002)	0.500(0.003)	0.500(0.008)	0.501(0.009)	0.484(0.013)	0.483(0.017)	0.484(0.027)	0.484(0.036)
		β_{02}	0.501(0.003)	0.500(0.006)	0.499(0.011)	0.498(0.017)	0.399(0.028)	0.399(0.035)	0.399(0.050)	0.399(0.072)
		β_{03}	0.707(0.001)	0.707(0.003)	0.707(0.005)	0.708(0.009)	1.022(0.013)	1.021(0.018)	1.022(0.028)	1.027(0.040)
		AEE(MSE×10)	0.279(0.006)	0.281(0.009)	0.287(0.015)	0.291(0.023)	0.434(1.469)	0.435(1.471)	0.443(1.474)	0.456(1.481)
(D3)	2500	β_{01}	0.500(0.002)	0.501(0.005)	0.501(0.009)	0.501(0.014)	0.484(0.019)	0.485(0.024)	0.485(0.039)	0.486(0.053)
		β_{02}	0.500(0.005)	0.499(0.009)	0.497(0.016)	0.496(0.027)	0.400(0.038)	0.399(0.048)	0.399(0.072)	0.391(0.097)
		β_{03}	0.707(0.002)	0.707(0.004)	0.708(0.008)	0.709(0.014)	1.022(0.018)	1.021(0.024)	1.021(0.040)	1.022(0.057)
		AEE(MSE×10)	0.280(0.011)	0.284(0.015)	0.293(0.026)	0.303(0.036)	0.435(1.467)	0.438(1.470)	0.454(1.478)	0.480(1.490)
	5000	β_{01}	0.500(0.002)	0.500(0.003)	0.501(0.006)	0.501(0.010)	0.483(0.013)	0.483(0.017)	0.482(0.025)	0.482(0.036)
		β_{02}	0.500(0.004)	0.500(0.006)	0.498(0.011)	0.496(0.018)	0.397(0.028)	0.397(0.035)	0.395(0.049)	0.391(0.070)
		β_{03}	0.707(0.001)	0.707(0.003)	0.708(0.006)	0.709(0.009)	1.021(0.014)	1.021(0.019)	1.020(0.029)	1.020(0.040)
		AEE(MSE×10)	0.279(0.006)	0.282(0.008)	0.288(0.015)	0.295(0.022)	0.434(1.471)	0.437(1.472)	0.444(1.476)	1.482(0.456)

Table A.14: Simulation results of the estimators for (M4) when $a = 1$ using our proposed method and the parametric method proposed in McMahan et al. (2016) when $\sigma^2(t) = (0.5t)^2$. Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE×10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $J \in \{250, 500\}$.

			Proposed Method				Parametric Method			
	J		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	250	β_{01}	0.500(0.009)	0.501(0.012)	0.501(0.014)	0.501(0.014)	0.483(0.062)	0.484(0.057)	0.483(0.048)	0.481(0.053)
		β_{02}	0.501(0.018)	0.498(0.022)	0.496(0.026)	0.496(0.028)	0.392(0.128)	0.394(0.116)	0.393(0.102)	0.399(0.100)
		β_{03}	0.706(0.007)	0.707(0.011)	0.708(0.013)	0.709(0.014)	1.027(0.056)	1.022(0.061)	1.017(0.059)	1.020(0.057)
		AEE(MSE×10)	0.290(0.106)	0.296(0.057)	0.301(0.043)	0.305(0.036)	0.507(1.443)	0.489(2.945)	0.471(7.458)	0.475(14.95)
	500	β_{01}	0.500(0.006)	0.501(0.008)	0.500(0.009)	0.501(0.009)	0.482(0.041)	0.482(0.039)	0.482(0.037)	0.484(0.033)
		β_{02}	0.501(0.012)	0.498(0.014)	0.498(0.016)	0.496(0.017)	0.396(0.088)	0.398(0.079)	0.397(0.071)	0.399(0.070)
		β_{03}	0.706(0.005)	0.708(0.007)	0.708(0.008)	0.709(0.009)	1.021(0.038)	1.023(0.041)	1.021(0.038)	1.020(0.041)
		AEE(MSE×10)	0.285(0.048)	0.290(0.032)	0.293(0.026)	0.295(0.023)	0.463(1.469)	0.459(2.943)	0.453(7.397)	0.449(14.84)
(D2)	250	β_{01}	0.500(0.013)	0.500(0.011)	0.501(0.014)	0.501(0.014)	0.482(0.060)	0.483(0.051)	0.482(0.052)	0.485(0.054)
		β_{02}	0.500(0.021)	0.500(0.021)	0.496(0.025)	0.496(0.025)	0.398(0.126)	0.399(0.108)	0.394(0.102)	0.397(0.095)
		β_{03}	0.706(0.011)	0.706(0.010)	0.709(0.012)	0.708(0.013)	1.021(0.057)	1.023(0.058)	1.023(0.061)	1.022(0.055)
		AEE(MSE×10)	0.292(0.099)	0.295(0.058)	0.302(0.044)	0.301(0.035)	0.494(1.449)	0.479(2.953)	0.484(7.457)	0.476(14.94)
	500	β_{01}	0.500(0.006)	0.500(0.008)	0.501(0.009)	0.501(0.009)	0.483(0.043)	0.484(0.037)	0.481(0.036)	0.483(0.035)
		β_{02}	0.501(0.012)	0.499(0.014)	0.498(0.016)	0.497(0.017)	0.401(0.087)	0.401(0.075)	0.395(0.070)	0.393(0.071)
		β_{03}	0.706(0.005)	0.707(0.007)	0.708(0.008)	0.708(0.009)	1.023(0.041)	1.021(0.042)	1.020(0.043)	1.024(0.041)
		AEE(MSE×10)	0.286(0.049)	0.290(0.033)	0.292(0.026)	0.294(0.021)	0.461(1.460)	0.451(2.945)	0.453(7.413)	0.460(14.85)
(D3)	250	β_{01}	0.500(0.009)	0.501(0.011)	0.500(0.014)	0.501(0.014)	0.484(0.059)	0.485(0.057)	0.487(0.052)	0.482(0.053)
		β_{02}	0.500(0.017)	0.499(0.020)	0.499(0.025)	0.496(0.027)	0.399(0.117)	0.407(0.103)	0.394(0.097)	0.391(0.099)
		β_{03}	0.706(0.006)	0.707(0.010)	0.708(0.012)	0.708(0.013)	1.022(0.058)	1.024(0.056)	1.019(0.058)	1.024(0.058)
		AEE(MSE×10)	0.290(0.094)	0.295(0.058)	0.299(0.043)	0.302(0.035)	0.490(1.454)	0.478(2.941)	0.475(7.436)	0.483(14.94)
	500	β_{01}	0.500(0.010)	0.500(0.008)	0.500(0.009)	0.501(0.010)	0.483(0.041)	0.485(0.039)	0.485(0.036)	0.483(0.036)
		β_{02}	0.500(0.014)	0.500(0.014)	0.497(0.016)	0.497(0.018)	0.399(0.087)	0.399(0.082)	0.395(0.074)	0.400(0.073)
		β_{03}	0.707(0.010)	0.707(0.007)	0.709(0.008)	0.708(0.009)	1.025(0.039)	1.024(0.044)	1.018(0.040)	1.019(0.041)
		AEE(MSE×10)	0.286(0.052)	0.288(0.033)	0.293(0.027)	0.294(0.023)	0.464(1.467)	0.460(2.950)	0.451(7.435)	0.450(14.88)

Table A.15: Simulation results of the estimators for (M4) when $a = 2$ using our proposed method and the parametric method proposed in McMahan et al. (2016) when $\sigma^2(t) = (0.5t)^2$. Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE×10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $N \in \{2500, 5000\}$.

			Proposed Method				Parametric Method			
	N		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	2500	β_{01}	0.500(0.001)	0.500(0.002)	0.500(0.005)	0.501(0.008)	−0.370(0.026)	−0.370(0.038)	−0.368(0.056)	−0.367(0.080)
		β_{02}	0.500(0.003)	0.500(0.004)	0.499(0.008)	0.498(0.013)	0.000(0.052)	0.001(0.070)	−0.003(0.110)	0.003(0.144)
		β_{03}	0.707(0.001)	0.707(0.002)	0.707(0.005)	0.707(0.008)	0.000(0.028)	0.000(0.041)	−0.003(0.062)	−0.002(0.092)
		AEE(MSE×10)	0.278(0.019)	0.280(0.026)	0.284(0.044)	0.290(0.064)	2.078(4.536)	2.075(4.542)	2.081(4.561)	2.073(4.591)
	5000	β_{01}	0.500(0.001)	0.500(0.002)	0.500(0.003)	0.500(0.005)	−0.371(0.010)	−0.370(0.026)	−0.372(0.039)	−0.373(0.057)
		β_{02}	0.500(0.002)	0.500(0.003)	0.499(0.005)	0.499(0.008)	0.000(0.038)	−0.002(0.052)	−0.001(0.075)	0.003(0.106)
		β_{03}	0.707(0.001)	0.707(0.002)	0.708(0.003)	0.708(0.005)	0.003(0.021)	0.003(0.029)	0.000(0.046)	0.000(0.066)
		AEE(MSE×10)	0.277(0.011)	0.279(0.016)	0.282(0.028)	0.285(0.041)	2.075(4.538)	2.077(4.541)	2.080(4.551)	2.077(4.566)
(D2)	2500	β_{01}	0.500(0.001)	0.500(0.002)	0.500(0.005)	0.501(0.008)	−0.371(0.026)	−0.371(0.036)	−0.374(0.056)	−0.373(0.080)
		β_{02}	0.500(0.003)	0.500(0.004)	0.499(0.008)	0.498(0.013)	0.000(0.052)	0.003(0.070)	0.005(0.106)	0.006(0.153)
		β_{03}	0.707(0.001)	0.707(0.002)	0.708(0.005)	0.708(0.008)	−0.001(0.028)	0.003(0.038)	0.000(0.062)	0.000(0.090)
		AEE(MSE×10)	0.278(0.019)	0.280(0.027)	0.285(0.045)	0.289(0.063)	2.079(4.537)	2.076(4.543)	2.076(4.561)	2.074(4.594)
	5000	β_{01}	0.500(0.001)	0.500(0.002)	0.500(0.003)	0.500(0.005)	−0.372(0.017)	−0.372(0.025)	−0.370(0.039)	−0.373(0.056)
		β_{02}	0.500(0.002)	0.500(0.003)	0.499(0.005)	0.498(0.008)	0.000(0.037)	0.001(0.049)	0.001(0.073)	−0.001(0.103)
		β_{03}	0.707(0.001)	0.707(0.002)	0.708(0.003)	0.708(0.005)	0.001(0.019)	0.002(0.027)	0.003(0.044)	0.002(0.065)
		AEE(MSE×10)	0.277(0.010)	0.279(0.015)	0.282(0.027)	0.286(0.040)	2.077(4.536)	2.076(4.539)	2.073(4.548)	2.078(4.563)
(D3)	2500	β_{01}	0.500(0.001)	0.500(0.002)	0.500(0.005)	0.500(0.008)	−0.371(0.025)	−0.369(0.038)	−0.368(0.055)	−0.367(0.082)
		β_{02}	0.500(0.003)	0.500(0.004)	0.499(0.008)	0.499(0.014)	0.001(0.052)	0.002(0.071)	0.006(0.106)	0.006(0.151)
		β_{03}	0.707(0.001)	0.707(0.002)	0.708(0.005)	0.708(0.008)	−0.002(0.028)	−0.003(0.039)	−0.003(0.063)	0.000(0.089)
		AEE(MSE×10)	0.278(0.019)	0.280(0.026)	0.284(0.044)	0.289(0.062)	2.079(4.529)	2.077(4.536)	2.071(4.553)	2.067(4.586)
	5000	β_{01}	0.500(0.001)	0.500(0.002)	0.500(0.003)	0.500(0.005)	−0.372(0.019)	−0.372(0.026)	−0.374(0.041)	−0.373(0.057)
		β_{02}	0.500(0.002)	0.500(0.003)	0.499(0.005)	0.499(0.008)	−0.004(0.037)	−0.004(0.050)	−0.003(0.078)	−0.002(0.107)
		β_{03}	0.707(0.001)	0.707(0.002)	0.707(0.003)	0.708(0.005)	0.001(0.021)	0.001(0.029)	0.000(0.048)	0.004(0.067)
		AEE(MSE×10)	0.278(0.011)	0.279(0.015)	0.282(0.027)	0.285(0.039)	2.082(4.537)	2.083(4.540)	2.084(4.550)	2.078(4.565)

Table A.16: Simulation results of the estimators for (M4) when $a = 2$ using our proposed method and the parametric method proposed in McMahan et al. (2016) when $\sigma^2(t) = (0.5t)^2$. Presented results include the sample mean and sample standard deviation (provided within the parenthesis) of the 500 estimates of $\boldsymbol{\beta}_0 = (\beta_{01} = 0.5, \beta_{02} = 0.5, \beta_{03} = 0.707)$, as well as the sample mean of 500 AEE's and MSE×10's (provided in parenthesis) across all considered pool sizes $c \in \{1, 2, 5, 10\}$ for $J \in \{250, 500\}$.

			Proposed Method				Parametric Method			
	J		$c = 1$	$c = 2$	$c = 5$	$c = 10$	$c = 1$	$c = 2$	$c = 5$	$c = 10$
(D1)	250	β_{01}	0.500(0.004)	0.501(0.006)	0.501(0.007)	0.501(0.007)	−0.367(0.083)	−0.369(0.080)	−0.374(0.080)	−0.374(0.082)
		β_{02}	0.501(0.009)	0.499(0.011)	0.499(0.012)	0.498(0.013)	0.001(0.162)	0.007(0.156)	0.012(0.153)	−0.011(0.151)
		β_{03}	0.707(0.004)	0.707(0.006)	0.707(0.007)	0.708(0.008)	−0.001(0.092)	0.005(0.099)	−0.002(0.092)	−0.005(0.102)
		AEE(MSE×10)	0.282(0.170)	0.287(0.093)	0.289(0.077)	0.290(0.066)	2.073(4.482)	2.065(9.094)	2.071(22.92)	2.098(45.97)
	500	β_{01}	0.500(0.003)	0.500(0.004)	0.501(0.004)	0.501(0.005)	−0.373(0.059)	−0.371(0.057)	−0.373(0.057)	−0.369(0.053)
		β_{02}	0.500(0.006)	0.500(0.007)	0.499(0.008)	0.498(0.008)	−0.005(0.115)	0.004(0.116)	0.003(0.114)	−0.003(0.105)
		β_{03}	0.707(0.003)	0.707(0.004)	0.707(0.005)	0.708(0.005)	−0.003(0.068)	−0.002(0.067)	−0.001(0.061)	−0.003(0.067)
		AEE(MSE×10)	0.281(0.078)	0.283(0.054)	0.284(0.048)	0.285(0.040)	2.088(4.508)	2.076(9.080)	2.078(22.78)	2.081(45.65)
(D2)	250	β_{01}	0.501(0.005)	0.500(0.006)	0.501(0.007)	0.500(0.008)	−0.368(0.082)	−0.367(0.079)	−0.364(0.081)	−0.374(0.078)
		β_{02}	0.500(0.009)	0.499(0.011)	0.498(0.012)	0.498(0.013)	0.010(0.163)	−0.008(0.155)	−0.008(0.153)	0.012(0.166)
		β_{03}	0.707(0.004)	0.707(0.006)	0.708(0.007)	0.708(0.008)	−0.001(0.087)	0.004(0.094)	−0.001(0.090)	0.003(0.095)
		AEE(MSE×10)	0.283(0.178)	0.287 (0.096)	0.289(0.077)	0.289(0.067)	2.067(4.481)	2.078(9.104)	2.079(22.90)	2.066(45.98)
	500	β_{01}	0.500(0.003)	0.501(0.004)	0.500(0.005)	0.500(0.005)	−0.375(0.059)	−0.372(0.055)	−0.374(0.056)	−0.372(0.054)
		β_{02}	0.500(0.006)	0.499(0.007)	0.499(0.008)	0.499(0.008)	−0.005(0.114)	0.004(0.113)	0.002(0.113)	−0.004(0.108)
		β_{03}	0.707(0.003)	0.707(0.004)	0.708(0.005)	0.708(0.005)	0.004(0.065)	−0.002(0.064)	0.001(0.065)	0.004(0.061)
		AEE(MSE×10)	0.281(0.078)	0.283(0.053)	0.285(0.044)	0.284(0.040)	2.083(4.494)	2.077(9.095)	2.078(22.82)	2.079(45.67)
(D3)	250	β_{01}	0.501(0.005)	0.501(0.006)	0.500(0.007)	0.500(0.007)	−0.363(0.085)	−0.371(0.079)	−0.369(0.082)	−0.368(0.078)
		β_{02}	0.499(0.010)	0.499(0.010)	0.498(0.013)	0.498(0.013)	−0.002(0.166)	−0.027(0.155)	−0.001(0.153)	−0.002(0.149)
		β_{03}	0.707(0.004)	0.707(0.006)	0.708(0.008)	0.708(0.008)	0.007(0.093)	−0.003(0.096)	0.000(0.090)	0.000(0.089)
		AEE(MSE×10)	0.284(0.173)	0.287(0.098)	0.289(0.075)	0.290(0.069)	2.065(4.472)	2.108(9.086)	2.078(22.92)	2.077(45.87)
	500	β_{01}	0.501(0.003)	0.500(0.004)	0.500(0.005)	0.500(0.005)	−0.371(0.056)	−0.370(0.057)	−0.374(0.055)	−0.373(0.056)
		β_{02}	0.500(0.006)	0.500(0.007)	0.499(0.008)	0.499(0.008)	−0.006(0.120)	−0.004(0.105)	−0.003(0.109)	−0.004(0.111)
		β_{03}	0.707(0.003)	0.707(0.004)	0.707(0.005)	0.708(0.005)	0.004(0.070)	−0.001(0.066)	−0.001(0.066)	0.001(0.065)
		AEE(MSE×10)	0.282(0.079)	0.283(0.056)	0.284(0.045)	0.285(0.044)	2.079(4.517)	2.083(9.062)	2.085(22.82)	2.083(45.69)

A.3 Bootstrapping from Section 2.5.2

For notational simplicity, we denote by $\boldsymbol{\mathcal{X}}_j$ the collection of covariates of individuals in the jth group, for $j = 1, \ldots, J$. The pooled data is then denoted by $\{(Z_j, \boldsymbol{\mathcal{X}}_j) : j = 1, \ldots, J\}$. Then we bootstrap the pairs $(Z_j, \boldsymbol{\mathcal{X}}_j)$'s for M times. At each time, we denote the bootstrap pooled data by $\{(Z_j^{(m)}, \boldsymbol{\mathcal{X}}_j^{(m)}) : j = 1, \ldots, J\}$. Applying our method on each bootstrap pooled data yields bootstrap estimates $\hat{\boldsymbol{\beta}}_{b:m} = (\hat{\beta}_{b:m,1}, \ldots, \hat{\beta}_{b:m,p})^{\mathrm{T}}$ and $\hat{\eta}_{b:m}(\cdot)$ for $m = 1, \ldots, M$. Then, one can estimate the standard error of $\hat{\beta}_j$ via the sample standard deviation of $\{\hat{\beta}_{b:1,j}, \ldots, \hat{\beta}_{b:M,j}\}$.

To illustrate the performance of the above bootstrapping method, we conducted a simulation study. One should note that, though bootstrapping is a powerful tool, its computational cost is huge, especially for simulation studies where 500 replications are needed. Thus, we only focused on a small proportion of the simulation studies that we have conducted in Section 2.4 of the Chapter 2. We hope this study could provide numerical evidences to support the use of it in more complex settings. In this simulation study, we repeat the simulation in Section 2.4 of the Chapter 2 under the combination of (D3), (M2), $J \in \{250, 500\}$ and $c = 2$. Under each, we independently generated 500 pooled data. On each pooled data, we applied our methodology to estimate $\boldsymbol{\beta}_0$ and $\eta_0(\cdot)$ and further used the above bootstrapping method to estimate the standard errors of our point estimates. The number of bootstrapping is $M = 500$. Table A.17 summarizes the simulation results. For making inferences, one can see that, as J increases from 250 to 500, the average of estimated standard errors (SE) are all in closer agreement with the sample standard deviation (SD) of our point estimates, and that all the empirical coverages (ECV) are increasing to the nominal level. These patterns suggest that the bootstrapping method is a good tool to estimate the standard errors for making statistical inferences based on our proposed estimators.

Table A.17: Presented results include the sample mean (Mean) and sample standard deviation (SD) of the estimates of $\boldsymbol{\beta}$ from 500 replications, as well as the mean of 500 estimated standard errors (each was obtained from 500 time bootstrapping) and the empirical coverage (ECV) of the 95% confidence intervals constructed using the estimated standard errors under the combination of (D3), (M2), $J \in \{250, 500\}$ and $c = 2$.

Truth		$J = 250$	$J = 500$
$\beta_{01} = 0.500$	Mean(SD)	0.515(0.126)	0.511(0.084)
	ECV(SE)	0.893(0.106)	0.926(0.079)
$\beta_{02} = 0.500$	Mean(SD)	0.491(0.126)	0.501(0.090)
	ECV(SE)	0.935(0.121)	0.926(0.086)
$\beta_{03} = 0.707$	Mean(SD)	0.670(0.113)	0.682(0.084)
	ECV(SE)	0.898(0.102)	0.906(0.074)

Appendix B

Chapter 3 Supplementary Materials

In this appendix, we provide a simulation study to demonstrate the impairment on estimation if ignoring the retesting outcomes (see Section 3.2). We provide a detailed derivation of the log-likelihood function $\ell(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \mathbf{X})$ and a simulation study to illustrate the computational advantage of our GEM algorithm over a direct maximization (see Section 3.3). The E-step of our GEM algorithm and the observed data information matrix $\mathcal{I}(\boldsymbol{\theta})$ are presented in closed forms (see Section 3.3). We provide additional simulation results for different settings of misclassification parameters, and extend the proposed method to model individual testing data, where due to identifiability issues, we keep the assay sensitivity and specificity for both infections as known (see Section 3.5). We further provide an additional analysis of the NPHL data (see Section 3.6). Lastly, additional simulation studies are provided to reveal the robustness of using the Gumbel copula and demonstrate the generalizability of our method to three infections (see Section 3.7).

B.1 Ignoring retesting outcomes

To show how the retesting information improves the estimation accuracy, we extend the GEM algorithm for the master pool testing responses; i.e., the testing outcomes solely from Stage 1. Same as the setting S2 of Table 3.2 (where retesting outcomes from Stage 2 are also considered), we consider

- $\boldsymbol{\beta}_1 = (-4, -2, 2, 0, 0, 0)^{\mathrm{T}}, \boldsymbol{\beta}_2 = (-5, -2, 0, -2, 0, 0)^{\mathrm{T}}$ and $\mathbf{x} = (1, x_1, \cdots, x_5)^{\mathrm{T}}$,

where $\mathbf{x}$ is simulated from $\mathcal{N}(\mathbf{0}, \mathbf{\Omega})$ with $[\mathbf{\Omega}]_{st} = 1$ if $s = t$ and $[\mathbf{\Omega}]_{st} = 0.5$ if $s \neq t$.

The number of individuals is $N = 3000$, the group size is $c \in \{2, 5, 10\}$, $S_{e:k} = S_{p:k} = 0.95$, for $k = 1, 2$, and we use the Gumbel copula with $\delta = 0.3$. Due to the lack of retesting information, keeping the misclassification parameters as unknown results in identifiability issue. Thus, we have to assume the true values of $S_{e:k}$'s and $S_{p:k}$'s are provided in advance (which is not realistic) and estimate only $\boldsymbol{\beta}_k$'s and δ.

The simulation is repeated 500 times at each c, and summary statistics are reported in Table B.1. By comparing to Table 3.2, one can clearly see that if ignore the retesting outcomes, the estimates exhibit larger bias and larger standard deviations. This evidence of inferiority becomes much clear as the group size increases.

Table B.1: Summary statistics of the 500 MLEs obtained under S2 and master pool testing, including the sample mean (Mean), the sample standard deviation (SD), the average of the estimated standard error (SE) and the empirical coverage (EC) of 95% confidence interval with $c = 2, 5, 10$. The prevalence (averaged over 500 repetitions) of the first and the second infections are 6.77% and 9.98%, respectively.

		$c = 2$		$c = 5$		$c = 10$	
# tests		1500		750		300	
	Truth	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)
β_{10}	-4	-4.04(0.25)	0.99(0.29)	-4.07(0.28)	0.99(0.32)	-4.15(0.37)	1.00(0.50)
β_{11}	-2	-2.01(0.20)	0.96(0.22)	-2.04(0.26)	0.97(0.27)	-2.07(0.34)	0.99(0.46)
β_{12}	2	2.03(0.20)	0.97(0.22)	2.05(0.26)	0.98(0.30)	2.09(0.35)	0.98(0.46)
β_{13}	0	-0.01(0.16)	0.94(0.15)	-0.02(0.21)	0.93(0.20)	-0.03(0.30)	0.95(0.28)
β_{14}	0	-0.01(0.14)	0.96(0.15)	0.01(0.20)	0.94(0.19)	-0.01(0.30)	0.95(0.28)
β_{15}	0	0.00(0.14)	0.96(0.15)	-0.01(0.19)	0.96(0.19)	-0.01(0.29)	0.96(0.28)
β_{20}	-5	-5.05(0.36)	0.97(0.39)	-5.11(0.42)	0.99(0.48)	-5.25(0.64)	1.00(0.88)
β_{21}	-2	-2.02(0.21)	0.96(0.22)	-2.05(0.27)	0.97(0.30)	-2.12(0.42)	0.97(0.49)
β_{22}	0	0.00(0.15)	0.96(0.15)	0.02(0.21)	0.93(0.19)	0.01(0.33)	0.95(0.30)
β_{23}	-2	-2.02(0.20)	0.96(0.22)	-2.05(0.26)	0.97(0.30)	-2.09(0.42)	0.97(0.49)
β_{24}	0	0.010(0.14)	0.96(0.15)	0.00(0.20)	0.95(0.19)	0.02(0.31)	0.95(0.30)
β_{25}	0	-0.01(0.16)	0.94(0.15)	0.00(0.20)	0.95(0.19)	-0.01(0.31)	0.95(0.30)
δ	0.3	0.30(0.09)	0.99(0.13)	0.33(0.13)	0.99(0.23)	0.39(0.20)	1.00(0.47)

B.2 Derivation of $\ell(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \mathbf{X})$

With the SHL pooled testing data $\boldsymbol{\mathcal{P}}$ and covariates $\mathbf{X}$, we can write the observed log-likelihood in the form of

$$\ell(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \mathbf{X}) = \sum_{j=1}^{J} \log \text{pr}(\boldsymbol{\mathcal{P}}_j|\mathbf{X}_j, \boldsymbol{\theta})\}, \tag{B.1}$$

where $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^{\text{T}}, \boldsymbol{\theta}_2^{\text{T}})^{\text{T}}$ and $\mathbf{X}_j$ collects $\{\mathbf{x}_{1j} \cdots, \mathbf{x}_{c_j j}\}$. It suffices to calculate $\text{pr}(\boldsymbol{\mathcal{P}}_j|\mathbf{X}_j, \boldsymbol{\theta})$ for all possible $\boldsymbol{\mathcal{P}}_j$. As described in Section 2, the $\boldsymbol{\mathcal{P}}_j$ takes one of two forms, either $\boldsymbol{Z}_j = (0,0)^{\text{T}}$ or $\boldsymbol{Z}_j \in \{(1,0)^{\text{T}}, (0,1)^{\text{T}}, (1,1)^{\text{T}}\}$ and $\boldsymbol{Y}_{1j}, \ldots, \boldsymbol{Y}_{c_j j}$. Hence, the probability in (B.1) can be decomposed by

$$\text{pr}(\boldsymbol{\mathcal{P}}_j|\mathbf{X}_j, \boldsymbol{\theta}) = I(Z_{j1} + Z_{j2} = 0)\text{pr}(Z_{j1} = 0, Z_{j2} = 0|\mathbf{X}_j, \boldsymbol{\theta}) \tag{B.2}$$

$$+ I(Z_{j1} + Z_{j2} > 0)\text{pr}(Z_{j1}, Z_{j2}, \boldsymbol{Y}_j|\mathbf{X}_j, \boldsymbol{\theta}), \tag{B.3}$$

where $\boldsymbol{Y}_j$ is the collection of $\boldsymbol{Y}_{1j}, \ldots, \boldsymbol{Y}_{c_j j}$ when $Z_{j1} + Z_{j2} > 0$. Before calculating (B.2) and (B.3), we denote

$$\mathcal{M}(a_1, a_2|b_1, b_2, \boldsymbol{\theta}_2) = \text{pr}(Y_{ij1} = a_1, Y_{ij2} = a_2|\widetilde{Y}_{ij1} = b_1, \widetilde{Y}_{ij2} = b_2, \boldsymbol{\theta}_2)$$

for a_k's, b_k's in $\{0, 1\}$. Under the assumptions listed in Section 2, we have

$$\begin{aligned}\mathcal{M}(a_1, a_2|b_1, b_2, \boldsymbol{\theta}_2) &= \text{pr}(Z_{j1} = a_1, Z_{j2} = a_2|\widetilde{Z}_{j1} = b_1, \widetilde{Z}_{j2} = b_2) \\ &= \prod_{k=1}^{2} S_{e:k}^{a_k b_k}(1 - S_{e:k})^{(1-a_k)b_k}(1 - S_{p:k})^{a_k(1-b_k)} S_{p:k}^{(1-a_k)(1-b_k)}. \end{aligned} \tag{B.4}$$

Now, we calculate $\text{pr}(Z_{j1} = 0, Z_{j2} = 0|\mathbf{X}_j, \boldsymbol{\theta})$ in (B.2). Using the Law of Total Probability and Assumption 3, we can write $\text{pr}(Z_{j1} = 0, Z_{j2} = 0|\mathbf{X}_j, \boldsymbol{\theta})$ by

$$\begin{aligned}&\sum_{b_1, b_2 \in \{0,1\}} \text{pr}(Z_{j1} = 0, Z_{j2} = 0|\widetilde{Z}_{j1} = b_1, \widetilde{Z}_{j2} = b_2, \boldsymbol{\theta}_2)\text{pr}(\widetilde{Z}_{j1} = b_1, \widetilde{Z}_{j2} = b_2|\mathbf{X}_j, \boldsymbol{\theta}_1) \\ = &\sum_{b_1, b_2 \in \{0,1\}} \mathcal{M}(0, 0|b_1, b_2, \boldsymbol{\theta}_2)\text{pr}(\widetilde{Z}_{j1} = b_1, \widetilde{Z}_{j2} = b_2|\mathbf{X}_j, \boldsymbol{\theta}_1). \end{aligned} \tag{B.5}$$

Because $\widetilde{Z}_{jk} = \max_{i=1}^{c_j} \widetilde{Y}_{ijk}$ and $\boldsymbol{Y}_{ij}|\mathbf{x}_{ij}$'s are independent binary random vectors with cell probabilities being $p_{ijy_1y_2}(\boldsymbol{\theta}_1)$'s for $y_1, y_2 \in \{0, 1\}$,

$$\begin{aligned}
&\mathrm{pr}(\widetilde{Z}_{j1} = 0, \widetilde{Z}_{j2} = 0|\mathbf{X}_j, \boldsymbol{\theta}_1) \\
&\qquad = \prod_{i=1}^{c_j} \mathrm{pr}(\widetilde{Y}_{ij1} = 0, \widetilde{Y}_{ij2} = 0|\mathbf{x}_{ij}, \boldsymbol{\theta}_1) = \prod_{i=1}^{c_j} p_{ij00}(\boldsymbol{\theta}_1) \qquad \text{(B.6)} \\
&\mathrm{pr}(\widetilde{Z}_{j1} = 1, \widetilde{Z}_{j2} = 0|\mathbf{X}_j, \boldsymbol{\theta}_1) \\
&\qquad = \mathrm{pr}(\widetilde{Z}_{j2} = 0|\mathbf{X}_j, \boldsymbol{\theta}_1) - \mathrm{pr}(\widetilde{Z}_{j1} = 0, \widetilde{Z}_{j2} = 0|\mathbf{X}_j, \boldsymbol{\theta}_1) \\
&\qquad = \prod_{i=1}^{c_j} \mathrm{pr}(\widetilde{Y}_{ij2} = 0|\mathbf{x}_{ij}, \boldsymbol{\theta}_1) - \prod_{i=1}^{c_j} p_{ij00}(\boldsymbol{\theta}_1) \\
&\qquad = \prod_{i=1}^{c_j} \{p_{ij00}(\boldsymbol{\theta}_1) + p_{ij10}(\boldsymbol{\theta}_1)\} - \prod_{i=1}^{c_j} p_{ij00}(\boldsymbol{\theta}_1) \qquad \text{(B.7)} \\
&\mathrm{pr}(\widetilde{Z}_{j1} = 0, \widetilde{Z}_{j2} = 1|\mathbf{X}_j, \boldsymbol{\theta}_1) = \prod_{i=1}^{c_j} \{p_{ij00}(\boldsymbol{\theta}_1) + p_{ij01}(\boldsymbol{\theta}_1)\} - \prod_{i=1}^{c_j} p_{ij00}(\boldsymbol{\theta}_1), \qquad \text{(B.8)}
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{pr}(\widetilde{Z}_{j1} = 1, \widetilde{Z}_{j2} = 1|\mathbf{X}_j, \boldsymbol{\theta}_1) = 1 &- \prod_{i=1}^{c_j} \{p_{ij00}(\boldsymbol{\theta}_1) + p_{ij01}(\boldsymbol{\theta}_1)\} \\
&- \prod_{i=1}^{c_j} \{p_{ij00}(\boldsymbol{\theta}_1) + p_{ij10}(\boldsymbol{\theta}_1)\} + \prod_{i=1}^{c_j} p_{ij00}(\boldsymbol{\theta}_1). \qquad \text{(B.9)}
\end{aligned}$$

Plugging (B.4) and (B.6)–(B.9) to (B.5) finishes the calculation of (B.2).

When $\boldsymbol{Z}_j \neq (0, 0)^{\mathrm{T}}$, the calculation of $\mathrm{pr}(Z_{j1}, Z_{j2}, \boldsymbol{Y}_j|\mathbf{X}_j, \boldsymbol{\theta})$ in (B.3) follows a similar pattern. We first rewrite it by

$$\begin{aligned}
&\sum_{b_1, b_2 \in \{0,1\}} \mathrm{pr}(Z_{j1}, Z_{j2}|\widetilde{Z}_{j1} = b_1, \widetilde{Z}_{j2} = b_2, \boldsymbol{\theta}_2) \mathrm{pr}(\widetilde{Z}_{j1} = b_1, \widetilde{Z}_{j2} = b_2, \boldsymbol{Y}_j|\mathbf{X}_j, \boldsymbol{\theta}) \\
= &\sum_{b_1, b_2 \in \{0,1\}} \mathcal{M}(Z_{j1}, Z_{j2}|b_1, b_2, \boldsymbol{\theta}_2) \mathrm{pr}(\widetilde{Z}_{j1} = b_1, \widetilde{Z}_{j2} = b_2, \boldsymbol{Y}_j|\mathbf{X}_j, \boldsymbol{\theta}). \qquad \text{(B.10)}
\end{aligned}$$

When $b_1 = b_2 = 0$, we have

$$\begin{aligned}
&\mathrm{pr}(\widetilde{Z}_{j1} = 0, \widetilde{Z}_{j2} = 0, \boldsymbol{Y}_j | \mathbf{X}_j, \boldsymbol{\theta}) \\
&\quad = \prod_{i=1}^{c_j} \mathrm{pr}(\widetilde{Y}_{ij1} = \widetilde{Y}_{ij2} = 0, Y_{ij1}, Y_{ij2} | \mathbf{x}_{ij}, \boldsymbol{\theta}) \\
&\quad = \prod_{i=1}^{c_j} \mathrm{pr}(Y_{ij1}, Y_{ij2} | \widetilde{Y}_{ij1} = 0, \widetilde{Y}_{ij2} = 0, \boldsymbol{\theta}_2) \mathrm{pr}(\widetilde{Y}_{ij1} = 0, \widetilde{Y}_{ij2} = 0 | \mathbf{x}_{ij}, \boldsymbol{\theta}_1) \\
&\quad = \prod_{i=1}^{c_j} \mathcal{M}(Y_{ij1}, Y_{ij2} | 0, 0, \boldsymbol{\theta}_2) p_{ij00}(\boldsymbol{\theta}_1). \qquad \text{(B.11)}
\end{aligned}$$

When $b_1 = 1$ and $b_2 = 0$, we have

$$\begin{aligned}
&\mathrm{pr}(\widetilde{Z}_{j1} = 1, \widetilde{Z}_{j2} = 0, \boldsymbol{Y}_j | \mathbf{X}_j, \boldsymbol{\theta}) \\
&\quad = \mathrm{pr}(\widetilde{Z}_{j2} = 0, \boldsymbol{Y}_j | \mathbf{X}_j, \boldsymbol{\theta}) - \mathrm{pr}(\widetilde{Z}_{j1} = 0, \widetilde{Z}_{j2} = 0, \boldsymbol{Y}_j | \mathbf{X}_j, \boldsymbol{\theta}) \\
&\quad = \prod_{i=1}^{c_j} \mathrm{pr}(\widetilde{Y}_{ij2} = 0, Y_{ij1}, Y_{ij2} | \mathbf{x}_{ij}, \boldsymbol{\theta}) - \prod_{i=1}^{c_j} \mathcal{M}(Y_{ij1}, Y_{ij2} | 0, 0, \boldsymbol{\theta}_2) p_{ij00}(\boldsymbol{\theta}_1) \\
&\quad = \prod_{i=1}^{c_j} \{\mathcal{M}(Y_{ij1}, Y_{ij2} | 0, 0, \boldsymbol{\theta}_2) p_{ij00}(\boldsymbol{\theta}_1) + \mathcal{M}(Y_{ij1}, Y_{ij2} | 1, 0, \boldsymbol{\theta}_2) p_{ij10}(\boldsymbol{\theta}_1)\} \\
&\qquad - \prod_{i=1}^{c_j} \mathcal{M}(Y_{ij1}, Y_{ij2} | 0, 0, \boldsymbol{\theta}_2) p_{ij00}(\boldsymbol{\theta}_1). \qquad \text{(B.12)}
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\mathrm{pr}(\widetilde{Z}_{j1} = 0, \widetilde{Z}_{j2} = 1, \boldsymbol{Y}_j | \mathbf{X}_j, \boldsymbol{\theta}) \\
&\quad = \prod_{i=1}^{c_j} \{\mathcal{M}(Y_{ij1}, Y_{ij2} | 0, 0, \boldsymbol{\theta}_2) p_{ij00}(\boldsymbol{\theta}_1) + \mathcal{M}(Y_{ij1}, Y_{ij2} | 0, 1, \boldsymbol{\theta}_2) p_{ij01}(\boldsymbol{\theta}_1)\} \\
&\qquad - \prod_{i=1}^{c_j} \mathcal{M}(Y_{ij1}, Y_{ij2} | 0, 0, \boldsymbol{\theta}_2) p_{ij00}(\boldsymbol{\theta}_1). \qquad \text{(B.13)}
\end{aligned}$$

Finally,

$$
\begin{aligned}
&\mathrm{pr}(\widetilde{Z}_{j1}=1,\,\widetilde{Z}_{j2}=1,\boldsymbol{Y}_j|\mathbf{X}_j,\boldsymbol{\theta})\\
&\quad=\mathrm{pr}(\boldsymbol{Y}_j|\mathbf{X}_j,\boldsymbol{\theta})-\sum_{b_1+b_2<2}\mathrm{pr}(\widetilde{Z}_{j1}=b_1,\widetilde{Z}_{j2}=b_2,\boldsymbol{Y}_j|\mathbf{X}_j,\boldsymbol{\theta})\\
&\quad=\prod_{i=1}^{c_j}\left\{\sum_{y_1,y_2\in\{0,1\}}\mathcal{M}(Y_{ij1},Y_{ij2}|y_1,y_2,\boldsymbol{\theta}_2)p_{ijy_1y_2}(\boldsymbol{\theta}_1)\right\}\\
&\qquad-\prod_{i=1}^{c_j}\{\mathcal{M}(Y_{ij1},Y_{ij2}|0,0,\boldsymbol{\theta}_2)p_{ij00}(\boldsymbol{\theta}_1)+\mathcal{M}(Y_{ij1},Y_{ij2}|1,0,\boldsymbol{\theta}_2)p_{ij10}(\boldsymbol{\theta}_1)\}\\
&\qquad-\prod_{i=1}^{c_j}\{\mathcal{M}(Y_{ij1},Y_{ij2}|0,0,\boldsymbol{\theta}_2)p_{ij00}(\boldsymbol{\theta}_1)+\mathcal{M}(Y_{ij1},Y_{ij2}|0,1,\boldsymbol{\theta}_2)p_{ij01}(\boldsymbol{\theta}_1)\}\\
&\qquad+\prod_{i=1}^{c_j}\mathcal{M}(Y_{ij1},Y_{ij2}|0,0,\boldsymbol{\theta}_2)p_{ij00}(\boldsymbol{\theta}_1).
\end{aligned}
\tag{B.14}
$$

Combining (B.11)–(B.14) with (B.10) finishes the calculation of (B.3) and thus completes the derivation of $\ell(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}},\mathbf{X})$.

B.3 A computational advantage of the GEM algorithm

As it mentioned in Section 3.3, the GEM algorithm can search for the MLE more efficiently (faster) than maximizing the likelihood directly. To illustrate this numerical advantage, we compare the computational cost of our GEM algorithm with that of maximizing the likelihood directly using the function "optim" in R. The simulation uses the setting S2, $N=3000$, $c=2$, $S_{e:k}=S_{p:k}=0.95$ for $k=1,2$, and the Gumbel copula with $\delta=0.3$. We generate 100 independent sets of the SHL pooled testing data. On each data set, we search for the MLE using both the GEM algorithm and the "optim" function in R. The initial values of both methods are chosen to be the same (as described in Section 3.5), and the stopping criteria are the same as well. We report summary statistics of the MLEs in Table B.2. We observe that the estimates obtained from two approaches are almost identical. However, the mean (standard deviation) of 100 computation times taken by the GEM algorithm is about 39.94 (19.29) seconds, much smaller than that of "optim" function which is 63.60 (58.28)

seconds. We believe this comparison is sufficient to demonstrate the computational advantage of the GEM algorithm.

Table B.2: Simulation results for parameter estimation using the GEM algorithm and the "optim" function. Reported are the sample means (Mean) and the sample standard deviations (SD) of the 100 MLEs.

		GEM algorithm	"optim" function
	Truth	Mean(SD)	Mean(SD)
β_{10}	-4	-4.05(0.20)	-4.04(0.20)
β_{11}	-2	-2.02(0.16)	-2.02(0.16)
β_{12}	2	2.02(0.15)	2.02(0.15)
β_{13}	0	0.00(0.11)	0.00(0.11)
β_{14}	0	0.00(0.11)	0.00(0.11)
β_{15}	0	0.01(0.11)	0.01(0.11)
β_{20}	-5	-5.06(0.29)	-5.05(0.29)
β_{21}	-2	-2.03(0.18)	-2.02(0.18)
β_{22}	0	0.01(0.11)	0.01(0.11)
β_{23}	-2	-2.03(0.18)	-2.02(0.18)
β_{24}	0	-0.02(0.13)	-0.02(0.13)
β_{25}	0	0.00(0.12)	0.00(0.12)
δ	0.3	0.30(0.05)	0.30(0.05)
$S_{e:1}$	0.95	0.95(0.02)	0.95(0.02)
$S_{e:2}$	0.95	0.95(0.02)	0.95(0.02)
$S_{p:1}$	0.95	0.95(0.01)	0.95(0.01)
$S_{p:2}$	0.95	0.95(0.01)	0.95(0.01)

B.4 Calculation of the E-step

B.4.1 Explicit formula of $\eta^{(d)}_{ijy_1y_2}$

For any $y_1, y_2 \in \{0,1\}$, $i = 1, \cdots, c_j$ and $j = 1, \cdots, J$, one can calculate that

$$\begin{aligned}\eta^{(d)}_{ijy_1y_2} &= \mathrm{pr}(\widetilde{Y}_{ij1} = y_1, \widetilde{Y}_{ij2} = y_2 | \boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta}^{(d)}) \\ &= \frac{\mathrm{pr}(\widetilde{Y}_{ij1} = y_1, \widetilde{Y}_{ij2} = y_2, \boldsymbol{\mathcal{P}}_j | \mathbf{X}_j, \boldsymbol{\theta}^{(d)})}{\sum_{y_1,y_2\in\{0,1\}} \mathrm{pr}(\widetilde{Y}_{ij1} = y_1, \widetilde{Y}_{ij2} = y_2, \boldsymbol{\mathcal{P}}_j | \mathbf{X}_j, \boldsymbol{\theta}^{(d)})}.\end{aligned}$$

It suffices to compute the $\mathrm{pr}(\widetilde{Y}_{ij1} = y_1, \widetilde{Y}_{ij2} = y_2, \boldsymbol{\mathcal{P}}_j | \mathbf{X}_j, \boldsymbol{\theta})$ for a generic $\boldsymbol{\theta}$ and $y_1, y_2 \in \{0,1\}$. Again, we consider two cases.

The first case is when $\boldsymbol{Z}_j = (0,0)^{\mathrm{T}}$. Then $\mathrm{pr}(\widetilde{Y}_{ij1} = y_1, \widetilde{Y}_{ij2} = y_2, \boldsymbol{\mathcal{P}}_j | \mathbf{X}_j, \boldsymbol{\theta}) = \mathrm{pr}(\widetilde{Y}_{ij1} = y_1, \widetilde{Y}_{ij2} = y_2, Z_{j1} = Z_{j2} = 0 | \mathbf{X}_j, \boldsymbol{\theta})$. Denote $\mathcal{G}_j = \{1, \ldots, c_j\}$. For $y_1 = 0$

and $y_2 = 0$, using the Law of Total Probability provides that

$$
\begin{aligned}
&\mathrm{pr}(\widetilde{Y}_{ij1} = \widetilde{Y}_{ij2} = Z_{j1} = Z_{j2} = 0 | \mathbf{X}_j, \boldsymbol{\theta}) \\
&= \sum_{a,b \in \{0,1\}} \left\{ \mathrm{pr}\left(\widetilde{Y}_{ij1} = \widetilde{Y}_{ij2} = Z_{j1} = Z_{j2} = 0 \middle| \max_{l \in \mathcal{G}_j \setminus \{i\}} \widetilde{Y}_{lj1} = a, \max_{l \in \mathcal{G}_j \setminus \{i\}} \widetilde{Y}_{lj2} = b, \mathbf{X}_j, \boldsymbol{\theta} \right) \right. \\
&\quad \left. \times \mathrm{pr}\left(\max_{l \in \mathcal{G}_j \setminus \{i\}} \widetilde{Y}_{lj1} = a, \max_{l \in \mathcal{G}_j \setminus \{i\}} \widetilde{Y}_{lj2} = b \middle| \mathbf{X}_j, \boldsymbol{\theta} \right) \right\} \\
&= \sum_{a,b \in \{0,1\}} \mathrm{pr}\left(Z_{j1} = 0, Z_{j2} = 0 \middle| \widetilde{Z}_{j1} = a, \widetilde{Z}_{j2} = b, \boldsymbol{\theta}_2 \right) p_{ij00}(\boldsymbol{\theta}_1) \gamma_{ijab}^{(-i)}(\boldsymbol{\theta}_1) \\
&= p_{ij00}(\boldsymbol{\theta}_1) \sum_{a,b \in \{0,1\}} \mathcal{M}(0,0|a,b,\boldsymbol{\theta}_2) \gamma_{ijab}^{(-i)}(\boldsymbol{\theta}_1), \qquad \text{(B.15)}
\end{aligned}
$$

where

$$
\gamma_{ijab}^{(-i)}(\boldsymbol{\theta}_1) = \mathrm{pr}\left(\max_{l \in \mathcal{G}_j \setminus \{i\}} \widetilde{Y}_{lj1} = a, \max_{l \in \mathcal{G}_j \setminus \{i\}} \widetilde{Y}_{lj2} = b \middle| \mathbf{X}_j \right),
$$

which can be calculated by

$$
\begin{aligned}
\gamma_{ij00}^{(-i)}(\boldsymbol{\theta}_1) &= \mathrm{pr}\left\{ \max_{l \in \mathcal{G}_j \setminus \{i\}} \widetilde{Y}_{lj1} = 0, \max_{l \in \mathcal{G}_j \setminus \{i\}} \widetilde{Y}_{lj2} = 0 \middle| \mathbf{X}_j \right\} = \prod_{l \in \mathcal{G}_j \setminus \{i\}} p_{lj00}(\boldsymbol{\theta}_1), \\
\gamma_{ij10}^{(-i)}(\boldsymbol{\theta}_1) &= \prod_{l \in \mathcal{G}_j \setminus \{i\}} \{ p_{lj00}(\boldsymbol{\theta}_1) + p_{lj10}(\boldsymbol{\theta}_1) \} - \gamma_{ij00}^{(-i)}(\boldsymbol{\theta}_1), \\
\gamma_{ij01}^{(-i)}(\boldsymbol{\theta}_1) &= \prod_{l \in \mathcal{G}_j \setminus \{i\}} \{ p_{lj00}(\boldsymbol{\theta}_1) + p_{lj01}(\boldsymbol{\theta}_1) \} - \gamma_{ij00}^{(-i)}(\boldsymbol{\theta}_1), \\
\gamma_{ij11}^{(-i)}(\boldsymbol{\theta}_1) &= 1 - \gamma_{ij00}^{(-i)}(\boldsymbol{\theta}_1) - \gamma_{ij01}^{(-i)}(\boldsymbol{\theta}_1) - \gamma_{ij10}^{(-i)}(\boldsymbol{\theta}_1).
\end{aligned}
$$

One can calculate

$$
\begin{aligned}
&\mathrm{pr}(\widetilde{Y}_{ij1} = 1, \widetilde{Y}_{ij2} = 0, Z_{j1} = Z_{j2} = 0 | \mathbf{X}_j, \boldsymbol{\theta}) \\
&\quad = \mathrm{pr}(Z_{j1} = Z_{j2} = 0, \widetilde{Y}_{ij1} = 1, \widetilde{Y}_{ij2} = 0, \max_{l \in \mathcal{G}_j \setminus \{i\}} \widetilde{Y}_{lj2} = 0 | \mathbf{X}_j, \boldsymbol{\theta}) \\
&\qquad + \mathrm{pr}(Z_{j1} = Z_{j2} = 0, \widetilde{Y}_{ij1} = 1, \widetilde{Y}_{ij2} = 0, \max_{l \in \mathcal{G}_j \setminus \{i\}} \widetilde{Y}_{lj2} = 1 | \mathbf{X}_j, \boldsymbol{\theta}) \\
&\quad = \mathcal{M}(0,0|1,0,\boldsymbol{\theta}_2) p_{ij10}(\boldsymbol{\theta}_1) \left\{ \gamma_{ij00}^{(-i)}(\boldsymbol{\theta}_1) + \gamma_{ij10}^{(-i)}(\boldsymbol{\theta}_1) \right\} \\
&\qquad + \mathcal{M}(0,0|1,1,\boldsymbol{\theta}_2) p_{ij10}(\boldsymbol{\theta}_1) \left\{ \gamma_{ij01}^{(-i)}(\boldsymbol{\theta}_1) + \gamma_{ij11}^{(-i)}(\boldsymbol{\theta}_1) \right\}. \qquad \text{(B.16)}
\end{aligned}
$$

Similarly,

$$\begin{aligned}
&\mathrm{pr}(\widetilde{Y}_{ij1}=0,\,\widetilde{Y}_{ij2}=1,Z_{j1}=Z_{j2}=0|\mathbf{X}_j,\boldsymbol{\theta})\\
&\quad=\mathrm{pr}(Z_{j1}=Z_{j2}=0,\widetilde{Y}_{ij1}=0,\widetilde{Y}_{ij2}=1,\max_{l\in\mathcal{G}_j\setminus\{i\}}\widetilde{Y}_{lj1}=0|\mathbf{X}_j,\boldsymbol{\theta})\\
&\qquad+\mathrm{pr}(Z_{j1}=Z_{j2}=0,\widetilde{Y}_{ij1}=0,\widetilde{Y}_{ij2}=1,\max_{l\in\mathcal{G}_j\setminus\{i\}}\widetilde{Y}_{lj1}=1|\mathbf{X}_j,\boldsymbol{\theta})\\
&\quad=\mathcal{M}(0,0|0,1,\boldsymbol{\theta}_2)p_{ij01}(\boldsymbol{\theta}_1)\left\{\gamma_{ij00}^{(-i)}(\boldsymbol{\theta}_1)+\gamma_{ij01}^{(-i)}(\boldsymbol{\theta}_1)\right\}\\
&\qquad+\mathcal{M}(0,0|1,1,\boldsymbol{\theta}_2)p_{ij01}(\boldsymbol{\theta}_1)\left\{\gamma_{ij10}^{(-i)}(\boldsymbol{\theta}_1)+\gamma_{ij11}^{(-i)}(\boldsymbol{\theta}_1)\right\}. \qquad \text{(B.17)}
\end{aligned}$$

Finally,

$$\mathrm{pr}(\widetilde{Y}_{ij1}=1,\widetilde{Y}_{ij2}=1,\boldsymbol{\mathcal{P}}_j)=(1-S_{e:1})(1-S_{e:2})p_{ij11}(\boldsymbol{\theta}_1). \qquad \text{(B.18)}$$

Combining (B.15)–(B.18) finishes the calculation of $\eta_{ijy_1y_2}^{(d)}$ when $\boldsymbol{\mathcal{P}}_j=\boldsymbol{Z}_j=(0,0)^{\mathrm{T}}$.

The second case is when $\boldsymbol{Z}_j\neq(0,0)^{\mathrm{T}}$ and we observe $\{Z_{j1},Z_{j2},\boldsymbol{Y}_{1j},\cdots,\boldsymbol{Y}_{c_jj}\}$. Again, we start the calculation of $\mathrm{pr}(\widetilde{Y}_{ij1}=y_1,\widetilde{Y}_{ij2}=y_2,\boldsymbol{\mathcal{P}}_j|\mathbf{X}_j,\boldsymbol{\theta})$ by considering $y_1=y_2=0$; i.e.,

$$\begin{aligned}
&\mathrm{pr}(\widetilde{Y}_{ij1}=0,\,\widetilde{Y}_{ij2}=0,Z_{j1},Z_{j2},\boldsymbol{Y}_{1j},\cdots,\boldsymbol{Y}_{c_jj}|\mathbf{X}_j,\boldsymbol{\theta})\\
&\quad=\sum_{b_1,b_2\in\{0,1\}}\Bigg\{\mathrm{pr}\Big(Z_{j1},Z_{j2}|\max_i\widetilde{Y}_{ij1}=b_1,\max_i\widetilde{Y}_{ij2}=b_2,\boldsymbol{\theta}_2\Big)\\
&\qquad\times\mathrm{pr}(\widetilde{Y}_{ij1}=0,\widetilde{Y}_{ij2}=0,Y_{ij1},Y_{ij2}|\mathbf{X}_j,\boldsymbol{\theta})\\
&\qquad\times\mathrm{pr}\Big(\boldsymbol{Y}_{1j},\ldots,\boldsymbol{Y}_{i-1,j},\boldsymbol{Y}_{i+1,j},\ldots,\boldsymbol{Y}_{c_jj},\\
&\qquad\qquad\max_{i\in\mathcal{G}_j\setminus\{i\}}\widetilde{Y}_{ij1}=b_1,\max_{i\in\mathcal{G}_j\setminus\{i\}}\widetilde{Y}_{ij2}=b_2|\mathbf{X}_j,\boldsymbol{\theta}\Big)\Bigg\}\\
&\quad=\mathcal{M}(Y_{ij1},Y_{ij2}|0,0,\boldsymbol{\theta}_2)p_{ij00}(\boldsymbol{\theta}_1)\sum_{b_1,b_2\in\{0,1\}}\mathcal{M}(Z_{j1},Z_{j2}|b_1,b_2,\boldsymbol{\theta}_2)\zeta_{ijb_1b_2}(\boldsymbol{\theta}),
\end{aligned} \qquad \text{(B.19)}$$

where

$$\zeta_{ijb_1b_2}(\boldsymbol{\theta})=\mathrm{pr}(\boldsymbol{Y}_{1j},\ldots,\boldsymbol{Y}_{i-1,j},\boldsymbol{Y}_{i+1,j},\ldots,\boldsymbol{Y}_{c_jj},\max_{i\in\mathcal{G}_j\setminus\{i\}}\widetilde{Y}_{ij1}=b_1,\max_{i\in\mathcal{G}_j\setminus\{i\}}\widetilde{Y}_{ij2}=b_2|\mathbf{X}_j,\boldsymbol{\theta})$$

which can be calculated by

$$\begin{aligned}
\zeta_{ij00}(\boldsymbol{\theta}) &= \prod_{l\in\mathcal{G}_j\setminus\{i\}} \mathcal{M}(Y_{lj1}, Y_{lj2}|0,0,\boldsymbol{\theta}_2)p_{lj00}(\boldsymbol{\theta}_1)\\
\zeta_{ij10}(\boldsymbol{\theta}) &= \prod_{l\in\mathcal{G}_j\setminus\{i\}} \{\mathcal{M}(Y_{lj1}, Y_{lj2}|0,0,\boldsymbol{\theta}_2)p_{lj00}(\boldsymbol{\theta}_1) + \mathcal{M}(Y_{lj1}, Y_{lj2}|1,0,\boldsymbol{\theta}_2)p_{lj10}(\boldsymbol{\theta}_1)\}\\
&\quad - \zeta_{ij00}(\boldsymbol{\theta})\\
\zeta_{ij01}(\boldsymbol{\theta}) &= \prod_{l\in\mathcal{G}_j\setminus\{i\}} \{\mathcal{M}(Y_{lj1}, Y_{lj2}|0,0,\boldsymbol{\theta}_2)p_{lj00}(\boldsymbol{\theta}_1) + \mathcal{M}(Y_{lj1}, Y_{lj2}|0,1,\boldsymbol{\theta}_2)p_{lj01}(\boldsymbol{\theta}_1)\}\\
&\quad - \zeta_{ij00}(\boldsymbol{\theta})\\
\zeta_{ij11}(\boldsymbol{\theta}) &= \prod_{l\in\mathcal{G}_j\setminus\{i\}} \left\{\sum_{y_1,y_2\in\{0,1\}} \mathcal{M}(Y_{lj1}, Y_{lj2}|y_1,y_2,\boldsymbol{\theta}_2)p_{ljy_1y_2}(\boldsymbol{\theta}_1)\right\} - \sum_{b_1+b_2<2} \zeta_{ijb_1b_2}(\boldsymbol{\theta}).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
&\mathrm{pr}(\widetilde{Y}_{ij1}=1,\, \widetilde{Y}_{ij2}=0, Z_{j1}, Z_{j2}, \boldsymbol{Y}_{1j}, \cdots, \boldsymbol{Y}_{c_jj}|\mathbf{X}_j, \boldsymbol{\theta})\\
&\quad= \mathcal{M}(Y_{ij1}, Y_{ij2}|1,0,\boldsymbol{\theta}_2)p_{ij10}(\boldsymbol{\theta}_1)\Big[\mathcal{M}(Z_{j1}, Z_{j2}|1,0,\boldsymbol{\theta}_2)\Big\{\zeta_{ij00}(\boldsymbol{\theta}_1)+\zeta_{ij10}(\boldsymbol{\theta}_1)\Big\}\\
&\qquad+ \mathcal{M}(Z_{j1}, Z_{j2}|1,1,\boldsymbol{\theta}_2)\Big\{\zeta_{ij01}(\boldsymbol{\theta}_1)+\zeta_{ij11}(\boldsymbol{\theta}_1)\Big\}\Big] \qquad \text{(B.20)}\\
&\mathrm{pr}(\widetilde{Y}_{ij1}=0,\, \widetilde{Y}_{ij2}=1, Z_{j1}, Z_{j2}, \boldsymbol{Y}_{1j}, \cdots, \boldsymbol{Y}_{c_jj}|\mathbf{X}_j, \boldsymbol{\theta})\\
&\quad= \mathcal{M}(Y_{ij1}, Y_{ij2}|0,1,\boldsymbol{\theta}_2)p_{ij01}(\boldsymbol{\theta}_1)\Big[\mathcal{M}(Z_{j1}, Z_{j2}|0,1,\boldsymbol{\theta}_2)\Big\{\zeta_{ij00}(\boldsymbol{\theta}_1)+\zeta_{ij01}(\boldsymbol{\theta}_1)\Big\}\\
&\qquad+ \mathcal{M}(Z_{j1}, Z_{j2}|1,1,\boldsymbol{\theta}_2)\Big\{\zeta_{ij10}(\boldsymbol{\theta}_1)+\zeta_{ij11}(\boldsymbol{\theta}_1)\Big\}\Big] \qquad \text{(B.21)}\\
&\mathrm{pr}(\widetilde{Y}_{ij1}=1,\, \widetilde{Y}_{ij2}=1, Z_{j1}, Z_{j2}, \boldsymbol{Y}_{1j}, \cdots, \boldsymbol{Y}_{c_jj}|\mathbf{X}_j, \boldsymbol{\theta})\\
&\quad= \mathcal{M}(Y_{ij1}, Y_{ij2}|1,1,\boldsymbol{\theta}_2)p_{ij11}(\boldsymbol{\theta}_1)\mathcal{M}(Z_{j1}, Z_{j2}|1,1,\boldsymbol{\theta}_2)\sum_{a,b\in\{0,1\}}\zeta_{ijab}(\boldsymbol{\theta}). \qquad \text{(B.22)}
\end{aligned}$$

Combining (B.19)–(B.22) finishes the calculation of $\eta^{(d)}_{ijy_1y_2}$ when $\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}}$.

B.4.2 Explicit formula of $\eta^{(d)}_{\mathcal{P},jk}$

Recall that $\eta^{(d)}_{\mathcal{P},jk} = \mathrm{pr}(\widetilde{Z}_{jk}=1|\boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta}^{(d)})$. It suffices to calculate

$$\eta_{j,ab} = \mathrm{pr}(\widetilde{Z}_{j1}=a, \widetilde{Z}_{j2}=b|\boldsymbol{\mathcal{P}}_j, \mathbf{X}_j, \boldsymbol{\theta}) = \frac{\mathrm{pr}(\widetilde{Z}_{j1}=a, \widetilde{Z}_{j2}=b, \boldsymbol{\mathcal{P}}_j|\mathbf{X}_j, \boldsymbol{\theta})}{\sum_{a,b\in\{0,1\}}\mathrm{pr}(\widetilde{Z}_{j1}=a, \widetilde{Z}_{j2}=b, \boldsymbol{\mathcal{P}}_j|\mathbf{X}_j, \boldsymbol{\theta})},$$

for any $j = 1, \cdots, J$, $a, b \in \{0, 1\}$ and $\boldsymbol{\theta}$ because $\eta^{(d)}_{\mathcal{P},j1} = \eta_{j,10} + \eta_{j,11}$ and $\eta^{(d)}_{\mathcal{P},j2} = \eta_{j,01} + \eta_{j,11}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(d)}$. Again, we calculate the $\text{pr}(\widetilde{Z}_{j1} = a, \widetilde{Z}_{j2} = b, \boldsymbol{\mathcal{P}}_j | \mathbf{X}_j, \boldsymbol{\theta})$ for $a, b \in \{0, 1\}$ under two scenarios.

The first one is when $\boldsymbol{\mathcal{P}}_j = (Z_{j1}, Z_{j2})^{\text{T}} = (0, 0)^{\text{T}}$. One can easily see that

$$\begin{aligned}\text{pr}(\widetilde{Z}_{j1} = a, \widetilde{Z}_{j2} = b, \boldsymbol{\mathcal{P}}_j | \mathbf{X}_j, \boldsymbol{\theta}) = \ & \text{pr}(\widetilde{Z}_{j1} = a, \widetilde{Z}_{j2} = b, Z_{j1} = 0, Z_{j2} = 0 | \mathbf{X}_j, \boldsymbol{\theta}) \\ & = \mathcal{M}(0, 0 | a, b, \boldsymbol{\theta}_2) \text{pr}(\widetilde{Z}_{j1} = a, \widetilde{Z}_{j2} = b | \mathbf{X}_j, \boldsymbol{\theta}_1), \quad \text{(B.23)}\end{aligned}$$

where $\text{pr}(\widetilde{Z}_{j1} = a, \widetilde{Z}_{j2} = b | \mathbf{X}_j, \boldsymbol{\theta}_1)$ has been computed in (B.6) – (B.9).

For the second case where $\boldsymbol{Z}_j \neq (0, 0)^{\text{T}}$ and $\boldsymbol{\mathcal{P}}_j = \{Z_{j1}, Z_{j2}, \boldsymbol{Y}_{1j}, \cdots, \boldsymbol{Y}_{c_j j}\}$, we have

$$\begin{aligned}\text{pr}(\widetilde{Z}_{j1} = a, \widetilde{Z}_{j2} = b, \boldsymbol{\mathcal{P}}_j | \mathbf{X}_j, \boldsymbol{\theta}) &= \text{pr}(\widetilde{Z}_{j1} = a, \widetilde{Z}_{j2} = b, Z_{j1}, Z_{j2}, \boldsymbol{Y}_j | \mathbf{X}_j, \boldsymbol{\theta}) \\ &= \mathcal{M}(Z_{j1}, Z_{j2} | a, b, \boldsymbol{\theta}_2) \text{pr}(\boldsymbol{Y}_j, \widetilde{Z}_{j1} = a, \widetilde{Z}_{j2} = b | \mathbf{X}_j, \boldsymbol{\theta}), \\ & \quad \text{(B.24)}\end{aligned}$$

where $\text{pr}(\boldsymbol{Y}_j, \widetilde{Z}_{j1} = a, \widetilde{Z}_{j2} = b | \mathbf{X}_j, \boldsymbol{\theta})$ has been computed in (B.11) – (B.14). Combining (B.23) and (B.24) completes the computation of $\eta^{(d)}_{\mathcal{P},jk}$.

B.5 Explicit calculation of $\mathcal{I}(\boldsymbol{\theta})$

Recall that in Section 3, we have presented

$$\mathcal{I}(\boldsymbol{\theta}) = -E\left\{ \frac{\partial^2 l_c(\boldsymbol{\theta} | \boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}}, \mathbf{X})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^{\text{T}}} \middle| \boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta} \right\} - \text{cov}\left\{ \frac{\partial l_c(\boldsymbol{\theta} | \boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}}, \mathbf{X})}{\partial \boldsymbol{\theta}} \middle| \boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta} \right\}, \quad \text{(B.25)}$$

where

$$\ell_c(\boldsymbol{\theta} | \boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}}, \mathbf{X}) = \ell_{c1}(\boldsymbol{\theta}_1 | \widetilde{\boldsymbol{Y}}, \mathbf{X}) + \ell_{c2}(\boldsymbol{\theta}_2 | \boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}})$$

with

$$\begin{aligned}\ell_{c1}(\boldsymbol{\theta}_1 | \widetilde{\boldsymbol{Y}}, \mathbf{X}) = \sum_{j=1}^{J} \sum_{i=1}^{c_j} \Big\{ & (1 - \widetilde{Y}_{ij1})(1 - \widetilde{Y}_{ij2}) \log p_{ij00}(\boldsymbol{\theta}_1) + \widetilde{Y}_{ij1}(1 - \widetilde{Y}_{ij2}) \log p_{ij10}(\boldsymbol{\theta}_1) \\ & + (1 - \widetilde{Y}_{ij1}) \widetilde{Y}_{ij2} \log p_{ij01}(\boldsymbol{\theta}_1) + \widetilde{Y}_{ij1} \widetilde{Y}_{ij2} \log p_{ij11}(\boldsymbol{\theta}_1) \Big\}\end{aligned}$$

$$
\begin{aligned}
\ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}},\widetilde{\boldsymbol{Y}}) = \sum_{j=1}^{J}\sum_{k=1}^{2}\Bigg[& \left\{\widetilde{Z}_{jk}Z_{jk} + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}})\sum_{i=1}^{c_j}\widetilde{Y}_{ijk}Y_{ijk}\right\}\log S_{e:k} \\
& + \left\{\widetilde{Z}_{jk}(1-Z_{jk}) + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}})\sum_{i=1}^{c_j}\widetilde{Y}_{ijk}(1-Y_{ijk})\right\}\log(1-S_{e:k}) \\
& + \left\{(1-\widetilde{Z}_{jk})(1-Z_{jk}) + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}})\sum_{i=1}^{c_j}(1-\widetilde{Y}_{ijk})(1-Y_{ijk})\right\}\log S_{p:k} \\
& + \left\{(1-\widetilde{Z}_{jk})Z_{jk} + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}})\sum_{i=1}^{c_j}(1-\widetilde{Y}_{ijk})Y_{ijk}\right\}\log(1-S_{p:k})\Bigg].
\end{aligned}
$$

In the following, we calculate the expectation term of (B.25) in Appendix B.5.1 and the covariance term in Appendix B.5.2.

B.5.1 The calculation of expectation term of (B.25)

One can easily calculate that

$$
\frac{\partial^2 \ell(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}},\widetilde{\boldsymbol{Y}},\boldsymbol{X})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}} = \begin{bmatrix} \frac{\partial^2 \ell_{c1}(\boldsymbol{\theta}_1|\widetilde{\boldsymbol{Y}},\boldsymbol{X})}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1^{\mathrm{T}}} & \mathbf{0} \\ \mathbf{0} & \frac{\partial^2 \ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}},\widetilde{\boldsymbol{Y}})}{\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_2^{\mathrm{T}}} \end{bmatrix}, \tag{B.26}
$$

where

$$
\begin{aligned}
\frac{\partial^2 \ell_{c1}(\boldsymbol{\theta}_1|\widetilde{\boldsymbol{Y}},\boldsymbol{X})}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1^{\mathrm{T}}} = \sum_{j=1}^{J}\sum_{i=1}^{c_j}\Bigg\{ & -\frac{(1-\widetilde{Y}_{ij1})(1-\widetilde{Y}_{ij2})}{p_{ij00}^2(\boldsymbol{\theta}_1)}\left\{\frac{\partial p_{ij00}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1}\right\}\left\{\frac{\partial p_{ij00}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1}\right\}^{\mathrm{T}} \\
& - \frac{\widetilde{Y}_{ij1}(1-\widetilde{Y}_{ij2})}{p_{ij10}^2(\boldsymbol{\theta}_1)}\left\{\frac{\partial p_{ij10}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1}\right\}\left\{\frac{\partial p_{ij10}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1}\right\}^{\mathrm{T}} \\
& - \frac{(1-\widetilde{Y}_{ij1})\widetilde{Y}_{ij2}}{p_{ij01}^2(\boldsymbol{\theta}_1)}\left\{\frac{\partial p_{ij01}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1}\right\}\left\{\frac{\partial p_{ij01}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1}\right\}^{\mathrm{T}} \\
& - \frac{\widetilde{Y}_{ij1}\widetilde{Y}_{ij2}}{p_{ij11}^2(\boldsymbol{\theta}_1)}\left\{\frac{\partial p_{ij11}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1}\right\}\left\{\frac{\partial p_{ij11}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1}\right\}^{\mathrm{T}} \\
& + \frac{(1-\widetilde{Y}_{ij1})(1-\widetilde{Y}_{ij2})}{p_{ij00}(\boldsymbol{\theta}_1)}\frac{\partial^2 p_{ij00}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1^{\mathrm{T}}} \\
& + \frac{\widetilde{Y}_{ij1}(1-\widetilde{Y}_{ij2})}{p_{ij10}(\boldsymbol{\theta}_1)}\frac{\partial^2 p_{ij10}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1^{\mathrm{T}}} \\
& + \frac{(1-\widetilde{Y}_{ij1})\widetilde{Y}_{ij2}}{p_{ij01}(\boldsymbol{\theta}_1)}\frac{\partial^2 p_{ij01}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1^{\mathrm{T}}} + \frac{\widetilde{Y}_{ij1}\widetilde{Y}_{ij2}}{p_{ij11}(\boldsymbol{\theta}_1)}\frac{\partial^2 p_{ij11}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1^{\mathrm{T}}}\Bigg\},
\end{aligned} \tag{B.27}
$$

and

$$\frac{\partial^2 \ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}},\widetilde{\boldsymbol{Y}})}{\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_2^{\mathrm{T}}} = \begin{bmatrix} \frac{\partial^2 \ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}},\widetilde{\boldsymbol{Y}})}{\partial S_{e:1}^2} & 0 & 0 & 0 \\ 0 & \frac{\partial^2 \ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}},\widetilde{\boldsymbol{Y}})}{\partial S_{e:2}^2} & 0 & 0 \\ 0 & 0 & \frac{\partial^2 \ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}},\widetilde{\boldsymbol{Y}})}{\partial S_{p:1}^2} & 0 \\ 0 & 0 & 0 & \frac{\partial^2 \ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}},\widetilde{\boldsymbol{Y}})}{\partial S_{p:2}^2} \end{bmatrix}$$

with

$$\begin{aligned} \frac{\partial^2 \ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}},\widetilde{\boldsymbol{Y}})}{\partial S_{e:k}^2} = \frac{1}{S_{e:k}^2(1-S_{e:k})^2} \Bigg\{ \sum_{j=1}^{J} \Bigg[\widetilde{Z}_{jk} \left\{ Z_{jk}(2S_{e:k}-1) - S_{e:k}^2 \right\} \\ + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}}) \sum_{i=1}^{c_j} \widetilde{Y}_{ijk} \left\{ Y_{ijk}(2S_{e:k}-1) - S_{e:k}^2 \right\} \Bigg] \Bigg\} \end{aligned} \tag{B.28}$$

$$\begin{aligned} \frac{\partial^2 \ell_{c2}(\boldsymbol{\theta}_2|\boldsymbol{\mathcal{P}},\widetilde{\boldsymbol{Y}})}{\partial S_{p:k}^2} = \frac{1}{S_{p:k}^2(1-S_{p:k})^2} \Bigg\{ \sum_{j=1}^{J} \Bigg[(1-\widetilde{Z}_{jk}) \left\{ (1-Z_{jk})(2S_{p:k}-1) - S_{p:k}^2 \right\} \\ + I(\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}}) \sum_{i=i}^{c_j} (1-\widetilde{Y}_{ijk}) \left\{ (1-Y_{ijk})(2S_{p:k}-1) - S_{p:k}^2 \right\} \Bigg] \Bigg\}. \end{aligned} \tag{B.29}$$

A quick inspection of (B.27)–(B.29) shows that, to compute (B.26), it suffices to calculate

$$\mathrm{pr}(\widetilde{Y}_{ij1} = y_1, \widetilde{Y}_{ij2} = y_2|\boldsymbol{\mathcal{P}},\mathbf{X},\boldsymbol{\theta}) \text{ and } \mathrm{pr}(\widetilde{Z}_{j1} = b_1, \widetilde{Z}_{j2} = b_2|\boldsymbol{\mathcal{P}},\mathbf{X},\boldsymbol{\theta})$$

for $y_1, y_2, b_1, b_2 \in \{0,1\}$. These calculations are derived explicitly in Appendix B.4.

B.5.2 The calculation of the covariance term of (B.25)

Note that first derivative of complete log-likelihood $\partial \ell_c(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}}, \mathbf{X})/\partial\boldsymbol{\theta}$ can be rewritten as

$$\begin{aligned}
&\sum_{j=1}^{J}\sum_{i=1}^{c_j}\Bigg[\frac{1}{p_{ij00}(\boldsymbol{\theta}_1)}\frac{\partial p_{ij00}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}} \\
&+ \widetilde{Y}_{ij1}(1-\widetilde{Y}_{ij2})\left\{\frac{1}{p_{ij10}(\boldsymbol{\theta}_1)}\frac{\partial p_{ij10}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}} - \frac{1}{p_{ij00}(\boldsymbol{\theta}_1)}\frac{\partial p_{ij00}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}}\right\} \\
&+ (1-\widetilde{Y}_{ij1})\widetilde{Y}_{ij2}\left\{\frac{1}{p_{ij01}(\boldsymbol{\theta}_1)}\frac{\partial p_{ij01}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}} - \frac{1}{p_{ij00}(\boldsymbol{\theta}_1)}\frac{\partial p_{ij00}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}}\right\} \\
&+ \widetilde{Y}_{ij1}\widetilde{Y}_{ij2}\left\{\frac{1}{p_{ij11}(\boldsymbol{\theta}_1)}\frac{\partial p_{ij11}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}} - \frac{1}{p_{ij00}(\boldsymbol{\theta}_1)}\frac{\partial p_{ij00}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}}\right\}\Bigg] \\
&+ \sum_{j=1}^{J}\sum_{k=1}^{2}\Bigg[\left\{\frac{\widetilde{Z}_{jk}(Z_{jk}-S_{e:k}) + \sum_{i\in\mathcal{I}_+}\widetilde{Y}_{ijk}(Y_{ijk}-S_{e:k})}{S_{e:k}(1-S_{e:k})}\right\}\frac{\partial S_{e:k}}{\partial\boldsymbol{\theta}} \\
&+ \left\{\frac{(1-\widetilde{Z}_{jk})(1-Z_{jk}-S_{p:k}) + \sum_{i\in\mathcal{I}_+}(1-\widetilde{Y}_{ijk})(1-Y_{ijk}-S_{p:k})}{S_{p:k}(1-S_{p:k})}\right\}\frac{\partial S_{p:k}}{\partial\boldsymbol{\theta}}\Bigg]. \quad \text{(B.30)}
\end{aligned}$$

For any $y_1, y_2 \in \{0,1\}$, let $\boldsymbol{p}_{y_1y_2}$ denote the collection of all individuals' cell probabilities for the infection status "y_1y_2"; i.e., $\boldsymbol{p}_{y_1y_2} = (p_{11y_1y_2}(\boldsymbol{\theta}_1), \cdots, p_{c_JJy_1y_2}(\boldsymbol{\theta}_1))^{\mathrm{T}}$, an N-dimensional vector. For any vector $\boldsymbol{p}$ of length m, we denote diag($\boldsymbol{p}$) by an $m\times m$-dimensional diagonal matrix with the diagonals being $\boldsymbol{p}$. Then diag$\{1/\boldsymbol{p}_{y_1y_2}\}$ is an $N\times N$ matrix and $\partial\boldsymbol{p}_{y_1y_2}/\partial\boldsymbol{\theta}^{\mathrm{T}}$ is an $N\times(2p+7)$ matrix. Furthermore, we denote $\widetilde{\boldsymbol{Y}}_{(k)} = (\widetilde{Y}_{11k}, \cdots, \widetilde{Y}_{c_JJk})^{\mathrm{T}}$, $\boldsymbol{Y}_{(k)} = (Y_{11k}, \cdots, Y_{c_JJk})^{\mathrm{T}}$, $\boldsymbol{c} = (\underbrace{c_1, \cdots, c_1}_{c_1}, \cdots, \underbrace{c_J \cdots, c_J}_{c_J})^{\mathrm{T}}$, $\boldsymbol{Z}_{(k)} = (\underbrace{Z_{1k}, \cdots, Z_{1k}}_{c_1}, \cdots, \underbrace{Z_{Jk}, \cdots, Z_{Jk}}_{c_J})^{\mathrm{T}}$, and

$$\widetilde{\boldsymbol{Z}}_{(k)} = (\underbrace{\widetilde{Z}_{1k}, \cdots, \widetilde{Z}_{1k}}_{c_1}, \cdots, \underbrace{\widetilde{Z}_{Jk}, \cdots, \widetilde{Z}_{Jk}}_{c_J})^{\mathrm{T}}.$$

Additionally, we let $\mathbf{0}_{N\times m}$ ($\mathbf{1}_{N\times m}$) represents an $N\times m$ zero (one) matrix and denote "$\odot$"("$\oslash$") by the Hadamard product (division), i.e., $[\boldsymbol{A}\odot\boldsymbol{B}]_{i,j} = [\boldsymbol{A}]_{i,j}[\boldsymbol{B}]_{i,j}$ and $[\boldsymbol{A}\oslash\boldsymbol{B}]_{i,j} = [\boldsymbol{A}]_{i,j}/[\boldsymbol{B}]_{i,j}$. With these notation, we are able to express (B.30) in a concise matrix form,

$$\frac{\partial \ell_c(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}}, \mathbf{X})}{\partial\boldsymbol{\theta}} = \boldsymbol{Q}^{\mathrm{T}}\boldsymbol{V} + \sum_{j=1}^{J}\sum_{i=1}^{c_j}\frac{1}{p_{ij00}(\boldsymbol{\theta}_1)}\frac{\partial p_{ij00}(\boldsymbol{\theta}_1)}{\partial\boldsymbol{\theta}}$$

where

$$\boldsymbol{Q} = \begin{bmatrix} \left\{\mathrm{diag}(\frac{1}{\boldsymbol{p}_{10}})\frac{\partial \boldsymbol{p}_{10}}{\partial \boldsymbol{\theta}^{\mathrm{T}}} - \mathrm{diag}(\frac{1}{\boldsymbol{p}_{00}})\frac{\partial \boldsymbol{p}_{00}}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right\}_{N\times(2p+7)} \\ \left\{\mathrm{diag}(\frac{1}{\boldsymbol{p}_{01}})\frac{\partial \boldsymbol{p}_{01}}{\partial \boldsymbol{\theta}^{\mathrm{T}}} - \mathrm{diag}(\frac{1}{\boldsymbol{p}_{00}})\frac{\partial \boldsymbol{p}_{00}}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right\}_{N\times(2p+7)} \\ \left\{\mathrm{diag}(\frac{1}{\boldsymbol{p}_{11}})\frac{\partial \boldsymbol{p}_{11}}{\partial \boldsymbol{\theta}^{\mathrm{T}}} - \mathrm{diag}(\frac{1}{\boldsymbol{p}_{00}})\frac{\partial \boldsymbol{p}_{00}}{\partial \boldsymbol{\theta}^{\mathrm{T}}}\right\}_{N\times(2p+7)} \\ \frac{1}{S_{e:1}(1-S_{e:1})}\left\{\boldsymbol{0}_{N\times(2p+3)}, \boldsymbol{1}_{N\times 1}, \boldsymbol{0}_{N\times 3}\right\} \\ \frac{1}{S_{e:2}(1-S_{e:2})}\left\{\boldsymbol{0}_{N\times(2p+4)}, \boldsymbol{1}_{N\times 1}, \boldsymbol{0}_{N\times 2}\right\} \\ \frac{1}{S_{p:1}(1-S_{p:1})}\left\{\boldsymbol{0}_{N\times(2p+5)}, \boldsymbol{1}_{N\times 1}, \boldsymbol{0}_{N\times 1}\right\} \\ \frac{1}{S_{p:2}(1-S_{p:2})}\left\{\boldsymbol{0}_{N\times(2p+6)}, \boldsymbol{1}_{N\times 1}\right\} \end{bmatrix}_{7N\times(2p+7)} \quad \text{and} \quad \boldsymbol{V} = \begin{bmatrix} \boldsymbol{V}_1 \\ \boldsymbol{V}_2 \\ \boldsymbol{V}_3 \\ \boldsymbol{V}_4 \\ \boldsymbol{V}_5 \\ \boldsymbol{V}_6 \\ \boldsymbol{V}_7 \end{bmatrix}_{7N\times 1}$$

with

$$\begin{aligned}
\boldsymbol{V}_1 &= \widetilde{\boldsymbol{Y}}_{(1)} \odot (1 - \widetilde{\boldsymbol{Y}}_{(2)}) \\
\boldsymbol{V}_2 &= (1 - \widetilde{\boldsymbol{Y}}_{(1)}) \odot \widetilde{\boldsymbol{Y}}_{(2)} \\
\boldsymbol{V}_3 &= \widetilde{\boldsymbol{Y}}_{(1)} \odot \widetilde{\boldsymbol{Y}}_{(2)} \\
\boldsymbol{V}_4 &= (\boldsymbol{Z}_{(1)} - S_{e:1}) \odot \widetilde{\boldsymbol{Z}}_{(1)} \oslash \boldsymbol{c} + I(\boldsymbol{Z}_{(1)} + \boldsymbol{Z}_{(2)} > 0) \odot (\boldsymbol{Y}_{(1)} - S_{e:1}) \odot \widetilde{\boldsymbol{Y}}_{(1)} \\
\boldsymbol{V}_5 &= (\boldsymbol{Z}_{(2)} - S_{e:2}) \odot \widetilde{\boldsymbol{Z}}_{(2)} \oslash \boldsymbol{c} + I(\boldsymbol{Z}_{(1)} + \boldsymbol{Z}_{(2)} > 0) \odot (\boldsymbol{Y}_{(2)} - S_{e:2}) \odot \widetilde{\boldsymbol{Y}}_{(2)} \\
\boldsymbol{V}_6 &= (1 - \boldsymbol{Z}_{(1)} - S_{p:1}) \odot (1 - \widetilde{\boldsymbol{Z}}_{(1)}) \oslash \boldsymbol{c} \\
&\quad + I(\boldsymbol{Z}_{(1)} + \boldsymbol{Z}_{(2)} > 0) \odot (1 - \boldsymbol{Y}_{(1)} - S_{p:1}) \odot (1 - \widetilde{\boldsymbol{Y}}_{(1)}) \\
\boldsymbol{V}_7 &= (1 - \boldsymbol{Z}_{(2)} - S_{p:2}) \odot (1 - \widetilde{\boldsymbol{Z}}_{(2)}) \oslash \boldsymbol{c} \\
&\quad + I(\boldsymbol{Z}_{(1)} + \boldsymbol{Z}_{(2)} > 0) \odot (1 - \boldsymbol{Y}_{(2)} - S_{p:2}) \odot (1 - \widetilde{\boldsymbol{Y}}_{(2)}).
\end{aligned}$$

Herein, for a vector $\boldsymbol{u} = (u_1, \ldots, u_m)^{\mathrm{T}}$, $I(\boldsymbol{u} > 0) = (I(u_1 > 0), \ldots, I(u_m > 0))^{\mathrm{T}}$.

Consequently, we have

$$\mathrm{cov}\left\{\frac{\partial \ell_c(\boldsymbol{\theta}|\boldsymbol{\mathcal{P}}, \widetilde{\boldsymbol{Y}}, \mathbf{X}, \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} | \boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta}\right\} = \boldsymbol{Q}^{\mathrm{T}} \mathrm{cov}(\boldsymbol{V}|\boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta}) \boldsymbol{Q}.$$

Elements of $\boldsymbol{Q}$ are easy to compute. We only show the derivation of the $\mathrm{cov}(\boldsymbol{V}|\boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta})$. By the symmetry of covariance matrix, it suffices to derive

$$\{\mathrm{cov}(\boldsymbol{V}_{l_1}, \boldsymbol{V}_{l_2}|\boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta}) : 1 \le l_1 \le l_2 \le 7\}.$$

Since all derivations are similar, we only present the calculation of $\text{cov}(\boldsymbol{V}_1, \boldsymbol{V}_1|\boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta})$, the elements of which are $\text{cov}(\widetilde{Y}_{ij1}(1-\widetilde{Y}_{ij2}), \widetilde{Y}_{i'j'1}(1-\widetilde{Y}_{i'j'2})|\boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta})$'s for all i, i', j, j'.

Clearly, if $j \neq j'$, $\text{cov}(\widetilde{Y}_{ij1}(1-\widetilde{Y}_{ij2}), \widetilde{Y}_{i'j'1}(1-\widetilde{Y}_{i'j'2})|\boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta}) = 0$ for any i and i'. This is because groups do not overlap with each other. We focus on

$$\begin{aligned}
&\text{cov}(\widetilde{Y}_{ij1}(1-\widetilde{Y}_{ij2}), \widetilde{Y}_{i'j1}(1-\widetilde{Y}_{i'j2})|\boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta}) \\
=\ &\text{pr}(\widetilde{Y}_{ij1}=1, \widetilde{Y}_{ij2}=0, \widetilde{Y}_{i'j1}=1, \widetilde{Y}_{i'j2}=0|\boldsymbol{\mathcal{P}}_j, \mathbf{X}_j, \boldsymbol{\theta}) \\
&- \text{pr}(\widetilde{Y}_{ij1}=1, \widetilde{Y}_{ij2}=0|\boldsymbol{\mathcal{P}}_j, \mathbf{X}_j, \boldsymbol{\theta})\text{pr}(\widetilde{Y}_{i'j1}=1, \widetilde{Y}_{i'j2}=0|\boldsymbol{\mathcal{P}}_j, \mathbf{X}_j, \boldsymbol{\theta})
\end{aligned}$$

for $j = 1, \ldots, J$. The two probabilities $\text{pr}(\widetilde{Y}_{ij1}=1, \widetilde{Y}_{ij2}=0|\boldsymbol{\mathcal{P}}_j, \mathbf{X}_j, \boldsymbol{\theta})$ and $\text{pr}(\widetilde{Y}_{i'j1}=1, \widetilde{Y}_{i'j2}=0|\boldsymbol{\mathcal{P}}_j, \mathbf{X}_j, \boldsymbol{\theta})$ have been calculated in Appendix B.4. Furthermore,

$$\begin{aligned}
&\text{pr}(\widetilde{Y}_{ij1}=1, \widetilde{Y}_{ij2}=0, \widetilde{Y}_{i'j1}=1, \widetilde{Y}_{i'j2}=0|\boldsymbol{\mathcal{P}}_j, \mathbf{X}_j, \boldsymbol{\theta}) \\
=\ &\frac{\text{pr}(\boldsymbol{\mathcal{P}}_j|\widetilde{Y}_{ij1}=1, \widetilde{Y}_{ij2}=0, \widetilde{Y}_{i'j1}=1, \widetilde{Y}_{i'j2}=0, \mathbf{X}_j, \boldsymbol{\theta})p_{ij10}(\boldsymbol{\theta}_1)p_{i'j10}(\boldsymbol{\theta}_1)}{\text{pr}(\boldsymbol{\mathcal{P}}_j|\mathbf{X}_j, \boldsymbol{\theta})}
\end{aligned}$$

where $\text{pr}(\boldsymbol{\mathcal{P}}_j|\mathbf{X}_j, \boldsymbol{\theta}) = \sum_{y_1, y_2 \in \{0,1\}} \text{pr}(\widetilde{Y}_{ij1}=y_1, \widetilde{Y}_{ij2}=y_2, \boldsymbol{\mathcal{P}}_j|\mathbf{X}_j, \boldsymbol{\theta})$ has also been computed in Appendix B.4. It remains to calculate $\text{pr}(\boldsymbol{\mathcal{P}}_j|\widetilde{Y}_{ij1}=1, \widetilde{Y}_{ij2}=0, \widetilde{Y}_{i'j1}=1, \widetilde{Y}_{i'j2}=0, \mathbf{X}_j, \boldsymbol{\theta})$. Again, we consider the two forms of $\boldsymbol{\mathcal{P}}_j$:

- Case 1. $\boldsymbol{\mathcal{P}}_j = \boldsymbol{Z}_j = (0,0)^{\text{T}}$.

$$\begin{aligned}
&\text{pr}(\boldsymbol{\mathcal{P}}_j|\widetilde{Y}_{ij1}=1, \widetilde{Y}_{ij2}=0, \widetilde{Y}_{i'j1}=1, \widetilde{Y}_{i'j2}=0, \mathbf{X}_j, \boldsymbol{\theta}) \\
=\ &\text{pr}(Z_{j1}=0, Z_{j2}=0|\widetilde{Y}_{ij1}=1, \widetilde{Y}_{ij2}=0, \widetilde{Y}_{i'j1}=1, \widetilde{Y}_{i'j2}=0, \mathbf{X}_j, \boldsymbol{\theta}) \\
=\ &\mathcal{M}(0,0|1,0,\boldsymbol{\theta}_2)\text{pr}\left\{\max_{l \in \mathcal{G}_j \setminus \{i,i'\}} \widetilde{Y}_{lj2} = 0 \middle| \mathbf{x}_{ij}, \boldsymbol{\theta}_1\right\} \\
&\qquad + \mathcal{M}(0,0|1,1,\boldsymbol{\theta}_2)\text{pr}\left\{\max_{l \in \mathcal{G}_j \setminus \{i,i'\}} \widetilde{Y}_{lj2} = 1 \middle| \mathbf{x}_{ij}, \boldsymbol{\theta}_1\right\} \\
=\ &\mathcal{M}(0,0|1,1,\boldsymbol{\theta}_2) \\
&\qquad + \{\mathcal{M}(0,0|1,0,\boldsymbol{\theta}_2) - \mathcal{M}(0,0|1,1,\boldsymbol{\theta}_2)\} \prod_{l \in \mathcal{G}_j \setminus \{i,i'\}} \{p_{lj10}(\boldsymbol{\theta}_1) + p_{lj00}(\boldsymbol{\theta}_1)\}.
\end{aligned}$$

- Case 2. $\boldsymbol{Z}_j \neq (0,0)^{\mathrm{T}}$ and $\boldsymbol{\mathcal{P}}_j = (Z_{j1}, Z_{j2}, \boldsymbol{Y}_{1j}, \cdots, \boldsymbol{Y}_{c_j j})$.

$$
\begin{aligned}
&\mathrm{pr}(\boldsymbol{\mathcal{P}}_j | \widetilde{Y}_{ij1} = 1, \widetilde{Y}_{ij2} = 0, \widetilde{Y}_{i'j1} = 1, \widetilde{Y}_{i'j2} = 0, \mathbf{X}_j, \boldsymbol{\theta}) \\
&= \sum_{b_2 \in \{0,1\}} \left[\mathrm{pr}\left\{ \boldsymbol{\mathcal{P}}_j \middle| \widetilde{Y}_{ij1} = 1, \widetilde{Y}_{ij2} = 0, \widetilde{Y}_{i'j1} = 1, \widetilde{Y}_{i'j2} = 0, \max_{l \in \mathcal{G}_j \setminus \{i,i'\}} \widetilde{Y}_{lj2} = b_2, \mathbf{X}_j, \boldsymbol{\theta} \right\} \right. \\
&\qquad \left. \times \mathrm{pr}\left\{ \max_{l \in \mathcal{G}_j \setminus \{i,i'\}} \widetilde{Y}_{lj2} = b_2 \middle| \mathbf{X}_j, \boldsymbol{\theta}_1 \right\} \right] \\
&= \mathrm{pr}(Y_{ij1}, Y_{ij2} | \widetilde{Y}_{ij1} = 1, \widetilde{Y}_{ij2} = 0, \boldsymbol{\theta}_2) \mathrm{pr}(Y_{i'j1}, Y_{i'j2} | \widetilde{Y}_{i'j1} = 1, \widetilde{Y}_{i'j2} = 0, \boldsymbol{\theta}_2) \\
&\quad \times \left[\mathrm{pr}(Z_{j1}, Z_{j2} | \widetilde{Z}_{j1} = 1, \widetilde{Z}_{j2} = 0, \boldsymbol{\theta}_2) \mathrm{pr}\left\{ \max_{l \in \mathcal{G}_j \setminus \{i,i'\}} \widetilde{Y}_{lj2} = 0 \middle| \mathbf{X}_j, \boldsymbol{\theta} \right\} \right. \\
&\qquad \left. + \mathrm{pr}(Z_{j1}, Z_{j2} | \widetilde{Z}_{j1} = 1, \widetilde{Z}_{j2} = 1, \boldsymbol{\theta}_2) \mathrm{pr}\left\{ \max_{l \in \mathcal{G}_j \setminus \{i,i'\}} \widetilde{Y}_{lj2} = 1 \middle| \mathbf{X}_j, \boldsymbol{\theta} \right\} \right] \\
&= \mathcal{M}(Y_{ij1}, Y_{ij2} | 1, 0, \boldsymbol{\theta}_2) \mathcal{M}(Y_{i'j1}, Y_{i'j2} | 1, 0, \boldsymbol{\theta}_2) \\
&\quad \times \left[\mathcal{M}(Z_{j1}, Z_{j2} | 1, 1, \boldsymbol{\theta}_2) + \{ \mathcal{M}(Z_{j1}, Z_{j2} | 1, 0, \boldsymbol{\theta}_2) - \mathcal{M}(Z_{j1}, Z_{j2} | 1, 1, \boldsymbol{\theta}_2) \} \right. \\
&\qquad \left. \times \prod_{l \in \mathcal{G}_j \setminus \{i,i'\}} \{ p_{lj10}(\boldsymbol{\theta}_1) + p_{lj00}(\boldsymbol{\theta}_1) \} \right].
\end{aligned}
$$

By now, we have completed the calculation of $\mathrm{cov}(\boldsymbol{V}_1, \boldsymbol{V}_1 | \boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta})$. The calculations of the remaining part of $\mathrm{cov}(\boldsymbol{V}_{l_1}, \boldsymbol{V}_{l_2} | \boldsymbol{\mathcal{P}}, \mathbf{X}, \boldsymbol{\theta})$ for $1 \leq l_1 \leq l_2 \leq 7$ follow a similar pattern and hence are omitted.

B.6 Results at different values of $S_{e:k}$'s and $S_{p:k}$'s

In Section 3.5, we present the simulation results under the case that all the assay sensitivity and specificity are 0.95. Herein, we report additional simulation results at different values of $S_{e:k}$'s and $S_{p:k}$'s. We consider following three combinations of testing errors, $R_1 : (S_{e:1}, S_{e:2}, S_{p:1}, S_{p:2})^{\mathrm{T}} = (0.90, 0.95, 0.93, 0.97)^{\mathrm{T}}$, $R_2 : (S_{e:1}, S_{e:2}, S_{p:1}, S_{p:2})^{\mathrm{T}} = (0.95, 0.95, 0.99, 0.99)^{\mathrm{T}}$ and $R_3 : (S_{e:1}, S_{e:2}, S_{p:1}, S_{p:2})^{\mathrm{T}} = (0.95, 0.95, 0.999, 0.999)^{\mathrm{T}}$. The first one is used to address the case when the sensitivity and specificity are different across infections. The later two are used to emulate the case when the true

assay specificity is close to one. We used the setting S2, $N = 3000$, $c = 2$, and the Gumbel copula with $\delta = 0.3$ to generate the SHL pooled testing data. The estimates from 500 replications are summarized in Table B.4. The results reinforce the conclusions from the original studies (the ones using 0.95 for sensitivity and specificity for both infections). One can observe that our estimate shows little bias, the estimated standard error is consistent with the sample standard deviation, and the empirical coverage is at the nominal level. One might note that for R_3, where the specificity for both infections are 0.999, the empirical coverage of one specificity doesn't reach the nominal level. However, our estimate still performs well in terms of exhibiting little bias. The variable selection results of each considered setting are also provided in Table B.3, from which we observe the same pattern as in Section 3.5 that the BIC criterion yields the smallest average prediction error and the highest SR value. In conclusion, we believe this simulation demonstrates that our methodology can handle different values of the assay sensitivity and specificity.

Table B.3: The average prediction error PE $\times$ 100 and the SR value (provided in parenthesis) of the MLE and the shrinkage estimates under the AIC, BIC, and ERIC tuning parameter selection criterion over 500 replications under $R_1 - R_3$ setting and the SHL pooling with $c = 2$.

	R_1	R_2	R_3
Estimate	PE×100(SR)	PE×100(SR)	PE×100(SR)
MLE	0.135(0.000)	0.111(0.000)	0.106(0.000)
AIC	0.099(0.434)	0.080(0.382)	0.074(0.432)
BIC	0.077(0.910)	0.060(0.882)	0.056(0.894)
ERIC	0.082(0.700)	0.065(0.726)	0.060(0.726)

Table B.4: Summary statistics of the 500 MLEs obtained under setting S2, R_1–R_3, and the SHL pooling with $c = 2$, including the sample mean (Mean), the sample standard deviation (SD), the average of the estimated standard errors (SE) and the empirical coverage (EC) of 95% confidence intervals. The prevalence (averaged over 500 repetitions) of the first and the second infections are 6.77% and 9.98%, respectively.

R_1				R_2				R_3			
	Truth	Mean(SD)	EC(SE)		Truth	Mean(SD)	EC(SE)		Truth	Mean(SD)	EC(SE)
β_{10}	-4	-4.03(0.24)	0.94(0.23)	β_{10}	-4	-4.02(0.17)	0.96(0.17)	β_{10}	-4	-4.05(0.16)	0.96(0.17)
β_{11}	-2	-2.02(0.18)	0.95(0.17)	β_{11}	-2	-2.01(0.14)	0.95(0.14)	β_{11}	-2	-2.03(0.14)	0.94(0.14)
β_{12}	2	2.03(0.18)	0.94(0.17)	β_{12}	2	2.01(0.15)	0.95(0.14)	β_{12}	2	2.02(0.14)	0.94(0.14)
β_{13}	0	-0.01(0.13)	0.94(0.13)	β_{13}	0	0.00(0.12)	0.94(0.11)	β_{13}	0	0.00(0.11)	0.94(0.11)
β_{14}	0	0.00(0.13)	0.95(0.13)	β_{14}	0	0.00(0.12)	0.94(0.11)	β_{14}	0	0.00(0.11)	0.94(0.11)
β_{15}	0	0.00(0.12)	0.96(0.13)	β_{15}	0	0.00(0.12)	0.94(0.11)	β_{15}	0	0.00(0.11)	0.95(0.11)
β_{20}	-5	-5.04(0.26)	0.96(0.26)	β_{20}	-5	-5.05(0.24)	0.96(0.24)	β_{20}	-5	-5.04(0.24)	0.97(0.24)
β_{21}	-2	-2.02(0.16)	0.96(0.16)	β_{21}	-2	-2.01(0.15)	0.97(0.15)	β_{21}	-2	-2.02(0.15)	0.96(0.15)
β_{22}	0	0.00(0.12)	0.93(0.12)	β_{22}	0	0.00(0.12)	0.94(0.12)	β_{22}	0	0.01(0.11)	0.95(0.11)
β_{23}	-2	-2.02(0.16)	0.96(0.16)	β_{23}	-2	-2.03(0.16)	0.95(0.16)	β_{23}	-2	-2.02(0.15)	0.96(0.15)
β_{24}	0	0.01(0.12)	0.95(0.12)	β_{24}	0	0.00(0.12)	0.95(0.12)	β_{24}	0	0.00(0.12)	0.94(0.11)
β_{25}	0	0.00(0.11)	0.96(0.12)	β_{25}	0	0.00(0.11)	0.96(0.12)	β_{25}	0	0.00(0.11)	0.96(0.11)
δ	0.3	0.29(0.07)	0.97(0.07)	δ	0.3	0.29(0.05)	0.96(0.06)	δ	0.3	0.29(0.05)	0.96(0.05)
$S_{e:1}$	0.90	0.90(0.03)	0.95(0.03)	$S_{e:1}$	0.95	0.95(0.02)	0.94(0.02)	$S_{e:1}$	0.95	0.95(0.01)	0.92(0.02)
$S_{e:2}$	0.95	0.95(0.01)	0.94(0.01)	$S_{e:2}$	0.95	0.95(0.01)	0.94(0.01)	$S_{e:2}$	0.95	0.95(0.01)	0.96(0.01)
$S_{p:1}$	0.93	0.93(0.01)	0.95(0.01)	$S_{p:1}$	0.99	0.95(0.00)	0.93(0.00)	$S_{p:1}$	0.999	0.998(0.001)	0.996(0.002)
$S_{p:2}$	0.97	0.97(0.00)	0.94(0.00)	$S_{p:2}$	0.99	0.95(0.00)	0.93(0.00)	$S_{p:2}$	0.999	0.999(0.001)	0.738(0.001)

B.7 Extension for individual testing

As it discussed in Section 3.5, if the N individuals are tested separately, we denote the covariates and the true infection statuses of the nth individual by $\mathbf{x}_n$ and $\widetilde{\boldsymbol{Y}}_n = (\widetilde{Y}_{n1}, \widetilde{Y}_{n2})^{\mathrm{T}}$, respectively, for $n = 1, \ldots, N$. The testing outcomes on the nth individual's specimen are denoted by $\boldsymbol{T}_n = (T_{n1}, T_{n2})^{\mathrm{T}}$; i.e., $T_{nk} = 1(0)$ means the individual tests positive (negative) for the kth infection. Because of potential identifiability issues, we solely estimate $\boldsymbol{\theta}_1$ from the individual testing data $\{(\boldsymbol{T}_n, \mathbf{x}_n) : n = 1, \ldots, N\}$ via the maximum likelihood estimation. The computation of the MLE of $\boldsymbol{\theta}_1$ is done via the following GEM algorithm.

B.7.1 The GEM algorithm

Similar to the GEM algorithm for the SHL pooling data, we treat the ture individual statuses as "missing"' data. Let $\boldsymbol{\mathcal{T}}$ be the collection of $\boldsymbol{T}_n$'s. The complete log-likelihood function of $\boldsymbol{\theta}_1$ can be written by

$$\begin{aligned}
&\ell_{c,IT}(\boldsymbol{\theta}_1|\boldsymbol{\mathcal{T}}, \widetilde{\boldsymbol{Y}}, \mathbf{X}) \\
&\quad = \sum_{n=1}^{N} \Big\{ \widetilde{Y}_{n1}\widetilde{Y}_{n2} \log p_{n11}(\boldsymbol{\theta}_1) + \widetilde{Y}_{n1}(1-\widetilde{Y}_{n2}) \log p_{n10}(\boldsymbol{\theta}_1) \\
&\qquad\qquad + (1-\widetilde{Y}_{n1})\widetilde{Y}_{n2} \log p_{n01}(\boldsymbol{\theta}_1) + (1-\widetilde{Y}_{n1})(1-\widetilde{Y}_{n2}) \log p_{n00}(\boldsymbol{\theta}_1) \Big\} \\
&\qquad + \sum_{n=1}^{N} \log \mathrm{pr}(\boldsymbol{T}_n|\widetilde{\boldsymbol{Y}}_n),
\end{aligned}$$

where $p_{n11}(\boldsymbol{\theta}_1) = \mathcal{C}\{g_1(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}_1), g_2(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}_2)|\delta\}$, $p_{n10}(\boldsymbol{\theta}_1) = g_1(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}_1) - p_{n11}(\boldsymbol{\theta}_1)$, $p_{n01}(\boldsymbol{\theta}_1) = g_2(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}_2) - p_{n11}(\boldsymbol{\theta}_1)$, $p_{n00}(\boldsymbol{\theta}_1) = 1 - p_{n11}(\boldsymbol{\theta}_1) - p_{n10}(\boldsymbol{\theta}_1) - p_{n01}(\boldsymbol{\theta}_1)$, and $\mathrm{pr}(\boldsymbol{T}_n|\widetilde{\boldsymbol{Y}}_n)$ solely depends on the testing error rates $\boldsymbol{\theta}_2$ which are assumed as known in this case. Thus,

for simplicity, we write

$$\ell_{c,IT}(\boldsymbol{\theta}_1|\boldsymbol{\mathcal{T}},\widetilde{\boldsymbol{Y}},\mathbf{X}) = \sum_{n=1}^{N}\Big\{\widetilde{Y}_{n1}\widetilde{Y}_{n2}\log p_{n11}(\boldsymbol{\theta}_1) + \widetilde{Y}_{n1}(1-\widetilde{Y}_{n2})\log p_{n10}(\boldsymbol{\theta}_1) + (1-\widetilde{Y}_{n1})\widetilde{Y}_{n2}\log p_{n01}(\boldsymbol{\theta}_1) + (1-\widetilde{Y}_{n1})(1-\widetilde{Y}_{n2})\log p_{n00}(\boldsymbol{\theta}_1)\Big\}.$$

At a current value of $\boldsymbol{\theta}_1^{(d)}$, the E-step calculates

$$\mathcal{Q}_{IT}(\boldsymbol{\theta}_1|\boldsymbol{\theta}_1^{(d)}) = E\{\ell_{c,IT}(\boldsymbol{\theta}_1|\boldsymbol{\mathcal{T}},\mathbf{X},\boldsymbol{\theta}_1^{(d)})\}.$$

It is easy to see that

$$\mathcal{Q}_{IT}(\boldsymbol{\theta}_1|\boldsymbol{\theta}_1^{(d)}) = \sum_{n=1}^{N}\sum_{y_1=0}^{1}\sum_{y_2=0}^{1}\mathrm{pr}(\widetilde{Y}_{n1}=y_1,\widetilde{Y}_{n2}=y_2|\boldsymbol{T}_n,\mathbf{x}_n,\boldsymbol{\theta}_1^{(d)})\log p_{ny_1y_2}(\boldsymbol{\theta}_1),$$

where, for $y_1, y_2 \in \{0,1\}$,

$$\begin{aligned}\mathrm{pr}(\widetilde{Y}_{n1}=y_1,&\widetilde{Y}_{n2}=y_2|\boldsymbol{T}_n,\mathbf{x}_n,\boldsymbol{\theta}_1) = \frac{\mathrm{pr}(\widetilde{Y}_{n1}=y_1,\widetilde{Y}_{n2}=y_2,\boldsymbol{T}_n)}{\sum_{y_1,y_2\in\{0,1\}}\mathrm{pr}(\widetilde{Y}_{n1}=y_1,\widetilde{Y}_{n2}=y_2,\boldsymbol{T}_n)}\\ &= \frac{\mathcal{M}(T_{n1},T_{n2}|y_1,y_2;\boldsymbol{\theta}_2)p_{ny_1y_2}(\boldsymbol{\theta}_1)}{\sum_{y_1,y_2\in\{0,1\}}\mathcal{M}(T_{n1},T_{n2}|y_1,y_2;\boldsymbol{\theta}_2)p_{ny_1y_2}(\boldsymbol{\theta}_1)}. \qquad \text{(B.31)}\end{aligned}$$

The following M-step updates $\boldsymbol{\theta}_1^{(d)}$ by $\boldsymbol{\theta}_1^{(d+1)} = (\boldsymbol{\beta}_1^{(d+1)^{\mathrm{T}}}, \boldsymbol{\beta}_2^{(d+1)^{\mathrm{T}}}, \delta^{(d+1)})^{\mathrm{T}}$ where

$$\begin{aligned}\boldsymbol{\beta}_1^{(d+1)} &= \underset{\boldsymbol{\beta}_1}{\mathrm{argmax}}\ \mathcal{Q}_{IT}(\boldsymbol{\beta}_1,\boldsymbol{\beta}_2^{(d)},\delta^{(d)}|\boldsymbol{\theta}_1^{(d)})\\ \boldsymbol{\beta}_2^{(d+1)} &= \underset{\boldsymbol{\beta}_2}{\mathrm{argmax}}\ \mathcal{Q}_{IT}(\boldsymbol{\beta}_1^{(d+1)},\boldsymbol{\beta}_2,\delta^{(d)}|\boldsymbol{\theta}_1^{(d)})\\ \delta^{(d+1)} &= \underset{\delta}{\mathrm{argmax}}\ \mathcal{Q}_{IT}(\boldsymbol{\beta}_1^{(d+1)},\boldsymbol{\beta}_2^{(d+1)},\delta|\boldsymbol{\theta}_1^{(d)}).\end{aligned}$$

Iterating between the E-step and M-step until a numerical convergence gives the MLE of $\boldsymbol{\theta}_1$.

B.7.2 Louis's method

The observed data information matrix under individual testing can also be calculated via Louis' method as

$$\mathcal{I}(\boldsymbol{\theta}_1) = -E\left\{\frac{\partial^2\ell_{c,IT}(\boldsymbol{\theta}_1|\boldsymbol{\mathcal{T}},\widetilde{\boldsymbol{Y}},\mathbf{X})}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1^{\mathrm{T}}}\middle|\boldsymbol{\mathcal{T}},\mathbf{X},\boldsymbol{\theta}_1\right\} - \mathrm{cov}\left\{\frac{\partial\ell_{c,IT}(\boldsymbol{\theta}_1|\boldsymbol{\mathcal{T}},\widetilde{\boldsymbol{Y}},\mathbf{X})}{\partial\boldsymbol{\theta}_1}\middle|\boldsymbol{\mathcal{T}},\mathbf{X},\boldsymbol{\theta}_1\right\}.$$

The calculation of the expectation term in $\mathcal{I}(\boldsymbol{\theta}_1)$ depends on the probabilities in (B.31). For the covariance term, we can have

$$\mathrm{cov}\left\{ \frac{\partial \ell_c(\boldsymbol{\theta}_1|\boldsymbol{\mathcal{T}}, \widetilde{\boldsymbol{Y}}, \mathbf{X})}{\partial \boldsymbol{\theta}_1} \middle| \boldsymbol{\mathcal{T}}, \mathbf{X}, \boldsymbol{\theta}_1 \right\} = \boldsymbol{Q}_{IT}^{\mathrm{T}} \mathrm{cov}(\boldsymbol{V}_{IT}|\boldsymbol{\mathcal{T}}, \mathbf{X}, \boldsymbol{\theta}_1) \boldsymbol{Q}_{IT},$$

where

$$\boldsymbol{Q}_{IT} = \begin{bmatrix} \left\{ \mathrm{diag}(\frac{1}{\boldsymbol{p}_{10}}) \frac{\partial \boldsymbol{p}_{10}}{\partial \boldsymbol{\theta}_1^{\mathrm{T}}} - \mathrm{diag}(\frac{1}{\boldsymbol{p}_{00}}) \frac{\partial \boldsymbol{p}_{00}}{\partial \boldsymbol{\theta}_1^{\mathrm{T}}} \right\}_{N \times (2p+3)} \\ \left\{ \mathrm{diag}(\frac{1}{\boldsymbol{p}_{01}}) \frac{\partial \boldsymbol{p}_{01}}{\partial \boldsymbol{\theta}_1^{\mathrm{T}}} - \mathrm{diag}(\frac{1}{\boldsymbol{p}_{00}}) \frac{\partial \boldsymbol{p}_{00}}{\partial \boldsymbol{\theta}_1^{\mathrm{T}}} \right\}_{N \times (2p+3)} \\ \left\{ \mathrm{diag}(\frac{1}{\boldsymbol{p}_{11}}) \frac{\partial \boldsymbol{p}_{11}}{\partial \boldsymbol{\theta}_1^{\mathrm{T}}} - \mathrm{diag}(\frac{1}{\boldsymbol{p}_{00}}) \frac{\partial \boldsymbol{p}_{00}}{\partial \boldsymbol{\theta}_1^{\mathrm{T}}} \right\}_{N \times (2p+3)} \end{bmatrix}_{3N \times (2p+3)},$$

and

$$\boldsymbol{V}_{IT} = \begin{bmatrix} \boldsymbol{V}_{1IT} \\ \boldsymbol{V}_{2IT} \\ \boldsymbol{V}_{3IT} \end{bmatrix} = \begin{bmatrix} \widetilde{\boldsymbol{Y}}_{(1)} \odot (1 - \widetilde{\boldsymbol{Y}}_{(2)}) \\ (1 - \widetilde{\boldsymbol{Y}}_{(1)}) \odot \widetilde{\boldsymbol{Y}}_{(2)} \\ \widetilde{\boldsymbol{Y}}_{(1)} \odot \widetilde{\boldsymbol{Y}}_{(2)} \end{bmatrix}_{3N \times 1}.$$

Again, the $\mathrm{cov}(\boldsymbol{V}_{IT}|\boldsymbol{\mathcal{T}}, \mathbf{X}, \boldsymbol{\theta}_1)$ follows a similar but simpler calculation than the one described in Appendix B.5.2.

B.7.3 Variable selection

Following Section 3.4, we let $\boldsymbol{\theta}_{\mathcal{A}} = (\beta_{11}, \cdots, \beta_{1p}, \beta_{21}, \cdots, \beta_{2p})^{\mathrm{T}}$ ($\hat{\boldsymbol{\theta}}_{\mathcal{A},IT}$) denote the sub-vector of $\boldsymbol{\theta}_1$ ($\hat{\boldsymbol{\theta}}_{1IT}$) associated with set $\mathcal{A}$ and $\hat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A},IT}$ denote the sub-matrix of $\mathcal{I}(\hat{\boldsymbol{\theta}}_1)$ indexed by $\mathcal{A} \times \mathcal{A}$, where $\mathcal{A} = \{2, \cdots, p+1, p+3, \cdots, 2p+2\}$ indexes all the slope coefficients. The shrinkage estimator of $\boldsymbol{\theta}_{\mathcal{A}}$ under individual testing is defined by

$$\widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda,IT} = \underset{\boldsymbol{\theta}_{\mathcal{A}}}{\mathrm{argmin}} \left\{ \frac{1}{2} (\hat{\boldsymbol{\theta}}_{\mathcal{A},IT} - \boldsymbol{\theta}_{\mathcal{A}})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A},IT} (\hat{\boldsymbol{\theta}}_{\mathcal{A},IT} - \boldsymbol{\theta}_{\mathcal{A}}) + \sum_{k=1}^{2} \lambda_k \sum_{i=1}^{p} \omega_{kj} |\beta_{kj}| \right\},$$

where $\omega_{kj} = |\hat{\beta}_{kj,IT}|^{-1}$ and $\lambda_1, \lambda_2 \geq 0$ are the tuning parameters which could be selected by minimizing the following three types of criteria:

- $\mathrm{BIC}(\lambda_1, \lambda_2) = (\hat{\boldsymbol{\theta}}_{\mathcal{A},IT} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda,IT})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A},IT} (\hat{\boldsymbol{\theta}}_{\mathcal{A},IT} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda,IT}) + \{df_{1,\lambda,IT} + df_{2,\lambda,IT}\} \log N$
- $\mathrm{AIC}(\lambda_1, \lambda_2) = (\hat{\boldsymbol{\theta}}_{\mathcal{A},IT} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda,IT})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A},IT} (\hat{\boldsymbol{\theta}}_{\mathcal{A},IT} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda,IT}) + 2\{df_{1,\lambda,IT} + df_{2,\lambda,IT}\}$

- $\text{ERIC}(\lambda_1, \lambda_2)$
  $= (\hat{\boldsymbol{\theta}}_{\mathcal{A},IT} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda,IT})^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\mathcal{A}\mathcal{A},IT} (\hat{\boldsymbol{\theta}}_{\mathcal{A},IT} - \widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda,IT}) + \sum_{k=1}^{2} df_{k,\lambda,IT} \log(N/\lambda_k)$

where $df_{k,\lambda,IT}$ is the number of slope coefficients selected by $\widetilde{\boldsymbol{\theta}}_{\mathcal{A},\lambda,IT}$ for the kth infection.

B.8 An additional analysis of the NHPL data

In the NHPL data analysis, we assigned individuals to groups of size 4 according to the analysis date. In this appendix section, we randomly assign individuals into groups. We let the group sizes vary from 2 to 10. At each group size, we repeat the process of generating the SHL testing responses and applying our method 500 times.

Box-plots of the 500 estimates at $c = 1, 2, \ldots, 10$ of regression coefficients for CT and NG, the copula parameter δ, and misclassification parameters $S_{e:k}$'s and $S_{p:k}$'s, are provided in Figures B.1–B.3. The case $c = 1$ corresponds to individual testing. The red solid horizontal line in each figure is the corresponding reference estimate obtained by using the "true" statuses (see Section 3.6).

As one can see that, when c varies from 2 to 10, the variation in the estimates at $c = 2$ is the smallest, indicating the optimal group size for estimation would be 2. Of course, in practice one has to consider other aspects, for example, the testing cost. Table B.5 includes the average number of tests for each c. We see that $c = 4$ yields the largest amount of cost savings. But no matter which group size $c \in \{2, \ldots, 10\}$ is used, estimates are always better than the one using individual testing in terms of that estimates at $c > 1$ are more aligned with the reference line and have smaller variation than the ones at $c = 1$.

Table B.5: The average number of tests for each c.

Individual testing	$c = 2$	$c = 3$	$c = 4$	$c = 5$	$c = 6$	$c = 7$	$c = 8$	$c = 9$	$c = 10$
14530	9750	8198	7791	7798	8007	8293	8616	8955	9296

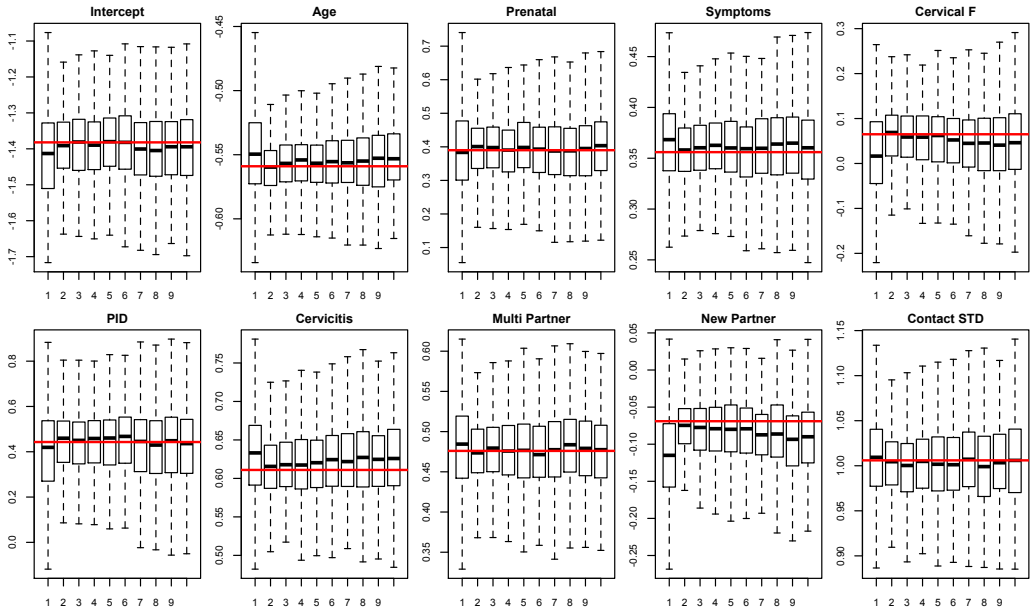

Figure B.1: Box-plots of the 500 estimates of regression coefficients for CT across $c \in \{1, \ldots, 10\}$. The solid lines in the figures denote the reference estimates.

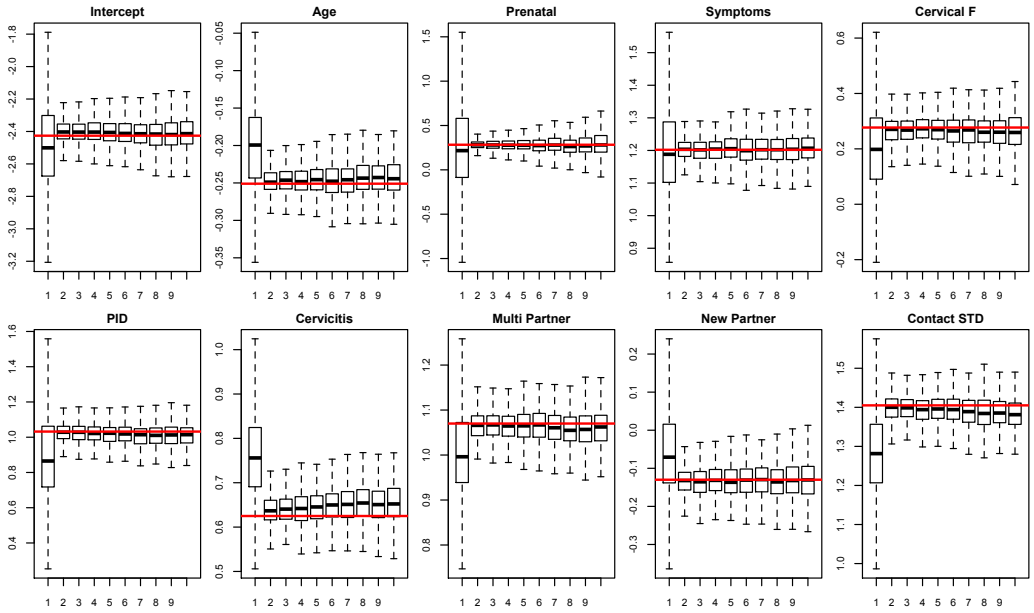

Figure B.2: Box-plots of the 500 estimates of regression coefficients for NG across $c \in \{1, \ldots, 10\}$. The solid lines in the figures denote the reference estimates.

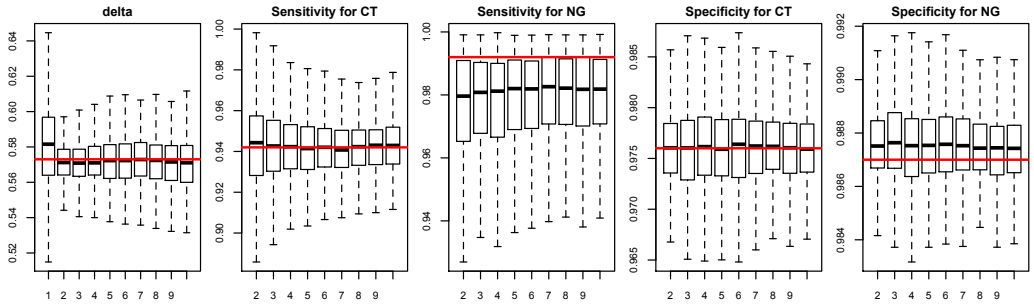

Figure B.3: Box-plots of the 500 estimates of copula parameter δ across $c \in \{1, \ldots, 10\}$ and misclassification parameters $S_{e:k}$'s and $S_{p:k}$'s across $c \in \{2, \ldots 10\}$. The solid lines in the figures denote the reference estimates.

B.9 Robustness of using Gumbel copulas

Recall that in Section 3.5, we used a Gumbel copula with copula parameter $\delta = 0.3$. To reveal the robustness property, we simulate data from either a Clayton or a Gaussian copula and estimate the parameters by using a Gumbel copula. More specifically, we used the Clayton copula $\mathcal{C}_c(u, v|\delta_c) = (u^{-\delta_c} + v^{-\delta_c} - 1)^{-1/\delta_c}$ with $\delta_c = 4.667$ and Gaussian copula with a correlation coefficient $\rho = 0.891$. Again, we consider the setting S2, $N = 3000$, the group size $c \in \{2, 5, 10\}$, and $S_{e:k} = S_{p:k} = 0.95$ for $k = 1, 2$.

The results by applying our proposed GEM algorithm with the use of a Gumbel copula on data simulated from the Clayton and Gaussian copula are presented in Table B.6 and Table B.7, respectively. Generally speaking, our estimation method derived under a Gumbel copula could perform well regardless of the true copula. In spite of the copula parameter, our estimators are almost on the target with a small bias. The estimation of the copula parameter is largely biased because the copula is misspecified. To examine the robustness of the variable selection, Table B.8 provides the average prediction error (PE) and selection rate (SR) of shrinkage estimators under each simulation setting. The definition of PE and SR are the same as the ones used in Section 3.5. By comparing to the (S2) results from Table 3.4, the misspecification of copula indeed increases the model prediction error. Notwithstanding, the

adaptive LASSO shrinkage estimators still yield a similar level of variable selection rate. The highest selection rate is again provided by BIC-type estimators and is still as high as 90% under each considered setting.

Table B.6: Summary statistics of the 500 MLEs obtained under setting S2, Clayton copula of $\delta_c = 4.667$ and the SHL pooling with $c = 2, 5, 10$, including the sample mean (Mean), the sample standard deviation (SD), the average of the estimated standard errors (SE) and the empirical coverage (EC) of 95% confidence intervals. The average number of tests (# of tests) under each protocol is also provided. The prevalence (averaged over 500 repetitions) of the first and the second infections are 6.77% and 9.97%, respectively.

		$c = 2$		$c = 5$		$c = 10$	
# tests		2498		2319		2683	
	Truth	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)
β_{10}	-4	-4.00(0.19)	0.95(0.19)	-3.99(0.20)	0.94(0.19)	-3.98(0.22)	0.92(0.20)
β_{11}	-2	-1.99(0.15)	0.94(0.15)	-1.98(0.16)	0.93(0.16)	-1.98(0.17)	0.93(0.17)
β_{12}	2	2.00(0.16)	0.92(0.15)	1.99(0.16)	0.93(0.16)	1.99(0.18)	0.92(0.17)
β_{13}	0	-0.01(0.11)	0.96(0.12)	-0.01(0.11)	0.96(0.12)	-0.01(0.12)	0.96(0.12)
β_{14}	0	0.00(0.11)	0.96(0.11)	0.00(0.12)	0.94(0.12)	0.00(0.12)	0.96(0.12)
β_{15}	0	0.00(0.12)	0.95(0.12)	0.00(0.12)	0.95(0.12)	0.00(0.12)	0.96(0.12)
β_{20}	-5	-5.06(0.25)	0.96(0.26)	-5.02(0.25)	0.96(0.27)	-5.00(0.28)	0.96(0.31)
β_{21}	-2	-2.02(0.15)	0.96(0.16)	-2.01(0.16)	0.96(0.17)	-2.01(0.17)	0.95(0.18)
β_{22}	0	0.00(0.12)	0.95(0.12)	0.00(0.12)	0.95(0.12)	0.00(0.13)	0.95(0.13)
β_{23}	-2	-2.02(0.16)	0.96(0.16)	-2.01(0.16)	0.96(0.17)	-2.01(0.17)	0.95(0.18)
β_{24}	0	0.00(0.12)	0.94(0.12)	0.00(0.12)	0.96(0.12)	0.01(0.12)	0.95(0.13)
β_{25}	0	-0.01(0.11)	0.97(0.12)	-0.01(0.11)	0.96(0.12)	-0.01(0.12)	0.96(0.13)
δ	0.3	0.19(0.03)	0.79(0.08)	0.19(0.03)	0.90(0.08)	0.20(0.04)	0.97(0.10)
$S_{e:1}$	0.95	0.95(0.02)	0.95(0.02)	0.95(0.02)	0.94(0.02)	0.95(0.02)	0.95(0.02)
$S_{e:2}$	0.95	0.95(0.01)	0.95(0.01)	0.95(0.01)	0.95(0.01)	0.95(0.02)	0.95(0.01)
$S_{p:1}$	0.95	0.95(0.01)	0.96(0.01)	0.95(0.01)	0.95(0.01)	0.95(0.01)	0.90(0.01)
$S_{p:2}$	0.95	0.95(0.01)	0.94(0.01)	0.95(0.01)	0.93(0.01)	0.95(0.01)	0.92(0.01)

Table B.7: Summary statistics of the 500 MLEs obtained under S2, Gaussian copula of $\rho = 0.891$ and the SHL pooling with $c = 2, 5, 10$, including the sample mean (Mean), the sample standard deviation (SD), the average of the estimated standard errors (SE) and the empirical coverage (EC) of 95% confidence intervals. The average number of tests (# of tests) under each protocol is also provided. The prevalence (averaged over 500 repetitions) of the first and the second infections are 6.77% and 9.97%, respectively.

		$c = 2$		$c = 5$		$c = 10$	
# tests		2485		2294		2660	
	Truth	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)	Mean(SD)	EC(SE)
β_{10}	-4	-4.03(0.19)	0.95(0.19)	-4.02(0.20)	0.95(0.20)	-4.02(0.21)	0.95(0.21)
β_{11}	-2	-2.01(0.15)	0.95(0.15)	-2.00(0.16)	0.95(0.16)	-2.01(0.17)	0.95(0.17)
β_{12}	2	2.02(0.16)	0.95(0.15)	2.01(0.17)	0.95(0.16)	2.01(0.17)	0.94(0.17)
β_{13}	0	0.00(0.11)	0.97(0.12)	-0.01(0.12)	0.95(0.12)	0.00(0.12)	0.95(0.12)
β_{14}	0	0.00(0.12)	0.94(0.12)	0.00(0.13)	0.93(0.12)	0.01(0.13)	0.95(0.13)
β_{15}	0	0.00(0.11)	0.96(0.12)	0.00(0.12)	0.96(0.12)	-0.01(0.12)	0.96(0.12)
β_{20}	-5	-5.01(0.27)	0.95(0.26)	-5.01(0.29)	0.94(0.28)	-5.01(0.32)	0.94(0.31)
β_{21}	-2	-2.01(0.17)	0.94(0.16)	-2.01(0.18)	0.94(0.17)	-2.01(0.19)	0.94(0.18)
β_{22}	0	0.01(0.12)	0.96(0.12)	0.01(0.12)	0.96(0.12)	0.01(0.13)	0.96(0.13)
β_{23}	-2	-2.01(0.16)	0.96(0.16)	-2.01(0.18)	0.94(0.17)	-2.00(0.19)	0.94(0.18)
β_{24}	0	0.01(0.12)	0.95(0.12)	0.01(0.12)	0.95(0.12)	0.00(0.13)	0.95(0.13)
β_{25}	0	-0.01(0.12)	0.96(0.12)	-0.01(0.12)	0.96(0.12)	-0.01(0.13)	0.96(0.13)
δ	0.3	0.23(0.05)	0.88(0.07)	0.24(0.05)	0.94(0.07)	0.24(0.06)	0.98(0.09)
$S_{e:1}$	0.95	0.95(0.02)	0.94(0.02)	0.95(0.02)	0.94(0.02)	0.95(0.02)	0.93(0.02)
$S_{e:2}$	0.95	0.95(0.02)	0.93(0.01)	0.95(0.02)	0.93(0.01)	0.95(0.02)	0.90(0.01)
$S_{p:1}$	0.95	0.95(0.01)	0.94(0.01)	0.95(0.01)	0.95(0.01)	0.95(0.01)	0.90(0.00)
$S_{p:2}$	0.95	0.95(0.01)	0.96(0.01)	0.95(0.01)	0.92(0.01)	0.95(0.01)	0.92(0.00)

Table B.8: The average prediction error PE $\times$ 100 and the SR value (provided in parenthesis) of the MLE and the shrinkage estimates under the AIC, BIC and ERIC tuning parameter selection criterion over 500 replications under Clayton and Gaussian copula setting across the SHL pooling with $c = 2, 5$ and 10.

		$c = 2$	$c = 5$	$c = 10$
Copula	Estimate	PE×100(SR)	PE×100(SR)	PE×100(SR)
Clayton	MLE	1.829(0.000)	1.839(0.000)	1.844(0.000)
	AIC	1.796(0.432)	1.807(0.432)	1.807(0.420)
	BIC	1.782(0.902)	1.791(0.888)	1.787(0.906)
	ERIC	1.789(0.746)	1.795(0.730)	1.788(0.736)
Gaussian	MLE	0.328(0.000)	0.339(0.000)	0.343(0.000)
	AIC	0.296(0.458)	0.306(0.390)	0.306(0.424)
	BIC	0.281(0.906)	0.287(0.902)	0.286(0.900)
	ERIC	0.291(0.740)	0.297(0.726)	0.294(0.726)

B.10 A simulation study for three infections

As it described in Section 3.7, our proposed methodology can be generalized to incorporate more than two infections. A simulation study is presented here to illustrate this generalizability. In this simulation, we specify $\mathcal{C}(u_1, u_2, u_3|\delta) = \exp\{-[(-\log u_1)^{1/\delta} + (-\log u_2)^{1/\delta} + (-\log u_3)^{1/\delta}]^{\delta}\}$ as a three-dimensional Gumbel copula with $\delta = 0.3$ to model the co-infection probability. Then, we consider the following simulation setting. Note that regression coefficients $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ and $\boldsymbol{\beta}_3$ are chosen to make the prevalence of each infection around 8%.

- $\boldsymbol{\beta}_1 = (-4, -2, 2, 0, 0, 0)^{\mathrm{T}}$, $\boldsymbol{\beta}_2 = (-5, -2, 0, -2, 0, 0)^{\mathrm{T}}$, and $\boldsymbol{\beta}_3$=$(-5, -3, 2, 0, 0, 0)^{\mathrm{T}}$
- $\mathbf{x} = (1, x_1, \cdots, x_5)^{\mathrm{T}}$, where $\mathbf{x}$ is simulated from $\mathcal{N}(\mathbf{0}, \boldsymbol{\Omega})$ with $[\boldsymbol{\Omega}]_{st} = 1$ if $s = t$ and $[\boldsymbol{\Omega}]_{st} = 0.5$ if $s \neq t$

In this simulation, we only consider the SHL protocol. The SHL pooled testing data is generated as follows. We consider a total of $N = 3000$ individuals need to be tested for three infections. To form the pool, we use a common group size denoted by c and $c \in \{2, 5, 10\}$. The total 3000 individuals are randomly assigned to J non-overlapping pools, where $J = N/c$ as the number of pools. Denote $\{ij\}$ the ith individual in the jth pool. We generate the covariates $\mathbf{x}_{ij}$, and calculate individual-level cell probabilities using the marginal logistic regression model and the three-dimensional Gumbel copula. Then, individual true infectious statuses $\widetilde{\boldsymbol{Y}}_{ij} = (\widetilde{Y}_{ij1}, \widetilde{Y}_{ij2}, \widetilde{Y}_{ij3})^{\mathrm{T}}$ are generated from a multinomial distribution with its associated cell probabilities. The true status of the jth pooled specimen for the kth infection can be calculated as $\widetilde{Z}_{jk} = \max_i \widetilde{Y}_{ijk}$.

To mimic the SHL protocol, we assume the assay testing sensitivity and specificity are 0.95 for all infections at both testing stages of the SHL protocol; i.e. $S_{e:k} = S_{p:k} = 0.95$, for $k = 1, 2, 3$. For the kth infection in the jth pool, we generate the pooled testing outcome Z_{jk} from a Bernoulli distribution with the suc-

cess probability of $S_{e:k}\widetilde{Z}_{jk} + (1 - S_{p:k})(1 - \widetilde{Z}_{jk})$. According to the SHL protocol, if $\max_k Z_{jk} = 1$, it proceeds to the second stage. In this situation, we generate the retesting outcome of the ith individual for kth infection Y_{ijk} from a Bernoulli distribution with the success probability of $S_{e:k}\widetilde{Y}_{ijk} + (1 - S_{p:k})(1 - \widetilde{Y}_{ijk})$. Finally, the SHL pooled testing data is a combination of all available $\boldsymbol{Z}_j = (Z_{j1}, Z_{j2}, Z_{j3})^{\mathrm{T}}$ and $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2}, Y_{ij3})^{\mathrm{T}}$. With the simulated data, we estimate the unknown parameters $\{\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \delta, S_{e:1}, S_{e:2}, S_{e:3}, S_{p:1}, S_{p:2}.S_{p:3}\}$ through a generalized three-infection GEM algorithm. The above process is repeated 500 times under all considered simulation settings.

The results are reported in Table B.9. One can see that the estimates exhibit little bias and the sample standard derivations are in a reasonable range. Though we did not use the Louis' method to estimate the standard errors nor perform marginal variable selection, we believe that these results are sufficient to demonstrate the generalizability of our GEM algorithm to calculate the MLEs of the regression coefficients, misclassification parameters and the copula parameter for the case of three infections.

Table B.9: Summary statistics of the 500 MLEs obtained under our simulation setting and the SHL pooling with $c = 2, 5, 10$, including the sample mean (Mean) and the sample standard deviation (SD). The average number of tests (# of tests) under each setting is also provided. The average prevalence of the first, the second and the third infections are 6.77%, 9.98%, and 5.89%, respectively.

		$c = 2$	$c = 5$	$c = 10$
# tests		2630	2441	2773
	Truth	Mean(SD)	Mean(SD)	Mean(SD)
β_{10}	-4	-4.04(0.21)	-4.04(0.21)	-4.05(0.23)
β_{11}	-2	-2.02(0.16)	-2.02(0.16)	-2.02(0.17)
β_{12}	2	2.02(0.16)	2.02(0.17)	2.02(0.17)
β_{13}	0	0.01(0.12)	0.01(0.12)	0.00(0.12)
β_{14}	0	0.00(0.11)	0.00(0.12)	0.00(0.13)
β_{15}	0	0.01(0.11)	0.01(0.12)	0.01(0.12)
β_{20}	-5	-5.08(0.28)	-5.09(0.29)	-5.10(0.32)
β_{21}	-2	-2.03(0.17)	-2.03(0.17)	-2.04(0.19)
β_{22}	0	-0.01(0.12)	0.00(0.12)	0.00(0.13)
β_{23}	-2	-2.04(0.16)	-2.05(0.17)	-2.05(0.19)
β_{24}	0	0.00(0.12)	-0.01(0.13)	0.00(0.13)
β_{25}	0	0.01(0.12)	0.01(0.13)	0.00(0.13)
β_{30}	-5	-5.08(0.27)	-5.09(0.28)	-5.07(0.30)
β_{31}	-3	-3.04(0.22)	-3.05(0.22)	-3.04(0.24)
β_{32}	2	2.03(0.18)	2.03(0.18)	2.02(0.17)
β_{33}	0	0.00(0.13)	0.01(0.14)	0.01(0.14)
β_{34}	0	0.01(0.12)	0.00(0.13)	0.01(0.13)
β_{35}	0	0.00(0.13)	0.00(0.13)	0.00(0.14)
δ	0.3	0.30(0.03)	0.30(0.03)	0.30(0.04)
$S_{e:1}$	0.95	0.95(0.02)	0.95(0.02)	0.95(0.02)
$S_{e:2}$	0.95	0.95(0.01)	0.95(0.02)	0.95(0.01)
$S_{e:3}$	0.95	0.95(0.02)	0.95(0.02)	0.95(0.02)
$S_{p:1}$	0.95	0.95(0.01)	0.95(0.01)	0.95(0.01)
$S_{p:2}$	0.95	0.95(0.01)	0.95(0.01)	0.95(0.01)
$S_{p:3}$	0.95	0.95(0.01)	0.95(0.01)	0.95(0.01)

Appendix C

Chapter 4 Supplementary Materials

C.1 Estimation for individual testing

In this appendix, we will demonstrate the estimation step when the N individuals are tested separately. We denote the covariates and the true infection statuses of the nth individual by $\mathbf{x}_n$ and $\widetilde{Y}_n$, respectively, for $n = 1, \ldots, N$. The testing outcome on the nth individual's specimen is denoted by T_n; i.e., $T_n = 1(0)$ means the individual tests positive (negative). The computation of the MLE of $\boldsymbol{\beta}$ is done via the following EM algorithm by treating the ture individual statuses as "missing" data.

Because misclassification parameters are assumed as known, the complete likelihood function can be written as

$$l_{c,IT}(\boldsymbol{\beta} \mid \widetilde{\boldsymbol{Y}}, \mathbf{x}) = \sum_{n=1}^{N} \left[(1 - \widetilde{Y}_n) \log\{1 - g(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta})\} + \widetilde{Y}_n \log g(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}) \right].$$

At a current value of $\boldsymbol{\beta}^{(d)}$, the E-step calculates $\mathcal{Q}_{IT}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(d)}) = E\{l_{c,IT}(\boldsymbol{\beta} \mid \widetilde{\boldsymbol{Y}}, \mathbf{x}, \boldsymbol{\beta}^{(d)})\}$, where

$$\mathcal{Q}_{IT}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(d)}) = \sum_{n=1}^{N} \Bigg[\mathrm{pr}(\widetilde{Y}_n = 0 \mid T_n, \mathbf{x}_n, \boldsymbol{\beta}^{(d)}) \log\{1 - g(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta})\} + \mathrm{pr}(\widetilde{Y}_n = 1 \mid T_n, \mathbf{x}_n, \boldsymbol{\beta}^{(d)}) \log g(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}) \Bigg].$$

It is easy to observe that

$$\mathrm{pr}(\widetilde{Y}_n = y \mid T_n, \mathbf{x}_n, \boldsymbol{\beta}^{(d)}) = \frac{\mathrm{pr}(\widetilde{Y}_n = y, T_n \mid \mathbf{x}_n, \boldsymbol{\beta}^{(d)})}{\sum_{y' \in \{0,1\}} \mathrm{pr}(\widetilde{Y}_n = y', T_n \mid \mathbf{x}_n, \boldsymbol{\beta}^{(d)})},$$

for $y = 0, 1$, and

$$\text{pr}(\widetilde{Y}_n = 0, T_n \mid \mathbf{x}_n, \boldsymbol{\beta}^{(d)}) = S_p^{1-T_n}(1 - S_p)^{T_n} \log\{1 - g(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}^{(d)})\}$$
$$\text{pr}(\widetilde{Y}_n = 1, T_n \mid \mathbf{x}_n, \boldsymbol{\beta}^{(d)}) = (1 - S_e)^{1-T_n} S_e^{T_n} \log g(\mathbf{x}_n^{\mathrm{T}}\boldsymbol{\beta}^{(d)}).$$

The M-step updates $\boldsymbol{\beta}^{(d)}$ via $\boldsymbol{\beta}^{(d+1)} = \underset{\boldsymbol{\beta}}{\text{argmax}} \mathcal{Q}_{IT}(\boldsymbol{\beta} \mid \boldsymbol{\beta}^{(d)})$. Iterating between the E-step and M-step until a numerical converge provides the MLE of $\boldsymbol{\beta}$.