5-2014

Nonparametric and semiparametric group testing regression models

Dewei Wang
Clemson University, wangdw1988@gmail.com

NONPARAMETRIC AND SEMIPARAMETRIC GROUP TESTING REGRESSION MODELS

A Dissertation
Presented to
the Graduate School of
Clemson University

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy
Mathematical Sciences

---

by
Dewei Wang
May 2014

Accepted by:
Dr. Colin M. Gallagher, Committee Co-Chair
Dr. Christopher S. McMahan, Committee Co-Chair
Dr. Karunarathna B. Kulasekera
Dr. Chanseok Park
Dr. Xiaoqian Sun

Abstract

This dissertation consists of three projects in the area of group testing. The method of group testing, through the use of pooling, has proven to be an efficient method of reducing the time and cost associated with screening for a binary characteristic of interest, such as infection status. The salient feature of group testing that provides for these gains in efficiency is that testing is performed on pooled specimens, rather than testing specimens one-by-one. In Chapter 1, we present a general introduction of group testing. Typically, the statistical literature surrounding group testing has investigated the implementation of pooled testing for the purposes of either case identification or estimation. In this dissertation, we mainly focuses on the estimation problem which involves the development of regression models that relate individual level covariates to testing responses observed from pooled specimens.

Primarily, the existing research in the area of estimation in group testing has focused on parametric regression models, where the shape of the link function is assumed as known and only a finite number of regression parameters has to be estimated. Recently, for the purpose of obviating the specification of the link function and increasing the flexibility of modeling, nonparametric group testing regression models have been studied. In Chapter 2, we propose a new nonparametric estimation procedure using a local likelihood approach. For easy illustration, in this part we consider the situation where each individual is assigned to exactly one pool and only this pooled specimen is tested. Further, we assume the assay used for screening is perfect. Both of these two assumptions will be relaxed in the rest chapters of this dissertation. We show that our proposed estimator enjoys an asymptotic normal distribution with the optimal nonparametric estimation rate. Finite sample performance of the method is exhibited via some simulated examples and a real data analysis.

To pursue a more suitable technique of modeling group testing data, in Chapter 3, we develop a general semiparametric framework which allows for the inclusion of only not one continuous covariate, but also multiple explanatory variables, all variants of decoding information, and imperfect testing. The asymp-

totic properties of our estimators are presented and guidance on finite sample implementation is provided. We illustrate the performance of our methods through simulation and by applying them to chlamydia and gonorrhea data collected by the Nebraska Public Health Laboratory as a part of the Infertility Prevention Project.

In Chapter 4, we focus on the evaluation of misclassification effect of testing pools which are constructed according to any types of group testing algorithms. The existing assumption regarding them are somehow restrictive. If they are invalid, the estimation procedure can lead to severely biased estimator. In this work, we relax previously made assumptions regarding testing error rates by acknowledging the underlying mechanistic structure of the diagnostic test being employed. For easy illustration of this methodology, we mainly concentrate in parametric regression methods and propose a general estimation framework that allows for the analysis of data arising from all group testing strategies. The finite sample performance of our proposed methodology are investigated through simulation and by applying our techniques to hepatitis B data from a study involving Irish prisoners. Through these studies, we show that our methods can result in more efficient parameter estimates, when compared to competing procedures that make use of individual level data, at a fraction of the cost of data collection.

Before proceeding to the main body of this dissertation, I would like to clarify that the notations defined in this work are self-contained in each separated chapter.

Dedication

This dissertation is dedicated to the love of my life

Chendi Jiang

Acknowledgments

No words could describe my gratitude to my advisors, Dr. Karunarathna B. Kulasekera, Dr. Colin M. Gallagher, and Dr. Christopher S. McMahan.

Table of Contents

Page

List of Tables

List of Figures

Chapter 1

Introduction

The origin of group testing is typically attributed to Dorfman (1943), which proposed the use of pooling as a means to reduce the time and cost associated with screening military inductees for syphilis during World War II. In general, group testing begins by collecting specimens (e.g., blood, urine, plasma, etc.) from individuals which are then physically combined to form a pooled specimen. The pooled specimen is then tested for the infection of interest and the observed response provides pertinent information pertaining to both estimation and classification; i.e., it provides evidence of whether or not the pool contains a positive member(s). Since its advent, group testing has been implemented for the purposes of screening for infectious diseases (Cardoso et al., 1998; Busch et al., 2005; Picher et al., 2005; Jirsa, 2008; Lewis et al., 2012; Van et al., 2012), discovering lead compounds in drug discovery (Remlinger et al., 2006), identifying rare mutations in genetics (Gastwirth, 2000), and detecting viral agents in the case of bioterrorism (Schmidt et al., 2005). These techniques have been used to screen millions of blood donations, both in the United States (US) and abroad, for the human immunodeficiency virus, hepatitis B virus, and hepatitis C virus (Hourfar et al., 2008; Stramer et al., 2013). Further, group testing is also routinely used to screen for a cadre of other infectious diseases; e.g., Lindan et al. (2005) notes that 12 percent of the medical screening labs in the US use pool testing for chlamydia screening.

In many infectious disease screening applications, it is of primary interest to diagnose each individual as either being positive or negative for the infection of interest. To facilitate this goal, the classification protocol presented in Dorfman (1943) suggested testing the initial master pooled specimens first. If a master pool tested negative then each contributing individual should be diagnosed as negative. On the other hand, if a master pool tested positive then it should be "decoded" by retesting each contributing specimen separately.

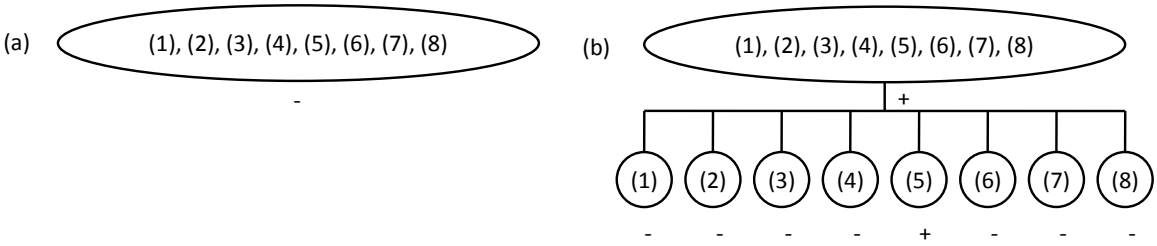


Figure 1.1: Possible outcomes of testing eight individuals via Dorfman decoding procedure. In (a), the master pool tests negative, and hence all of them are diagnosed as negative. In (b), the master pool tests positive. Then each individual is retested separately. In this case, only the fifth individual is diagnosed as positive.

This testing protocol is commonly referred to as Dorfman decoding (also see Figure 1.1). Due to its simplicity, Dorfman decoding has been widely implemented in practice. Since this seminal work, many variants of Dorfman's decoding algorithm have been proposed in an effort to reduce testing cost and/or increase classification accuracy. For example, Litvak et al. (1994) studied the halving algorithm (see Figure 1.2), which also starts with testing the initial master pool. If it tests negative, like Dorfman decoding, all the individuals are diagnosed as negative. However, whenever a pooled specimen tests positive, instead of retesting individuals separately, it proceeds to randomly assign their specimens into two smaller pools of equal size and then test these two new pooled specimens until every individual is diagnosed as either positive or negative. For other

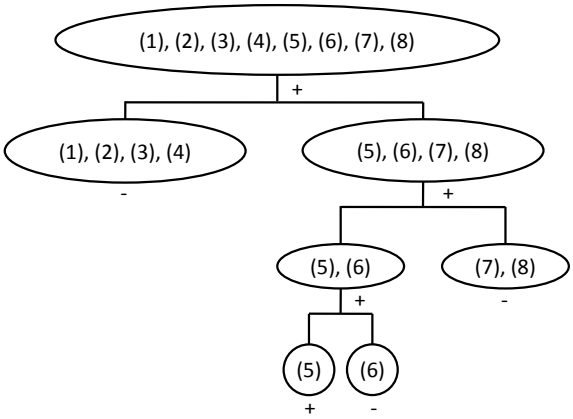


Figure 1.2: A possible outcome of testing eight individuals via halving. It starts with testing the master pool. During the process, each pool that tests positive is then divided into two equally-sized pools which are tested one-by-one until all the individuals are diagnosed as either positive or negative. In this case, only the fifth individual is diagnosed as positive.

testing algorithms, such as array testing (Phatarfod & Sudbury, 1994), see Kim et al. (2007) for a thorough review. In addition, group testing strategies have also been developed for the purposes of preserving anonymity in estimation studies (Hammick & Gastwirth, 1994) and for quality control purposes (Gastwirth & Johnson,

1994; Johnson & Pearson, 1999). In virtually all of the aforementioned situations, the associated group testing algorithm may require that a number of the individuals' specimens be assigned to multiple pools and/or be tested individually.

The other goal of group testing is for estimation. The use of group testing techniques, as a cost effective data collection mechanism, for conducting inference was first proposed by Thompson (1962), and has since received a great deal of attention in the statistical literature. Many of the earlier works in this area use group testing data to estimate population level characteristics, such as the proportion of infected individuals; for a review see Bilder & Tebbs (2005) and the references therein. More recently, authors have developed binary regression models which relate pool response data to individual-level covariate information through a specified link function; see Vansteelandt et al. (2000), Bilder & Tebbs (2009), Chen et al. (2009), and Huang & Tebbs (2009). To obviate the specification of the link function, Delaigle & Meister (2011) proposed the first nonparametric binary regression technique for group testing data that allow for the incorporation of a single continuous explanatory variable. They introduced a local moment estimator and showed that its asymptotic squared error enjoys the optimal rate. However, this method ignores the heterogeneity in the variance of the group testing responses. Delaigle & Hall (2012) extended this method to the case of homogeneous grouping; i.e., instead of randomly assigning individuals to each group, specimens of individuals who have similar covariate information are pooled together. By doing this, we can gain more information about the underlying population and hence produce an estimator with a smaller asymptotic squared error. However, this pooling strategy is not quite commonly in practice. In Chapter 2, we focus on the random grouping mechanism and propose a new nonparametric estimator of the regression function based on a locally weighted likelihood approach. Pointwise asymptotic normality of our estimator is established. Further, simulation results show that our methodology is as good if not better than the existing procedure in terms of comparing the mean squared error in prediction of the estimates.

All of the aforementioned regression methods proceed under the assumption that each of the individuals are assigned to exactly one pool, and make use of the testing responses observed from assaying these pools to perform inference. Therefore, these regression techniques can not be used to analyze data arising from classification studies. Merging the goals of estimation and classification, Xie (2001) and Zhang et al. (2013) allow for the incorporation of additional retesting information gained from decoding positive pools. Further, Zhang et al. (2013) illustrated that regression parameter estimates obtained from incorporating decoding information are more efficient than those based on individual level testing data, when the assay being used is imperfect. That is to say, these authors were able to show that more precise inference can be realized

through the analysis of group testing data than can be obtained through the use of individual level testing data, and at a fraction of the data collection cost. These two works are proceeded through the use of traditional parametric models, such as a logistic model. A natural question would be how to develop a nonparametric model to incorporate classification for estimation. Considering all the nonparametric methods mentioned above, one drawback is that they mainly considered the case where only one continuous covariate is available. To cover multiple covariates, a straightforward extension is through the use of multivariate kernel functions. However, this approach suffers from the so-called "curse-of-dimensionality"; i.e., convergence rate of the estimator decreases exponentially with rise in the number of covariates. In Chapter 3, we propose a general framework for modeling all variants of group testing data while allowing for the incorporation of multivariate covariates and accounting for the misclassification effects. The new model can be viewed as a generalization of the traditional single index model. The single index model was proposed by Ichimura (1993). It bridges the gap between parametric and nonparametric modeling; i.e., it keeps the interpretability of the parametric model and the flexibility of a nonparametric method while avoiding the "curse-of-dimensionality". Due to these, single index models are classified as semiparametric and have gained a lot of popularity during the past two decades; see Härdle et al. (1993), Klein & Spady (1993), Xia et al. (2002), Xia (2006), Zhu & Xue (2006), Wang et al. (2010), Cui et al. (2011) and the references therein. Unlike these literature, we do not have a response available for each individual, instead, we only have the availability of the high-structured testing responses obtained form assaying pools of individuals.

The last chapter of this dissertation mainly concerns with the evaluation of assay measurement error. In the statistical literature, there are two measurements of the testing error rates of an assay; i.e., sensitivity and specificity. Sensitivity (specificity) is defined as the probability that a specimen tests positive (negative) given that it is truly positive (negative). In all the aforemetioned studies, these two rates are commonly assumed as known constants both of which do not depend on the pool size. However, in many applications, a diagnose result of a specimen is based on a measurement of its concentration level of a certain biological marker. If the measured concentration is above (below) a pre-determinated threshold, the specimen is diagnosed as positive (negative). In group testing, many individuals' specimens are physically mixed together. One very possible situation is that a positive specimen can be easily diluted by many other negative ones. This dilution effect can highly affect the diagnostic accuracy. McMahan et al. (2013) developed a method to evaluate pool specific testing error rates for the regression analysis of testing responses of initial master pools. In this chapter, we generalized this idea to allow for testing responses obtained from all variants of group testing algorithms.

Chapter 2

A New Nonparametric Group Testing Regression Estimation

2.1 Introduction

Group (pooled) testing arises frequently in scientific studies. Pooling specimens for the purpose of estimating the prevalence of disease has proven to be an efficient method of reducing time and cost associated with sampling. For example, rather than testing blood specimens collected from individuals separately, group testing specifies that the specimens are first pooled and the resulting pooled specimen is then tested for the existence of the characteristic. This type of testing has also been used in pollution detection (Nagi & Raggi, 1972; Wahed et al., 2006) and contamination and toxicity studies (Lennon, 2007).

In group testing studies, experimenters often collect data on auxiliary variables that are easy and cost effective to measure. In most of these studies, the probability curve $p(x) = \text{pr}(T = 1|X = x)$ is of interestwhere T is the binary response and X is a covariate. Delaigle & Meister (2011) proposed a nonparametric estimator of $p(x)$ when the grouping mechanism is homogeneous, i.e. groups are constructed using similar values of the covariate. In practice, constructing pools in this fashion may not be feasible. In this article we consider the case where individuals are grouped randomly with observed binary responses are of the form T_j^*, $j = 1, ..., J$ where $T_j^* = \max_{1 \le i \le n_j} T_{ij}$, where T_{ij} is the status of the ith individual in the jth pool. T_{ij}s are not observed although all the accompanying covariates $X_{ij}, i = 1, ..., n_j$, $j = 1, ..., J$ are measured. Parametric analysis of binary data of this type has been addressed by Vansteelandt et al. (2000),

Bilder & Tebbs (2009), Chen et al. (2009), and Huang & Tebbs (2009) among others. A thorough literature review followed by a nonparamtric estimation method for $p(x)$ based on a method of moment argument is presented by Delaigle & Meister (2011). They obtained an expression for the pointwise asymptotic mean square error of their estimator and provided a detailed bandwidth selection method.

In this Chapter we address the estimation of $p(x)$ using a likelihood approach. We examine a localization procedure that provides an asymptotically normal estimator of $p(x)$ which maintains high finite sample accuracy. Our numerical results show that, for the examined examples, the proposed procedure has the same type of finite sample properties compared with Delaigle & Meister (2011) when group sizes are equal and the regression function p is smooth. However, when the group sizes are not equal, our method appears to have better finite sample properties compared with theirs. The same trend seems to be true when $p(x)$ is more fluctuant. It is noteworthy that the function $p(x)$ may not be smooth in all situations. For example, when examining the probability of an adverse reaction based on a drug dosage, the reaction probability can sharply increase or even jump at certain dosage thresholds. In addition, unequal grouping is not uncommon when one uses individuals in clusters of units in a system. For example, one may consider each class as a group when there are multiple schools of different sizes and levels in a school system from which the data are collected.

The remainder of this Chapter is organized as follows. In Section 2.2 we describe our procedure and state the main asymptotic results. Section 2.3 is devoted to a simulation study and a real data analysis followed by a short discussion. All the proofs are listed in the Appendix A.

2.2 A Semi-Local Likelihood Method

We describe the proposed estimator followed by its properties in this section. We assume $(T_{ij}, X_{ij}), i = 1, ..., n_j; j = 1, ..., J$ are i.i.d. random vectors. In what follows we assume that all of the covariates, X_{ij}s, and the pool testing responses T_j^*s, as defined in the previous section, are available. For any fixed x, and a user defined finite bandwidth h, we define $I_x = [x - h, x + h]$ and $Z_{ij} = X_{ij} I_x(X_{ij})$, where $I_x(X_{ij}) = 1$ if $X_{ij} \in I_x$, and $I_x(X_{ij}) = 0$ otherwise. Then the mixed pdf of the Zs is given by

$$f_Z(z) = \begin{cases} \int_{I_x^c} f(u)du & \text{if } z = 0 \\ f(z) & \text{if } z \in I_x \setminus 0 \end{cases},$$

where I_x^c is the complement of the set I_x^c, $I_x \setminus 0$ is the set I_x excluding 0 and $f(\cdot)$ is the density function of an X. Then $T_{ij} \mid Z_{ij} = z$ is a Bernoulli random variable with pr$(T_{ij} = 0|Z_{ij} = z) = r(z)$ where

$$r(z) = \begin{cases} r_1 & \text{if } z = 0 \\ q(z) & \text{if } z \in I_x \setminus 0 \end{cases},$$

with $r_1 = \int_{I_x^c} q(u)f(u)du / \int_{I_x^c} f(u)du$ and $q(z) = 1 - p(z)$. It is easy to see that $0 < r_1 \leq \sup_x q(x)$, and $r_1 \to q_*$ where $q_* = E[q(X)]$, as $h \to 0$. Note that $r(z)$ can also be written as

$$r(z) = r_1^{1-I_x(z)} \times q(z)^{I_x(z)}.$$

Now, we can write the log-likelihood of T_j^*, $j = 1, ..., J$, conditional on Z_{ij}s as

$$\frac{1}{N}\sum_{j=1}^{J}\Big\{(1-T_j^*)\sum_{i=1}^{n_j}\log r(Z_{ij}) + T_j^*\log\Big[1-\exp\Big(\sum_{i=1}^{n_j}\log r(Z_{ij})\Big)\Big]\Big\}.$$

For small h and a fixed x, a Taylor expansion gives the following approximation

$$\begin{aligned} \log r(Z_{ij}) &\approx I_x(X_{ij})g(x) + I_x(X_{ij})g'(x)(X_{ij}-x) + (1-I_x(X_{ij}))\log r_1 \\ &= I_x(X_{ij})\theta_1 + I_x(X_{ij})(X_{ij}-x)\theta_2 + (1-I_x(X_{ij}))\theta_3, \end{aligned} \tag{2.1}$$

where $g(\cdot) = \log q(\cdot), \theta_1 = g(x), \theta_2 = g'(x), \theta_3 = \log r_1$. Define $\theta = (\theta_1, \theta_2, \theta_3)^\top$, $\mathbf{X}_j = (X_{1j}, \ldots, X_{n_j j})^\top$ and $\tilde{\mathbf{X}}_j = \sum_{i=1}^{n_j}(I_x(X_{ij}), I_x(X_{ij})(X_{ij}-x), 1-I_x(X_{ij}))^\top$. Equation (2.1) provides a local linear approximation of $\log r(\cdot)$ using the X_{ij} in I_x, in the event that no X_{ij} in I_x, then no local linear approximation would be performed. Then, we can write the local log-likelihood as

$$l(\theta) = \frac{1}{N}\sum_{j=1}^{J}\Big\{(1-T_j^*)\theta^\top\tilde{\mathbf{X}}_j + T_j^*\log\Big[1-\exp\Big(\theta^\top\tilde{\mathbf{X}}_j\Big)\Big]\Big\}\omega_h(\mathbf{X}_j, x), \tag{2.2}$$

where $\omega_h(\mathbf{X}_j, x) = \prod_{i=1}^{n_j} K_h(X_{ij}-x)^{\delta_x(X_{ij})}$, $\delta_x(X_{ij}) = I_x(X_{ij}) / \sum_{i=1}^{n_j} I_x(X_{ij})$, which is defined to be 0 if the denominator is 0, and $K_h(\cdot) = h^{-1}K(\cdot/h)$ for a symmetric and continuous density function $K(\cdot)$.

Note that if $p(\cdot)$ has sufficient smoothness, we can use a local polynomial approximation for $g(x)$ in (2.1) and estimate the derivatives of g up to a desired order. However, since in practice the order of the smoothness of $p(\cdot)$ is usually unknown and the local linear estimator behaves better than the local constant

estimator (Fan & Gijbels, 1996), we present the local linear approximation case here.

The Hessian matrix of $l(\theta)$ is given by

$$l''(\theta) = -\frac{1}{N}\sum_{j=1}^{J}\frac{T_j^* \exp(\tilde{\mathbf{X}}_j^\top\theta)}{(1-\exp(\tilde{\mathbf{X}}_j^\top\theta))^2}\tilde{\mathbf{X}}_j\tilde{\mathbf{X}}_j^\top\omega_h(\mathbf{X}_j, x).$$

Since $l''(\theta)$ is negative definite with probability 1 when $N \to \infty$, the local log-likelihood (2.2) has a unique maximizer with respect to θ with probability 1. Let $\hat{\theta}$ be the maximizer of l. Then the first component of $\hat{\theta}$, $\hat{g}(x)$, is our proposed estimator of $g(x)$. Subsequently, our estimator of $p(x)$ is given by $1-\exp(\hat{g}(x))$.

Remark 2.2.1. *The log-likelihood of T_j^*s conditional on X_{ij}s instead of Z_{ij}s is*

$$\frac{1}{N}\sum_{j=1}^{J}\Big\{(1-T_j^*)\sum_{i=1}^{n_j}\log q(X_{ij}) + T_j^*\log\Big[1-\exp\Big(\sum_{i=1}^{n_j}\log q(X_{ij})\Big)\Big]\Big\}.$$

One could suggest to estimate $\log q(x)$ *by applying a local Poisson function, i.e.,* $q(X_{ij}) \approx \exp(\theta_1+\theta_2(X_{ij}-x))$ *and then maximizing the following local log-likelihood with respect to* (θ_1, θ_2)

$$\begin{aligned}\tilde{l}(\theta_1,\theta_2) \quad = \quad & \frac{1}{N}\sum_{j=1}^{J}\Big\{(1-T_j^*)\sum_{i=1}^{n_j}(\theta_1+\theta_2(X_{ij}-x)) \\ & +T_j^*\log\Big[1-\exp\Big(\sum_{i=1}^{n_j}(\theta_1+\theta_2(X_{ij}-x))\Big)\Big]\Big\}\prod_{i=1}^{n_j}K_h(X_{ij}-x).\end{aligned}$$

When group sizes are larger than one, the product of kernel functions acts like a multivariate kernel which results in a degraded estimation rate (Fan & Gijbels, 1996). Moreover, if we take $K(\cdot)$ to be a kernel function of compact support, such as the Epanechnikov kernel, once one $K_h(X_{ij}-x)$ is zero, the whole product part is zero which impacts the contribution of other X_{ij}s with nonzero values of $K_h(X_{ij}-x)$. Our truncated version rectifies this problem by counting every X_{ij} in the neighborhood I_x. Thus, the use of Z_{ij}s is more informative. One might argue to use $\omega_h(\boldsymbol{X}_j, x) = K(\|\boldsymbol{X}_j - x\|/h)$ in place of $\prod_{i=1}^{n_j}K_h(X_{ij}-x)$ above. However, this still acts like a multivariate kernel limiting its use.

2.3 Asymptotic Properties

To present the large sample properties of our estimator, we assume several mild regularity conditions:

Condition 2.1. $\sup_j n_j < \infty$.

Condition 2.2. $\log q(x)$ *has bounded second order derivative in a neighborhood of* x*, and* f *is positive and continuous in that neighborhood.*

Condition 2.3. $Nh \to \infty$ *and* Nh^5 *is bounded.*

This first condition is also used in Delaigle & Meister (2011). The next two are commonly used conditions on smoothness. Further, we introduce some notations. Under Condition 2.1, suppose there are only K different group sizes, denoted by $n^{(1)}, \ldots, n^{(K)}$. Let J_k be the number of groups of size $n^{(k)}$ and $\lim_{N\to\infty} n^{(k)} J_k/N = \gamma_k$. Then $\sum_{k=1}^K \gamma_k = 1$. For easy notation, we suppose the data are ordered as follows: the first J_1 groups are of size n_1, the next J_2 groups are of size n_2, and so on until the last J_K groups are of size n_K. Now, let $a = \sum_{k=1}^K \gamma_k V_{k1}$, $b = \sum_{k=1}^K \gamma_k (n^{(k)} - 1) V_{k1}$, $c = \sum_{k=1}^K \gamma_k (n^{(k)} - 1)^2 V_{k1}$, $d = \sum_{k=1}^K \gamma_k V_{k0}$, and $e = \sum_{k=1}^K \gamma_k V_{k2}$, where

$$\begin{aligned} V_{k0} &= n^{(k)} \exp(E_{k0})/(1 - \exp(E_{k0})), \\ V_{k1} &= f(x) \exp(E_{k1})/(1 - \exp(E_{k1})), \\ V_{k2} &= [\exp(E_{k1})\theta_2^* f(x)/(1 - \exp(E_{k1}))^2 + \exp(E_{k1}) f'(x)/(1 - \exp(E_{k1}))], \end{aligned}$$

with $E_{km} = m\theta_1^* + (n^{(k)} - m) \log q_*$. Further denote

$$V_0 = \begin{pmatrix} a\mu_0 & 0 & b\mu_0 \\ 0 & a\mu_2 & 0 \\ b\mu_0 & 0 & c\mu_0 + d \end{pmatrix}, V_1 = \begin{pmatrix} a\nu_0 & 0 & b\nu_0 \\ 0 & a\nu_2 & 0 \\ b\nu_0 & 0 & c\nu_0 \end{pmatrix} \text{ and } \text{Bias}_\theta = V_0^{-1} b_\theta,$$

where $b_\theta = 2^{-1} g^{(2)}(x) \cdot (a\mu_2 h^2, e\mu_4 h^3, b\mu_2 h^2)^\top$, $\mu_i = \int_{-1}^1 u^i K(u) du$ and $\nu_i = \int_{-1}^1 u^i K^2(u) du$.

The first theorem below provides the consistency of the estimator. The second theorem provide the large sample distribution of $\hat{\theta}$.

Theorem 2.3.1. *Under Conditions 2.1–2.3, we have*

$$H(\hat{\theta} - \theta^*) \to_p 0,$$

where $H = diag\{1, h, 1\}$*,* $\theta^* = (\theta_1^*, \theta_2^*, \theta_3^*)^\top$ *is the value of* θ *calculated by the true probability curve* $p(x)$*, and* $\to_p$ *means converges in probability.*

Proof. See Appendix A.1 □

Theorem 2.3.2. *Under the same conditions of Theorem 2.3.1,*

$$\sqrt{Nh}(H(\hat{\theta} - \theta^*) - \mathit{Bias}_\theta) \to_d N(0, V_0^{-1} V_1 V_0^{-1}),$$

where $\to_d$ *means converges in distribution.*

Proof. See Appendix A.2 □

For any vector γ, let $[\gamma]_1$ be its first element, and for any matrix Γ, let $[\Gamma]_{11}$ be its $(1,1)$th element. Then, we have

$$\sqrt{Nh}(\hat{\theta}_1 - \theta_1^* - [\text{Bias}_\theta]_1) \to_d N(0, [V_0^{-1} V_1 V_0^{-1}]_{11}).$$

Our estimate of $p(x)$ is $\hat{p}(x) = 1 - \exp(\hat{\theta}_1)$. Then $\hat{p}(x) - p(x) = -[\exp(\hat{\theta}_1) - \exp(\theta_1^*)]$. Simply following the delta method, we have the asymptotic properties of $\hat{p}(x)$.

Corollary 2.3.1. *Under conditions of Theorem 2.3.2, we have*

$$\sqrt{Nh}(\hat{p}(x) - p(x) - B(x)) \to_d N(0, V(x)),$$

where $B(x) = -(1 - p(x))[\mathit{Bias}_\theta]_1$ *and* $V(x) = [1 - p(x)]^2 [V_0^{-1} V_1 V_0^{-1}]_{11}$ *with*

$$\begin{aligned} [\mathit{Bias}_\theta]_1 &= \frac{g^{(2)}(x)\mu_2}{2\mu_0} h^2, \\ [V_0^{-1} V_1 V_0^{-1}]_{11} &= \frac{\nu_0}{\mu_0} \Big[\frac{c\mu_0 + d}{ac\mu_0^2 + ad\mu_0 - b^2\mu_0^2} - \frac{b^2 d \mu_0^2}{(ac\mu_0^2 + ad\mu_0 - b^2\mu_0^2)^2} \Big]. \end{aligned}$$

Proof. The proof of this corollary simply follows applying the delta method to the results of Theorem 2.3.2, and hence is omitted. □

2.4 Empirical Studies

In this section we provide a simulation study followed by the analysis of a real data set to illustrate our proposed method.

2.4.1 Bandwidth Selection

It is well known that bandwidth selection is crucial in nonparametric estimation. To save computational cost, we follow Delaigle & Meister (2011) to investigate a plug-in method. Based on Theorem 2.3.2, we can write $B_\theta(x) = [\text{Bias}_\theta]_1$ and $V_\theta(x) = [V_0^{-1}V_1V_0^{-1}]_{11}$ to emphasize the dependence of these quantities on x , which are the bias and the asymptotic variance of estimating θ_1 by $\hat{\theta}_1$. A reasonable way to pick the bandwidth h is by minimizing a weighted "asymptotic mean integrated squared error" given by

$$\text{AMISE}(h) = \int \Big[B_\theta^2(u) + \frac{V_\theta(u)}{Nh}\Big]\omega(u)du,$$

with respect to h for a suitable weight function w. Here, we take $w(u) = f(u)$. This gives

$$\text{AMISE} = \frac{\mu_2^2}{4\mu_0^2}B_\theta h^4 + \frac{V_\theta^*}{Nh},$$

where $B_\theta = \int g^{(2)}(x)f(x)dx$, $V_\theta^* = \int V_\theta(x)f(x)dx$. Then, the optimal bandwidth is given by

$$h^* = (V_\theta^*\mu_0^2/B_\theta\mu_2^2)^{-1/5}N^{-1/5}.$$

However, h^* can not be directly calculated since B_θ and V_θ^* are unknown. We can use

$$\hat{h} = (\hat{V}_\theta\mu_0^2/\hat{B}_\theta\mu_2^2)^{-1/5}N^{-1/5},$$

by replacing V_θ^* and B_θ with the estimates $\hat{V}_\theta^*$ and $\hat{B}_\theta$ given below.

We denote G_i as the number of groups of size $\geq i$, where $i = 1, \ldots, \max_j n_j$. For each fixed i, we pick $X_{i,j}, j = 1, \ldots, G_i$ from each group and denote the order statistics by $X_{i,(1)} < X_{i,(2)} < \cdots < X_{i,(G_i)}$. For a given estimator $\hat{V}_\theta(X_{i,(j)})$ of $V_\theta(X_{i,(j)})$, let $\hat{V}_i = \sum_{j=1}^{G_i-1}\hat{V}_\theta(X_{i,(j)})\hat{f}(X_{i,(j)})(X_{i,(j+1)} - X_{i,(j)})$, where $\hat{f}(x)$ is a kernel density estimate of $f(x)$. Then we can estimate V_θ^* by

$$\hat{V}_\theta^* = \sum_{i=1}^{\max_j n_j} w_i\hat{V}_i,$$

where $w_i = \sqrt{G_i}/\sum_{l=1}^{\max_j n_j}\sqrt{G_l}$. Now, it suffices to find a $\hat{V}_\theta(X_{i,(j)})$. We start by deriving a consistent

estimate $\hat{q}_*$ of q_* by maximizing the full likelihood of T_j^*s given by

$$\prod_{j=1}^{J}\{T_j^*(1-q_*^{n_j})+(1-T_j^*)q_*^{n_j}\}.$$

Note that γ_k (defined in the Appendix) can be estimated by the proportion of the groups of size n_k among all the groups. In estimating the ratio $q(X_{i,(j)})/(1-q_*^{n^{(k)}-1}q(X_{i,(j)}))$ which appears in V_{k1} (defined in the Appendix), we use $J_k^{-1}\sum_{j:n_j=n^{(k)}}(1-T_j^*)$ to estimate the denominator since it is required to be less than 1. Using the arguments of Delaigle & Meister (2011), the numerator can be estimated by $N\hat{\mu}^*\hat{q}_*^{-n_j}(1-T_j^*)/\sum_{j=1}^{J}n_j\hat{q}_*^{n_j-1}$ where $\hat{\mu}^*=N^{-1}\sum_{j=1}^{J}n_j(1-T_j^*)$. Furthermore, B_θ can be estimated nonparametrically by

$$\hat{B}_\theta=\sum_{i=1}^{\max_j n_j}G_i^{-1}w_i\sum_{j=1}^{G_i}\{\hat{g}_i^{(2)}(X_{ij})\}^2,$$

where the construction of $\hat{g}_i^{(2)}(x)$ is similar to Delaigle & Meister (2011) which is omitted here.

It is well known that nonparametric estimators are in general not stable near boundaries. We replace $\hat{B}_\theta$ and $\hat{V}_\theta$ by weighted versions (Gasser et al., 1991) as $\hat{B}_\theta=\sum_{i=1}^{\max_j n_j}G_i^{-1}w_i\sum_{j=1}^{G_i}\{\hat{g}_i^{(2)}(X_{ij})\}^2\omega_B(X_{ij})$ and $\hat{V}_\theta=\sum_{j=1}^{G_i-1}\hat{V}_\theta(X_{i,(j)})\hat{f}(X_{i,(j)})(X_{i,(j+1)}-X_{i,(j)})\omega_V(X_{i,(j)})$, where $\omega_B(x)$ and $\omega_V(x)$ are two weight functions. Our suggestion is to take $\omega_B(x)=1_{(q_{0.1},q_{0.9})}(x)$ and $\omega_V(x)=1_{(q_{0.3},q_{0.7})}(x)$, where $1_{(a,b)}(x)$ is the indicator function (it equals to 1 if $a\le x\le b$; otherwise 0), and q_α is the α quantile of all the X_{ij}s.

2.4.2 Numerical Simulation

Our numerical studies were conducted to check the finite sample performance of the proposed semi-local likelihood estimator of $p(x)$. We considered the following models each with $X\sim U[-1,1]$ and $X\sim N(0,0.5^2)$.

Model 2.1. $p(x)=\{\sin(3\pi x/2)+1.2\}/[20+360x^2\{sign(x)+1\}]$;

Model 2.2. $p(x)=\sin^2(\pi(x-1)/2)\cos^2(1.5\pi(x-1))/6$;

Model 2.3. $p(x)=\cos^2(\pi x)/8$;

Model 2.4. $p(x)=\cos^2(\pi x)/16+x^2/20$.

The first model is similar to the model used in Delaigle & Meister (2011). The others are designed to have relatively high fluctuant structure. For each model above, we considered both $N=5000$ and $10,000$.

The group sizes for equal group size case were $n_j = 5$ or 10. For the unequal group sizes case n_js were randomly and uniformly chosen from $\{1, \ldots, 5\}$ or $\{1, \ldots, 10\}$. We simulated 200 random samples of $\{(X_{ij}, T_j^*), i = 1, \ldots, n_j, j = 1, \ldots, J\}$ for each setting of N, n_j, p and the distribution of X, where $N = \sum_{j=1}^{J} n_j$, $T_j^* = \max_{1 \le i \le n_j} T_{ij}$ and T_{ij}s are generated according to a Bernoulli distribution with success probability $p(X_{ij})$. The bandwidth, h, was selected using the procedure outlined in Section 2.4.1. Based on this h, the estimator $\hat{p}(x)$, written LL (local likelihood estimator), of $p(x)$ was calculated. We also applied the method from Delaigle & Meister (2011). These authors provided four ways for selecting the bandwidth, ROT, ROT$_{\omega_0}$, PI$_{\omega_1}$ and PI$_{\omega_0}$. The kernel $K(\cdot)$ was taken to be the standard normal density in all cases and resulting estimates were truncated to be in $[0, 1]$ since $p(x)$ is a probability curve. We compared our estimate with each of their four estimates based on the integrated squared error ISE $= \int_a^b \{\hat{p}(x) - p(x)\}^2 dx \approx M^{-1}(b-a) \sum_{i=1}^{M} \{\hat{p}(t_i) - p(t_i)\}^2$ for 200 replications, where $[a, b]$ is the interval of interest, and $\{t_i, i = 1, \ldots, M\}$ is an even partition of $[a, b]$. Furthermore, to get a feel for the pointwise behavior of each estimator, we calculated the following pointwise mean square error ratio (PMSER),

$$\text{PMSER}(t_i) = \frac{\sum_{k=1}^{200} \{\hat{p}_k(t_i) - p(t_i)\}^2}{\sum_{k=1}^{200} \{\tilde{p}_k(t_i) - p(t_i)\}^2}, i = 1, ..., M,$$

where $\hat{p}_k$ is our estimator of p for the kth sample and the $\tilde{p}_k$ denotes the estimators proposed by Delaigle & Meister (2011).

In Tables 2.1–2.3 below we provide a subset of our findings. The average and the standard deviation of the 200 ISEs corresponding to each estimator and the proportion of PMSER(t_i) values < 1 among all the t_is for $M = 300$ for $N = 10000$ are also given. The results for $N = 5000$ followed an almost identical pattern and are therefore not presented here. Additionally, global integrated squared errors GISE $= \int_a^b \{\bar{p}(x) - p(x)\}^2 dx$ were compared, where $\bar{p}(x) = \sum_{k=1}^{200} \hat{p}_k(x)/200$ which is referred to as the average curve for each method.

From Tables 2.1 and 2.2 we can see that all means and standard deviations of 200 replications using our method are smaller than the corresponding values for the methods in Delaigle & Meister (2011) for the case of unequal groups. Moreover, the proportion of PMSER value below 1 is great than 50% in all such cases. For the case of equal groups, a summary is presented in Table 2.3. The average ISE values and the pointwise mean square error values indicate that the two methods are very similar in the case of equal group sizes. In comparing the GISEs, our method seems to outperform the moment type estimator in all examined cases, a few results listed in Table 2.4. Plots of the averaged estimates of $p(x)$, $\bar{p}(x)$, for all models

Table 2.1: Simulation results for Models 2.1-2.4 when group sizes are unequal and X follows uniform, $N = 10^4$. The presented results are: $10^4\times$ MISE ($10^4\times$stdev, proportion of PMSER < 1).

n_j	Model	LL	ROT	ROT$_{\omega_0}$	PI$_{\omega_1}$	PI$_{\omega_0}$
1-5	2.1	1.56 (.76)	3.63 (1.22, .76)	4.21 (1.20, .77)	2.35 (.94, .66)	1.95 (.89, .67)
	2.2	7.92 (2.21)	23.7 (1.43, .80)	24.4 (1.34, .80)	14.9 (2.15, .90)	11.8 (2.32, .92)
	2.3	3.61 (1.82)	28.6 (4.79, .90)	30.7 (4.25, .91)	6.72 (2.60, .71)	4.56 (2.15, .72)
	2.4	2.53 (1.16)	7.13 (1.29, .80)	7.65 (1.34, .81)	3.04 (1.23, .63)	2.62 (1.21, .63)
1-10	2.1	2.60 (1.40)	4.61 (1.66, .62)	5.12 (1.60, .65)	2.97 (1.38, .58)	2.80 (1.40, .65)
	2.2	8.93 (3.01)	26.0 (2.30, .76)	26.6 (2.26, .76)	15.8 (3.28, .87)	12.3 (3.32, .88)
	2.3	6.13 (2.86)	30.5 (5.52, .87)	32.8 (4.61, .87)	9.65 (3.59, .71)	7.64 (3.37. .71)
	2.4	4.37 (2.19)	8.89 (2.24, .74)	9.47 (2.26, .76)	4.70 (2.26, .56)	4.52 (2.34, .61)

Table 2.2: Simulation results for Models 2.1–2.4 when group sizes are unequal and X follows normal, $N = 10^4$. The presented results are: $10^4\times$ MISE ($10^4\times$stdev, proportion of PMSER < 1).

n_j	Model	LL	ROT	ROT$_{\omega_0}$	PI$_{\omega_1}$	PI$_{\omega_0}$
1-5	2.1	3.57 (1.37)	8.80 (1.87, .87)	9.40 (1.68, .87)	4.58 (1.53, .85)	3.88 (1.49, .66)
	2.2	19.7 (5.41)	33.1 (5.29, .82)	33.9 (5.40, .82)	23.0 (5.08, .77)	20.9 (5.40, .70)
	2.3	15.1 (4.64)	40.7 (2.44, .95)	41.3 (1.90, .94)	22.0 (4.52, .88)	17.7 (4.64, .62)
	2.4	8.26 (4.52)	14.6 (3.90, .74)	15.0 (3.75, .74)	8.79 (4.33, .56)	8.44 (4.52, .55)
1-10	2.1	3.47 (1.84)	9.17 (2.68, .82)	9.75 (2.56, .82)	4.26 (2.18, .81)	3.86 (2.19, .80)
	2.2	19.0 (4.19)	30.9 (3.65, .84)	31.6 (3.79, .82)	21.3 (3.27, .74)	19.3 (3.35, .63)
	2.3	11.2 (3.90)	41.5 (3.06, .90)	41.9 (2.77, .90)	17.7 (4.30, .87)	14.6 (4.18, .87)
	2.4	4.99 (2.54)	12.53 (3.01, .83)	13.0 (2.9, .83)	6.28 (2.76, .79)	5.69 (2.55, .76)

Table 2.3: Simulation results for Models 2.1 and 2.3 when group sizes are equal, $N = 10^4$. U and N denote uniform and normal, respectively. The presented results are: $10^4\times$ MISE ($10^4\times$stdev, proportion of PMSER < 1).

n_j	Model	$f(x)$	LL	ROT	ROT$_{\omega_0}$	PI$_{\omega_1}$	PI$_{\omega_0}$
5	2.1	U	2.22 (1.08)	3.87 (1.23, .50)	4.43 (1.22, .53)	2.67 (1.0, .43)	2.28 (1.0, .44)
		N	4.59 (1.51)	8.92 (2.08, .78)	9.54 (1.94, .79)	5.22 (1.55, .58)	4.46 (1.47, .23)
	2.3	U	5.56 (2.57)	29.8 (5.06, .87)	31.9 (4.38, .88)	8.76 (3.36, .59)	5.91 (2.79, .50)
		N	19.8 (5.27)	41.2 (2.74, .91)	41.7 (2.47, .91)	26.3 (4.42, .81)	21.9 (4.58, .49)
10	2.1	U	3.89 (2.04)	5.28 (1.90, .40)	5.72 (1.80, .42)	3.71 (1.81, .34)	3.48 (1.91, .34)
		N	5.94 (3.13)	10.2 (3.33, .72)	10.7 (3.27, .74)	6.01 (3.08, .47)	5.56 (3.13, .31)
	2.3	U	11.7 (5.73)	32.8 (5.6, .77)	34.8 (4.99, .79)	13.9 (5.42, .50)	11.3 (5.38, .43)
		N	16.9 (6.84)	42.8 (5.0, .85)	43.3 (4.74, .86)	24.5 (5.94, .80)	20.9 (6.21, .76)

Table 2.4: $10^4\times$ GISE for Models 2.1–2.4 when all $n_j = 10$, X is normal and $N = 10^4$.

Model	LL	ROT	ROT_{ω_0}	PI_{ω_1}	PI_{ω_0}
2.1	2.58	8.56	9.15	2.87	3.55
2.2	19.9	30.0	30.8	20.0	21.9
2.3	9.03	39.4	40.1	15.2	19.4
2.4	2.21	11.0	11.6	3.31	4.31

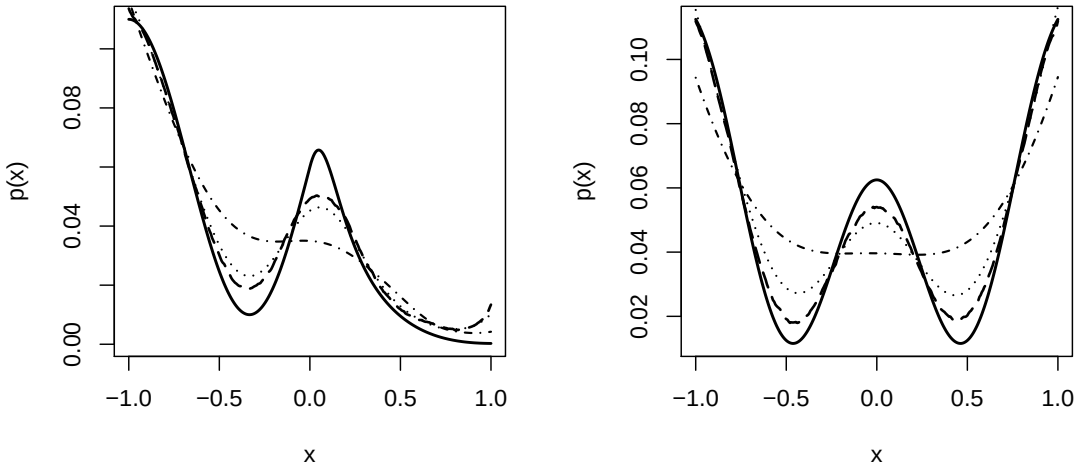


Figure 2.1: Average Curves: LL $(-\ -\ -)$, best between ROT and ROT_{ω_0} $(-\cdot-\cdot-)$, best between PI_{ω_1} and PI_{ω_0} $(\cdots)$. Left to right: Model 2.1, $X \sim U[-1,1]$, $n_j = 10$, $N = 10000$; Model 2.4, $X \sim U[-1,1]$, $n_j \sim U\{1,\ldots,10\}$, $N = 5000$.

reveal (see Figure 2.1 for Models 2.1 and 2.4) that our estimator appears to be significantly less biased over almost the entire support of X. When compared to the estimator proposed in Delaigle & Meister (2011), it is worthwhile to point out that the bias in their estimators becomes more prominent when $p(x)$ is less smooth or more fluctuant. This suggests that our method generally outperforms those proposed in Delaigle & Meister (2011) both globally and locally.

2.4.3 Real Data Analysis

We also applied our method to two real data sets from 1999-2000 in NHANES study which were previously analyzed by Delaigle & Meister (2011) and are available at (`www.cdc.gov/nchs/nhanes/nhanes1999-2000/nhanes99_00.htm`). The first data set contained two variables: the age variable X and the test result T_{HBc} which is a binary response taking values 0 and 1 indicating that the antibody to hepatitis B virus core antigen is absent or present in the patient's serum or plasma, respectively. The sample size was 7121, and X ranged from 6 to 85 years after removing the individuals with missing X or T_{HBc}. The second data set contained the age variable X, and a response variable $T_{\text{CL}} = 0$ or 1 indicating the absence or presence of genital chlamydia trachomatis infection in the urine of the patient, respectively. After removing the missing values, X ranged from 12 to 40 years, and the sample size was 2042. Our goal is to estimate the following two conditional probability curves: $p_{\text{HBc}}(x) = \text{pr}(T_{\text{HBc}=1}|X = x)$ and $p_{\text{CL}}(x) = \text{pr}(T_{\text{CL}=1}|X = x)$.

To evaluate the performance of our method, in each case, we first applied the local linear estimation based on all the (X, Y). The resulting estimator is denoted by $\tilde{p}$ and is treated as our reference curve. Then we artificially pooled the data randomly assigning individuals to groups of size $n_j \sim U\{1, 2\}$, $n_j \sim U\{1, \ldots, 5\}$, or $n_j \sim U\{1, \ldots, 10\}$. In each of these aforementioned cases, we calculated our estimator $\hat{p}$ using the individual level covariates and the simulated pool responses. This process was then repeated 200 times for both infections on pooling strategy. The average curve along with a two standard deviation pointwise confidence bands based on the 200 replications are presented in Figure 2.2. Here the lower band was truncated at 0. From these graphs it appears that there is a large degree of agreement between our estimator and the reference estimator.

Delaigle & Meister (2011) evaluated their estimator using the estimates corresponding to quantiles of the ISD values, and the estimate corresponding to the median ISD value showed boundary bias. Since we have established the asymptotic normality of our estimator, we prefer to use the average of the estimates with pointwise confidence bands in assessing the estimation accuracy. The average of the 200 estimates shows

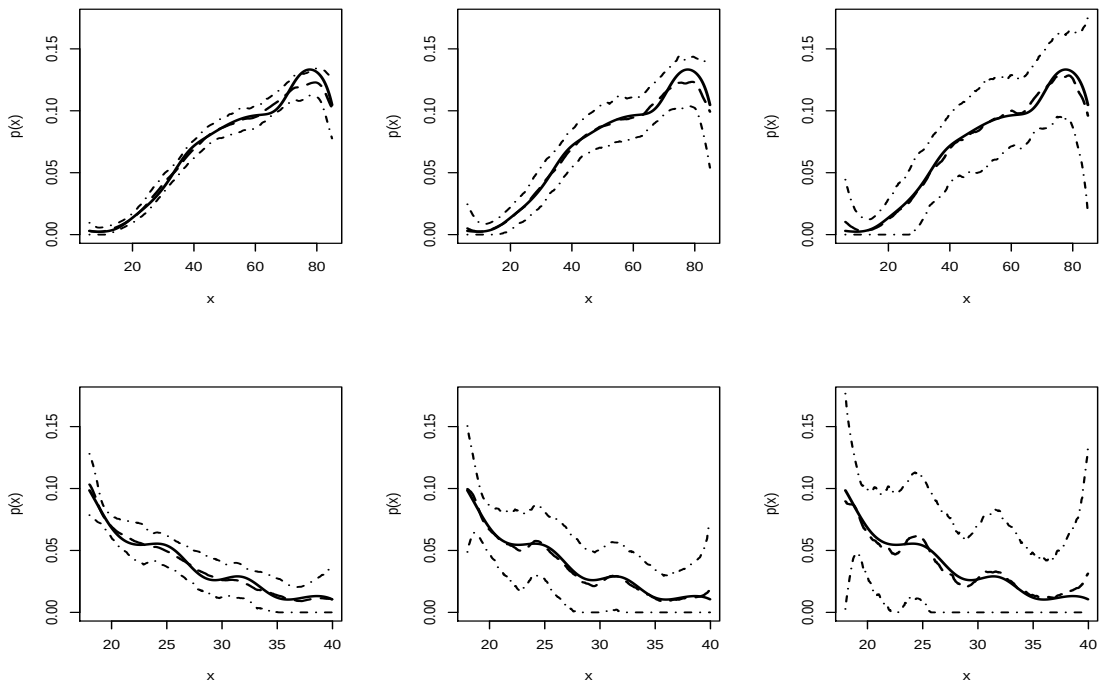


Figure 2.2: NHANES study: Average curves $(- \ - \ -)$ and confidence bands $(- \cdot - \cdot -)$ for T_{HBC} (Top) and T_{CL} (Bottom). Left to right: $n_j \sim U\{1, 2\}$, $n_j \sim U\{1, \ldots, 5\}$, $n_j \sim U\{1, \ldots, 10\}$.

minimal boundary bias and the ideal curve is well within the point-wise confidence bounds.

2.5 Discussion

We have provided an effective way of estimating the regression function $\text{pr}(Y = 1|X = x)$ based on group data. Our estimator seems to perform well in all possible sampling situations for a variety of model functions. The proposed bandwidth selection procedure seems to provide very satisfactory estimation results. An interesting extension of these ideas would be to test the equality of the regression curves for different populations.

Chapter 3

Semiparametric Group Testing Regression Models

3.1 Introduction

Group testing, also known as pooled testing, was first proposed by Dorfman (1943) as a means to reduce the cost associated with screening World War II inductees for syphilis. In order to reduce testing expenditure, Dorfman suggested that pooled specimens, formed from combining blood samples collected from individuals, be tested for the presence of syphilis. If the initial pool, also referred to as a master pool, tested negative then all contributing men could be declared negative at the expense of only one test. Alternatively, positive master pools would be resolved by retesting each of the contributing specimens one-by-one. Since this seminal work, many variants of Dorfman's decoding strategy have been proposed in an effort to further reduce screening cost or increase classification accuracy; for a review see Kim et al. (2007).

In addition to being used for case identification, pooling techniques have also been implemented for the purposes of estimation, predominantly in the context of estimating population level characteristics; see Bilder & Tebbs (2005) for a review. More recently, authors have developed binary regression models which relate pool response data to individual-level covariate information through a specified link function; see Vansteelandt et al. (2000), Bilder & Tebbs (2009), Chen et al. (2009), and Huang & Tebbs (2009). To obviate the specification of the link function, Delaigle & Meister (2011), Delaigle & Hall (2012), and Wang et al. (2013) proposed nonparametric binary regression techniques for group testing data that allow for the

incorporation of a single continuous explanatory variable. Delaigle & Meister (2011) discussed extensions of their approach which allow for multiple covariates via a multivariate kernel function. However, due to the curse of dimensionality this approach may not be suitable for evaluating multiple explanatory variables. The aforementioned regression methods were designed to model data arising from master pool testing only; i.e., these methods cannot incorporate information gained from decoding positive pools. To our knowledge, the only binary regression models that allow for the incorporation of decoding information were proposed by Xie (2001) and Zhang et al. (2013), and were developed under parametric assumptions.

Since its advent, group testing has been successfully implemented for the purposes of screening for a variety of infectious diseases (Lewis et al., 2012; Van et al., 2012), and has found applications in areas such as genetics (Gastwirth, 2000), drug discovery (Remlinger et al., 2006), medical entomology (Venette et al., 2002), veterinarian science (Munoz-Zanzi et al., 2000), and plant pathology (Venette et al., 2002). The group testing strategy implemented varies according to the goals of the study and often does not conclude with master pool testing. Consequently, in this chapter we propose a general regression methodology for modeling test responses obtained from all group testing algorithms that allows for the incorporation of multiple covariates and accounts for imperfect testing. Unlike the aforementioned parametric methods, our semiparametric model enjoys the modeling flexibility of nonparametric procedures, but is not subject to the curse of dimensionality when multiple predictors are available. We develop hypothesis testing methods for evaluating the significance of potential predictors based on the asymptotic properties of our proposed estimators. Through simulation, we illustrate that our methodology can more reliably evaluate potential predictors when compared to analogous parametric methods.

Our methodology falls broadly into the class of single index models, which have attracted much attention in the statistical literature over the past few decades; see Ichimura (1993), Härdle et al. (1993), Klein & Spady (1993), Xia et al. (2002), Xia (2006), Zhu & Xue (2006), Cui et al. (2011) and the references therein. Though similar, there exists a fundamental difference between our method and those previously proposed in the literature. Specifically, all existing single index models require that a response be available for each individual, while in contrast our method requires only the availability of the responses obtained from testing pools of individuals. Therefore, the complex data structure resulting from group testing algorithms cannot be handled by any of the existing single index techniques.

3.2 Models and Methodology

3.2.1 Modeling Assumptions and General Estimation Procedure

In what follows, we propose a general modeling framework for data arising from any group testing algorithm. Our proposed methodology can be greatly simplified under two of the most common group testing algorithms, master pool testing and Dorfman decoding, as is illustrated in the subsequent sections. Consider implementing a group testing algorithm to screen N individuals for a binary characteristic of interest, such as infection status. In general, this process begins by randomly assigning each of the individuals to exactly one of J initial groups of size c_j. Let $\mathcal{G}_j = \{1, \ldots, c_j\}$ be a collection of indices identifying the c_j individuals assigned to the jth group. Within the jth group, screening is performed according to the protocol outlined by the specified group testing algorithm, resulting in K_j testing responses Y_{jl}, for $l = 1, \ldots, K_j$. We let $Y_{jl} = 1$ indicate that the lth pool tested positive, and $Y_{jl} = 0$ otherwise. We identify the individuals in the jth group whose specimens were pooled and tested by the lth assay by the set $\mathcal{P}_{jl} \subseteq \mathcal{G}_j$, and we define $Z_{jl} = (Y_{jl}, \mathcal{P}_{jl})$. For notational convenience, we collect all of the observed testing data associated with the jth group into the set $Z_j = \{Z_{j1}, \ldots, Z_{jK_j}\}$, and we assume throughout that $Z_j \perp Z_{j'}$ for all $j \neq j'$, where $\perp$ denotes statistical independence.

Let T_{ij} denote the true status of the ith individual in the jth group, where $T_{ij} = 1$ indicates that the individual is positive, and $T_{ij} = 0$ otherwise. For modeling purposes, we assume that $X_{ij} = (X_{ij1}, \ldots, X_{ijp})^{\mathrm{T}}$, a p-dimensional vector of covariates, is available for each individual and that the random vectors (T_{ij}, X_{ij}) are independent and identically distributed. In order to relate the individuals' true statuses to their predictor variables, we proceed under the single index generalization; i.e., we assume that $\text{pr}(T_{ij} = 1 \mid X_{ij} = x) = p(x^{\mathrm{T}}\beta)$, where $p(\cdot)$ is an unknown smooth probability curve and $\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a p-dimensional vector of regression parameters. To ensure identifiability, as with all single index models, we assume that the support of the covariate vectors, $\mathbb{X}$, is a bounded convex set with at least one interior point and the parameter space of β is $\mathcal{B} = \{\beta = (\beta_1, \ldots, \beta_p)^{\mathrm{T}} : \|\beta\| = 1, \beta_1 > 0\}$, where $\|\beta\|$ denotes the Euclidean norm of β (Lin & Kulasekera, 2007). If one observed T_{ij}, for $i = 1, \ldots, c_j$ and $j = 1, \ldots, J$, then standard single index estimation procedures could be employed to estimate $p(\cdot)$ and β, but when the assay being used is imperfect and the testing responses are based on pooled assessments the individuals' true statuses are latent and these techniques are inapplicable.

To account for imperfect testing, we let S_e and S_p denote the sensitivity and specificity, respectively, of the assay being employed; i.e., S_e is the probability that a specimen will test positive given it is truly

positive and S_p is the probability that a specimen will test negative given it is truly negative. We assume that S_e and S_p are known, constant, and independent of the pool size. Further, we assume that given the true status of the pools being tested $Y_{jl} \perp Y_{jl'}$, for $l \neq l'$. These assumptions are common in the group testing literature; see, Xie (2001), Kim et al. (2007), and Zhang et al. (2013).

Using the testing error rates and these assumptions we now relate the observed testing outcomes to the true underlying statuses of the specimens being tested. To accomplish this, we let $\mathcal{Z}(c)$ denote the set of all possible outcomes resulting from screening a group of size c according to a specific group testing algorithm. Likewise, we define the set of all possible true statuses for the individuals assigned to a group of size c to be $\mathcal{T}(c)$. The conditional probability of observing any $Z = \{(Y_1, \mathcal{P}_1), \ldots, (Y_K, \mathcal{P}_K)\} \in \mathcal{Z}(c)$ given any $T = (T_1, \ldots, T_c) \in \mathcal{T}(c)$ can be calculated as

$$M(Z, T, c) = \text{pr}(\mathcal{P}) \prod_{l=1}^{K} \left\{ S_e^{Y_l \tilde{Y}_l} (1 - S_e)^{(1-Y_l)\tilde{Y}_l} (1 - S_p)^{Y_l(1-\tilde{Y}_l)} S_p^{(1-Y_l)(1-\tilde{Y}_l)} \right\},$$

where $\tilde{Y}_l = \max_{i \in \mathcal{P}_l} T_i$ and $\mathcal{P} = \{\mathcal{P}_1, \ldots, \mathcal{P}_K\}$. The probability $\text{pr}(\mathcal{P})$ accounts for the randomness, if any, in the pooling protocol of the group testing algorithm. In Appendix B.1 we provide a derivation of $M(Z, T, c)$ and illustrate how $\text{pr}(\mathcal{P})$ should be evaluated.

In what follows we relate the observed testing outcomes arising from a group testing algorithm to the individual-level covariate information. Through an application of the law of total probability it is easy to show that the conditional probability of observing Z_j given β, $p(\cdot)$, and $\mathcal{X}_j$ can be expressed as

$$\mathcal{R}\{Z_j; \mathcal{X}_j, \beta, p(\cdot)\} = \sum_{T \in \mathcal{T}(c_j)} M(Z_j, T, c_j) \prod_{i=1}^{c_j} p(X_{ij}^{\mathrm{T}}\beta)^{T_i} \{1 - p(X_{ij}^{\mathrm{T}}\beta)\}^{1-T_i}, \tag{3.1}$$

where $\mathcal{X}_j = (X_{1j}, \ldots, X_{c_j j})^{\mathrm{T}}$. To derive (3.1) we proceed under the assumption that the observed testing outcomes are independent of the measured covariates, given the individuals' true statuses. Thus, the full conditional log-likelihood of $\{(Z_1, \mathcal{X}_1), \ldots, (Z_J, \mathcal{X}_J)\}$ can be expressed as

$$l\{\beta, p(\cdot)\} = \sum_{j=1}^{J} \log \mathcal{R}\{Z_j; \mathcal{X}_j, \beta, p(\cdot)\}.$$

If $p(\cdot)$ were known, an estimate of β could be obtained as the maximizer of $l\{\beta, p(\cdot)\}$. Thus, the primary challenge of fitting our model is to account for the dependence between the infinite-dimensional parameter $p(\cdot)$ and the finite-dimensional parameter β. To explicitly acknowledge this dependence, we write $p(\cdot)$ as $p_\beta(\cdot)$,

and again point out that an estimate of β could be obtained as the maximizer of $l\{\beta, p_\beta(\cdot)\}$, if $p_\beta(\cdot)$ were known. In order to estimate the regression parameters, we propose to replace the unknown function $p_\beta(\cdot)$ by a consistent estimator, $\hat{p}_\beta(\cdot)$, so that our estimator of β can be obtained as $\hat{\beta} = \text{argmax}_{\beta \in \mathcal{B}}\, l\{\beta, \hat{p}_\beta(\cdot)\}$.

As previously stated, traditional single index techniques are not applicable in this context, because the individuals' statuses are latent. To circumvent this, we propose to make use of the individuals' diagnosed statuses. To this end, let D_{ij} denote the diagnosed status of the ith individual in the jth group, such that $D_{ij} = 1$ indicates a positive diagnosis, and $D_{ij} = 0$ otherwise. Typically, an individual's diagnosed status is determined based on the observed testing outcomes and the specified testing protocol; i.e., $D_{ij} = \Lambda(i, Z_j)$, where Λ is a decision function unique to the group testing algorithm being implemented. Define $\mathcal{F}_{ij}(t, \mu) = \text{pr}(D_{ij} = 1 \mid T_{ij} = t)$, which can be calculated as

$$\mathcal{F}_{ij}(t, \mu) = \sum_{Z \in \mathcal{Z}_i(c_j)} \sum_{T \in \mathcal{T}(c_j)} I(T_i = t) M(Z, T, c_j) \prod_{k \neq i} \left\{ \mu^{1-T_k} (1-\mu)^{T_k} \right\},$$

where $\mu = \text{pr}(T_{ij} = 0)$ and $\mathcal{Z}_i(c) = \{z \in \mathcal{Z}(c) : \Lambda(i; z) = 1\}$; i.e., $\mathcal{Z}_i(c)$ is the set of all possible testing outcomes which would result in the ith individual in a group of size c being diagnosed positive. The quantities $\mathcal{F}_{ij}(1, \mu)$ and $1 - \mathcal{F}_{ij}(0, \mu)$ are commonly referred to as the pooling sensitivity and specificity, respectively, and under specific group testing algorithms these measures of testing accuracy have nice analytic forms; see Kim et al. (2007).

In order to develop an estimator of $p_\beta(\cdot)$, we consider the conditional probability that an individual will be diagnosed positive, given the linear predictor $X_{ij}^{\mathrm{T}}\beta$, which can be expressed as

$$E(D_{ij} \mid X_{ij}^{\mathrm{T}}\beta = u) = a_{ij}(\mu) + b_{ij}(\mu) p_\beta(u), \tag{3.2}$$

where $a_{ij}(\mu) = \mathcal{F}_{ij}(0, \mu)$ and $b_{ij}(\mu) = \mathcal{F}_{ij}(1, \mu) - \mathcal{F}_{ij}(0, \mu)$. The unknowns in (3.2) are μ and $p_\beta(\cdot)$. Since μ is the unconditional probability that an individual is truly negative, one could obtain an estimator, $\hat{\mu}$, of this parameter by maximizing the full log-likelihood

$$l_p(\mu) = \sum_{j=1}^{J} \log \left(\sum_{T \in \mathcal{T}(c_j)} \left[M(Z_j, T, c_j) \prod_{i=1}^{c_j} \left\{ \mu^{1-T_i} (1-\mu)^{T_i} \right\} \right] \right), \tag{3.3}$$

with respect to μ; i.e., $\hat{\mu} = \text{argmax}_\mu\, l_p(\mu)$. Then, based on equation (3.2), we can obtain a local linear kernel

estimator of $p_\beta(\cdot)$ at a given point u by minimizing

$$\sum_{j=1}^{J}\sum_{i=1}^{c_j}\left[D_{ij}-a_{ij}(\hat{\mu})-b_{ij}(\hat{\mu})\left\{p_\beta(u)+p'_\beta(u)(X_{ij}^{\mathrm{T}}\beta-u)\right\}\right]^2 K_h(X_{ij}^{\mathrm{T}}\beta-u), \tag{3.4}$$

with respect to $\{p_\beta(u), p'_\beta(u)\}^{\mathrm{T}}$, where $p'_\beta(\cdot)$ denotes the first derivative of $p_\beta(\cdot)$, h is a user defined bandwidth, $K(\cdot)$ is a symmetric kernel density function, and $K_h(\cdot)=h^{-1}K(\cdot/h)$. We define $\{\hat{p}_\beta(u), \hat{p}'_\beta(u)\}^{\mathrm{T}}$, the minimizer of (3.4), to be our estimator of $\{p_\beta(u), p'_\beta(u)\}^{\mathrm{T}}$. More explicitly, we could write $\hat{p}_\beta(u)$ and $\hat{p}'_\beta(u)$ as

$$\hat{p}_\beta(u)=\frac{\hat{T}_{N0}(u,\beta)\hat{S}_{N2}(u,\beta)-\hat{T}_{N1}(u,\beta)\hat{S}_{N1}(u,\beta)}{\hat{S}_{N0}(u,\beta)\hat{S}_{N2}(u,\beta)-\hat{S}^2_{N1}(u,\beta)}, \tag{3.5}$$

$$\hat{p}'_\beta(u)=\frac{\hat{T}_{N1}(u,\beta)\hat{S}_{N0}(u,\beta)-\hat{T}_{N0}(u,\beta)\hat{S}_{N1}(u,\beta)}{\hat{S}_{N0}(u,\beta)\hat{S}_{N2}(u,\beta)-\hat{S}^2_{N1}(u,\beta)}, \tag{3.6}$$

where

$$\begin{aligned}
\hat{T}_{Nl}(u,\beta)=&N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}\{D_{ij}-a_{ij}(\hat{\mu})\}b_{ij}(\hat{\mu})\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,u;l),\\
\hat{S}_{Nl}(u,\beta)=&N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}b^2_{ij}(\hat{\mu})\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,u;l),\\
\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,u;l)=&K_h(X_{ij}^{\mathrm{T}}\beta-u)\left(\frac{X_{ij}^{\mathrm{T}}\beta-u}{h}\right)^l.
\end{aligned}$$

This closed form of $\hat{p}_\beta(u)$ can help us greatly improve the computational efficiency of our method. Consequently, our final estimators can be expressed as

$$\hat{\beta}=\text{argmax}_{\beta\in\mathcal{B}}\, l\left\{\beta,\hat{p}_\beta(\cdot)\right\},\quad \hat{p}(u)=\hat{p}_{\hat{\beta}}(u). \tag{3.7}$$

Note that (3.4) is not the standard form of the local sum of squares, because the diagnosed statuses are correlated and $\hat{\mu}$ is a random term that depends on the observed testing data. Despite these differences, in Section 3.3 we show that our approach efficiently estimates β and $p(\cdot)$. In the following two sections, we outline the formulas necessary to implement our regression methodology under master pool testing and Dorfman decoding. A more detailed illustration is provided in Appendix B.2.

3.2.2 Estimation under Master Pool Testing

The testing protocol under master pool testing specifies that specimens collected from individuals belonging to a common group be combined to form a single master pool which is subsequently assayed; i.e., the testing data available for modeling is $Z_j = \{(Y_{j1}, \mathcal{P}_{j1})\}$, where $\mathcal{P}_{j1} = \mathcal{G}_j$. If $Y_{j1} = 0$, then all individuals in this group are diagnosed as negative, whereas $Y_{j1} = 1$ indicates that at least one individual is at risk. Thus, we define $D_{ij} = \Lambda(i, Z_j) = Y_{j1}$. Under master pool testing, the log-likelihood (3.3) reduces to

$$l_p(\mu) = \sum_{j=1}^{J} (1 - Y_{j1}) \log p_{j0} + Y_{j1} \log(1 - p_{j0}),$$

where $p_{j0} = 1 - S_e - \delta_{c_j}$ and $\delta_c = (1 - S_e - S_p)\mu^c$. Similarly, a series of simple arguments provide that $a_{ij}(\mu) = S_e + \delta_{c_j - 1}$ and $b_{ij}(\mu) = S_e - a_{ij}(\mu)$. Finally, for the jth group the observed testing data Z_j belongs to the set $\{(0, \mathcal{P}_{j1}), (1, \mathcal{P}_{j1})\}$, and the conditional probability outlined in (3.1) associated with either of these outcomes is $\mathcal{R}\{(0, \mathcal{P}_{j1}); \mathcal{X}_j, \beta, p(\cdot)\} = 1 - S_e - \delta_0 \prod_{i=1}^{c_j} \{1 - p(X_{ij}^{\mathrm{T}}\beta)\}$ or $\mathcal{R}\{(1, \mathcal{P}_{j1}); \mathcal{X}_j, \beta, p(\cdot)\} = 1 - \mathcal{R}\{(0, \mathcal{P}_{j1}); \mathcal{X}_j, \beta, p(\cdot)\}$. The estimators defined in (3.7) are then obtained as described in Section 3.2.1.

3.2.3 Estimation under Dorfman Decoding

Dorfman decoding proceeds in a similar fashion to master pool testing, with the key difference that positive pools are resolved by retesting all contributing individuals one-by-one. Consequently, Z_j can take two forms, the first being $Z_j = \{(Y_{j1}, \mathcal{P}_{j1})\}$, where $Y_{j1} = 0$ and $\mathcal{P}_{j1} = \mathcal{G}_j$, denoting that the master pool tested negative. The second occurs when the master pool test is positive; i.e., $Y_{j1} = 1$ and $\mathcal{P}_{j1} = \mathcal{G}_j$, in which case $Z_j = \{(Y_{j1}, \mathcal{P}_{j1}), \ldots, (Y_{jK_j}, \mathcal{P}_{jK_j})\}$ where $K_j = c_j + 1$ and $\mathcal{P}_{jl} = \{l - 1\}$, for $l = 2, \ldots, K_j$. The ith individual's diagnosed status is determined to be $D_{ij} = \Lambda(i, Z_j) = 1$ if and only if $Y_{j1} = 1$ and $Y_{j,i+1} = 1$, $D_{ij} = \Lambda(i, Z_j) = 0$ otherwise; i.e., a positive diagnosis requires both the master pool and individual level test to be positive. Under Dorfman testing, the log-likelihood (3.3) reduces to

$$l_p(\mu) = \sum_{j=1}^{J} \left\{ I(Y_{j1} = 0) \log p_{j0} + \sum_{k=0}^{c_j} I\left(Y_{j1} = 1, \sum_{l=2}^{c_j+1} Y_{jl} = k \right) \log p_{j1k} \right\}.$$

where $p_{j1k} = \delta_{c_j}(1 - S_p)^k S_p^{c_j - k} + S_e(S_e + \delta_1)^k (1 - S_e - \delta_1)^{c_j - k}$, $p_{j0} = 1 - S_e - \delta_{c_j}$, and $\delta_c = (1 - S_e - S_p)\mu^c$. Similarly, simple arguments yield $a_{ij}(\mu) = (1 - S_p)^2 \mu^{c_j - 1} + S_e(1 - S_p)(1 - \mu^{c_j - 1})$ and

$b_{ij}(\mu) = S_e^2 - a_{ij}(\mu)$.

The approach described in Section 3.2.2 can be used to calculate the probability that the jth master pool will test negative; i.e., in this case we have that $\mathcal{R}\{(0, \mathcal{P}_{j1}); \mathcal{X}_j, \beta, p(\cdot)\} = 1 - S_e - \delta_0 \prod_{i=1}^{c_j}\{1 - p(X_{ij}^{\mathrm{T}}\beta)\}$. To express the probability of the other testing outcomes, we define $\mathcal{I}_{j1} = \{i \in \mathcal{G}_j : D_{ij} = 1\}$ and $\mathcal{I}_{j0} = \{i \in \mathcal{G}_j : D_{ij} = 0\}$; i.e., the sets $\mathcal{I}_{j1}$ and $\mathcal{I}_{j0}$ identify the $k = |\mathcal{I}_{j1}|$ and $c_j - k = |\mathcal{I}_{j0}|$ individuals in the jth group that were diagnosed as positive and negative, respectively. Thus, for other testing outcomes $\mathcal{R}\{Z_j; \mathcal{X}_j, \beta, p(\cdot)\}$ is

$$\sum_{k_1=0}^{k}\sum_{k_0=0}^{c_j-k} S_e^{k_1+I(k_1+k_0>0)}(1-S_e)^{k_0} S_p^{c_j-k-k_0}(1-S_p)^{k-k_1+I(k_1+k_0=0)} \prod_{l=0}^{1} \mathrm{pr}(\mathcal{S}_{jl} = k_l), \tag{3.8}$$

where $\mathcal{S}_{jl} = \sum_{i\in\mathcal{I}_{jl}} T_{ij}$. The probabilities in (3.8) are conditional on the unknown parameters and predictor variables, so $\mathcal{S}_{j1}$ and $\mathcal{S}_{j0}$ are the sum of independent and non-identically distributed Bernoulli random variables; i.e., $\mathcal{S}_{j1}$ and $\mathcal{S}_{j0}$ each follow a Poisson binomial distribution. The estimators defined in (3.7) are then obtained as described in Section 3.2.1.

3.3 Asymptotic Properties

We assume that $J \to \infty$ as $N \to \infty$ while group sizes remain finite. This is reasonable since in practice the group sizes are naturally bounded by implementation considerations. Further, this assumption is common in the group testing literature; see Delaigle & Meister (2011). We denote the range of c_j by $\{c^{(1)}, \ldots, c^{(M)}\}$. More explicitly, for all pooled observations there exists an m such that $c_j = c^{(m)}$. Further, for each m we let J_m denote the number of groups having size $c^{(m)}$, and assume that $J_m c^{(m)}/N \to \gamma_m$ as $N \to \infty$; i.e., γ_m represents the proportion of individuals assigned to groups of size $c^{(m)}$.

Theorem 3.3.1 provides the asymptotic properties of our proposed estimators $\hat{\beta}$ and $\hat{p}(\cdot)$. These properties holds under the following regularity conditions.

Condition 3.1. *The functions $d_\beta(u) = E(X \mid X^{\mathrm{T}}\beta = u)$ and $p_\beta(u)$ have bounded and continuous second order derivatives.*

Condition 3.2. *The density function of $X^{\mathrm{T}}\beta$ is bounded away from zero and satisfies a Lipschitz condition of order 1 on $\{u = x^{\mathrm{T}}\beta : x \in \mathbb{X}\}$.*

Condition 3.3. *The bandwidth $h = CN^{-1/5}$ for some constant $C > 0$, and $K(\cdot)$ is a bounded and symmetric density function with bounded first derivative.*

Condition 3.4. *The function $M(\cdot, \cdot, \cdot)$ is bounded away from 0.*

Condition 3.5. *The equation $\beta^{\mathrm{T}}\Omega\beta = 0$ has the unique root $\beta = \beta_0$ in $\mathcal{B}$.*

Conditions 3.1–3.3 are common in the single index literature. The Lipschitz condition in Condition 3.2 allows for discrete predictor variables. Condition 3.4 is easily satisfied when the assay is imperfect, as long as $0.5 < S_e, S_p < 1$. This also assures that the denominator in Ω is bounded away from 0. Condition 3.5 guarantees that the matrix $\mathcal{J}_0^{\mathrm{T}}\Omega\mathcal{J}_0$ is positive definite.

Further, in order to succinctly present these results we let $\beta_0 = (\beta_{01}, \beta_0^{(1)\mathrm{T}})^{\mathrm{T}}$ and $p_0(\cdot)$ denote the true unknown parameters, where $\beta_0^{(1)} = (\beta_{02}, \ldots, \beta_{0p})^{\mathrm{T}}$. We define

$$\Omega_c = c^{-1} \sum_{z \in \mathcal{Z}(c)} E\left[\mathcal{R}^{-1}\left\{z; \mathcal{X}^{(c)}, \beta_0, p_0(\cdot)\right\} \sum_{i=1}^{c} \{P_i(z,1,c) - P_i(z,0,c)\}^2 p_0'^2(X_i^{\mathrm{T}}\beta_0)\Gamma(X_i)\right],$$

where $\mathcal{X}^{(c)} = (X_1, \ldots, X_c)^{\mathrm{T}}$, $\Gamma(X) = \{X - E(X \mid X^{\mathrm{T}}\beta_0\}\{X - E(X^{\mathrm{T}} \mid X^{\mathrm{T}}\beta_0)\}^{\mathrm{T}}$, and $P_i(z,t,c) = \mathrm{pr}\{Z = z \mid T_i = t, \mathcal{X}^{(c)}, \beta_0, p_0(\cdot)\}$. Finally, we define $\Omega = \sum_{m=1}^{M} \gamma_m \Omega_{c^{(m)}}$ which plays an integral role in the asymptotic variance covariance matrix of $\hat{\beta}$. Under a specific testing protocol, e.g., master pool testing or Dorfman decoding, the above expression for Ω can be more explicit. To illustrate this fact, in Appendix B.3 we provide distinct versions of Ω for the methodology described in Sections 3.2.2 and 3.2.3. Using the above expressions we now give our main result.

Theorem 3.3.1. *Under Conditions 3.1–3.5, we have that*

$$N^{1/2}(\hat{\beta} - \beta_0) \to N(0, \Sigma)$$

in distribution, where $\Sigma = \mathcal{J}_0(\mathcal{J}_0^{\mathrm{T}}\Omega\mathcal{J}_0)^{-1}\mathcal{J}_0^{\mathrm{T}}$, $\mathcal{J}_0$ is the functional value of $\partial B(\beta^{(1)})/\partial\beta^{(1)}$ evaluated at $\beta^{(1)} = \beta_0^{(1)}$, and $B(\beta^{(1)}) = ([1 - \|\beta^{(1)}\|^2]^{1/2}, \beta^{(1)\mathrm{T}})^{\mathrm{T}}$. Further,

$$\sup_{x \in \mathbb{X}} \left|\hat{p}(x^{\mathrm{T}}\hat{\beta}) - p_0(x^{\mathrm{T}}\beta_0)\right|^2 = O_p\left\{(\log N)/(Nh)\right\}.$$

Proof. The proof is in Appendix B.6. □

The consistency rate for estimating $p_0(\cdot)$ is the same rate demonstrated for kernel smoothing es-

timators in a univariate nonparametric regression context; see Mach & Silverman (1982). The estimator $\hat{\mu}$ is a maximum likelihood estimator, its asymptotic normality follows from standard arguments and hence is omitted.

Theorem 3.3.1 suggests that large sample inference is possible once a good estimator $\hat{\Sigma}$ of Σ is obtained. To this end, Appendix B.4 gives an extension of a plug-in estimator of Σ that was originally proposed by Wang et al. (2010). Using $\hat{\beta}$ and $\hat{\Sigma}$ one can conduct Wald type inference (Wald , 1943); i.e., at the significance level α, a confidence interval for β_{0r} can be constructed as

$$\hat{\beta}_{0r} \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}_r N^{-1/2} \quad (r = 1, \ldots, p),$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal distribution and $\hat{\sigma}_r^2$ is the rth diagonal element of $\hat{\Sigma}$. Further, for $r < p$ one may also perform hypothesis tests of the form,

$$H_0 : \beta_{0q_1} = \cdots = \beta_{0q_r} = 0 \quad \text{versus} \quad H_1 : \text{not all } \beta_{0q_1}, \ldots, \beta_{0q_r} \text{ equal } 0,$$

using the test statistic $R_N = N(D\hat{\beta})^{\mathrm{T}}(D\hat{\Sigma}D^{\mathrm{T}})^{-1}D\hat{\beta}$, where D is a $r \times p$ matrix such that $D\beta_0 = (\beta_{0q_1}, \ldots, \beta_{0q_r})^T$. Given the results in Theorem 3.3.1, we have that under the null hypothesis R_N converges in distribution to a chi-square random variable having r degrees of freedom. Consequently, at the significance level α one would reject the null hypothesis if $R_N > \chi_r^2(1-\alpha)$, where $\chi_r^2(a)$ is the ath quantile of a chi-square distribution having r degrees of freedom.

3.4 Numerical Analysis

A simulation study was conducted to assess the finite sample performance of our methodology. This study considered the following three underlying true regression models:

Model 3.1. $p_0(u) = 1/\{1 + \exp(4 - 2u)\}$;

Model 3.2. $p_0(u) = \exp(-5u^2 - 1.5)$;

Model 3.3. $p_0(u) = [\sin\{\pi(u - 0.3)\} + 1.3]/[10 + 20(u - 0.3)^2\{sign(u - 0.3) + 1\}]$,

where $u = X^{\mathrm{T}}\beta_0$. Model 3.1 provides a situation under which a logistic link is appropriate, and Models 3.2 and 3.3 emulate the gonorrhea and chlamydia data studied in Section 3.5. For each of the above models

we considered a vector of predictors of the form $X = (X_1, X_2, X_3)^{\mathrm{T}}$, where X_1 follows a standard normal distribution, while X_2 and X_3 each follow a Bernoulli distribution with success probabilities 0.4 and 0.3, respectively. The regression parameters were specified to be $\beta_0 = (\beta_{01}, \beta_{02}, \beta_{03})^{\mathrm{T}} = \{1/3, (8/9 - \delta^2)^{1/2}, \delta\}^{\mathrm{T}}$, where $\delta = \{0, 0.1, 0.2, 0.3, 0.4\}$.

We set $N = 10000$ and considered a common group size $c_j = c$ for all $j = 1, \ldots, J$, where $J = N/c$ and $c \in \{1, 2, 5, 10\}$. The setting $c = 1$ corresponds to individual level testing. In order to generate group testing data, we first generated individual level data; i.e., for each of the N individuals we generated the pair (T_{ij}, X_{ij}). Specifically, the predictor vector X_{ij} was simulated according to the distributions described above and T_{ij} was subsequently determined according to a Bernoulli(p_{ij}) distribution, where $p_{ij} = p_0(X_{ij}^{\mathrm{T}}\beta_0)$. To create group testing data, we then simulated the screening of the N individuals according to both master pool testing and Dorfman decoding, chosen due to their popularity. To allow for testing errors, we generated testing responses using $S_e = 0.93$ and $S_p = 0.99$. Under both master pool testing and Dorfman decoding, this data generating process was repeated 500 times for each model and configuration of (c, δ).

For each of the group testing data sets we estimated the regression parameter β_0 and the link function $p_0(\cdot)$ using the methodology outlined in Section 3.2. To implement our approach we specified $K(\cdot)$ to be the Gaussian kernel, and selected the bandwidth in a similar fashion to the method proposed in Härdle et al. (1993). Specifically, the bandwidth $\tilde{h}$ was chosen such that $(\tilde{\beta}, \tilde{h})$ is the maximizer of $\mathrm{CV}(\beta, h) = \sum_{j=1}^{J} \log \mathcal{R}\{Z_j; \mathcal{X}_j, \beta, \hat{p}_{\beta}^{(-j)}(\cdot)\}$, where $\hat{p}_{\beta}^{(-j)}(u)$ denotes the leave-one-out estimator of $p_{\beta}(u)$ obtained from minimizing (3.4) when the information pertaining to the jth pool is omitted. For comparative purposes, we also implemented the parametric methods proposed in Vansteelandt et al. (2000) and Zhang et al. (2013) for master pool testing and Dorfman decoding, respectively, under the assumption the link function is logistic.

Table 3.1 provides summary statistics of the 500 estimates of β_0 obtained by our methodology, across all considered models and settings of c, under Dorfman decoding, when δ =0.1. Our approach exhibits little, if any, evidence of bias and the average standard errors are in agreement with the sample standard deviation of the parameter estimates. The empirical coverage probabilities for 95% confidence intervals are predominantly at their nominal level. Further, the parameter estimates obtained from analyzing group testing data can be as, if not more, efficient than the estimates based on individual level data; i.e., in most cases the estimators have smaller variances when $c > 1$. This suggests that more precise inference can be obtained from analyzing group testing decoding data, when compared to individual level testing information, and at a fraction of the cost of data collection, similar findings were reported in Zhang et al. (2013).

Table 3.1: Summary of simulation results for data arising from Dorfman decoding. BIAS and SD, empirical bias ($\times 10^3$) and standard deviation ($\times 100$) of the 500 estimates; SE, average standard error ($\times 100$); COV, empirical coverage probability ($\times 100$) for nominal 95% confidence interval; EMSE, average mean squared error of prediction ($\times 10^4$); RE, ratio of EMSE of the parametric model to the EMSE of our semiparametric model.

	Parameter	Measure	$c = 1$	$c = 2$	$c = 5$	$c = 10$
Model 3.1	β_{01}	BIAS(SD)	8.7(3.5)	9.0(3.1)	6.5(3.3)	7.1(3.1)
		COV(SE)	93.6(3.5)	94.4(3.2)	95.3(3.2)	95.8(3.3)
	β_{02}	BIAS(SD)	−5.1(1.4)	−4.7(1.3)	−4.4(1.3)	−4.4(1.3)
		COV(SE)	96.0(1.4)	96.2(1.3)	96.2(1.3)	96.8(1.4)
	β_{03}	BIAS(SD)	−1.9(5.2)	−4.2(4.9)	−1.6(5.4)	−4.1(5.6)
		COV(SE)	94.4(5.3)	94.6(5.0)	93.2(5.1)	92.6(5.3)
	$p_0(x\beta_0)$	EMSE(RE)	1.31(0.37)	1.25(0.35)	1.28(0.38)	1.27(0.39)
	Percentage reduction in testing			37.3 %	52.5 %	43.6 %
Model 3.2	β_{01}	BIAS(SD)	1.5(1.4)	0.7(1.4)	0.5(1.4)	1.9(1.4)
		COV(SE)	93.0(1.4)	95.3(1.4)	93.8(1.4)	95.1(1.4)
	β_{02}	BIAS(SD)	−1.2(0.6)	−1.0(0.6)	−0.6(0.6)	−1.1(0.6)
		COV(SE)	93.4(0.6)	94.5(0.6)	93.6(0.6)	96.2(0.6)
	β_{03}	BIAS(SD)	−0.7(3.4)	1.2(3.2)	−2.8(3.2)	−2.6(3.1)
		COV(SE)	93.0(3.0)	92.3(2.9)	92.6(2.9)	92.9(3.0)
	$p_0(x\beta_0)$	EMSE(RE)	1.25(25.33)	1.09(29.83)	1.18(27.43)	1.18(27.24)
	Percentage reduction in testing			31.9 %	41.9 %	29.7 %
Model 3.3	β_{01}	BIAS(SD)	7.6(2.5)	8.9(2.4)	8.5(2.4)	7.5(2.4)
		COV(SE)	92.4(2.5)	92.8(2.4)	92.3(2.5)	93.0(2.5)
	β_{02}	BIAS(SD)	−3.7(1.0)	−4.4(1.0)	−4.3(1.0)	−4.0(1.0)
		COV(SE)	93.8(1.0)	92.8(1.0)	94.3(1.0)	93.0(1.0)
	β_{03}	BIAS(SD)	−1.7(3.7)	−0.2(3.9)	1.5(3.6)	1.3(4.0)
		COV(SE)	92.4(3.6)	92.4(3.5)	94.0(3.6)	92.7(3.6)
	$p_0(x\beta_0)$	EMSE(RE)	1.61(13.80)	1.46(15.18)	1.46(15.19)	1.57(14.19)
	Percentage reduction in testing			34.8 %	47.4 %	36.7 %

Table 3.1 also provides the average mean squared error of prediction, where we define $\text{MSE}\{\hat{\beta}, \hat{p}(\cdot)\} = N^{-1} \sum_{j=1}^{J} \sum_{i=1}^{c_j} \{\hat{p}(X_{ij}^{\mathrm{T}}\hat{\beta}) - p_0(X_{ij}^{\mathrm{T}}\beta_0)\}^2$ to be the mean squared error of prediction for a given data set. This measure suggests that our methodology can more accurately estimate the link function, using decoding data, than the analogous method that makes use of individual level testing information. Table 3.1 provides the ratio of the average mean squared error of prediction for the parametric and our semiparametric model. We see that when the true underlying model is logistic the average mean squared error of prediction of our approach is roughly three times larger than that of the parametric model which assumes a logistic link. In contrast, when the true model is not logistic the average mean squared error of prediction associated with the parametric model can be up to thirty times greater than that of our methodology.

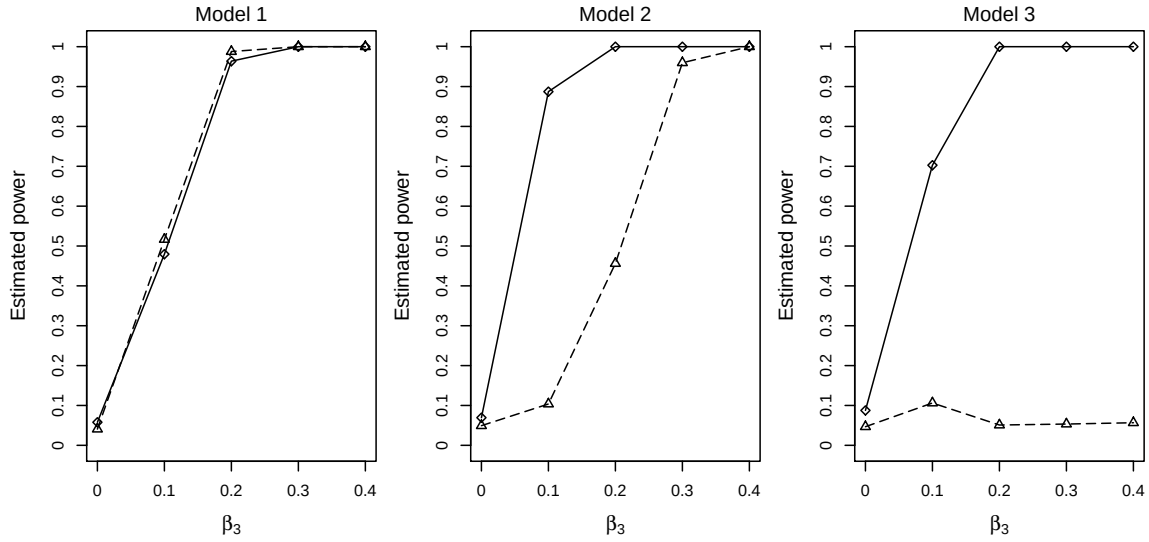


Figure 3.1: Estimated power curves under Dorfman decoding: the solid and dashed curves correspond to our approach and the parametric techniques, respectively.

We conducted a power analysis of the hypothesis test for β_{03}, using the estimates resulting from our regression procedures and the methodology outlined in Section 3.3 to perform the test of $H_0 : \beta_{03} = 0$ versus $H_1 : \beta_{03} \neq 0$, at the $\alpha = 0.05$ significance level. The same analysis was also performed for each data set using the aforementioned parametric models, again assuming a logistic link. The hypothesis testing results were used to construct power curves for our semiparametric approach and the competing parametric model, across all considered configurations. The power curves corresponding to data arising from Dorfman decoding when $c = 5$ are presented in Figure 3.1. Under both the semiparametric and parametric models the hypothesis testing procedure suggested in Section 3.3 maintains its correct size across all considered settings. The estimated power curves under Model 3.1 are very similar, with the parametric model having slightly more power. This suggests that our methodology performs almost as well as the parametric model which assumes the correct link function. If the link function is misspecified under the parametric model these methods lose the power to detect significant predictor variables, a feature not shared by our approach.

The results presented in Table 3.1 and Figure 3.1 are based on analyzing data arising from Dorfman decoding, and the parameter estimates summarized in Table 3.1 correspond to the case in which $\delta = 0.1$. The analogous table and figure for master pool testing are provided in the Appendix B.5. Under both group testing algorithms, summaries of the parameter estimates pertaining to other considered values of δ were practically identical and power curves constructed for the other values of c resulted in the same conclusions.

Consequently, these additional results were omitted for brevity.

3.5 Application to Chlamydia and Gonorrhea Data

In this section we illustrate our methodology using chlamydia and gonorrhea data collected by the Nebraska Public Health Laboratory. This laboratory tests patients individually for the presence of these bacterial infections, whereas other such laboratories have adopted group testing strategies; e.g., the Iowa Hygienic Laboratory uses a Dorfman type algorithm (Jirsa, 2008) to screen for these sexually transmitted diseases. The data we consider consist of individual level testing responses obtained from assaying urine specimens collected from $N = 7310$ female patients. In addition to these testing responses we also have access to several predictor variables: namely, X_1, standardized age; X_2, a binary variable indicating the presence of symptoms, with 1 indicating symptoms were present; and X_3, a binary variable indicating the purpose of screening, with 1 indicating family planning. Using these data, we are able to artificially construct group testing data, treating the testing responses available in the data set as the individuals' true infection statuses. We then assigned each of the individuals to a group of size c based on their specimen arrival date, where $c \in \{1, 2, 5, 10\}$. Dorfman decoding was implemented to screen the groups for both diseases, where testing responses for chlamydia and gonorrhea were simulated using the sensitivities 0.947 and 0.913 and specificities 0.989 and 0.993, respectively. These specifications were chosen to emulate the protocol and assay currently used by the Iowa Hygienic Laboratory. This process was repeated 500 times for each value of c and our model was fit to each resulting data set.

Table 3.2: Summary of results for data arising from Dorfman decoding: MEAN, mean ($\times 100$) of the 500 estimates; SE, average standard error ($\times 100$).

	Parameter	Measure	$c = 1$	$c = 2$	$c = 5$	$c = 10$
Chlamydia	β_{01}	MEAN(SE)	$81.7(6.3)$	$82.9(6.4)$	$82.7(6.1)$	$82.6(6.2)$
	β_{02}	MEAN(SE)	$-41.3(9.2)$	$-40.4(9.5)$	$-41.4(9.1)$	$-39.9(9.3)$
	β_{03}	MEAN(SE)	$38.8(14.7)$	$37.7(15.3)$	$36.8(14.8)$	$37.9(14.7)$
	Percentage reduction in testing			$34.0\ \%$	$45.7\ \%$	$35.4\ \%$
Gonorrhea	β_{01}	MEAN(SE)	$47.6(5.1)$	$47.7(2.4)$	$48.1(2.5)$	$47.1(2.8)$
	β_{02}	MEAN(SE)	$-70.0(7.8)$	$-69.8(3.6)$	$-70.0(3.8)$	$-71.2(4.3)$
	β_{03}	MEAN(SE)	$50.4(11.3)$	$53.1(5.7)$	$52.5(5.8)$	$51.3(6.3)$
	Percentage reduction in testing			$45.9\ \%$	$71.0\ \%$	$74.0\ \%$

Table 3.2 provides a summary of the parameter estimates obtained from analyzing the Dorfman

decoding data. The regression parameter estimates obtained by our methodology are similar across all values of c, and in many situations exhibit less variability than the estimates based on the artificial individual level data; i.e., when $c = 1$. Figure 3.2 provides 0.025, 0.5, and 0.975 pointwise quantile curves of the 500 estimated regression functions obtained from analyzing the Dorfman decoding data when $c = 1, 2, 5$, and 10. The estimated regression curves based on the group testing data exhibit less variability when compared to those based on individual screening data. These results indicate that through group testing the screening cost for chlamydia and gonorrhea can be reduced by up to 45.7% and 74.0%, respectively, while providing more precise inference.

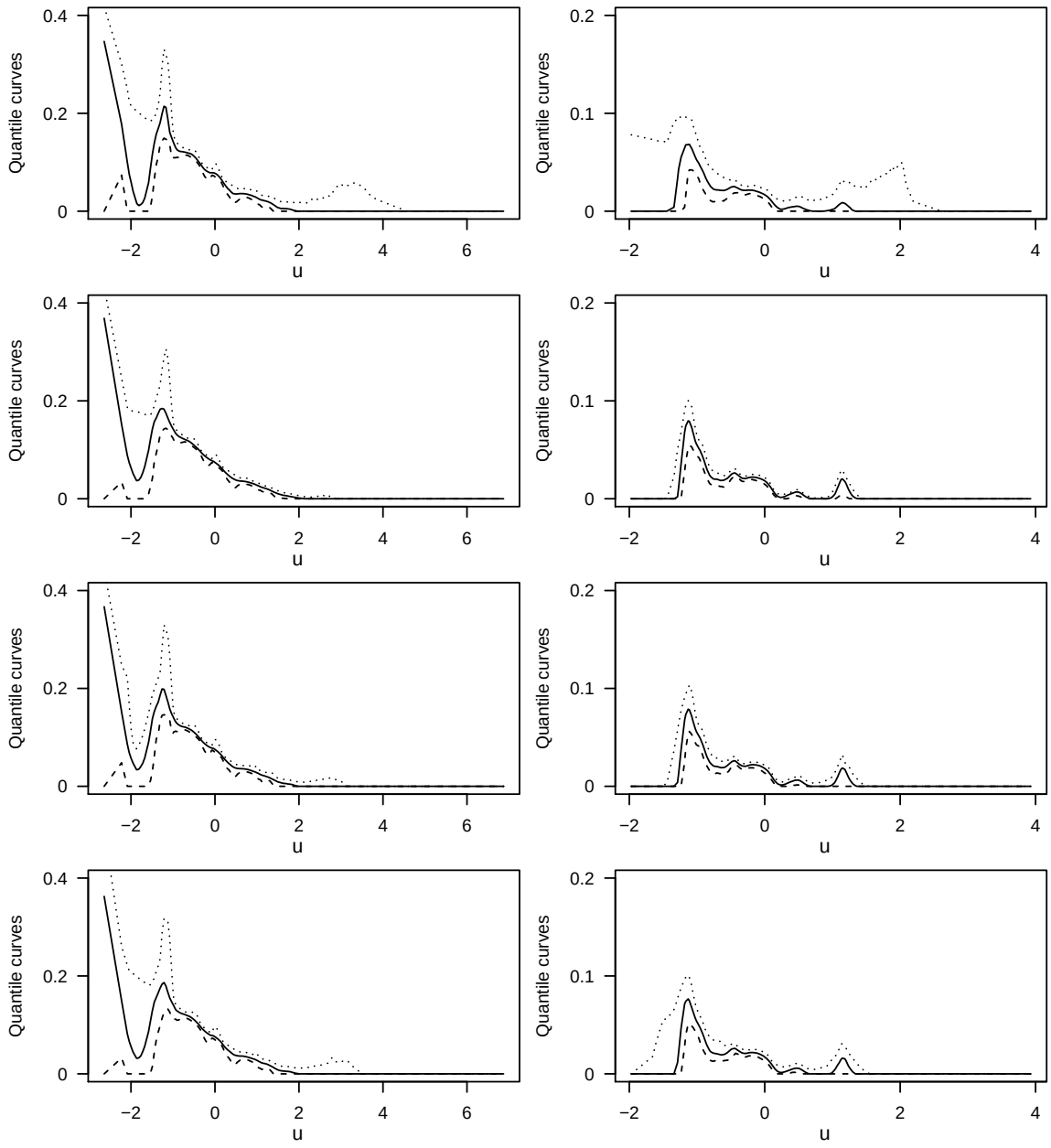


Figure 3.2: Pointwise quantile curves as a function of the linear predictor u. Left column: chlamydia data, right column: gonorrhea data. The four rows from top to bottom correspond to $c = 1, 2, 5$, and 10, respectively. The dashed, solid, and dotted lines correspond to the 0.025, 0.5, 0.975 quantiles, respectively.

Chapter 4

Parametric Group Testing Regression Models with Pool Dilution Effects

4.1 Introduction

To account for assay measurement error, most existing literature including the group testing regression techniques in previous chapters proceed under simplifying assumptions; i.e., they assume that the testing error rates, sensitivity and specificity, are known, constant, and are functionally independent of the pool size. These assumptions naturally come under scrutiny, when one considers the underlying mechanistic structure of an assay. In general, a diagnostic test measures the concentration of a biological marker (biomarker) within a specimen, and its binary response indicates whether or not this measurement exceeds a predetermined threshold. Consequently, the composition of a pooled specimen, in terms of the number of positive and negative individuals, plays a key role in determining the testing error rates. For example, a specimen that might test positive when tested individually, may be "diluted" past the assay's threshold of detection when pooled with multiple negative specimens. Acknowledging this structure, McMahan et al. (2013) proposed a method of identifying pool specific testing error rates based on the distributions of the latent biomarker concentration levels of the individuals. Further, these authors demonstrated that proceeding under the traditional assumptions, when they are invalid, may lead to severely biased estimation. It is important to note that the methodology outlined in McMahan et al. (2013) was developed solely for the regression analysis of testing responses obtained from pools; i.e., it does not allow for the incorporation of decoding/retesting information.

With increasing health care costs, many agencies, such as those previously mentioned, have adopted group testing for the dual purposes of estimation and classification. These organizations are therefore privy to group testing data that includes decoding information, and thus, could greatly benefit from the methodological development of binary regression techniques that can incorporate the same. Consequently in this chapter, we propose a general binary regression model that allows for the incorporation of information that may arise from all variants of group testing schemes, to include decoding algorithms. To appropriately account for assay measurement error we extend the methodology presented in McMahan et al. (2013). Through numerical studies we identify settings in which our methods result in more efficient estimates, when compared to those based on individual level testing data. We also illustrate that competing group testing regression methods that proceed under the traditional assumptions may result in severely biased inference.

The remainder of this chapter is organized as follows. In Section 4.2 we propose a general binary regression framework which can handle any type of data that arise from group testing schemes. To account for assay measurement error we then derive explicit expressions that relate the observed testing outcomes to the underlying biomarker concentration levels that are being measured. In Section 4.3 we investigate the finite sample performance of our proposed methodology and compare our approach to existing modeling techniques. To further illustrate the performance of our proposed procedure, we also apply our regression methodology to hepatitis B data in Section 4.4. We conclude with a summary discussion in Section 4.5.

4.2 General Notation and Methodology

We consider the situation in which group testing is to be implemented for the purposes of screening N individuals for a binary characteristic of interest, such as infection status. In general, this process begins by collecting specimens (e.g., blood, urine, etc.) from individuals and assigning each of these specimens to exactly one of J groups of size n_j, for $j = 1, ..., J$. Within each group, these specimens are then screened according to a group testing strategy; e.g, Dorfman decoding, halving (Litvak et al., 1994), or array testing (Phatarfod & Sudbury, 1994). Depending on the goal of the study, this process may necessitate that a given individual be involved in several testing outcomes. For example, in the classification problem the testing of pooled and/or individual specimens continues until each subject can be diagnosed as either positive or negative.

To formalize our notation in this context, we begin by defining $\mathcal{G}_j = \{1, ..., n_j\}$ to be the collection of indices corresponding to the specimens assigned to the jth group, such that for each of the K_j observed

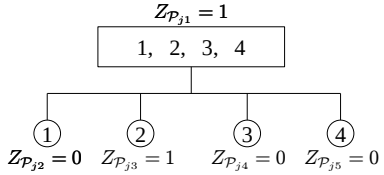
testing responses associated with this group we may identify the individuals involved by $\mathcal{P}_{jl} \subseteq \mathcal{G}_j$, for $l = 1, ..., K_j$, where we use l as a testing index. More explicitly, $\mathcal{P}_{jl}$ corresponds to the individuals in the jth group whose specimens were pooled and assayed by the lth test. We assume that under the selected group testing scheme, that each individual in the jth group should be tested at least once (i.e., $\cup_{l=1}^{K_j} \mathcal{P}_{jl} = \mathcal{G}_j$) and that pooling specimens across groups does not occur. On the other hand, we allow for the situation in which a specimen may belong to multiple pools within a given group (i.e., we do not require $\mathcal{P}_{jl} \cap \mathcal{P}_{jl'} = \emptyset$ for all l and l') and we do not restrict attention to schemes that begin with master pool testing (i.e., we do not mandate that $\mathcal{G}_j \in \{\mathcal{P}_{j1}, ..., \mathcal{P}_{jK_j}\}$). Let $Z_{\mathcal{P}_{jl}}$ denote the binary response observed from assaying the pool formed from amalgamating the $\mathcal{P}_{jl}$ individual specimens, such that $Z_{\mathcal{P}_{jl}} = 1$ indicates that the pool tested positive, $Z_{\mathcal{P}_{jl}} = 0$ otherwise. We collect all of the observed testing outcomes associated with the jth group into the binary vector $\boldsymbol{Z}_j = (Z_{\mathcal{P}_{j1}}, \ldots, Z_{\mathcal{P}_{jK_j}})^{\mathrm{T}}$. Consequently, $\boldsymbol{Z}_j$ is a correlated binary vector that cannot be divided into two independent sub-vectors (otherwise, one could treat this group as two separate groups). Additionally, we assume that $\boldsymbol{Z}_j$ and $\boldsymbol{Z}_{j'}$ are independent, for all $j \neq j'$, which we believe to be reasonable because we do not allow for pooling specimens across groups. For the purpose of clarity, Table 4.1 provides three simple examples to illustrate the use and flexibility of our set notation.

Let $T_{ij} = 1$ denote that the ith individual in the jth group is truly positive, $T_{ij} = 0$ otherwise, for $i = 1, ..., n_j$ and $j = 1, ..., J$. We assume throughout that the statuses T_{ij} are independent random variables. For notational convenience, we collect all of the statuses associated with the jth group into the binary vector $\boldsymbol{T}_j = (T_{1j}, \ldots, T_{n_j j})^{\mathrm{T}}$. It is important to note that when the assay being used is imperfect T_{ij} is unobservable, even under individual testing. For modeling purposes, we assume that the infection probability for the ith individual in the jth group is related to the linear predictor $\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}$ through a monotone and differentiable link function $\eta(\cdot)$, where $\boldsymbol{x}_{ij} = (1, x_{ij1}, \ldots, x_{ijp})^{\mathrm{T}}$ is a $(p+1)$-dimensional vector of covariates and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is the corresponding vector of regression parameters; i.e., $\mathrm{pr}(T_{ij} = 1|\boldsymbol{x}_{ij}) = \eta^{-1}(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})$. We also assume that conditional on the true statuses of the individuals, the observed testing responses are independent of all measured covariates.

We again emphasize that in the presence of diagnostic testing error, observing $\boldsymbol{Z}_j$ is not equivalent to observing $\boldsymbol{T}_j$, even if the goal of the testing algorithm is to classify each individual as either positive or negative. Further, it is well known that ignoring these discrepancies when performing inference may lead to severely biased estimation; e.g., see McMahan et al. (2013). Hence, it is necessary for experimenters to incorporate the effect of imperfect testing into the modeling process. To accomplish this task, we first let $\mathcal{T}_j$ denote the collection of all possible outcomes of $\boldsymbol{T}_j$. Then for any $\boldsymbol{t}_j \in \mathcal{T}_j$ and for any possible realization

Table 4.1: Illustration of Notation

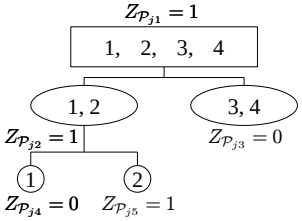
Example 1: Dorfman testing



Stage 1: Specimens collected from subjects $\mathcal{G}_j = \{1, 2, 3, 4\}$ are assigned to a master pool, which tests positive. The testing response is denoted by $Z_{\mathcal{P}_{j1}} = 1$, where $\mathcal{P}_{j1} = \{1, 2, 3, 4\}$.

Stage 2: Dorfman retesting reverts to individual testing, resulting in the four additional testing responses;
$Z_{\mathcal{P}_{j2}} = 0$, where $\mathcal{P}_{j2} = \{1\}$,
$Z_{\mathcal{P}_{j3}} = 1$, where $\mathcal{P}_{j3} = \{2\}$,
$Z_{\mathcal{P}_{j4}} = 0$, where $\mathcal{P}_{j4} = \{3\}$,
$Z_{\mathcal{P}_{j5}} = 0$, where $\mathcal{P}_{j5} = \{4\}$.

Example 2: Three-stage halving

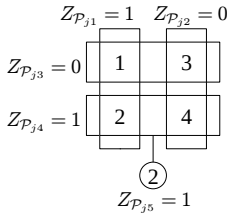


Stage 1: Exactly the same as **Stage 1** under Dorfman testing.

Stage 2: Halving then divides the positive master pool into two equally sized sub-pools, which are then tested, resulting in the two additional testing responses;
$Z_{\mathcal{P}_{j2}} = 1$, where $\mathcal{P}_{j2} = \{1, 2\}$,
$Z_{\mathcal{P}_{j3}} = 0$, where $\mathcal{P}_{j3} = \{3, 4\}$.

Stage 3: Of the sub-pools, one tests negative (requiring no further testing) and one tests positive. The positive sub-pool is decoded by retesting each specimen belonging to it individually, resulting in the two additional testing responses; $Z_{\mathcal{P}_{j4}} = 0$, where $\mathcal{P}_{j4} = \{1\}$,
$Z_{\mathcal{P}_{j5}} = 1$, where $\mathcal{P}_{j5} = \{2\}$.

Example 3: Array testing

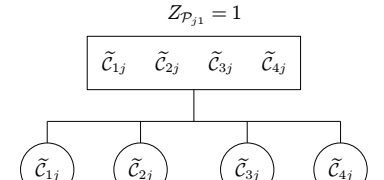


Stage 1: Specimens collected from subjects $\mathcal{G}_j = \{1, 2, 3, 4\}$ are assigned to a 2×2 array, row (column) pools are formed from combining specimens that share a common row (column). Row and column pools are tested, resulting in four testing responses;
$Z_{\mathcal{P}_{j1}} = 1$, where $\mathcal{P}_{j1} = \{1, 2\}$,
$Z_{\mathcal{P}_{j2}} = 0$, where $\mathcal{P}_{j2} = \{3, 4\}$,
$Z_{\mathcal{P}_{j3}} = 0$, where $\mathcal{P}_{j3} = \{1, 3\}$,
$Z_{\mathcal{P}_{j4}} = 1$, where $\mathcal{P}_{j4} = \{2, 4\}$.

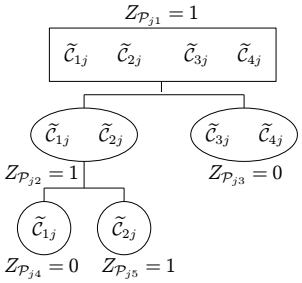
Stage 2: Specimens belonging to the intersection of positive rows and columns are retested individually, resulting in one additional testing response;
$Z_{\mathcal{P}_{j5}} = 1$, where $\mathcal{P}_{j5} = \{2\}$.

Underlying Structure: Relating Testing Outcomes to Latent Biomarker Levels

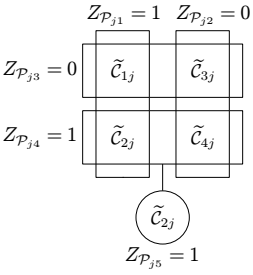
Let specimens 1, 2, 3, and 4 (above) have biomarker levels $\widetilde{\mathcal{C}}_{1j} = 1$, $\widetilde{\mathcal{C}}_{2j} = 7$, $\widetilde{\mathcal{C}}_{3j} = 1$, and $\widetilde{\mathcal{C}}_{4j} = 1$, respectively.
Assume that the test being employed measures biomarker levels without error and that the threshold is $t(c) = 2$, for all c.



Testing responses:
$Z_{\mathcal{P}_{j1}} = I\{4^{-1} \sum_{i \in \mathcal{P}_{j1}} \widetilde{\mathcal{C}}_{ij} > t(4)\} = 1$
$Z_{\mathcal{P}_{j2}} = I\{\widetilde{\mathcal{C}}_{1j} > t_0\} = 0$
$Z_{\mathcal{P}_{j3}} = I\{\widetilde{\mathcal{C}}_{2j} > t_0\} = 1$
$Z_{\mathcal{P}_{j4}} = I\{\widetilde{\mathcal{C}}_{3j} > t_0\} = 0$
$Z_{\mathcal{P}_{j5}} = I\{\widetilde{\mathcal{C}}_{4j} > t_0\} = 0$



Testing responses:
$Z_{\mathcal{P}_{j1}} = I\{4^{-1} \sum_{i \in \mathcal{P}_{j1}} \widetilde{\mathcal{C}}_{ij} > t(4)\} = 1$
$Z_{\mathcal{P}_{j2}} = I\{2^{-1} \sum_{i \in \mathcal{P}_{j2}} \widetilde{\mathcal{C}}_{ij} > t(2)\} = 1$
$Z_{\mathcal{P}_{j3}} = I\{2^{-1} \sum_{i \in \mathcal{P}_{j3}} \widetilde{\mathcal{C}}_{ij} > t(2)\} = 0$
$Z_{\mathcal{P}_{j4}} = I\{\widetilde{\mathcal{C}}_{1j} > t_0\} = 0$
$Z_{\mathcal{P}_{j5}} = I\{\widetilde{\mathcal{C}}_{2j} > t_0\} = 1$



Testing responses:
$Z_{\mathcal{P}_{j1}} = I\{2^{-1} \sum_{i \in \mathcal{P}_{j1}} \widetilde{\mathcal{C}}_{ij} > t(2)\} = 1$
$Z_{\mathcal{P}_{j2}} = I\{2^{-1} \sum_{i \in \mathcal{P}_{j2}} \widetilde{\mathcal{C}}_{ij} > t(2)\} = 0$
$Z_{\mathcal{P}_{j3}} = I\{2^{-1} \sum_{i \in \mathcal{P}_{j3}} \widetilde{\mathcal{C}}_{ij} > t(2)\} = 0$
$Z_{\mathcal{P}_{j4}} = I\{2^{-1} \sum_{i \in \mathcal{P}_{j4}} \widetilde{\mathcal{C}}_{ij} > t(2)\} = 1$
$Z_{\mathcal{P}_{j5}} = I\{\widetilde{\mathcal{C}}_{2j} > t_0\} = 1$

$\boldsymbol{z}_j$ of $\boldsymbol{Z}_j$, we define $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j) = \text{pr}(\boldsymbol{Z}_j = \boldsymbol{z}_j \mid \boldsymbol{T}_j = \boldsymbol{t}_j)$. By an application of the Law of Total Probability, one can relate the observed testing outcomes, given the observed pooling structure, to the individual level covariates as follows

$$\begin{aligned}
\text{pr}(\boldsymbol{Z}_j = \boldsymbol{z}_j | \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_j j}) &= \sum_{\boldsymbol{t}_j \in \mathcal{T}_j} \text{pr}(\boldsymbol{Z}_j = \boldsymbol{z}_j \mid \boldsymbol{T}_j = \boldsymbol{t}_j, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_j j}) \text{pr}(\boldsymbol{T}_j = \boldsymbol{t}_j | \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_j j}) \\
&= \sum_{\boldsymbol{t}_j \in \mathcal{T}_j} \text{pr}(\boldsymbol{Z}_j = \boldsymbol{z}_j \mid \boldsymbol{T}_j = \boldsymbol{t}_j) \prod_{i=1}^{n_j} \text{pr}(\boldsymbol{T}_{ij} = \boldsymbol{t}_{ij} | \boldsymbol{x}_{ij}) \\
&= \sum_{\boldsymbol{t}_j \in \mathcal{T}_j} \left\{ M_j(\boldsymbol{z}_j, \boldsymbol{t}_j) \prod_{i=1}^{n_j} \{ t_{ij} \eta^{-1}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta}) + (1 - t_{ij})[1 - \eta^{-1}(\boldsymbol{x}_{ij}^\top \boldsymbol{\beta})] \} \right\}.
\end{aligned}$$

To emphasize the dependence of the above probability on the unknown regression coefficients, $\boldsymbol{\beta}$, we write

$$\text{pr}(\boldsymbol{Z}_j = \boldsymbol{z}_j | \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_j j}) = \mathcal{R}(\boldsymbol{z}_j, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_j j}, \boldsymbol{\beta}).$$

Using the observed data $\{(\boldsymbol{z}_j, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_j j}), j = 1, ..., J\}$, we can express the observed data log-likelihood as

$$l(\boldsymbol{\beta}) = \sum_{j=1}^{J} \log \mathcal{R}(\boldsymbol{z}_j, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_j j}, \boldsymbol{\beta}). \tag{4.1}$$

If each $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j)$ were known, then one could easily estimate $\boldsymbol{\beta}$ by directly maximizing (4.1) with respect to $\boldsymbol{\beta}$. In the next section we provide details on how to evaluate $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j)$ based on the underlying characteristics of the assay being employed.

4.2.1 Evaluation of Misclassification

To account for imperfect testing and pool dilution effects, we generalize the methodology described in Wein & Zenios (1996), Zenios & Wein (1998), and McMahan et al. (2013). Following the work of these authors, we proceed under the standard convention that a diagnostic test classifies a specimen as positive (negative) if its measured biomarker concentration is above (below) a predetermined threshold. For generality, we allow the assay threshold, which we denote by $t(c)$, for a pool to vary with the number of specimens, say c, of which it is comprised. Typically, the specification of $t(c)$ has proceeded in one of two fashions. In particular, Tu et al. (1994) specified that $t(c) = t_0$ for any pool size c, where t_0 is the assay threshold under individual testing. This approach has also been implemented in the infectious disease screening literature; e.g., see Currie et al. (2004). Alternatively, to account for the effect of pooling Vansteelandt et al. (2005)

specified that $t(c) = t_0/c$. In either case, we first derive a closed-form expression for $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j)$ in terms of the relevant biomarker distributions under an arbitrary thresholding strategy. We then investigate the effect of both of the aforementioned thresholding strategies on inference in Sections 4.3 and 4.4.

To this end, we define $\widetilde{\mathcal{C}}_{ij}$ to be the true biomarker concentration level for the ith individual in the jth group, and we assume that conditional on the individual's true status that $\widetilde{\mathcal{C}}_{ij}|T_{ij} \sim f_{\widetilde{\mathcal{C}}|T_{ij}} = T_{ij} f_{\widetilde{\mathcal{C}}^+} + (1 - T_{ij}) f_{\widetilde{\mathcal{C}}^-}$, where $f_{\widetilde{\mathcal{C}}^+}$ and $f_{\widetilde{\mathcal{C}}^-}$ are the probability density functions for the biomarker concentration levels of the infected and uninfected individuals, respectively. We initially assume that these biomarker distributions are known, this assumption is later relaxed in Section 4.4. For notational convenience, we define $\widetilde{\boldsymbol{\mathcal{C}}}_j = (\widetilde{\mathcal{C}}_{1j}, \ldots, \widetilde{\mathcal{C}}_{n_j j})^{\mathrm{T}}$, for each j, to be the collection of the true biomarker levels for the n_j individuals assigned to the jth group. To account for the underlying structure of the assay being employed we are left to relate $\widetilde{\boldsymbol{\mathcal{C}}}_j$ to the testing outcomes $\boldsymbol{Z}_j$.

When pooled assessments are being made, we assume that the true biomarker concentration of the pool is the arithmetic average of the biomarker concentrations of the individual specimens contributing to the pool; i.e., letting $\widetilde{\mathcal{C}}_{\mathcal{P}_{jl}}$ denote the biomarker concentration for the pool consisting of the $\mathcal{P}_{jl}$ individuals we assume that $\widetilde{\mathcal{C}}_{\mathcal{P}_{jl}} = c_{jl}^{-1} \sum_{i \in \mathcal{P}_{jl}} \widetilde{\mathcal{C}}_{ij}$, where c_{jl} denotes the cardinality of the set $\mathcal{P}_{jl}$. We view this assumption to be reasonable, as long as the individual specimens being pooled are of equal volume. Additionally, this assumption is common among the biomarker pooling literature (Liu & Schisterman, 2003; Liu et al., 2004; Mumford et al., 2006; Bondell et al., 2007; Vexler et al., 2008) and has previously been assumed in the group testing estimation literature (Wein & Zenios, 1996; Zenios & Wein, 1998; McMahan et al., 2013). To simplify this relationship, we define the design vector associated with the test of the pool consisting of the $\mathcal{P}_{jl}$ individuals to be $\boldsymbol{D}_{\mathcal{P}_{jl}} = c_{jl}^{-1} \mathbf{1}_{\mathcal{P}_{jl}}$, where $\mathbf{1}_{\mathcal{P}_{jl}}$ is a n_j-dimensional binary vector whose $\mathcal{P}_{jl}$th components are 1, and all others are 0. Using this notation we can express the pool biomarker concentration levels as $\widetilde{\mathcal{C}}_{\mathcal{P}_{jl}} = \boldsymbol{D}_{\mathcal{P}_{jl}}^{\mathrm{T}} \widetilde{\boldsymbol{\mathcal{C}}}_j$. It is important to point out, that in the presence of measurement error each $\widetilde{\mathcal{C}}_{\mathcal{P}_{jl}}$ is unobservable.

We now derive our expression for $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j)$ in terms of the aforementioned biomarker distributions. To account for assay measurement error, we let $\mathcal{C}_{\mathcal{P}_{jl}}$ denote the error laden measurement of $\widetilde{\mathcal{C}}_{\mathcal{P}_{jl}}$, and we assume that conditional on the true biomarker concentration levels that $\mathcal{C}_{\mathcal{P}_{jl}} \overset{ind}{\sim} f_{\mathcal{C}|\widetilde{\mathcal{C}}_{\mathcal{P}_{jl}}}$, for $l = 1, ..., K_j$ and $j = 1, ..., J$. Thus, the observed testing responses, under our classification rule, are given by $Z_{\mathcal{P}_{jl}} = I\{\mathcal{C}_{\mathcal{P}_{jl}} > t(c_{jl})\}$. For purposes of clarity, we provide a simple illustration of how testing responses are derived in this context in Table 4.1. Noting this relationship, we are able to write the probabilities associated

with the observed testing outcomes in terms of the measured biomarker concentrations; e.g.,

$$\text{pr}(Z_{\mathcal{P}_{jl}} = z_{\mathcal{P}_{jl}}) = \text{pr}\{\mathcal{C}_{\mathcal{P}_{jl}} \in A(z_{\mathcal{P}_{jl}}, c_{jl})\},$$

where $A(z,c) = z \cdot \{u : u > t(c)\} + (1-z) \cdot \{u : u \leq t(c)\}$ and $z \in \{0,1\}$; i.e., $A(0,c) = \{u : u \leq t(c)\}$ and $A(1,c) = \{u : u > t(c)\}$. Using this relationship we can express $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j)$ as follows

$$\begin{aligned} M_j(\boldsymbol{z}_j, \boldsymbol{t}_j) &= \text{pr}(Z_{\mathcal{P}_{j1}} = z_{\mathcal{P}_{j1}}, \dot{;} Z_{\mathcal{P}_{jK_j}} = z_{\mathcal{P}_{jK_j}} \mid \boldsymbol{T}_j = \boldsymbol{t}_j) \\ &= \text{pr}\{\mathcal{C}_{\mathcal{P}_{j1}} \in A(z_{\mathcal{P}_{j1}}, c_{j1}), ..., \mathcal{C}_{\mathcal{P}_{jK_j}} \in A(z_{\mathcal{P}_{jK_j}}, c_{jK_j}) \mid \boldsymbol{T}_j = \boldsymbol{t}_j\} \\ &= \text{pr}\{\boldsymbol{\mathcal{C}}_j \in \boldsymbol{A}(\boldsymbol{z}_j, \mathbf{c}_j) \mid \boldsymbol{T}_j = \boldsymbol{t}_j\}, \end{aligned}$$

where $\boldsymbol{\mathcal{C}}_j = (\mathcal{C}_{\mathcal{P}_{j1}}, ..., \mathcal{C}_{\mathcal{P}_{jK_j}})^{\text{T}}$, $\boldsymbol{A}(\boldsymbol{z}_j, \mathbf{c}_j) = A(z_{\mathcal{P}_{j1}}, c_{j1}) \times A(z_{\mathcal{P}_{j2}}, c_{j2}) \times \cdots \times A(z_{\mathcal{P}_{jK_j}}, c_{jK_j})$, and $\mathbf{c}_j = (c_{j1}, ..., c_{jK_j})^{\text{T}}$. Based on the probability density functions $f_{\widetilde{\mathcal{C}}|T_{ij}}$ and $f_{\mathcal{C}|\widetilde{\mathcal{C}}_{\mathcal{P}_{jl}}}$, we have that the conditional probability density function of $\boldsymbol{\mathcal{C}}_j$ given $\boldsymbol{T}_j = \boldsymbol{t}_j$ is

$$f_{\boldsymbol{\mathcal{C}}_j|\boldsymbol{T}_j=\boldsymbol{t}_j}(\boldsymbol{u}) = \int \prod_{l=1}^{K_j} f_{\mathcal{C}|\widetilde{\mathcal{C}}_{\mathcal{P}_{jl}}=\boldsymbol{D}^{\text{T}}_{\mathcal{P}_{jl}}\boldsymbol{y}}(u_l) \prod_{i=1}^{n_j} f_{\widetilde{\mathcal{C}}|T_{ij}=t_{ij}}(y_{ij}) d\boldsymbol{y}, \tag{4.2}$$

where $\boldsymbol{u} = (u_1, \ldots, u_{K_j})^{\text{T}}$ and $\boldsymbol{y} = (y_{1j}, \ldots, y_{n_j j})^{\text{T}}$. Finally,

$$M_j(\boldsymbol{z}_j, \boldsymbol{t}_j) = \int_{\boldsymbol{A}(\boldsymbol{z}_j, \mathbf{c}_j)} f_{\boldsymbol{\mathcal{C}}_j|\boldsymbol{T}_j=\boldsymbol{t}_j}(\boldsymbol{u}) d\boldsymbol{u}. \tag{4.3}$$

When no retesting is performed, the above expression is equivalent to the results presented in McMahan et al. (2013) for evaluating the assay sensitivity and specificity associated with master pool testing.

One should note, that the integral in (4.2) is multidimensional if $n_j > 1$, and so is the integral in (4.3) if the individuals in the jth group are involved in more than one test (i.e., $\boldsymbol{Z}_j$ is non-scalar). In general these integrals may be difficult to evaluate analytically, but this challenge is easily overcome using Monte Carlo techniques, as will be illustrated in Section 4.4. It is possible to obtain a closed form expression of $f_{\boldsymbol{\mathcal{C}}_j|\boldsymbol{T}_j=\boldsymbol{t}_j}$, if we assume that $\widetilde{\mathcal{C}}_{ij}|T_{ij} = 1 \sim N(\mu_+, \sigma^2_+)$, $\widetilde{\mathcal{C}}_{ij}|T_{ij} = 0 \sim N(\mu_-, \sigma^2_-)$, and $\mathcal{C}|\widetilde{\mathcal{C}} \sim N(\widetilde{\mathcal{C}}, \tau^2)$. Although a special case, these distributional assumptions are common among the pooled biomarker literature (e.g., see Faraggi et al., 2003; Liu & Schisterman, 2003; Liu et al., 2004; Mumford et al., 2006). Under the assumption of normality, we have that $\widetilde{\boldsymbol{\mathcal{C}}}_j|\boldsymbol{T}_j = \boldsymbol{t}_j \sim N(\boldsymbol{\mu}(\boldsymbol{t}_j), \boldsymbol{\Sigma}(\boldsymbol{t}_j))$, where $\boldsymbol{\mu}(\boldsymbol{t}_j) = (\mathbf{1} - \boldsymbol{t}_j)\mu_- + \boldsymbol{t}_j\mu_+$ and $\boldsymbol{\Sigma}(\boldsymbol{t}_j) = \sigma^2_-\text{diag}\{\mathbf{1} - \boldsymbol{t}_j\} + \sigma^2_+\text{diag}\{\boldsymbol{t}_j\}$. Here, for a k-dimensional vector $\boldsymbol{a}$, we let $\text{diag}\{\boldsymbol{a}\}$ denote a

$k \times k$ diagonal matrix whose diagonal elements are $\boldsymbol{a}$. We define the matrix $\boldsymbol{D}_j = (\boldsymbol{D}_{\mathcal{P}_{j1}}, \ldots, \boldsymbol{D}_{\mathcal{P}_{jK_j}})$, so that the vector of true concentration levels of the pools associated with $\boldsymbol{Z}_j$ can be expressed as $\boldsymbol{D}_j^{\mathrm{T}}\widetilde{\boldsymbol{\mathcal{C}}}_j$. Noting that $\boldsymbol{D}_j^{\mathrm{T}}\widetilde{\boldsymbol{\mathcal{C}}}_j|\boldsymbol{T}_j = \boldsymbol{t}_j \sim N(\boldsymbol{D}_j^{\mathrm{T}}\boldsymbol{\mu}(\boldsymbol{t}_j), \boldsymbol{D}_j^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{t}_j)\boldsymbol{D}_j)$, it is easy to show that

$$\boldsymbol{\mathcal{C}}_j|\boldsymbol{T}_j = \boldsymbol{t}_j \sim N\left(\boldsymbol{D}_j^{\mathrm{T}}\boldsymbol{\mu}(\boldsymbol{t}_j), \boldsymbol{D}_j^{\mathrm{T}}\boldsymbol{\Sigma}(\boldsymbol{t}_j)\boldsymbol{D}_j + \tau^2\boldsymbol{I}_{K_j}\right),$$

where $\boldsymbol{I}_{K_j}$ is a $K_j \times K_j$ identity matrix. Thus, under this special case, it is easy to calculate $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j)$ using standard statistical software.

4.2.2 Maximum Likelihood Approach

Using the observed data $\{(\boldsymbol{z}_j, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_jj}), j = 1, ..., J\}$, one can estimate $\boldsymbol{\beta}$ by maximizing (4.1) directly after using (4.3) to evaluate $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j)$, for all $\boldsymbol{t}_j \in \mathcal{T}_j$. We denote the resulting maximum likelihood estimator (MLE) as $\widehat{\boldsymbol{\beta}}$. The standard theoretical properties for MLEs hold for $\widehat{\boldsymbol{\beta}}$ under the assumption that the group sizes remain finite so that $J \to \infty$ as $N \to \infty$. This assumption is common among the group testing literature, and we view it to be reasonable because in practice pool sizes are typically bounded by implementation considerations. The variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ can be estimated by the negative inverse Hessian of (4.1) evaluated at $\widehat{\boldsymbol{\beta}}$, and can be used to conduct Wald type inference (Wald , 1943).

One may note, that the evaluation of $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j)$ over all $\boldsymbol{t}_j \in \mathcal{T}_j$ could pose a significant computational burden, especially if n_j is large. To obviate this difficulty, we point out that $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j)$ is free of $\boldsymbol{\beta}$ and can therefore be calculated before numerical optimization routines are implemented. To further alleviate this computational burden we have developed efficient algorithms for computing these terms under two of the most popular group testing schemes: Dorfman decoding and three-stage halving; these algorithms are provided in Appendix C.1. In conjunction with the aforementioned algorithms, we have had little difficulty implementing a quasi-Newton optimization routine in R for the purposes of identifying the MLE. Depending on the complexity of the group testing strategy, it may not be feasible to directly maximize the observed data log-likelihood using numerical techniques. In these situations, our methodology can still be implemented through the use of an expectation maximization (EM) algorithm, which we also provide in Appendix C.2.

4.3 Numerical Analysis

In this section, we illustrate the performance of our proposed methodology through simulation, and compare our results to those obtained from more traditional group testing regression techniques. These traditional methods generally proceed under the assumption that the testing error rates, sensitivity (S_e) and specificity (S_p), are known, constant, and do not depend on the pool size. More explicitly, the testing error rates are the same for all pool sizes, to include individual level testing. The sensitivity (specificity) of an assay is typically defined to be the probability that the assay will classify a specimen as positive (negative) given it is truly positive (negative). To incorporate retesting information, authors have made the further assumption that the testing outcomes for pools (individuals) are independent given their true statuses. Under these assumptions, the conditional probability of observing $\boldsymbol{z}_j$, given the individuals' true latent statuses $\boldsymbol{t}_j$, can be expressed as

$$M_j(\boldsymbol{z}_j, \boldsymbol{t}_j) = \prod_{l=1}^{K_j} \left\{ S_e^{z_{\mathcal{P}_{jl}} \widetilde{z}_{\mathcal{P}_{jl}}} (1-S_e)^{(1-z_{\mathcal{P}_{jl}})\widetilde{z}_{\mathcal{P}_{jl}}} (1-S_p)^{z_{\mathcal{P}_{jl}}(1-\widetilde{z}_{\mathcal{P}_{jl}})} S_p^{(1-z_{\mathcal{P}_{jl}})(1-\widetilde{z}_{\mathcal{P}_{jl}})} \right\}, \tag{4.4}$$

where $\widetilde{z}_{\mathcal{P}_{jl}} = I\{\sum_{i \in \mathcal{P}_{jl}} t_{ij} > 0\}$ is the true status of the pool being tested. Substituting the above expression into (4.1) and maximizing directly results in obtaining an estimate of $\boldsymbol{\beta}$, under these more traditional assumptions.

4.3.1 Data Generation and Model Fitting

In this study, we consider the following models:

Model 4.1. $logit\{pr(T_{ij} = 1 \mid x_{ij1})\} = \beta_0 + \beta_1 x_{ij1}$; $\boldsymbol{\beta} = (\beta_0, \beta_1)^{\mathrm{T}} = (-3, 2)^{\mathrm{T}}$,

Model 4.2. $logit\{pr(T_{ij} = 1 \mid x_{ij1})\} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij1}^2$; $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}} = (-3, 1, 0.5)^{\mathrm{T}}$,

Model 4.3. $logit\{pr(T_{ij} = 1 \mid x_{ij1}, x_{ij2})\} = \beta_0 + \beta_1 x_{ij1} + \beta_2 x_{ij2}$; $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}} = (-3, 2, 1)^{\mathrm{T}}$,

where $x_{ij1} \sim \mathrm{N}(0, 0.75^2)$ and $x_{ij2} \sim \mathrm{Bernoulli}(0.1)$. These model choices emulate situations in which group testing could be employed and provide for mean prevalences ranging from 8–10 percent. These models were also studied in McMahan et al. (2013). Normal distributions were chosen for the individual biomarker concentrations; specifically, $\widetilde{\mathcal{C}}|T = 1 \sim N(2, \sigma_+^2)$ and $\widetilde{\mathcal{C}}|T = 0 \sim N(0.1, 0.3^2)$, where $\sigma_+ \in \{0.8, 0.9, 1\}$. To account for assay measurement error, we specified the conditional distribution of the measured concentration levels to be $\mathcal{C}|\widetilde{\mathcal{C}} \sim N(\widetilde{\mathcal{C}}, 0.02^2)$. The assay threshold was chosen to be $t_0 = 0.7$, so that the specificity

under individual testing would be $S_p = 0.977$, while the sensitivities would be $S_e = 0.948, 0.926, 0.903$ corresponding to $\sigma_+ = 0.8, 0.9, 1$, respectively.

In this study, we specified $N = 3600$ and considered a common initial group size of n; i.e., $n_j = n$, for all $j = 1, \ldots, J$, where $n \in \{2, 4, 6\}$ and $J = N/n$. We then randomly generated the individual level covariates $\boldsymbol{x}_{ij}$, for $i = 1, \ldots, n$ and $j = 1, \ldots, J$, which were then used to calculate the individuals' infection probabilities, p_{ij}, according to the aforementioned models. It is worth while to point out, that randomly simulating covariate values in this fashion, and subsequently the infection probabilities, is equivalent to practitioners randomly assigning subjects to groups. Each individual's true status, T_{ij}, was then determined according to a Bernoulli distribution having success probability p_{ij}. The corresponding biomarker concentration level, $\widetilde{\mathcal{C}}_{ij}$, was then independently generated according to either $\widetilde{\mathcal{C}}_{ij}|T_{ij} = 1 \sim N(2, \sigma_+^2)$ or $\widetilde{\mathcal{C}}_{ij}|T_{ij} = 0 \sim N(0.1, 0.03^2)$, depending on the value of T_{ij}. This process was repeated 500 times for each model and configuration of (σ_+, n) resulting in $27,000$ independent data sets of the form $\{(\widetilde{\mathcal{C}}_{1j}, ..., \widetilde{\mathcal{C}}_{nj}, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{nj}), j = 1, ..., J\}$.

In what follows, we describe how group testing data was generated based on the individual level biomarker data. As one might expect, the data structure is highly dependent on the particular group testing strategy being employed. For the purposes of our simulation, presented herein, we have opted to investigate three of the most common strategies: master pool testing (MT), Dorfman testing (DT), and three-stage halving (TH). In order to levy pool diagnoses, we considered two methods of specifying the assay threshold for pooled specimens; i.e., we considered letting $t(c) = t_0$ and $t(c) = t_0/c$ as was suggested in Currie et al. (2004) and (Vansteelandt et al., 2005), respectively. Under MT, all individuals within a given group are pooled together and the pool is tested, with no further testing being implemented. Thus, the testing response vector, under MT, for the jth group is given by $\boldsymbol{Z}_j^{MT} = Z_{\mathcal{P}_{j1}}$, where $\mathcal{P}_{j1} = \mathcal{G}_j = \{1, ..., n\}$, and is determined by $Z_{\mathcal{P}_{j1}} = I\{\mathcal{C}_{\mathcal{P}_{j1}} > t(n)\}$, where $\mathcal{C}_{\mathcal{P}_{j1}} \sim N(\widetilde{\mathcal{C}}_{\mathcal{P}_{j1}}, 0.02^2)$ and $\widetilde{\mathcal{C}}_{\mathcal{P}_{j1}} = n^{-1} \sum_{i=1}^{n} \widetilde{\mathcal{C}}_{ij}$. Master pool testing, unlike DT and TH, is not a decoding algorithm; i.e., it does not levy a diagnosis for each individual.

The first decoding algorithm that we consider is DT, which specifies that a group whose master pool test is negative requires no further screening, but in those cases where the master pool test is positive the group is resolved by retesting each subject individually. Therefore, under DT the testing response vector for the jth group is identical to that under MT, if the master pool test is negative. Alternatively, if the master pool test is positive the testing response vector is given by $\boldsymbol{Z}_j^{DT} = (Z_{\mathcal{P}_{j1}}, Z_{\mathcal{P}_{j2}}, ..., Z_{\mathcal{P}_{jK_j}})^{\mathrm{T}}$, where $Z_{\mathcal{P}_{j1}}$ is determined as discussed above, $K_j = n + 1$, and $\mathcal{P}_{jl} = \{l - 1\}$, for $l = 2, ..., K_j$. The response $Z_{\mathcal{P}_{jl}}$, for $l = 2, ..., K_j$, corresponds to individually testing the $(l-1)$th subject and is determined according to

$Z_{\mathcal{P}_{jl}} = I\{\mathcal{C}_{\mathcal{P}_{jl}} > t_0\}$, where $\mathcal{C}_{\mathcal{P}_{jl}} \sim N(\widetilde{\mathcal{C}}_{ij}, 0.02^2)$ and $i = l - 1$.

The second decoding algorithm we consider is TH, which is very similar to DT with the exception of an additional decoding stage before reverting to individual testing. Under TH when a positive master pool response is observed the positive group is randomly divided into two equally sized subgroups and these subgroups are tested using the threshold $t(n/2)$. If a subgroup tests negative then testing is complete, alternatively if a subgroup tests positive then all contributing subjects are retested individually. For brevity, we have chosen not to explicitly describe the construction of the response vector for TH, but it follows a similar methodology as that described above. Using these group testing strategies, we are able to create the following group testing data $\{(\boldsymbol{z}_j^{MT}, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{nj}), j = 1, ..., J\}$, $\{(\boldsymbol{z}_j^{DT}, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{nj}), j = 1, ..., J\}$, and $\{(\boldsymbol{z}_j^{TH}, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{nj}), j = 1, ..., J\}$ corresponding to MT, DT, and TH, respectively, for each of the biomarker data sets, $\{(\widetilde{\mathcal{C}}_{1j}, ..., \widetilde{\mathcal{C}}_{nj}, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{nj}), j = 1, ..., J\}$.

The regression methodology discussed in Section 4.2 was applied to each of the group testing data sets. To implement these techniques, we calculated $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j)$ using the closed form expression presented in Section 4.2.1, that is available under the assumption of normality. In Appendix C.3, we illustrate how one could approximate these quantities using Monte Carlo techniques, when the assumption of normality is not valid. We then maximized (4.1) directly using a Quasi-Newton optimization routine available in R. For the purposes of comparison, we also fit the regression models that proceed under the more traditional assumptions using the reformulation presented in (4.4) and the appropriate individual S_e and S_p levels. Additionally, for each of the biomarker data sets we also generated subject level testing responses (i.e., $n = 1$) and fit the individual data model.

4.3.2 Simulation Results

Table C.3 provides summary statistics of the 500 estimates of $\boldsymbol{\beta}$ obtained from Model 4.2 for all considered group testing algorithms under the two thresholding strategies, when $n \in \{2, 4, 6\}$ and $\sigma_+ = 1$. From these results, we see that the maximum likelihood estimates of $\boldsymbol{\beta}$ obtained from our regression methodology show little, if any, evidence of bias, across all considered configurations. The same cannot always be said for the estimates obtained from the traditional group testing regression models. More specifically, the parameter estimates obtained by the two different regression methodologies almost agree at $n = 2$, but as n increases the estimates obtained by the traditional methods tend to become more biased, and in some cases severely so (e.g., when $n = 6$). It is worth while to point out that the bias in the parameter estimates obtained by the traditional regression techniques are less pronounced when the assay threshold is allowed to vary with

the size of the pool being tested (i.e., $t(c) = t_0/c$). These findings, likely explain why the estimated coverage probabilities for the traditional modeling approach tend to be incongruously small given the specified confidence level, while those associated with our techniques remain at their nominal level, regardless of the group size. One will also notice, that the standard deviation of the estimated regression coefficients, from our proposed method, are predominantly in agreement with the corresponding average standard errors, suggesting that the variance-covariance matrix of $\widehat{\boldsymbol{\beta}}$ is being estimated correctly. One will also note that as n increases, so does the variability in $\boldsymbol{\beta}$, this is an expected phenomenon because the number of testing responses, used to estimate $\boldsymbol{\beta}$, decrease as n increases. However, this effect is attenuated for the data collected by the decoding algorithms (DT and TH), which is explained by the addition of the retesting information associated with decoding positive pools. Figure 4.1 provides plots of the average estimated regression functions for Model 4.2 under both thresholding strategies, when $n \in \{2, 4, 6\}$, $\sigma_+ = 1$, and for all considered group testing algorithms. This figure reinforces the main findings discussed above; i.e., the regression curves estimated by our methods are on target and tend to capture the true underlying model. Alternatively, the regression curves estimated under the traditional assumptions exhibit a great deal of bias, especially for larger group sizes.

In Figure 4.2 we provide evidence that our proposed methodology for analyzing group testing data may result in estimates of $\boldsymbol{\beta}$ that are less variable, when compared to the parameter estimates resulting from the individual data model (i.e., $n = 1$). Further, these estimates can be obtained at a fraction of the data collection cost incurred by testing individuals one-by-one. Specifically, in this figure we provide plots of the percentage reduction in testing cost obtained through the use of DT and TH for screening when compared to testing individuals separately, across all considered configurations when $\sigma_+ = 1$. We also provide the relative efficiency, which we define to be the ratio between the MSE of the estimates obtained from analyzing group testing data and the MSE of the estimates resulting from the individual data model. These results suggest that if the group size n is sensibly chosen, then the estimates obtained from data collected by a group testing decoding algorithm can be more efficient (less variable) than those obtained from individual level data, and at roughly 65% and 80% of the cost of testing under the threshold strategies $t(c) = t_0$ and $t(c) = t_0/c$, respectively.

In Appendix C.4, we provide a complete summary of our simulation results across all considered group testing strategies and values of n, when $\sigma_+ = 1$. The results under other considered settings of σ_+ were practically identical and are therefore omitted. In addition to the numerical study discussed above, we have also performed simulations that allow for different biomarker distributional assumptions (e.g., gamma, Weibull, and log-normal), the evaluation of $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j)$ through Monte Carlo techniques, and the use of dif-

Table 4.2: Simulation results for Model 4.2 having regression parameters $\boldsymbol{\beta} = (-3, 1, 0.5)^{\mathrm{T}}$. Presented results include the sample mean (Mean) and standard deviation (SD) of the 500 estimates of $\boldsymbol{\beta}$, when $n \in \{2, 4, 6\}$ and $\sigma_+ = 1$. The average standard error (SE) and estimated 95% Wald coverage probabilities (Cov) are also provided. Assuming a 99% confidence level for the coverage probabilities, the margin of error is 0.03. Estimates outside this margin of error are shown in bold. Note, MT, DT, and TH denote individual testing, master pool testing, Dorfman testing, and three-stage halving, respectively.

When $t(c) = t_0$:			Acknowledging the Dilution Effect			Traditional Approach		
	n	Measure	MT	DT	TH	MT	DT	TH
$\widehat{\beta}_0$	2	Mean(Cov)	$-3.01(0.95)$	$-3.01(0.94)$	$--(--)$	$-3.52(\mathbf{0.05})$	$-3.15(\mathbf{0.72})$	$--(--)$
		SD(SE)	$0.12(0.12)$	$0.11(0.11)$	$--(--)$	$0.17(0.16)$	$0.10(0.10)$	$--(--)$
	4	Mean(Cov)	$-3.03(0.96)$	$-3.02(0.94)$	$-3.01(0.96)$	$-4.49(\mathbf{0.00})$	$-3.83(\mathbf{0.00})$	$-3.75(\mathbf{0.00})$
		SD(SE)	$0.20(0.19)$	$0.14(0.14)$	$0.14(0.14)$	$0.43(0.43)$	$0.16(0.15)$	$0.14(0.13)$
	6	Mean(Cov)	$-3.11(0.96)$	$-3.03(0.95)$	$-3.02(0.95)$	$-7.56(\mathbf{0.14})$	$-4.93(\mathbf{0.00})$	$-4.54(\mathbf{0.00})$
		SD(SE)	$0.34(0.33)$	$0.19(0.19)$	$0.19(0.19)$	$3.90(3.10)$	$0.45(0.36)$	$0.24(0.21)$
$\widehat{\beta}_1$	2	Mean(Cov)	$1.03(0.94)$	$1.02(0.95)$	$--(--)$	$1.38(\mathbf{0.87})$	$0.99(\mathbf{0.91})$	$--(--)$
		SD(SE)	$0.19(0.17)$	$0.12(0.12)$	$--(--)$	$0.40(0.35)$	$0.12(0.11)$	$--(--)$
	4	Mean(Cov)	$1.09(0.94)$	$1.03(0.94)$	$1.03(0.96)$	$1.97(\mathbf{0.88})$	$1.10(\mathbf{0.91})$	$0.99(\mathbf{0.92})$
		SD(SE)	$0.41(0.39)$	$0.18(0.17)$	$0.15(0.16)$	$0.93(0.89)$	$0.26(0.22)$	$0.18(0.17)$
	6	Mean(Cov)	$1.25(0.95)$	$1.05(0.97)$	$1.05(0.96)$	$5.24(0.94)$	$1.85(\mathbf{0.79})$	$1.17(0.94)$
		SD(SE)	$0.81(0.75)$	$0.26(0.24)$	$0.23(0.23)$	$5.68(4.51)$	$0.88(0.67)$	$0.40(0.33)$
$\widehat{\beta}_2$	2	Mean(Cov)	$0.48(0.95)$	$0.49(0.95)$	$--(--)$	$0.32(\mathbf{0.86})$	$0.45(0.94)$	$--(--)$
		SD(SE)	$0.14(0.13)$	$0.10(0.10)$	$--(--)$	$0.24(0.20)$	$0.09(0.09)$	$--(--)$
	4	Mean(Cov)	$0.44(0.96)$	$0.48(0.97)$	$0.49(0.95)$	$-0.08(\mathbf{0.72})$	$0.29(\mathbf{0.72})$	$0.32(\mathbf{0.69})$
		SD(SE)	$0.27(0.28)$	$0.14(0.14)$	$0.14(0.14)$	$0.42(0.40)$	$0.16(0.14)$	$0.14(0.12)$
	6	Mean(Cov)	$0.37(0.93)$	$0.47(0.95)$	$0.48(0.93)$	$-1.21(\mathbf{0.88})$	$-0.04(\mathbf{0.60})$	$0.19(\mathbf{0.66})$
		SD(SE)	$0.52(0.47)$	$0.21(0.20)$	$0.21(0.20)$	$2.05(1.64)$	$0.40(0.31)$	$0.23(0.19)$
When $t(c) = t_0/c$:			Acknowledging the Dilution Effect			Traditional Approach		
	n	Measure	MT	DT	TH	MT	DT	TH
$\widehat{\beta}_0$	2	Mean(Cov)	$-3.04(0.96)$	$-3.01(0.95)$	$--(--)$	$-2.22(\mathbf{0.00})$	$-2.46(\mathbf{0.00})$	$--(--)$
		SD(SE)	$0.17(0.17)$	$0.11(0.11)$	$--(--)$	$0.08(0.08)$	$0.08(0.08)$	$--(--)$
	4	Mean(Cov)	$-3.12(0.97)$	$-3.01(0.95)$	$-3.00(0.96)$	$-1.89(\mathbf{0.00})$	$-2.32(\mathbf{0.00})$	$-2.11(\mathbf{0.00})$
		SD(SE)	$0.29(0.29)$	$0.11(0.11)$	$0.11(0.11)$	$0.09(0.10)$	$0.07(0.08)$	$0.07(0.08)$
	6	Mean(Cov)	$-3.16(0.97)$	$-3.01(0.96)$	$-3.01(0.94)$	$-1.81(\mathbf{0.00})$	$-2.37(\mathbf{0.00})$	$-2.10(\mathbf{0.00})$
		SD(SE)	$0.39(0.42)$	$0.11(0.11)$	$0.11(0.11)$	$0.12(0.13)$	$0.07(0.08)$	$0.07(0.08)$
$\widehat{\beta}_1$	2	Mean(Cov)	$1.09(0.96)$	$1.01(0.97)$	$--(--)$	$0.67(\mathbf{0.11})$	$0.82(\mathbf{0.42})$	$--(--)$
		SD(SE)	$0.30(0.30)$	$0.12(0.13)$	$--(--)$	$0.10(0.10)$	$0.08(0.09)$	$--(--)$
	4	Mean(Cov)	$1.21(0.94)$	$1.02(0.97)$	$1.01(0.95)$	$0.59(\mathbf{0.17})$	$0.84(\mathbf{0.53})$	$0.76(\mathbf{0.20})$
		SD(SE)	$0.65(0.59)$	$0.13(0.13)$	$0.13(0.12)$	$0.13(0.14)$	$0.08(0.09)$	$0.08(0.08)$
	6	Mean(Cov)	$1.29(0.93)$	$1.01(0.96)$	$1.02(0.96)$	$0.62(\mathbf{0.47})$	$0.86(\mathbf{0.67})$	$0.80(\mathbf{0.38})$
		SD(SE)	$0.90(0.85)$	$0.13(0.13)$	$0.13(0.13)$	$0.21(0.21)$	$0.09(0.09)$	$0.09(0.09)$
$\widehat{\beta}_2$	2	Mean(Cov)	$0.46(0.95)$	$0.49(0.96)$	$--(--)$	$0.43(\mathbf{0.90})$	$0.46(0.94)$	$--(--)$
		SD(SE)	$0.18(0.19)$	$0.10(0.10)$	$--(--)$	$0.08(0.09)$	$0.07(0.08)$	$--(--)$
	4	Mean(Cov)	$0.42(0.92)$	$0.49(0.95)$	$0.49(0.95)$	$0.41(0.93)$	$0.47(0.96)$	$0.44(0.92)$
		SD(SE)	$0.35(0.32)$	$0.10(0.10)$	$0.10(0.10)$	$0.11(0.12)$	$0.07(0.08)$	$0.07(0.08)$
	6	Mean(Cov)	$0.39(0.94)$	$0.49(0.96)$	$0.49(0.96)$	$0.43(\mathbf{0.98})$	$0.49(0.97)$	$0.46(0.93)$
		SD(SE)	$0.47(0.44)$	$0.10(0.10)$	$0.10(0.10)$	$0.16(0.18)$	$0.08(0.08)$	$0.08(0.08)$

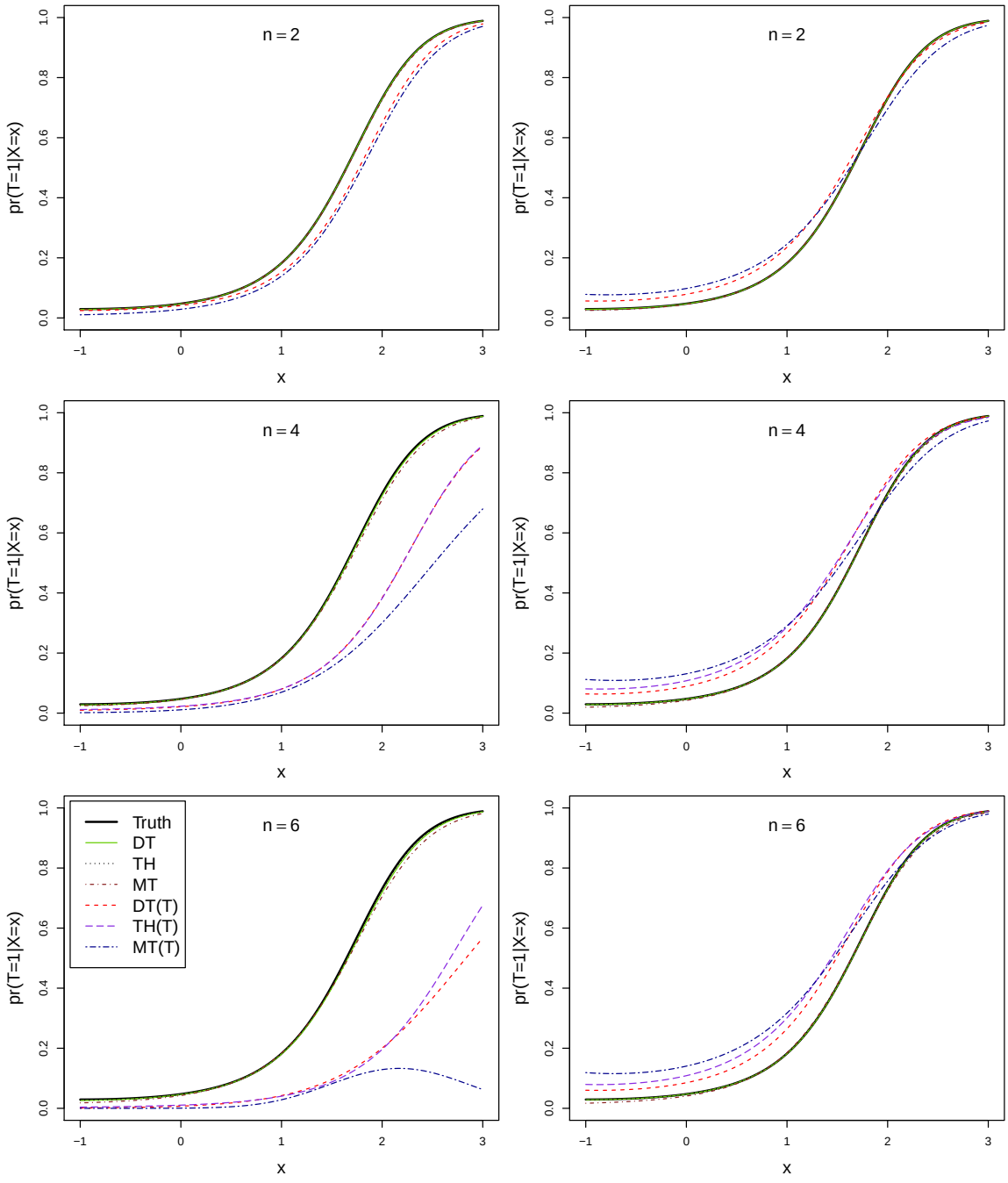


Figure 4.1: Plots of the estimated regression functions averaged over 500 data sets for Model 4.2 when $\sigma_+ = 1$ and $n \in \{2, 4, 6\}$. We use DT(T), TH(T), and MT(T) to denote the results obtained under the traditional modeling assumptions for the group testing algorithms DT, TH, and MT, respectively. The panels on the left and right of the figure correspond to thresholding strategies $t(c) = t_0$ and $t(c) = t_0/c$, respectively.

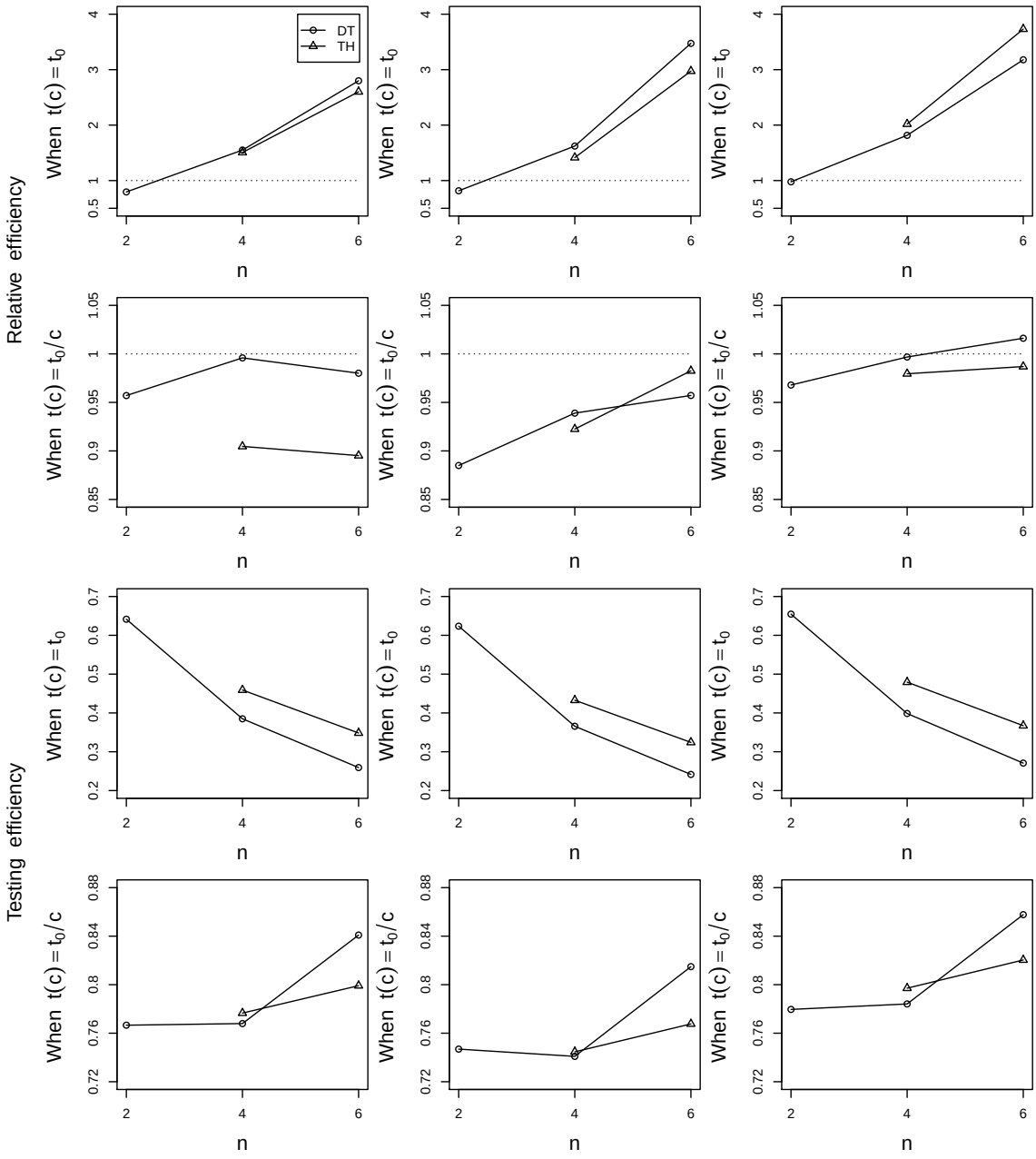


Figure 4.2: Simulation results concerning the efficiency of the parameter estimates obtained from modeling group testing data, resulting from different algorithms. Presented are results for Model 4.1 (left), Model 4.2 (middle), and Model 4.3 (right) across all considered group sizes (n), when $\sigma_+ = 1$. We define the testing efficiency to be the ratio between the average number of tests performed by a group testing algorithm and the number of tests required to conduct individual level testing; i.e., N. The relative efficiency is defined to be the ratio between MSE($\widehat{\boldsymbol{\beta}}$) obtained from modeling group testing data and the MSE($\widehat{\boldsymbol{\beta}}$) from modeling individual level testing data, where MSE($\widehat{\boldsymbol{\beta}}$) $= tr\{E[(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}}]\}$.

ferent group testing algorithms. We have also investigated the characteristics of our regression methodology under the situation in which the biomarker distributions are unknown and have to be estimated. The findings from these alternate studies are congruous with the results presented herein, and we have therefore opted not to include them.

4.4 Irish HBV Data

To further illustrate our methodology, we apply our techniques to the hepatitis B data previously analyzed by McMahan et al. (2013). This data set was originally compiled in an effort to assess the prevalence of antibodies to the hepatitis B virus (HBV) in Irish prisoners, for further details see (Allwright et al., 200). In the original study, oral fluid specimens were collected from each individual and were subsequently tested through the use of the Murex ICE enzyme immunoassay, the observed optical density (OD) readings from this process were then recorded. All positive results were then confirmed using an in-house radioimmunoassay. The data set also provides a diagnosed status for each of the individuals, which we will treat as their true infection status in this study. At the time of specimen collection, covariate information (e.g., age, gender, drug use, etc.) from participating subjects was collected via voluntary questionnaire. After the individuals with missing predictor variables and/or testing information were removed, we are left with a sample of size $N = 1098$. Specifically, there were complete records for 60 HBV-positive and 1038 HBV-negative individuals. The main purpose of this study is to compare the performance of our group testing regression methods to those that proceed under the more traditional assumptions.

One will note that the above information was collected on the individual level; i.e., for all N individuals we have access to their OD reading and covariate information. Using this information we are able to artificially construct group testing data. Proceeding in this fashion allows us to assess the performance of our methodology across a wide variety of settings (e.g., various group sizes, grouping schemes, and thresholding strategies), which would not be possible otherwise. Additionally, this approach, which allows for comparisons between estimates obtained from group testing and individual level data, has become common practice in the group testing regression literature (e.g., see Delaigle & Hall, 2012). To create group testing data we first note that the OD readings, which were available to us, are simply a measurement of the underlying antibody concentration levels. Thus, as in McMahan et al. (2013), we assume that the observed OD readings are linearly related to the true antibody concentration levels and were measured without error. Subsequently, we may determine the OD reading for a pool formed from combining the $\mathcal{P}_{jl}$ individuals by

$\widetilde{\mathcal{C}}_{\mathcal{P}_{jl}} = c_{jl}^{-1} \sum_{i \in \mathcal{P}_{jl}} \widetilde{\mathcal{C}}_{ij}$. Notice, we use the $\widetilde{\mathcal{C}}$ notation defined in Section 4.2.1 to represent the OD readings. Testing outcomes for pools were then determined by $Z_{\mathcal{P}_{jl}} = I\{\widetilde{\mathcal{C}}_{\mathcal{P}_{jl}} > t(c_{jl})\}$, where we considered the thresholding strategies $t(c) = t_0$ and $t(c) = t_0/c$. However, the assay threshold for individual level testing (t_0) was not provided to us. Thus, to choose a reasonable value of t_0, we first partition all 1098 OD readings into two sets $\text{OD}^+ = \{\widetilde{\mathcal{C}}_i : T_i = 1\}$ and $\text{OD}^- = \{\widetilde{\mathcal{C}}_i : T_i = 0\}$. We then select t_0 to minimize the discrepancies between the individuals' true statuses and their diagnosed statuses based on the OD readings; i.e.,

$$t_0 = \arg\max_t \left\{ \sum_{\widetilde{\mathcal{C}}_i \in \text{OD}^+} I(\widetilde{\mathcal{C}}_i > t) + \sum_{\widetilde{\mathcal{C}}_i \in \text{OD}^-} I(\widetilde{\mathcal{C}}_i < t) \right\}.$$

In this study we did not have access to the underlying distribution of the OD readings for the positive and negative individuals, which we denote by $f_{\widetilde{\mathcal{C}}^+}$ and $f_{\widetilde{\mathcal{C}}^-}$, respectively. Consequently, we estimated these distributions through the use of training data. Specifically, density estimation proceeded under the assumption that the OD readings followed a parametric model and we considered three such models: gamma, Weibull, and log-normal. Two training data sets were created by randomly sampling 10 observations from OD^+ and 44 observations from OD^-. Using the training data, we obtained the estimates $\widehat{f}_{\widetilde{\mathcal{C}}^+}$ and $\widehat{f}_{\widetilde{\mathcal{C}}^-}$ of $f_{\widetilde{\mathcal{C}}^+}$ and $f_{\widetilde{\mathcal{C}}^-}$, respectively, through the use of maximum likelihood techniques. In order to fit the traditional regression models we calculated the sensitivity and specificity of individual level testing to be $S_e = \int_{t_0}^{\infty} \widehat{f}_{\widetilde{\mathcal{C}}^+}(x)dx$ and $S_p = \int_{\infty}^{t_0} \widehat{f}_{\widetilde{\mathcal{C}}^-}(x)dx$, respectively. To implement our regression methodology, we calculated $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j)$ as follows

$$M_j(\boldsymbol{z}_j, \boldsymbol{t}_j) = \int \prod_{l=1}^{K_j} I\{\boldsymbol{D}_{\mathcal{P}_{jl}}^{\mathrm{T}} \boldsymbol{y} \in A(z_{\mathcal{P}_{jl}}, c_{jl})\} \prod_{i=1}^{n_j} \widehat{f}_{\widetilde{\mathcal{C}}}(y_{ij} \mid T_{ij} = t_{ij}) d\boldsymbol{y},$$

where $\widehat{f}_{\widetilde{\mathcal{C}}}(\cdot \mid T) = T\widehat{f}_{\widetilde{\mathcal{C}}^+}(\cdot) + (1-T)\widehat{f}_{\widetilde{\mathcal{C}}^-}(\cdot)$. It is worth while to point out that the aforementioned integral is difficult to compute analytically, consequently we used the Monte Carlo techniques described in Appendix C.3 to approximate it.

To make our comparisons, we consider the following simple second-order logistic model

$$\text{logit}\{\text{pr}(T_{ij} = 1 \mid x_{ij})\} = \beta_0 + \beta_1 x_{ij} + \beta_2 x_{ij}^2,$$

where x_{ij} denotes the age of the ith individual in the jth group. After the removal of the training data, there were $N = 1044$ observations remaining. We again specify a common group size n, where $n \in \{1, 2, 4, 6\}$, and we assigned each of the N individuals to one of the $J = N/n$ groups. In this study we considered both homogeneous and random grouping schemes; i.e., individuals of a common age were grouped together

(homogeneous grouping) or individuals were randomly assigned to groups (random grouping). The former grouping strategy has been shown to result in more efficient parameter estimation when compared to the latter (e.g., see Bilder & Tebbs, 2009; Delaigle & Hall, 2012), however homogeneous grouping is not always feasible in practical applications. The group testing strategies chosen for this study were MT and DT, and group testing data was subsequently generated in a similar fashion to the methods described in Section 4.3.1. For each of the group testing data sets, we estimated the regression parameters under our methodology. In order to compare our approach to existing techniques we also estimated the regression parameters for each data set using the group testing regression models which proceed under the traditional modeling assumptions. Further, to provide a standard by which comparisons can be made we also fit the individual data model (i.e., $n = 1$). This process was repeated 1000 times for each pool size, with a new training data set being selected each time.

In order to assess misclassification error rates of the different thresholding strategies we compared the individuals' true statuses to the diagnosed statuses obtained from the two group testing decoding algorithms. A summary of these results across all of the considered testing configurations is provided in Appendix C.5. From these comparisons, we found that the thresholding strategy $t(c) = t_0$ resulted in an extremely high false negative rate; i.e., under this strategy many of the truly positive individuals were classified as negative. Consequently, we have chosen to focus our attention on the data arising from the thresholding strategy $t(c) = t_0/c$, which resulted in misclassification rates similar to that of individual level testing. Table C.4 provides a summary of the 1000 estimates of $\boldsymbol{\beta}$ across all considered configurations under our selected thresholding strategy. Figure 4.3 provides plots of the estimated regression functions averaged over the 1000 replications across all considered configurations under the same thresholding strategy and random grouping. From these results, one will first note that the estimates obtained by our regression methodology appear to be more reliable when compared to the estimates resulting from the more traditional regression techniques, across all considered configurations. These results reinforce the main findings discussed in Section 4.3.2. Specifically, Figure 4.3 illustrates that the traditional regression methodology tends to drastically overestimate the age-specific probabilities of HBV infection for larger group sizes (e.g., $n = 6$), while the estimates from our method remain in agreement with the results from the individual level data. These trends can also be observed in the summary of the estimates of $\boldsymbol{\beta}$ provided in Table C.4. The discrepancies between the estimates obtained by our method and those resulting from the individual data model, are likely explained by the error introduced by having to estimate $f_{\widetilde{\mathcal{C}}^+}$ and $f_{\widetilde{\mathcal{C}}^-}$.

Table 4.3: Irish HBV data: Presented results include the sample mean (standard deviation) of the 1000 maximum likelihood estimates of $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^{\mathrm{T}}$, across all considered configurations under the thresholding strategy $t(c) = t_0/c$. Note, IT, DT, and TH denote individual testing, Dorfman testing, and three-stage halving, respectively.

Lognormal: Under IT, the summary of the estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$ are -2.92(0.06), 1.09(0.11), and -0.31(0.06), respectively.

		Random Grouping				Homogeneous Grouping			
		Our Approach		Traditional Approach		Our Approach		Traditional Approach	
	n	DT	TH	DT	TH	DT	TH	DT	TH
$\widehat{\beta}_0$	2	$-2.90(0.06)$	$--(--)$	$-2.88(0.06)$	$--(--)$	$-2.90(0.05)$	$--(--)$	$-2.88(0.06)$	$--(--)$
	4	$-2.87(0.08)$	$-2.86(0.07)$	$-2.75(0.07)$	$-2.75(0.07)$	$-2.84(0.09)$	$-2.84(0.09)$	$-2.70(0.07)$	$-2.69(0.07)$
	6	$-2.92(0.07)$	$-2.87(0.07)$	$-2.53(0.09)$	$-2.50(0.09)$	$-2.89(0.08)$	$-2.87(0.07)$	$-2.30(0.07)$	$-2.26(0.09)$
$\widehat{\beta}_1$	2	$1.13(0.11)$	$--(--)$	$1.13(0.11)$	$--(--)$	$1.14(0.10)$	$--(--)$	$1.14(0.10)$	$--(--)$
	4	$1.10(0.11)$	$1.11(0.11)$	$1.09(0.11)$	$1.10(0.11)$	$1.10(0.10)$	$1.12(0.10)$	$1.07(0.09)$	$1.09(0.10)$
	6	$1.10(0.11)$	$1.09(0.11)$	$1.09(0.12)$	$1.08(0.12)$	$1.10(0.11)$	$1.12(0.11)$	$0.80(0.07)$	$0.83(0.08)$
$\widehat{\beta}_2$	2	$-0.32(0.06)$	$--(--)$	$-0.32(0.06)$	$--(--)$	$-0.32(0.06)$	$--(--)$	$-0.32(0.06)$	$--(--)$
	4	$-0.31(0.06)$	$-0.31(0.06)$	$-0.31(0.06)$	$-0.31(0.06)$	$-0.32(0.05)$	$-0.33(0.06)$	$-0.32(0.05)$	$-0.33(0.06)$
	6	$-0.31(0.06)$	$-0.31(0.06)$	$-0.32(0.06)$	$-0.32(0.06)$	$-0.31(0.06)$	$-0.33(0.06)$	$-0.28(0.05)$	$-0.30(0.05)$

Gamma: Under IT, the summary of the estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$ are -2.90(0.06), 1.09(0.10), and -0.31(0.06), respectively.

		Random Grouping				Homogeneous Grouping			
		Our Approach		Traditional Approach		Our Approach		Traditional Approach	
	n	DT	TH	DT	TH	DT	TH	DT	TH
$\widehat{\beta}_0$	2	$-2.88(0.06)$	$--(--)$	$-2.86(0.06)$	$--(--)$	$-2.87(0.06)$	$--(--)$	$-2.86(0.06)$	$--(--)$
	4	$-2.82(0.09)$	$-2.81(0.08)$	$-2.73(0.07)$	$-2.72(0.07)$	$-2.78(0.10)$	$-2.78(0.09)$	$-2.68(0.07)$	$-2.66(0.07)$
	6	$-2.88(0.08)$	$-2.83(0.07)$	$-2.50(0.08)$	$-2.46(0.09)$	$-2.85(0.11)$	$-2.82(0.09)$	$-2.27(0.07)$	$-2.22(0.08)$
$\widehat{\beta}_1$	2	$1.13(0.11)$	$--(--)$	$1.13(0.11)$	$--(--)$	$1.14(0.10)$	$--(--)$	$1.15(0.11)$	$--(--)$
	4	$1.09(0.11)$	$1.11(0.11)$	$1.09(0.11)$	$1.11(0.11)$	$1.09(0.10)$	$1.10(0.10)$	$1.08(0.10)$	$1.09(0.10)$
	6	$1.10(0.10)$	$1.10(0.11)$	$1.09(0.12)$	$1.09(0.12)$	$1.10(0.11)$	$1.12(0.10)$	$0.80(0.07)$	$0.83(0.07)$
$\widehat{\beta}_2$	2	$-0.32(0.06)$	$--(--)$	$-0.32(0.06)$	$--(--)$	$-0.33(0.06)$	$--(--)$	$-0.33(0.06)$	$--(--)$
	4	$-0.31(0.06)$	$-0.32(0.06)$	$-0.31(0.06)$	$-0.32(0.06)$	$-0.32(0.06)$	$-0.32(0.05)$	$-0.33(0.06)$	$-0.33(0.05)$
	6	$-0.31(0.06)$	$-0.31(0.06)$	$-0.32(0.06)$	$-0.32(0.06)$	$-0.32(0.06)$	$-0.33(0.06)$	$-0.29(0.05)$	$-0.30(0.05)$

Weibull: Under IT, the summary of the estimates $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$ are -2.88(0.06), 1.09(0.10), and -0.30(0.05), respectively.

		Random Grouping				Homogeneous Grouping			
		Our Approach		Traditional Approach		Our Approach		Traditional Approach	
	n	DT	TH	DT	TH	DT	TH	DT	TH
$\widehat{\beta}_0$	2	$-2.86(0.06)$	$--(--)$	$-2.84(0.07)$	$--(--)$	$-2.86(0.06)$	$--(--)$	$-2.84(0.07)$	$--(--)$
	4	$-2.81(0.09)$	$-2.81(0.09)$	$-2.71(0.08)$	$-2.70(0.08)$	$-2.78(0.11)$	$-2.78(0.10)$	$-2.66(0.07)$	$-2.65(0.08)$
	6	$-2.86(0.10)$	$-2.82(0.10)$	$-2.48(0.09)$	$-2.45(0.09)$	$-2.81(0.15)$	$-2.80(0.12)$	$-2.25(0.08)$	$-2.20(0.09)$
$\widehat{\beta}_1$	2	$1.12(0.10)$	$--(--)$	$1.13(0.11)$	$--(--)$	$1.13(0.10)$	$--(--)$	$1.14(0.10)$	$--(--)$
	4	$1.09(0.11)$	$1.11(0.11)$	$1.09(0.11)$	$1.10(0.11)$	$1.08(0.10)$	$1.11(0.10)$	$1.07(0.10)$	$1.10(0.10)$
	6	$1.09(0.11)$	$1.10(0.11)$	$1.09(0.12)$	$1.09(0.12)$	$1.08(0.11)$	$1.12(0.11)$	$0.80(0.07)$	$0.84(0.08)$
$\widehat{\beta}_2$	2	$-0.32(0.05)$	$--(--)$	$-0.32(0.06)$	$--(--)$	$-0.32(0.06)$	$--(--)$	$-0.32(0.06)$	$--(--)$
	4	$-0.31(0.06)$	$-0.32(0.06)$	$-0.31(0.06)$	$-0.32(0.06)$	$-0.32(0.06)$	$-0.33(0.06)$	$-0.33(0.06)$	$-0.33(0.06)$
	6	$-0.31(0.06)$	$-0.31(0.06)$	$-0.32(0.06)$	$-0.32(0.06)$	$-0.32(0.06)$	$-0.33(0.06)$	$-0.29(0.05)$	$-0.30(0.05)$

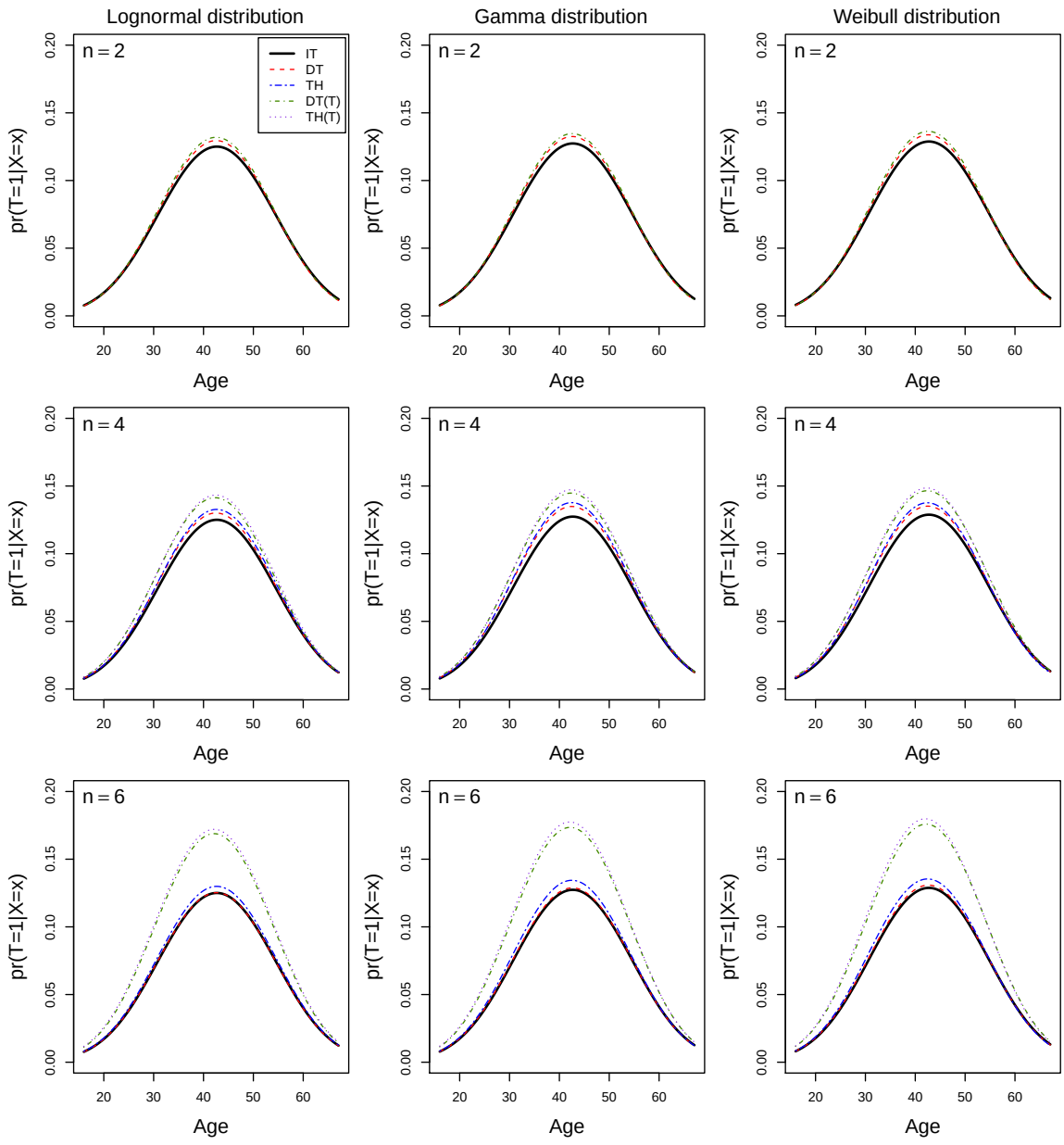


Figure 4.3: Irish HBV data. Plots of the estimated regression functions averaged over the 1000 data sets across all considered configurations under the thresholding strategy $t(c) = t_0/c$ and random grouping. From left to right, the figures present the regression estimates corresponding to the assumption that the OD readings follow a log-normal, gamma, and Weibull distribution. We use DT(T) and TH(T) to denote the results obtained under the traditional modeling assumptions for the group testing algorithms DT and TH, respectively.

4.5 Discussion

In this chapter, we have generalized the group testing regression methodology proposed in McMahan et al. (2013), to allow for the incorporation of testing information obtained from all group testing strategies. We have also illustrated that regression methods which operate under the more traditional assumptions can result in extremely biased inference, when the assumptions are in fact invalid. Through numerical studies, we have been able to show that our proposed techniques can result in more efficient parameter estimates, when compared to those based on individual level data, at a fraction of the cost of data collection.

Appendices

Appendix A Technical proofs related to Chapter 2

A.1 Proof of Theorem 2.3.1

Let $\alpha = H(\theta - \theta^*)$, $\hat{\alpha} = H(\hat{\theta} - \theta^*)$ and $\tilde{\mathbf{U}}_j = H^{-1}\tilde{\mathbf{X}}_j$. Put

$$\begin{aligned} l(\alpha) &= \frac{1}{N}\sum_{j=1}^{J}\Big\{(1-T_j^*)(\tilde{\mathbf{X}}_j^\top\theta^* + \tilde{\mathbf{U}}_j^\top\alpha) + T_j^*\log\Big[1-\exp(\tilde{\mathbf{X}}_j^\top\theta^* + \tilde{\mathbf{U}}_j^\top\alpha)\Big]\Big\}\omega_h(\mathbf{X}_j, x) \\ &= \sum_{k=1}^{K}\frac{J_k}{N}\cdot\frac{1}{J_k}\sum_{j=J_{k-1}+1}^{J_{k-1}+J_k} l_j(\alpha;k), \end{aligned}$$

where $J_0 = 0$ and $l_j(\alpha;k)$ is the kernel weighted likelihood corresponding to a pooled data of size $n^{(k)}$. Since $l(\alpha)$ is strictly concave, it is sufficient to show that, for any given $\eta > 0$, there exists a small constant ε, such that

$$\liminf_N P\Big\{\sup_{\|\alpha\|=\varepsilon} l(\alpha) < l(0)\Big\} = 1-\eta.$$

By Taylor's expansion around the origin, for any α with $\|\alpha\| = \varepsilon$,

$$l(\alpha) - l(0) = l'(0)^\top\alpha + \frac{1}{2}\alpha^\top l''(0)\alpha + R(\alpha'), \tag{A.1}$$

with α' lying between α and 0, and where

$$R(\alpha') = \frac{1}{6}\sum_{j,k,l}\alpha_j'\alpha_k'\alpha_l'\frac{\partial^3 l(\alpha')}{\partial\alpha_j\partial\alpha_k\partial\alpha_l}.$$

First, since for fixed k, $l_j(\alpha;k)$, $j = J_{k-1}+1,\ldots,J_{k-1}+J_k$ are i.i.d., we have

$$l'(0) = \sum_{k=1}^{K}\frac{J_k}{N}\cdot\frac{1}{J_k}\sum_{j=J_{k-1}+1}^{J_{k-1}+J_k} l_j'(0;k) \to_p \sum_{k=1}^{K}\frac{\gamma_k}{n^{(k)}}E[l_j'(0;k)], \tag{A.2}$$

where

$$l_j'(0;k) = \left(1 - \frac{T_j^*}{1-\exp(\tilde{\mathbf{X}}_j^\top\theta^*)}\right)\tilde{\mathbf{U}}_j\omega_h(\mathbf{X}_j, x).$$

We know $E[l_j'(0;k)] = E_{\mathbf{Z}_j}\{E[l_j'(0;k)|\mathbf{Z}_j]\}$ with $\mathbf{Z}_j = (Z_{1j},\ldots,Z_{n^{(k)}j})^\top$. It is easy to see that

$$E[l_j'(0;k)|\mathbf{Z}_j = 0] = 0.$$

When some of the Z_{ij}'s are not zero, local linear approximation provides that

$$E\Big[T_j^*|\mathbf{Z}_j\Big] = 1 - \exp(\tilde{\mathbf{X}}_j^\top \theta^*) - \exp(\tilde{\mathbf{X}}_j^\top \theta^*)\frac{g^{(2)}(x)}{2}\sum_{i=1}^{n^{(k)}} I_x(X_{ij})(X_{ij}-x)^2\{1+o(1)\}.$$

Applying Taylor's expansion, $E[l_j'(0;k)]$ can be written as

$$E_{\mathbf{Z}_j}\Big[\frac{g^{(2)}(x)}{2}\cdot\Big\{\frac{\exp(A)}{1-\exp(A)} + \frac{\exp(A)\theta_2^*}{(1-\exp(A))^2}\cdot B\{1+o(1)\}\Big\}\tilde{\mathbf{U}}_j h^{-\sum_{i=1}^{n^{(k)}} I_x(X_{ij})} C\Big],$$

where $A = \sum_{i=1}^{n^{(k)}}\{I_x(X_{ij})\theta_1^* + (1-I_x(X_{ij}))\theta_3^*\}$, $B = \sum_{i=1}^{n^{(k)}} I_x(X_{ij})(X_{ij}-x)$ and $C = \sum_{i=1}^{n^{(k)}}(X_{ij}-x)^2\{1+o(1)\}$. Let M_m be the event that only m of the $I_x(X_{ij})$s are zero (since Xs are i.i.d, without loss of generality, we assume $I_x(X_{1j}) = \cdots = I_x(X_{mj}) = 0$). Then conditioning on M_ms,

$$E[l_j'(0;k)] = \frac{g^{(2)}(x)}{2}\sum_{m=1}^{n^{(k)}}\binom{n^{(k)}}{m}P_x^{n^{(k)}-m} I_k^{(m)},$$

where P_x is the probability of an X falling out of I_x, i.e., $P_x = \int_{I_x^c} f(u)du$, and

$$\begin{aligned} I_k^{(m)} &= \int_{x-h}^{x+h}\cdots\int_{x-h}^{x+h}\Big\{\frac{\exp(A_m)}{1-\exp(A_m)} + \frac{\exp(A_m)\theta_2^*}{(1-\exp(A_m))^2}\cdot B_m\{1+o(1)\}\Big\} \\ &\quad \times C_m\begin{pmatrix} m \\ B_m h^{-1} \\ n^{(k)}-m \end{pmatrix}\frac{1}{h}\prod_{i=1}^{m} f(X_{ij})K\Big(\frac{X_{ij}-x}{h}\Big)^{1/m} dX_{1j}\cdots dX_{mj}, \end{aligned}$$

with $A_m = m\theta_1^* + (n^{(k)}-m)\theta_3^*$, $B_m = \sum_{i=1}^m (X_{ij}-x)$, and $C_m = \sum_{i=1}^m (X_{ij}-x)^2\{1+o(1)\}$. By Conditions 2.2 and 2.3, $h\to 0$. Then $\theta_3^* \to \log q_*$ and $P_x \to 1$. We can write

$$I_k^{(1)} = \Big(\mu_2 V_{k1}h^2, \mu_4 V_{k2}h^3, (n^{(k)}-1)\mu_2 V_{k1}h^2\Big)^\top.$$

Simple integration provides that, for $m>1$, $I_k^{(m)} = o(I_k^{(1)})$. Hence

$$E[l_j'(0;k)] = \frac{n^{(k)}g^{(2)}(x)I_k^{(1)}}{2}\{1+o(1)\}. \tag{A.3}$$

By the assumption $n^{(k)}J_k/N \to \gamma_k$ and $(A.2)$, we can conclude that

$$l'(0) = b_\theta + o_p(1) = o_p(1). \tag{A.4}$$

Thus, with probability tending to 1,

$$|l'(0)^\top \alpha| \leq \varepsilon^3. \tag{A.5}$$

For $l''(0)$, similarly

$$l''(0) = \sum_{k=1}^{K} \frac{J_k}{N} \cdot \frac{1}{J_k} \sum_{j=J_{k-1}+1}^{J_{k-1}+J_k} l_j''(0;k) \to_p \sum_{k=1}^{K} \frac{\gamma_k}{n^{(k)}} E[l_j''(0;k)], \tag{A.6}$$

where

$$l_j''(0;k) = \frac{T_j^* \exp(\tilde{\mathbf{X}}_j^\top \theta^*)}{(1-\exp(\tilde{\mathbf{X}}_j^\top \theta^*))^2} \tilde{\mathbf{U}}_j \tilde{\mathbf{U}}_j^\top \omega_h(\mathbf{X}_j, x).$$

When $\mathbf{Z}_j = 0$, $\tilde{\mathbf{U}}_j \tilde{\mathbf{U}}_j^\top = \text{diag}\{0, 0, n^{(k)2}\}$ and $\omega_h(\mathbf{X}_j, x) = 1$. Thus, as $h \to 0$,

$$E[l_j''(0;k)|\mathbf{Z}_j = 0] \to \text{diag}\Big\{0, 0, \frac{n^{(k)2}\exp(E_{k0})}{(1-\exp(E_{k0})}\Big\}.$$

When some of Z_{ij}'s are not zero, using the same argument as above, we have

$$E[l_j''(0;k)] = n^{(k)} \begin{pmatrix} V_{k1}\mu_0 & 0 & (n^{(k)}-1)V_{k1}\mu_0 \\ 0 & V_{k1}\mu_2 & 0 \\ (n^{(k)}-1)V_{k1}\mu_0 & 0 & (n^{(k)}-1)^2 V_{k1}\mu_0 + V_{k0} \end{pmatrix} \{1+o(1)\}.$$

By $(A.6)$,

$$l''(0) = -V_0 + o_p(1). \tag{A.7}$$

Let $\lambda_{\min}(V_0)$ be the smallest eigenvalue of V_0. Since V_0 is positive definite, $\lambda_{\min}(V_0)$ is a positive number. Thus, with probability tending to 1,

$$\alpha^\top l''(0)\alpha \leq -\lambda_{\min}(V_0)\varepsilon^2. \tag{A.8}$$

Similarly, we can find that

$$|R(\alpha)| = \varepsilon^3 O_p(1). \tag{A.9}$$

Substituting $(A.5), (A.8)$ and $(A.9)$ into $(A.1)$, its sign is completely decided by the term of ε^2 when ε is small enough. This completes the proof of Theorem 1.

A.2 Proof of Theorem 2.3.2

Continuing to use the notation introduced in the proof of Theorem 1, by Taylor's expansion, we have $0 = l'(\hat{\alpha}) = l'(0) + l''(0)\hat{\alpha} + O_p(\|\hat{\alpha}\|^2)$. Hence, by $(A.7)$,

$$\hat{\alpha} = -\{-V_0 + o_p(1)\}^{-1} l'(0). \tag{A.10}$$

It suffices to establish the asymptotic normality of $l'(0)$. By $(A.4)$, $E[l'(0)] = b_\theta + o(1)$. For Var$[l'(0)]$, we have

$$\text{Var}[l'(0)] = N^{-1} \sum_{k=1}^{K} (J_k/N) \text{Var}[l'_j(0;k)].$$

Since Var$[l'_j(0;k)] = E[l'_j(0;k) l'_j(0;k)^\top] - E[l'_j(0;k)] E[l'_j(0;k)]^\top$, and $(A.3)$ shows the rate of $E[l'_j(0;k)]$, we only need to find $E[l'_j(0;k) l'_j(0;k)^\top]$ where

$$l'_j(0;k) l'_j(0;k)^\top = \left(1 - \frac{T_j^*}{1 - \exp(\tilde{\mathbf{X}}_j^\top \theta^*)}\right)^2 \tilde{\mathbf{U}}_j \tilde{\mathbf{U}}_j^\top \omega_h^2(\mathbf{X}_j, x).$$

Using similar argument as in the proof of Theorem 1, we have

$$\text{Var}[l'_j(0;k)] = n^{(k)} \frac{V_{k1}}{h} \begin{pmatrix} \nu_0 & 0 & (n^{(k)} - 1)\nu_0 \\ 0 & \nu_2 & 0 \\ (n^{(k)} - 1)\nu_0 & 0 & (n^{(k)} - 1)^2 \nu_0 \end{pmatrix} \{1 + o(1)\}.$$

Combining with the assumption that $J_k/N \to \gamma_k / n^{(k)}$, we have

$$\text{Var}[l'(0)] = N^{-1} h^{-1} V_1 + o(N^{-1} h^{-1}). \tag{A.11}$$

By Cauchy-Schwarz inequality, V_1 is a singular matrix only when $K = 1$. To make the notation consistent we treat a constant as a degraded normal random variable with mean being itself and variance being 0. Applying the Cramér-Wold device, we need to show that for any constant vector $b \neq 0$,

$$\sqrt{Nh}\{b^\top l'(0) - b^\top E l'(0)\} \to_D N\{0, b^\top V_1 b\}. \tag{A.12}$$

When $K = 1$ and b is linear to $(-(n_1 - 1), 0, 1)^\top$, Var$[\sqrt{Nh}b^\top l'(0)] \to b^\top V_1 b = 0$. Otherwise $b^\top V_1 b$ is a positive number. By the equation in (A.2) and for any fixed k, $b^\top l'_j(0; k)$s are identical and independently distributed, the normality of $(A.12)$ of $b^\top l'(0)$ follows from the central limit theory combining with $(A.3)$ and $(A.11)$. Consequently, by $(A.10)$, it completes the proof.

Appendix B Technical arguments and additional simulation results related to Chapter 3

B.1 A general formula of $M(\cdot,\cdot,\cdot)$

We first provide the detailed derivation of $M(Z,T,C)$ and the illustration of how to evaluate $\text{pr}(\mathcal{P})$ which are mentioned in Section 3.2.1. Generally speaking, a group testing algorithm consists of several stages of screening. At each stage, it randomly selects a set of pools based on the information obtained at the previous stage, then it tests these pools. This process continues until the stopping rule is met. Thus, to evaluate $M(\cdot,\cdot,\cdot)$, one need consider two random processes. The first one comes from the measurement accuracy of the assay, the other one is due to the random selection of pools for next stage of screening given all the information obtained by the current stage. We denote the testing outcome as $Z=\{Z_l=(Y_l,\mathcal{P}_l), l=1,\ldots,K\}$ where we assume that Z_l occurs no later than Z_{l+1}. For any $z=\{z_l=(y_l,\rho_l), l=1,\ldots,K\}\in\mathcal{Z}(c)$ and $t=(t_1,\ldots,t_c)^{\mathrm{T}}\in\mathcal{T}(c)$, we have

$$\begin{aligned}
M(z,t,c) =&\text{pr}(Z=z\mid T=t)\\
=&\text{pr}\{(Y_l,\mathcal{P}_l)=(y_l,\rho_l), l=1,\ldots,K\mid T=t\}\\
=&\prod_{l=1}^{K}\{\text{pr}(\mathcal{P}_l=\rho_l\mid Z_1=z_1,\ldots,Z_{l-1}=z_{l-1},T=t)\\
&\times\text{pr}(Y_l=y_l\mid\mathcal{P}_l=\rho_l,T=t,Z_1=z_1,\ldots,Z_{l-1}=z_{l-1})\}\\
=&\text{pr}(\mathcal{P})\prod_{l=1}^{K}\text{pr}(Y_l=y_l\mid\mathcal{P}_l=\rho_l,T=t,Z_1=z_1,\ldots,Z_{l-1}=z_{l-1})
\end{aligned}$$

where $\text{pr}(\mathcal{P})=\prod_{l=1}^{K}\text{pr}(\mathcal{P}_l=\rho_l\mid Z_1=z_1,\ldots,Z_{l-1}=z_{l-1},T=t)$ and $\mathcal{P}=\{\mathcal{P}_1,\ldots,\mathcal{P}_l\}$.

Given all the individuals' true statues $T=t$, the true underlying status of any pool $\mathcal{P}_l$ formed by these individual's specimens, i.e., $\tilde{Y}_l=\max_{i\in\mathcal{P}_l}T_i$, is given by $\tilde{Y}_l=y_l=\max_{i\in\rho_l}t_i$. Then, we have

$$\begin{aligned}
&\text{pr}(Y_l=y_l\mid\mathcal{P}_l=\rho_l,T=t,Z_1=z_1,\ldots,Z_{l-1}=z_{l-1})\\
=&\text{pr}(Y_l=y_l\mid\mathcal{P}_l=\rho_l,\tilde{Y}_l=\tilde{y}_l)\\
=&S_e^{y_l\tilde{y}_l}(1-S_e)^{(1-y_l)\tilde{y}_l}(1-S_p)^{y_l(1-\tilde{y}_l)}S_p^{(1-y_l)(1-\tilde{y}_l)}.
\end{aligned}$$

Now, we obtain the formula presented in our manuscript, i.e.,

$$M(z,t,c) = \text{pr}(\mathcal{P}) \prod_{l=1}^{K} \left\{ S_e^{y_l \tilde{y}_l} (1-S_e)^{(1-y_l)\tilde{y}_l} (1-S_p)^{y_l(1-\tilde{y}_l)} S_p^{(1-y_l)(1-\tilde{y}_l)} \right\}.$$

We would like to point out that pr($\mathcal{P}$) purely evaluates how likely these pools were selected at each stage given all the test results from previous stages. Thus, the evaluation of pr($\mathcal{P}$) neither depends on the individual true states T nor on the unknown parameters $\{\beta, p(\cdot)\}$. If the group testing algorithm is deterministic, i.e., there is no random selection of pools involved in the screening process, pr($\mathcal{P}$) $= 1$. This type of testing algorithm includes but not limited to master pool testing, Dorfman decoding and array testing. For example, in Dorfman decoding, it always starts with testing the master pool at the first stage. If it tests negative, the screening process stops. Otherwise, the next stage of screening proceeds to retest every individual one-by-one. Thus, pr($\mathcal{P}$) is always 1. If the group testing algorithm involves random selection of pools, pr($\mathcal{P}$) is just a product of probabilities which evaluates how likely to arrange individuals into subpools. For instance, in halving algorithm, if a pool $\{1, \ldots, 4\}$ tests positive, it randomly divides this pool into two halves and then testing each half. Thus, the probability of selecting $\{1, 3\}$ and $\{2, 4\}$ as the two halves given the pool $\{1, \ldots, 4\}$ tests positive is just $1/3$.

B.2 Detailed illustration of Sections 3.2.2 and 3.2.3

we would like first to present a detailed calculation of p_{j0}, $a_{ij}(\mu)$, $b_{ij}(\mu)$, and $\mathcal{R}\{(0, \mathcal{G}_j); \mathcal{X}_j, \beta, p(\cdot)\}$ under master pool testing. Since T_{ij}s are identical and independent Bernoulli random variables with probability of success being $1-\mu$, we have

$$\begin{aligned} p_{j0} =& \text{pr}(Y_{j1} = 0 \mid \max_i T_{ij} = 1)\text{pr}(\max_i T_{ij} = 1) \\ &+ \text{pr}\{Y_{j1} = 0 \mid \max_i T_{ij} = 0)\text{pr}(\max_i T_{ij} = 0) \\ =& 1 - S_e - \delta_{c_j}. \end{aligned} \tag{B.1}$$

Similarly, to calculate $a_{ij}(\mu)$ and $b_{ij}(\mu)$, we have

$$\begin{aligned}\mathcal{F}_{ij}(\mu) =&\text{pr}(D_{ij}=1 \mid T_{ij}=0)\\ =&\text{pr}(Y_{j1}=1 \mid T_{ij}=0, \max_{k\neq i} T_{ij}=0)\text{pr}(\max_{k\neq i} T_{ij}=0)\\ &+\text{pr}(Y_{j1}=1 \mid T_{ij}=0, \max_{k\neq i} T_{ij}=1)\text{pr}(\max_{k\neq i} T_{ij}=1)\\ =&S_e+\delta_{c_j-1},\end{aligned}$$

and $\mathcal{F}_{ij}(1,\mu) = \text{pr}(Y_{j1}=1 \mid T_{ij}=1) = S_e$. Consequently we obtain $a_{ij}(\mu) = S_e + \delta_{c_j-1}$ and $b_{ij}(\mu) = \mathcal{F}_{ij}(1,\mu) - \mathcal{F}_{ij}(0,\mu) = -\delta_{c_j-1}$. The calculation of $\mathcal{R}\{(0,\mathcal{P}_{j1}); \mathcal{X}_j, \beta, p(\cdot)\}$ follows the spirit of the calculation of p_{j0}. The only difference is that it accounts for the covariate effect, i.e.,

$$\mathcal{R}\{(0,\mathcal{P}_{j1}); \mathcal{X}_j, \beta, p(\cdot)\} = 1 - S_e - (1-S_e-S_p)\prod_{i=1}^{c_j}\{1-p(X_{ij}^{\mathrm{T}}\beta)\}. \tag{B.2}$$

Then we present details of how we obtained the simplified expression of $l_p(\mu)$, $a_{ij}(\mu)$, $b_{ij}(\mu)$, and $l\{\beta, p(\cdot)\}$ under Dorfman decoding. For easy illustration, we start with $\mathcal{R}\{z; \mathcal{X}_j, \beta, p(\cdot)\}$ for $z \in \mathcal{Z}(c_j)$. When $z = (0, \mathcal{G}_j)$, we have $\mathcal{R}\{(0,\mathcal{G}_j); \mathcal{X}_j, \beta, p(\cdot)\}$ exactly as (B.2). For other $z \in \mathcal{Z}(c_j)$, one could express z as $z = \{(y_{j1}=1, \mathcal{G}_j), (y_{jl}, \{l-1\}), l = 2, \ldots, c_j+1\}$; i.e., the master pool tests positive, and then individuals are retested one-by-one. In this case, the individual diagnosis for the ith individual is $y_{j,i+1}$. Thus, we have $\mathcal{I}_{j1} = \{i \in \mathcal{G}_j : y_{j,i+1} = 1\}$, $\mathcal{I}_{j0} = \{i \in \mathcal{G}_j : y_{j,i+1} = 0\}$. Following the notation introduced in the manuscript, let $\mathcal{S}_{j1} = \sum_{i\in\mathcal{I}_{j1}} T_{ij}$ and $\mathcal{S}_{j0} = \sum_{i\in\mathcal{I}_{j0}} T_{ij}$. Then

$$\begin{aligned}&\mathcal{R}\{z; \mathcal{X}_j, \beta, p(\cdot)\} = \text{pr}\{Z_j = z \mid \mathcal{X}_j, \beta, p(\cdot)\}\\ =&\text{pr}(Z_j = z \mid \mathcal{S}_{j1}=0, \mathcal{S}_{j0}=0)\text{pr}\{\mathcal{S}_{j1}=0, \mathcal{S}_{j0}=0 \mid \mathcal{X}_j, \beta, p(\cdot)\}\\ &+\sum_{\substack{k_1=0\\ k_1+k_0\neq 0}}^{k}\sum_{k_0=0}^{c_j-k}\text{pr}(Z_j = z \mid \mathcal{S}_{j1}=k_1, \mathcal{S}_{j0}=k_0)\text{pr}\{\mathcal{S}_{j1}=k_1, \mathcal{S}_{j0}=k_0 \mid \mathcal{X}_j, \beta, p(\cdot)\}.\end{aligned}$$

Under the assumption of testing errors, one can easily see that

$$\text{pr}(Z_j = z \mid \mathcal{S}_{j1}=0, \mathcal{S}_{j0}=0) = (1-S_p)(1-S_p)^k S_p^{c_j-k}.$$

When $k_1 + k_0 \neq 0$, there exists at least one individual which is truly positive. Hence, the true underlying

status of pool $\mathcal{G}_j$ is positive; i.e., $\tilde{Y}_{j1} = 1$. Then,

$$\begin{aligned}\text{pr}(Z_j = z \mid \mathcal{S}_{j1} = k_1, \mathcal{S}_{j0} = k_0) =& \text{pr}(Y_{jl} = 1 \mid \mathcal{S}_{j1} = k_1, \mathcal{S}_{j0} = k_0) \\ & \times \text{pr}(Y_{jl} = y_{jl}, l = 2, \ldots, c_j + 1 \mid \mathcal{S}_{j1} = k_1, \mathcal{S}_{j0} = k_0) \\ =& S_e \times S_e^{k_1}(1 - S_p)^{k-k_1}(1 - S_p)^{k_0} S_p^{c_j - k_0}.\end{aligned}$$

It is important to note that given individual covariates and the unknown parameters, T_{ij}, $i = 1, \ldots, c_j$, are independent Bernoulli random variables with probability of success being $\{p(X_{ij}^{\mathrm{T}}\beta)\}$. Moreover, we have $\mathcal{I}_{j1} \cap \mathcal{I}_{j0} = \emptyset$; i.e., one individual cannot be diagnosed as both negative and positive. We have that, given the unknown parameters, $\mathcal{S}_{j1} \mid \mathcal{X}_j$ and $\mathcal{S}_{j0} \mid \mathcal{X}_j$ are independent Poisson binomial random variables. Consequently, $\text{pr}\{\mathcal{S}_{j1} = k_1, \mathcal{S}_{j0} = k_0 \mid \mathcal{X}_j, \beta, p(\cdot)\} = \prod_{l=0}^{1} \text{pr}\{\mathcal{S}_{jl} = k_l \mid \mathcal{X}_j, \beta, p(\cdot)\}$ for any k_1 and k_2. Thus, $\mathcal{R}\{z; \mathcal{X}_j, \beta, p(\cdot)\}$ could be simplified as

$$\sum_{k_1=0}^{k} \sum_{k_0=0}^{c_j - k} \Bigg[S_e^{k_1 + I(k_1 + k_0 > 0)} (1 - S_e)^{k_0} S_p^{c_j - k - k_0} (1 - S_p)^{k - k_1 + I(k_1 + k_0 = 0)} \prod_{l=0}^{1} \text{pr}\{\mathcal{S}_{jl} = k_l \mid \mathcal{X}_j, \beta, p(\cdot)\} \Bigg], \tag{B.3}$$

It is worthwhile to point out that, the calculation of a Poisson binomial probability can be easily done through the method introduced in Wang (1993).

Now, we illustrate the calculation of $l_p(\mu)$. Note that $p_{j0} = \text{pr}[Z_j = \{(0, \mathcal{G}_j)\}]$ is the same as in (B.1), it suffices to calculate p_{j1k}. Since, herein we view T_{ij}s as identical and independent Bernoulli random variables with probability of success being $1 - \mu$, p_{j1k} is actually the probability of $Z_j = z$ for $z = \{(1, \mathcal{G}_j), (y_{jl}, \{l - 1\}), l = 2, \ldots, c_j + 1\} \in \mathcal{Z}(c_j)$ if $\sum_{l=2}^{c_j+1} y_{jl} = k$. The calculation of p_{j1k} simply follows (B.3), but one should replace $\text{pr}\{\mathcal{S}_{jl} = k_l \mid \mathcal{X}_j; \beta, p(\cdot)\}$ by $\text{pr}(\mathcal{S}_{jl} = k_l)$; i.e., covariate effects should not be considered. Thus, $\text{pr}(\mathcal{S}_{jl} = k_l)$ is a simple binomial probability statement. It leads us to

$$p_{j1k} = \delta_{c_j}(1 - S_p)^k S_p^{c_j - k} + S_e\{S_e + \delta_1\}^k \{1 - S_e - \delta_1\}^{c_j - k}.$$

Further, for $a_{ij}(\mu)$ and $b_{ij}(\mu)$, we have

$$\mathcal{F}_{ij}(0, \mu) = \text{pr}(Y_{j1} = 1, Y_{j,i+1} = 1 \mid T_{ij} = 0) = \qquad (1 - S_p)^2 \mu^{c_j - 1} + S_e(1 - S_p)(1 - \mu^{c_j - 1})$$

and

$$\mathcal{F}_{ij}(1,\mu) = \text{pr}(Y_{j1} = 1, Y_{j,i+1} = 1 \mid T_{ij} = 1) = S_e^2.$$

Consequently, $a_{ij}(\mu) = (1-S_p)^2\mu^{c_j-1} + S_e(1-S_p)(1-\mu^{c_j-1})$ and $b_{ij}(\mu) = S_e^2 - a_{ij}(\mu)$.

B.3 Expressions of Ω under master pool testing and Dorfman decoding

As presented in our manuscript, we have

$$\Omega_c = c^{-1}\sum_{z\in\mathcal{Z}(c)} E\left[\mathcal{R}^{-1}\left\{z;\mathcal{X}^{(c)},\beta_0,p_0(\cdot)\right\}\sum_{i=1}^{c}\{P_i(z,1,c)-P_i(z,0,c)\}^2 p_0'^2(X_i^{\text{T}}\beta_0)\Gamma(X_i)\right],$$

where $\mathcal{X}^{(c)} = (X_1,\ldots,X_c)^{\text{T}}$, $\Gamma(X) = \{X - E(X \mid X^{\text{T}}\beta_0\}\{X - E(X^{\text{T}} \mid X^{\text{T}}\beta_0)\}^{\text{T}}$, and $P_i(z,t,c) = \text{pr}\{Z = z \mid T_i = t, \mathcal{X}^{(c)},\beta_0,p_0(\cdot)\}$. Further $\Omega = \sum_{m=1}^{M}\gamma_m\Omega_{c^{(m)}}$. Note that, the only thing that remains unclear in calculating Ω_c is how to evaluate $P_i(z,t,c)$. A general formula could be derived through the use of the Law of Total Probability; i.e.,

$$\begin{aligned}
P_i(z,t,c) =& \text{pr}(Z = z \mid T_i = t, \mathcal{X}^{(c)},\beta_0,p_0(\cdot)) \\
=& \sum_{t\in\mathcal{T}(c)} \text{pr}(Z = z \mid T = t, T_i = t)\text{pr}\{T = t \mid \mathcal{X}^{(c)},\beta_0,p_0(\cdot)\} \\
=& \sum_{T\in\mathcal{T}(c_j);T_i=t} M(z,T,c_j)\prod_{i=1}^{c_j}\left[p_0(X_{ij}^{\text{T}}\beta_0)^{T_i}\{1-p_0(X_{ij}^{\text{T}}\beta_0)\}^{1-T_i}\right]. \qquad \text{(B.4)}
\end{aligned}$$

This formula is similar to the general expression of $\mathcal{R}\{z;\mathcal{X}^{(c)};\beta_0,p_0(\cdot)\}$ presented in § 2.1, i.e.,

$$\mathcal{R}\{z;\mathcal{X}^{(c)},\beta,p(\cdot)\} = \sum_{T\in\mathcal{T}(c_j)} M(z,T,c)\prod_{i=1}^{c_j}\left[p_0(X_{ij}^{\text{T}}\beta_0)^{T_i}\{1-p_0(X_{ij}^{\text{T}}\beta_0)\}^{1-T_i}\right]. \qquad \text{(B.5)}$$

However, these two expressions involve summation over the sample space $\mathcal{T}(c)$ which can be extremely large if c is large. As in § 2.2 and § 2.3 of our manuscript, these expressions can be greatly simplified under master pool testing and Dorfman decoding, respectively.

Under master pool testing, Z takes two forms $(1,\mathcal{G})$ or $(0,\mathcal{G})$ where $\mathcal{G} = \{1,\ldots,c\}$. As in the

manuscript, (B.5) could be simplified as

$$\begin{aligned}\mathcal{R}\{(0,\mathcal{G});\mathcal{X}^{(c)},\beta_0,p_0(\cdot)\} =&1-S_e-(1-S_e-S_p)\prod_{i=1}^{c}\{1-p_0(X_i^{\mathrm{T}}\beta_0)\},\\ \mathcal{R}\{(1,\mathcal{G});\mathcal{X}^{(c)},\beta_0,p_0(\cdot)\} =&S_e+(1-S_e-S_p)\prod_{i=1}^{c}\{1-p_0(X_i^{\mathrm{T}}\beta_0)\}.\end{aligned}$$

Further, we have (B.4) as

$$\begin{aligned}P_i\{(0,\mathcal{G}),1,c\} =&1-S_e,\\ P_i\{(0,\mathcal{G}),0,c\} =&S_p\prod_{r\neq i}\{1-p_0(X_r^{\mathrm{T}}\beta_0)\}+(1-S_e)\left[1-\prod_{r\neq i}\{1-p_0(X_r^{\mathrm{T}}\beta_0)\}\right]\\ =&(1-S_e)-(1-S_e-S_p)\prod_{r\neq i}\{1-p_0(X_r^{\mathrm{T}}\beta_0)\},\\ P_i\{(1,\mathcal{G}),1,c\} =&S_e,\\ P_i\{(1,\mathcal{G}),0,c\} =&(1-S_p)\prod_{r\neq i}\{1-p_0(X_r^{\mathrm{T}}\beta_0)\}+S_e\left[1-\prod_{r\neq i}\{1-p_0(X_r^{\mathrm{T}}\beta_0)\}\right]\\ =&S_e+(1-S_e-S_p)\prod_{r\neq i}\{1-p_0(X_r^{\mathrm{T}}\beta_0)\}.\end{aligned}$$

Under Dorfman decoding, for any $z\in\mathcal{Z}(c)$, z could either take the form of $z=(0,\mathcal{G})$ or $z=\{(1,\mathcal{G}),(y_l,\{l-1\}),l=2,\ldots,c+1\}$. When $z=(0,\mathcal{G})$, the calculation of $\mathcal{R}\{(0,\mathcal{G});\mathcal{X}^{(c)},\beta_0,p_0(\cdot)\}$ and $P_i\{(0,\mathcal{G}),t,c)\}$ for $t=0,1$, are the same as the above ones under master pool testing. For $z=\{(y_1=1,\mathcal{G}),(y_l,\{l-1\}),l=2,\ldots,c+1\}$, we have the $\mathcal{I}_1=\{i\in\mathcal{G}:y_{i+1}=1\}$ and $\mathcal{I}_0=\{i\in\mathcal{G}:y_{i+1}=0\}$. Denote $k=|\mathcal{I}_1|$, $c-k=|\mathcal{I}_0|$, $\mathcal{S}_1=\sum_{i\in\mathcal{I}_1}T_i$ and $\mathcal{S}_0=\sum_{i\in\mathcal{I}_0}T_i$. Then, similarly as in (3.8),

$$\begin{aligned}\mathcal{R}\{z;\mathcal{X}^{(c)},\beta_0,p_0(\cdot)\}=\sum_{k_1=0}^{k}\sum_{k_0=0}^{c-k}\Big[&S_e^{k_1+I(k_1+k_0>0)}(1-S_e)^{k_0}S_p^{c_j-k}(1-S_p)^{k-k_1+I(k_1+k_0=0)}\\ &\times\prod_{l=0}^{1}\mathrm{pr}\{\mathcal{S}_l=k_l\mid\mathcal{X}^{(c)},\beta_0,p_0(\cdot)\}\Big].\end{aligned}$$

For $P_i(z,t,c)$ for $t=0,1$, We first consider the case where $z=\{(1,\mathcal{G}),(0,\{i\}),i=1,\ldots,c\}$, i.e., the

master pool tests positive and all individuals retest negative. Then

$$
\begin{aligned}
P_i(z,1,c) =& (1-S_e)^2 \prod_{r\neq i}\{S_p + (1-S_e-S_p)p_0(X_r^{\mathrm{T}}\beta_0)\},\\
P_i(z,0,c) =& (1-S_e)S_p^c \prod_{r\neq i}\{1-p_0(X_r^{\mathrm{T}}\beta_0)\}\\
&+ S_e \sum_{k_0=1}^{c-1}(1-S_e)^{k_0}S_p^{c-k_0}\mathrm{pr}\left\{\sum_{r\neq i}T_r = k_0 \mid \mathcal{X}^{(c)},\beta_0,p_0(\cdot)\right\}.
\end{aligned}
$$

Similarly, when $z = \{(1,\mathcal{G}),(1,\{i\}), i = 1,\ldots,c\}$, i.e., the master pool tests positive and all individuals retest positive. Then

$$
\begin{aligned}
P_i(z,1,c) =& S_e^2 \prod_{r\neq i}\{1-S_p-(1-S_e-S_p)p_0(X_r^{\mathrm{T}}\beta_0)\},\\
P_i(z,0,c) =& (1-S_p)^{c+1} \prod_{r\neq i}\{1-p_0(X_{rj}^{\mathrm{T}}\beta_0)\}\\
&+ S_e \sum_{k_1=1}^{c-1} S_e^{k_1}(1-S_p)^{c-k_1}\mathrm{pr}\left\{\sum_{r\neq i}T_r = k_1 \mid \mathcal{X}^{(c)},\beta_0,p_0(\cdot)\right\}.
\end{aligned}
$$

For the remaining $z \in \mathcal{Z}(c)$, we consider two situations, i.e., $i \in \mathcal{I}_{j1}$ or $i \in \mathcal{I}_0$. When $i \in \mathcal{I}_1$,

$$
\begin{aligned}
P_i(z,1,c) =& S_e^2 \prod_{r\in\mathcal{I}_1\setminus\{i\}}\{1-S_p-(1-S_e-S_p)p_0(X_r^{\mathrm{T}}\beta_0)\}\\
&\times \prod_{r\in\mathcal{I}_{j0}}\{S_p+(1-S_e-S_p)p_0(X_r^{\mathrm{T}}\beta_0)\},\\
P_i(z,0,c) =& (1-S_p)^{k+1}S_p^{c-k}\prod_{r\neq i}\{1-p_0(X_r^{\mathrm{T}}\beta_0)\}\\
&+ S_e \sum_{k_1=0}^{k-1}\sum_{k_0=0}^{c-k}\Bigg[S_e^{k_1}(1-S_p)^{k-k_1}(1-S_e)^{k_0}S_p^{c-k_0}I(k_1+k_0>0)\\
&\times \mathrm{pr}\Bigg\{\sum_{r\in\mathcal{I}_1\setminus\{i\}}T_r = k_1 \mid \mathcal{X}^{(c)},\beta_0,p_0(\cdot)\Bigg\}\mathrm{pr}\Bigg\{\sum_{r\in\mathcal{I}_0}T_r = k_0 \mid \mathcal{X}^{(c)},\beta_0,p_0(\cdot)\Bigg\}\Bigg].
\end{aligned}
$$

And when $i \in \mathcal{I}_0$,

$$
\begin{aligned}
P_i(z,1,c) =& S_e(1-S_e) \prod_{r\in\mathcal{I}_1} \{1 - S_p - (1-S_e-S_p)p_0(X_r^{\mathrm{T}}\beta_0)\} \\
&\times \prod_{r\in\mathcal{I}_0\setminus\{i\}} \{S_p + (1-S_e-S_p)p_0(X_r^{\mathrm{T}}\beta_0)\}, \\
P_i(z,0,c) =& (1-S_p)^{k+1} S_p^{c-k} \prod_{r\neq i} \{1 - p_0(X_r^{\mathrm{T}}\beta_0)\} \\
&+ S_e \sum_{k_1=0}^{k} \sum_{k_0=0}^{c-1-k} \Bigg[S_e^{k_1}(1-S_p)^{k-k_1}(1-S_e)^{k_0} S_p^{c-k_0} I(k_1+k_0>0) \\
&\times \mathrm{pr}\Big\{ \sum_{r\in\mathcal{I}_1} T_r = k_1 \mid \mathcal{X}^{(c)}, \beta_0, p_0(\cdot) \Big\} \mathrm{pr}\Big\{ \sum_{r\in\mathcal{I}_0\setminus\{i\}} T_r = k_0 \mid \mathcal{X}^{(c)}, \beta_0, p_0(\cdot) \Big\} \Bigg].
\end{aligned}
$$

B.4 A plug-in estimator of Σ

For any $z \in \mathcal{Z}(c_j)$, we define

$$
\begin{aligned}
\Omega\{z; \mathcal{X}_j, \beta_0, p_0(\cdot), p_0'(\cdot), d_{\beta_0}(\cdot)\} = \mathcal{R}^{-2}\{z; \mathcal{X}_j, \beta_0, p_0(\cdot)\} \sum_{i=1}^{c_j} \Big[\Delta_i^2\{z; \mathcal{X}_j, \beta_0, p_0(\cdot)\} p_0'^2(X_{ij}^{\mathrm{T}}\beta_0) \\
\{X - d_{\beta_0}(X^{\mathrm{T}}\beta_0)\}\{X - d_{\beta_0}(X^{\mathrm{T}}\beta_0)\}^{\mathrm{T}} \Big],
\end{aligned}
$$

where $\Delta_i\{Z_j; \mathcal{X}_j, \beta_0, p_0(\cdot)\} = \mathrm{pr}\{Z_j = z \mid T_{ij} = 1, \mathcal{X}_j, \beta_0, p_0(\cdot)\} - \mathrm{pr}\{Z_j = z \mid T_{ij} = 0, \mathcal{X}_j, \beta_0, p_0(\cdot)\}$. Through an application of the law of large numbers, it is easy to see that

$$
N^{-1} \sum_{j=1}^{J} \Omega\{Z_j; \mathcal{X}_j, \beta_0, p_0(\cdot), p_0'(\cdot), d_{\beta_0}(\cdot)\} = \Omega + o_p(1). \tag{B.6}
$$

Thus, we define our estimator of Ω as

$$
\hat{\Omega} = N^{-1} \sum_{j=1}^{J} \Omega\{Z_j; \mathcal{X}_j, \hat{\beta}, \hat{p}_{\hat{\beta}}(\cdot), \hat{p}'_{\hat{\beta}}(\cdot), \hat{d}_{\hat{\beta}}(\cdot)\}, \tag{B.7}
$$

where $\hat{\beta}$ is our estimator of β_0; $\hat{p}_\beta(u)$ and $\hat{p}'_\beta(u)$ are estimators of $p_\beta(u)$ and $p'_\beta(u)$, respectively, as defined in the manuscript; and $\hat{d}_\beta(u)$ is an estimator of $d_\beta(u)$ and will be defined later.

Our estimator of $d_\beta(x^{\mathrm{T}}\beta)$ is defined as

$$\hat{d}_\beta(u) = \frac{\hat{D}_{N0}(u,\beta)\hat{S}_{N2}(u,\beta) - \hat{D}_{N1}(u,\beta)\hat{S}_{N1}(u,\beta)}{\hat{S}_{N2}(u,\beta)\hat{S}_{N0}(u,\beta) - \hat{S}^2_{N1}(u,\beta)}, \tag{B.8}$$

where

$$\hat{D}_{Nl}(u,\beta) = N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j} b^2_{ij}(\hat{\mu})X_{ij}\mathcal{K}_h(X^{\mathrm{T}}_{ij}\beta, u; l).$$

According to the proof of Lemma B.4, we have

$$\begin{aligned}
\sup_{x\in\mathbb{X}} |\hat{p}_{\hat{\beta}}(x^{\mathrm{T}}\hat{\beta}) - p_0(x^{\mathrm{T}}\beta_0)| &= o_p(1),\\
\sup_{x\in\mathbb{X}} |\hat{p}'_{\hat{\beta}}(x^{\mathrm{T}}\hat{\beta}) - p'_0(x^{\mathrm{T}}\beta_0)| &= o_p(1),\\
\text{and } \sup_{x\in\mathbb{X}} \|\hat{d}_{\hat{\beta}}(x^{\mathrm{T}}\hat{\beta}) - d_{\beta_0}(x^{\mathrm{T}}\beta_0)\| &= o_p(1).
\end{aligned}$$

Consequently, we have

$$\hat{\Omega} = N^{-1}\sum_{j=1}^{J}\Omega\{Z_j; \mathcal{X}_j, \beta_0, p_0(\cdot), p'_0(\cdot), d_{\beta_0}(\cdot)\} + o_p(1).$$

Combining with (B.6) and (B.7), it proves that $\hat{\Omega}$ is a consistent estimator of Ω.

We then estimate $\mathcal{J}_0$ by $\hat{\mathcal{J}}_0$ where $\hat{\mathcal{J}}_0$ is the functional value of $\partial B(\beta^{(1)})\partial\beta^{(1)}$ at $\beta^{(1)} = (\hat{\beta}_2, \ldots, \hat{\beta}_p)^{\mathrm{T}}$. By the consistency of $\hat{\beta}$, it is easy to see that $\hat{\mathcal{J}}_0$ converges in probability to $\mathcal{J}_0$ as $N \to \infty$. Finally, we obtain our consistent plug-in estimator of Σ as

$$\hat{\Sigma} = \hat{\mathcal{J}}_0(\hat{\mathcal{J}}^{\mathrm{T}}_0\hat{\Omega}\hat{\mathcal{J}}_0)^{-1}\hat{\mathcal{J}}^{\mathrm{T}}_0.$$

B.5 Simulation results under master pool testing

Table B.1 summarized the behavior of our 500 estimators of β_0 under master pool testing for Model 3.1–3.3 when $\delta = 0.1$. From these results, we see that the estimates of β_0 are generally on target and exhibit little evidence of bias. As c becomes larger, testing expenditure reduces significantly. However, as a trade-off, the estimation efficiency of $\hat{\beta}$ decreases and the variability of estimating the link increases. This phenomenon is expected since the number of pool responses on which the estimates are based, $J = N/c$, decreases as c

Table B.1: Summary of simulation results for data arising from Dorfman decoding: BIAS and SD, empirical bias ($\times 10^3$) and standard deviation ($\times 100$) of the 500 estimates; SE, average standard error ($\times 100$); COV, empirical coverage probability ($\times 100$) for nominal 95% confidence interval; EMSE, average mean squared error of prediction ($\times 10^4$); RE, ratio of EMSE of the parametric model to the EMSE of our semiparametric model.

	Parameter	Measure	$c = 1$	$c = 2$	$c = 5$	$c = 10$
Model 3.1	β_{01}	BIAS(SD)	8.7(3.5)	14.4(4.1)	34.2(6.2)	59.7(9.8)
		COV(SE)	93.6(3.5)	93.7(4.3)	92.2(6.6)	89.5(10.0)
	β_{02}	BIAS(SD)	−5.1(1.4)	−8.3(1.7)	−21.5(2.9)	−46.2(5.8)
		COV(SE)	96.0(1.4)	95.5(1.8)	95.8(3.1)	93.6(5.5)
	β_{03}	BIAS(SD)	−1.9(5.2)	−4.1(6.7)	−0.7(10.6)	6.1(16.4)
		COV(SE)	94.4(5.3)	92.9(6.4)	90.0(9.5)	88.9(14.7)
	$p_0(x\beta_0)$	EMSE(RE)	1.31(0.37)	1.77(0.34)	4.14(0.31)	9.74(0.29)
	Percentage reduction in testing			50.0 %	80.0 %	90.0 %
Model 3.2	β_{01}	BIAS(SD)	1.5(1.4)	3.5(1.9)	9.1(3.6)	24.7(7.1)
		COV(SE)	93.0(1.4)	95.3(2.0)	94.9(3.7)	92.9(6.8)
	β_{02}	BIAS(SD)	−1.2(0.6)	−2.3(0.8)	−8.0(1.6)	−21.4(3.8)
		COV(SE)	93.4(0.6)	96.2(0.8)	96.4(1.7)	95.2(3.3)
	β_{03}	BIAS(SD)	−0.7(3.4)	−2.4(4.4)	6.4(7.7)	−11.5(13.5)
		COV(SE)	93.0(3.0)	90.2(3.9)	92.5(7.1)	88.8(12.2)
	$p_0(x\beta_0)$	EMSE(RE)	1.25(25.33)	2.37(13.09)	6.40(5.39)	16.53(2.32)
	Percentage reduction in testing			50.0 %	80.0 %	90.0 %
Model 3.3	β_{01}	BIAS(SD)	7.6(2.5)	12.0(3.1)	27.4(5.4)	54.9(10.1)
		COV(SE)	92.4(2.5)	93.6(3.4)	93.1(5.8)	91.7(9.6)
	β_{02}	BIAS(SD)	−3.7(1.0)	−6.5(1.3)	−17.5(2.9)	−45.3(7.4)
		COV(SE)	93.8(1.0)	94.4(1.4)	95.9(2.7)	94.7(5.7)
	β_{03}	BIAS(SD)	−1.7(3.7)	0.1(5.4)	−2.9(10.2)	−1.7(16.8)
		COV(SE)	92.4(3.6)	92.9(5.1)	91.2(8.8)	88.4(14.3)
	$p_0(x\beta_0)$	EMSE(RE)	1.61(13.80)	2.73(8.26)	6.06(3.89)	14.39(1.88)
	Percentage reduction in testing			50.0 %	80.0 %	90.0 %

increases. It is not surprising that the standard deviation of the estimates tends to increase with the pool size. This also affects our estimator of the covariance matrix Σ. Consequently, when c increases, the estimated 95% coverage decreases.

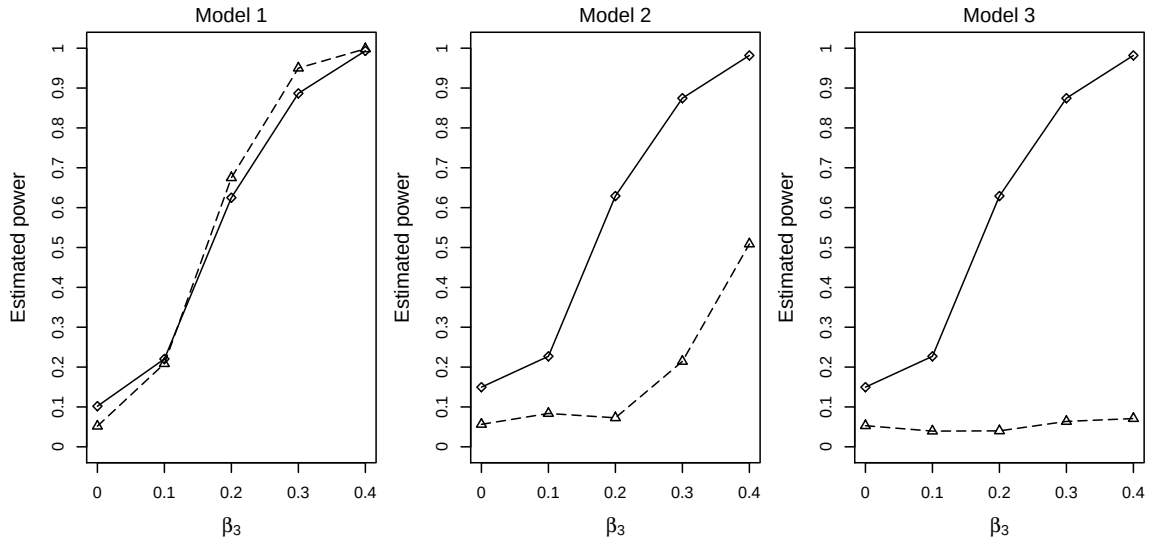


Figure B.1: Estimated power curves under master pool testing: the solid and dashed curves correspond to our approach and the parametric techniques, respectively.

Figure B.1 presents the estimated power curves corresponding to data arising from master pool testing when $c = 5$. From these results, one will first notice that under Model 3.1 the estimated power curves for our semiparametric approach and the competing parametric modeling are very similar, with the parametric model having slightly more power. This trend is similar as the one under Dorfman decoding. It suggests that our methodology performs almost as well as the "oracle" approach (i.e., the parametric model which assume the correct link function). On the other hand, if the link function is misspecified under the parametric model (e.g. see Models 3.2 and 3.3), these methods lose the power to detect significant predictor variables, a feature not shared by our approach; a same trend observed under Dorfman decoding. However, unlike Dorfman decoding, master pool testing cannot gain decoding information from positive pools. It greatly affects the accuracy in the size study. Thus, if one prefers a more accurate estimator, collecting data through Dorfman decoding may be a better choice than solely testing master pools.

B.6 Proof of Theorem 3.3.1

B.6.1 A Brief description of the proofs

In the following, we denote $a_N = O_P(b_N)$ if a_N/b_N is bounded in probability, $a_N = o_P(b_N)$ if a_N/b_N converges to zero in probability. Since the function $B(\beta^{(1)}) = \beta$ is a one-to-one mapping from $\mathcal{B}^{(1)} = \{\beta^{(1)} \in \mathbb{R}^{p-1} : \|\beta^{(1)}\| < 1\}$ to $\mathcal{B}$, $\hat{\beta}$ can be viewed as $\hat{\beta} = B(\hat{\beta}^{(1)})$ where $\hat{\beta}^{(1)}$ is the maximizer of $l\{B(\beta^{(1)}), \hat{p}_{B(\beta^{(1)})}(\cdot)\}$ in $\mathcal{B}^{(1)}$. Denote $\hat{G}(\beta^{(1)})$ as the partial derivative of $l\{B(\beta^{(1)}), \hat{p}_{B(\beta^{(1)})}(\cdot)\}$ with respect to $\beta^{(1)}$. It could be written as

$$\hat{G}(\beta^{(1)}) = \mathcal{J}_\beta^{\mathrm{T}} \sum_{j=1}^{J} \mathcal{R}^{-1}\{Z_j; \mathcal{X}_j, \beta, \hat{p}_\beta(\cdot)\} \sum_{i=1}^{c_j} \Delta_i \{Z_j; \mathcal{X}_j, \beta, \hat{p}_\beta(\cdot)\} \hat{p}_\beta^{(1)}(X_{ij}^{\mathrm{T}}\beta).$$

where $\mathcal{J}_\beta = \partial B(\beta^{(1)})/\partial\beta^{(1)}$, $\Delta_i \{z_j; \mathcal{X}_j, \beta, \hat{p}_\beta(\cdot)\} = \mathrm{pr}\{Z_j = z_j \mid T_{ij} = t, \mathcal{X}_j, \beta, \hat{p}_\beta(\cdot)\} - \mathrm{pr}\{Z_j = z_j \mid T_{ij} = 0, \mathcal{X}_j, \beta, \hat{p}_\beta(\cdot)\}$ and $\hat{p}_\beta^{(1)}(X^{\mathrm{T}}\beta) = \partial\hat{p}_\beta(X^{\mathrm{T}}\beta)/\partial\beta$. An asymptotically equivalent version of $\hat{G}$ could be written as

$$G(\beta^{(1)}) = \mathcal{J}_\beta^{\mathrm{T}} \sum_{j=1}^{J} \mathcal{R}^{-1}\{Z_j; \mathcal{X}_j, \beta, p_\beta(\cdot)\} \sum_{i=1}^{c_j} \Big[\Delta_i \{Z_j; \mathcal{X}_j, \beta, p_\beta(\cdot)\} \\ p'_\beta(X_{ij}^{\mathrm{T}}\beta) \left\{X_{ij} - d_\beta(X_{ij}^{\mathrm{T}}\beta)\right\}\Big].$$

We derive that

$$\sup_{X\in\mathcal{X}, \beta^{(1)}\in\mathcal{B}_N^{(1)}} |\hat{p}_\beta(X^{\mathrm{T}}\beta) - p_0(X^{\mathrm{T}}\beta_0)| = O_p(\{\log N/(Nh)\}^{1/2}) \tag{B.9}$$

and

$$\sup_{\beta^{(1)}\in\mathcal{B}_N^{(1)}} \|\hat{G}(\beta^{(1)}) - G(\beta_0^{(1)}) + N\mathcal{J}_0^{\mathrm{T}}\Omega\mathcal{J}_0(\beta - \beta_0)\| = o_p(N^{1/2}), \tag{B.10}$$

where $\mathcal{B}_N^{(1)} = \{\beta^{(1)} \in \mathcal{B}^{(1)} : \|\beta^{(1)} - \beta_0^{(1)}\| \le CN^{-1/2}\}$ for some constant $C > 0$. Equation (B.10) implies that

$$\left\|\hat{\beta}^{(1)} - \beta_0^{(1)}\right\| = O_p(N^{-1/2}), \tag{B.11}$$

Consequently,

$$\hat{G}(\hat{\beta}^{(1)}) = G(\beta_0^{(1)}) - \mathcal{J}_0^{\mathrm{T}}\Omega\mathcal{J}_0(\hat{\beta}^{(1)} - \beta_0^{(1)}) + o_p(N^{1/2}).$$

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

$$N^{1/2}(\hat{\beta}^{(1)} - \beta_0^{(1)}) = (\mathcal{J}_0^{\mathrm{T}}\Omega\mathcal{J}_0)^{-1}\{N^{1/2}G(\beta_0^{(1)})\} + o_p(1).$$

Then, the asymptotic normality of $\hat{\beta}$ follows Central Limit Theorem and Slutsky's Theorem applied to the righthand side of the above equation. Combining (B.9) with (B.11) gives

$$\sup_{X\in\mathbb{X}} \left|\hat{p}(X^{\mathrm{T}}\hat{\beta}) - p_0(X^{\mathrm{T}}\beta_0)\right| = O_p(\{\log N/(Nh)\}^{1/2}),$$

which completes the proof of Theorem 3.3.1.

In the next section we prove equations (B.9), (B.10), and (B.11). Lemmas B.1-B.3 below are used to obtain the bounds for the centralized rth moments of $\hat{q}_\beta$ and $\hat{q}_\beta^{(1)}$ given in Propositions B.1-B.2. These two propositions are then used to obtain Lemma B.4 which proves (B.9). Then combining (B.9) with Lemma B.5 below, we prove (B.10) in Proposition B.3. Finally, we show (B.11) in Lemma B.6.

B.6.2 Detailed proofs

Before proceeding to the detailed proofs, we would like to introduce some notation. 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 $a_N/b_N = O(1)$; $a_N \overset{a.s.}{\to} a$ if a_N converges almost surely to a; and $a_N = \mathcal{O}_r(b_N)$, if $E(|a_N|^r) = O(b_N^r)$. $E_T(X)$ denotes the conditional expectation of X given T. By Cauchy-Schwartz inequality, we have $\mathcal{O}_r(a_N)\mathcal{O}_r(b_N) = \mathcal{O}_{r/2}(a_N b_N)$. We further denote the summation over all the groups with size $c^{(K)}$ by $\sum_{|j|=c^{(K)}}$. Then $\sum_{j=1}^{J}\sum_{i=1}^{c_j}$ can be written as $\sum_{m=1}^{M}\sum_{i=1}^{c^{(K)}}\sum_{|j|=c^{(K)}}$. A term of the form $\sum_{|j|=c^{(K)}} A_j$ means A_js are from groups of size $c^{(K)}$. For example, in $\sum_{|j|=c^{(K)}} D_{ij}$, D_{ij} is the diagnosis result of the ith individual in a group of size $c^{(K)}$.

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

$$E\left(\left|\sum_{i=1}^{n} X_i\right|^r\right) \simeq \sum_{i=1}^{n} E\left(|X_i|^r\right) + \left|\sum_{i=1}^{n} E(X_i)\right|^r + \left\{\sum_{i=1}^{n} E(X_i^2)\right\}^{r/2}. \tag{B.12}$$

For the proof of (B.12) we refer to Petrov (1995).

Lemma B.1. *Under Condition 3.4, we have* $\hat{\mu} \xrightarrow{a.s.} \mu_0$ *as* $N \to \infty$*, and for* $r \geq 2$*,*

$$E\left\{\sup_{ij} |\eta_{ij}(\hat{\mu}) - \eta_{ij}(\mu_0)|^r\right\} = O(N^{-r/2})$$

if $\sup_{ij} \sup_{u\in[0,1]} |\eta'_{ij}(u)|$ *is bounded.*

Proof. Condition 3.4 guarantees that $\text{pr}(Z_j = z)$, for any $z \in \mathcal{Z}(c_j)$, are bounded away from 0 when $\mu \in [0,1]$. By the uniform law of large number, $l(\mu)$ converges almost surely to $E[l(\mu)]$ uniformly in $\mu \in [0,1]$. Consequently, $\hat{\mu} = \arg\max_\mu l(\mu) \xrightarrow{a.s.} \mu_0 = \arg\max_\mu E[l(\mu)]$. To show the rate of rth moment convergence, we notice that

$$\hat{\mu} - \mu_0 = \{-N^{-1}l''(\bar{\mu})\}^{-1} N^{-1} l'(\mu_0)$$

where $\bar{\mu}$ is between $\hat{\mu}$ and μ_0. By $\hat{\mu} \xrightarrow{a.s.} \mu_0$, we have that $-N^{-1}l''(\bar{\mu})$ converges almost surely to a positive number $\mathcal{I}(\mu_0)$, i.e., when N is large, $\{-N^{-1}l''(\bar{\mu})\}^{-1}$ is bounded almost surely. On the other hand, we have $N^{-1/2}l'(\mu_0)$ converges in distribution to $N(0, \mathcal{I}(\mu_0)^{-1})$. By the continuous mapping theorem, $|N^{-1/2}l'(\mu_0)|^r$ converges in distribution to $|N(0,\mathcal{I}(\mu_0)^{-1})|^r$. Hence,

$$E\,|\hat{\mu} - \mu_0|^r = O(N^{-r/2}).$$

Then the moment convergence rate on η_{ij} follows through a Taylor expansion. □

Note that, Lemma B.1 holds for $a_{ij}(\cdot)$, $b_{ij}(\cdot)$, $a_{ij}(\cdot)b_{ij}(\cdot)$, and $b^2_{ij}(\cdot)$.

Lemma B.2. *For any* $r \geq 2$*, if* $\eta_{ij}(\cdot)$ *and* ω_{ij} *satisfy that* $E\{\sup_{ij} |\eta_{ij}(\hat{\mu}) - \eta_{ij}(\mu_0)|^{2r}\} = O(N^{-r})$ *and* $N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j} |\omega_{ij}| = \mathcal{O}_{2r}(w_N)$*, then*

$$\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{c_j} [\{\eta_{ij}(\hat{\mu}) - \eta_{ij}(\mu_0)\}\omega_{ij}] = \mathcal{O}_r(N^{-1/2}w_N).$$

Proof. Using the Cauchy-Schwartz inequality, we have

$$
\begin{aligned}
& E\left(\left|\frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{c_j}[\{\eta_{ij}(\hat{\mu})-\eta_{ij}(\mu_0)\}\omega_{ij}]\right|^r\right) \\
\leq & \left[E\left\{\sup_{ij}|\eta_{ij}(\hat{\mu})-\eta_{ij}(\mu_0)|^{2r}\right\}\times E\left\{\left(N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}|\omega_{ij}|\right)^{2r}\right\}\right]^{1/2} \\
= & O(N^{-r/2}).
\end{aligned}
$$

□

Lemma B.3. *Under Conditions 3.1–3.4, for any $\beta \in \mathcal{B}$, we have*

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

for some constant C, and further

$$
\sup_{x\in\mathbb{X},\beta\in\mathcal{B}}|\hat{p}_\beta(x^{\mathrm{T}}\beta)-p_\beta(x^{\mathrm{T}}\beta)| \overset{a.s.}{\to} 0.
$$

Proof. Explicit expressions of $\hat{p}_\beta(u)$ and $\hat{p}'_\beta(u)$ are provided in (3.5) and (3.6), respectively. Replacing $\hat{\mu}$ with μ_0, we denote

$$
\begin{aligned}
T_{Nl}(u,\beta) =& N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}\{D_{ij}-a_{ij}(\mu_0)\}b_{ij}(\mu_0)\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,u;l), \\
S_{Nl}(u,\beta) =& N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}b_{ij}^2(\mu_0)\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,u;l).
\end{aligned}
$$

Since D_{ij}s are correlated but independent across j, we could rewrite $T_{Nl}(x^{\mathrm{T}}\beta,\beta)$ as

$$
\begin{aligned}
T_{Nl}(x^{\mathrm{T}}\beta,\beta) &= \sum_{m=1}^{M}\frac{c^{(K)}J_m}{N}\cdot\frac{1}{c^{(K)}}\sum_{i=1}^{c^{(K)}}T_{Nlmi}(x^{\mathrm{T}}\beta,\beta), \\
S_{Nl}(x^{\mathrm{T}}\beta,\beta) &= \sum_{m=1}^{M}\frac{c^{(K)}J_m}{N}\cdot\frac{1}{c^{(K)}}\sum_{i=1}^{c^{(K)}}S_{Nlmi}(x^{\mathrm{T}}\beta,\beta),
\end{aligned}
$$

where $T_{Nlmi}=\sum_{|j|=c^{(K)}}T_{Nlmij}, T_{Nlmij}(x^{\mathrm{T}}\beta,\beta)=J_m^{-1}\{D_{ij}-a_{ij}(\mu_0)\}b_{ij}(\mu_0)\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,x^{\mathrm{T}}\beta;l)$, $S_{Nlmi}=$

$\sum_{|j|=c^{(K)}} S_{Nlmij}$, and $S_{Nlmij}(x^{\mathrm{T}}\beta,\beta) = J_m^{-1} b_{ij}^2(\mu_0)\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta, x^{\mathrm{T}}\beta; l)$. Using (B.12),

$$E_{X^{\mathrm{T}}\beta}\left[T_{Nlmi(X^{\mathrm{T}}\beta,\beta)} - E_{X^{\mathrm{T}}\beta}\{T_{Nlmi(X^{\mathrm{T}}\beta,\beta)}\}\right]^r \simeq \sum_{|j|=c^{(K)}} E\left(|T_{Nlmij}|^r\right) + \left\{\sum_{|j|=c^{(K)}} E(T_{Nlmij}^2)\right\}^{r/2}.$$

Noting that $J_m \simeq N$, we have

$$\begin{aligned} E_{X^{\mathrm{T}}\beta}\left(|T_{Nlmij}|^r\right) =& \frac{1}{J_m^r h^r}\int [\{D_{ij} - a_{ij}(\mu_0)\}b_{ij}(\mu_0)]^r K^r\left(\frac{u - X^{\mathrm{T}}\beta}{h}\right)\left(\frac{u - X^{\mathrm{T}}\beta}{h}\right)^{lr} f_{X^{\mathrm{T}}\beta}(u)du \\ =& O(N^{-r}h^{1-r}). \end{aligned}$$

Consequently, $E_{X^{\mathrm{T}}\beta}\left[T_{Nlmi(X^{\mathrm{T}}\beta,\beta)} - E_{X^{\mathrm{T}}\beta}\{T_{Nlmi(X^{\mathrm{T}}\beta,\beta)}\}\right]^r = O(N^{1-r}h^{1-r}) + O(N^{-r/2}h^{-r/2}) = O(h^{2r})$. Therefore,

$$T_{Nlmi}(X^{\mathrm{T}}\beta,\beta) = E_{X^{\mathrm{T}}\beta}\left\{T_{Nlmij}(X^{\mathrm{T}}\beta,\beta)\right\} + \mathcal{O}_r(h^2), \tag{B.13}$$

$$S_{Nlmi}(X^{\mathrm{T}}\beta,\beta) = E_{X^{\mathrm{T}}\beta}\left\{S_{Nlmij}(X^{\mathrm{T}}\beta,\beta)\right\} + \mathcal{O}_r(h^2). \tag{B.14}$$

For any X being independent with X_{ij}s or being one of the X_{ij}s, based on (B.12) and by the boundedness of $\mathbb{X}$ and $\mathcal{B}$,

$$\begin{aligned} & E\left[\left\{N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}\left|D_{ij}\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta, X^{\mathrm{T}}\beta; l)\right|\right\}^{2r}\right] \\ & \simeq E_{X^{\mathrm{T}}\beta}\left[\left\{N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}\left|\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta, X^{\mathrm{T}}\beta; l)\right|\right\}^{2r}\right] \\ \simeq & \sum_{j=1}^{J}\sum_{i=1}^{c_j} E_{X^{\mathrm{T}}\beta}\left\{\left|N^{-1}\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta, X^{\mathrm{T}}\beta; l)\right|^r\right\} + \left|E_{X^{\mathrm{T}}\beta}\left\{|\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta, X^{\mathrm{T}}\beta; l)|\right\}\right|^r \\ & + \left[\sum_{j=1}^{J}\sum_{i=1}^{c_j} E\left\{\left|N^{-1}\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta, X^{\mathrm{T}}\beta; l)\right|^2\right\}\right]^{r/2} \\ = & O(N^{1-r}h^{1-r}) + O(1) + O(N^{-r/2}h^{-r/2}) = O(1). \end{aligned}$$

Thus, by Lemma B.2, we have

$$\hat{T}_{Nl}(X^{\mathrm{T}}\beta,\beta) = T_{Nl}(X^{\mathrm{T}}\beta,\beta) + \mathcal{O}_r(N^{-1/2}),$$
$$\hat{S}_{Nl}(X^{\mathrm{T}}\beta,\beta) = S_{Nl}(X^{\mathrm{T}}\beta,\beta) + \mathcal{O}_r(N^{-1/2}).$$

Combining these with (B.13) and (B.14) provides that

$$\hat{T}_{Nl}(X^{\mathrm{T}}\beta,\beta) = E_{X^{\mathrm{T}}\beta}\{T_{Nl}(X^{\mathrm{T}}\beta,\beta)\} + \mathcal{O}_r(h^2),$$
$$\hat{S}_{Nl}(X^{\mathrm{T}}\beta,\beta) = E_{X^{\mathrm{T}}\beta}\{S_{Nl}(X^{\mathrm{T}}\beta,\beta)\} + \mathcal{O}_r(h^2).$$

Then, similar to the proof of expression (A.10) in Zhu & Xue (2006), we have

$$\sup_{x\in\mathbb{X},\beta\in\mathcal{B}} \left| \hat{T}_{Nl}(x^{\mathrm{T}}\beta,\beta) - \left\{ \sum_{m=1}^{M} \frac{\gamma_m}{c^{(K)}} \sum_{i=1}^{c^{(K)}} b_{ij}^2(\mu_0) \right\} p_\beta(x^{\mathrm{T}}\beta) f_\beta(x^{\mathrm{T}}\beta)\pi_l \right| \overset{a.s.}{\to} 0,$$
$$\sup_{x\in\mathbb{X},\beta\in\mathcal{B}} \left| \hat{S}_{Nl}(x^{\mathrm{T}}\beta,\beta) - \left\{ \sum_{m=1}^{M} \frac{\gamma_m}{c^{(K)}} \sum_{i=1}^{c^{(K)}} b_{ij}^2(\mu_0) \right\} f_\beta(x^{\mathrm{T}}\beta)\pi_l \right| \overset{a.s.}{\to} 0,$$

where $f_{X^{\mathrm{T}}\beta}$ is the density of $X^{\mathrm{T}}\beta$ and $\pi_l = \int K(t)t^l dt$. Finally, the proof follows Condition 3.2. □

Proposition B.1. *Under Conditions 3.1–3.4, we have, for any $\beta \in \mathcal{B}$ and $r \geq 2$,*

$$\hat{p}_\beta(X_{ij}^{\mathrm{T}}\beta) = p_\beta(X_{ij}^{\mathrm{T}}\beta) + \mathcal{O}_r(h^2)$$

and

$$\hat{p}'_\beta(X_{ij}^{\mathrm{T}}\beta) = p'_\beta(X_{ij}^{\mathrm{T}}\beta) + \mathcal{O}_r(h),$$

over all (i,j)s.

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

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

where

$$\hat{H}_{Nl}(u,\beta) = N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}\big[\{D_{ij}-a_{ij}(\hat{\mu})\}b_{ij}(\hat{\mu}) - b_{ij}^2(\hat{\mu})p_\beta(u) \\ -b_{ij}^2(\hat{\mu})p'_\beta(u)(X_{ij}^{\mathrm{T}}\beta - u)\big]\,\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta, u; l).$$

Using the results presented in Lemma B.3, it suffices to show that $\hat{H}_{Nl}(X^{\mathrm{T}}\beta,\beta) = \mathcal{O}_s(h^l)$ where $s = 2r$. Similarly as the proof in Lemma B.3, it can be shown that

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

where $H_{Nl}(u,\beta)$ is the version of $\hat{H}_{Nl}(u,\beta)$ by replacing $\hat{\mu}$ with μ_0. Thus, it leaves us to show that $H_{Nl}(X^{\mathrm{T}}\beta,\beta) = \mathcal{O}_s(h^l)$. To this end, we rewrite it as

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

where $H_{Nlmi}(X^{\mathrm{T}}\beta,\beta) = \sum_{|j|=c^{(K)}} H_{Nlmij}$ with $H_{Nlmij} = J_m^{-1}\{D_{ij} - a_{ij}(\mu_0)\}b_{ij}(\mu_0) - b_{ij}^2(\mu_0)p_\beta(X^{\mathrm{T}}\beta) - b_{ij}^2(\mu_0)p'_\beta(X^{\mathrm{T}}\beta)(X_{ij}^{\mathrm{T}}\beta - X^{\mathrm{T}}\beta)\}\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta, X^{\mathrm{T}}\beta; l)$. By (B.12), for $s = 2r \geq 2$, we have

$$E_{X^{\mathrm{T}}\beta}\left\{|H_{Nlmi}(X^{\mathrm{T}}\beta,\beta)|^s\right\} \simeq \left|\sum_{|j|=c^{(K)}} E_{X^{\mathrm{T}}\beta}\left\{H_{Nlmij}(X^{\mathrm{T}}\beta,\beta)\right\}\right|^s \tag{B.15}$$

$$+\sum_{j=1}^{J_m} E_{X^{\mathrm{T}}\beta}\left\{|H_{Nlmij}(X^{\mathrm{T}}\beta,\beta)|^s\right\} + \left[\sum_{j=1}^{J_m} E_{X^{\mathrm{T}}\beta}\left\{H_{Nlmij}^2(X^{\mathrm{T}}\beta,\beta)\right\}\right]^{r/2}. \tag{B.16}$$

Simple Taylor expansion provides that $\sum_{|j|=c^{(K)}} E_{X^{\mathrm{T}}\beta}\left\{H_{Nlmij}(X^{\mathrm{T}}\beta,\beta)\right\} = O(h^2)$ which implies that the term (B.15) is also of order $O(h^{2s})$. Further, note that

$$\begin{aligned} E_{X^{\mathrm{T}}\beta}\left(|H_{Nlmij}|^s\right) =& \frac{h}{J_m^s h^s}\int[\{D_{ij}-a_{ij}(\mu_0)\}b_{ij}(\mu_0) - b_{ij}^2(\mu_0)p_\beta(u) - b_{ij}^2(\mu_0)p'_\beta(u)(u - X^{\mathrm{T}}\beta)]^s \\ &\times h^{-1}K^s\left(\frac{u - X^{\mathrm{T}}\beta}{h}\right)\left(\frac{u - X^{\mathrm{T}}\beta}{h}\right)^{ls} f_{X^{\mathrm{T}}\beta}(u)du \\ =& O(N^{-s}h^{1-s}). \end{aligned}$$

Therefore, the term (B.16) is of order $O(h^{2s})$. Consequently, $E_{X^{\mathrm{T}}\beta}\left\{|H_{Nlmi}(X^{\mathrm{T}}\beta,\beta)|^s\right\} = O(h^{2s})$. More-

over, by the boundedness of $\mathbb{X}$ and $\mathcal{B}$, we can conclude that

$$H_{Nlmi}(X^{\mathrm{T}}\beta, \beta) = \mathcal{O}_s(h^2).$$

which completes the proof. □

Proposition B.2. *Let $\mathcal{B}_N^{(1)} = \{\beta^{(1)} \in \mathcal{B}^{(1)} : \|\beta^{(1)} - \beta_0^{(1)}\| \leq CN^{-1/2}\}$ for some constant $C > 0$. Under Conditions 3.1–3.4, for any $\beta^{(1)} \in \mathcal{B}_N^{(1)}$ and $r \geq 2$, we have*

$$\left\| \hat{p}_\beta^{(1)}(X_{ij}^{\mathrm{T}}\beta) - p'_\beta(X_{ij}^{\mathrm{T}}\beta)\left(X_{ij} - d_\beta(X_{ij}^{\mathrm{T}}\beta)\right) \right\| = \mathcal{O}_r(h)$$

over all (i, j)s.

Remark B.1. *This proposition indicates that $\partial p_\beta(X^{\mathrm{T}}\beta)/\partial\beta \neq Xp'_\beta(X^{\mathrm{T}}\beta)$, which is reasonable, since we cannot ignore the dependence of p_β on β. It is worthwhile to point out that Proposition B.1 holds for any $\beta \in \mathcal{B}$, however, Proposition B.2 requires β to be in a root-N neighborhood of β_0.*

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

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

where

$$\begin{aligned} \hat{R}_{Nl}(X^{\mathrm{T}}\beta, \beta) =& N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j} b_{ij}(\hat{\mu})\left\{D_{ij} - a_{ij}(\hat{\mu}) - b_{ij}(\hat{\mu})\hat{p}_\beta(X^{\mathrm{T}}\beta)\right. \\ &\left. - b_{ij}(\hat{\mu})\hat{p}'_\beta(X^{\mathrm{T}}\beta)(X_{ij}^{\mathrm{T}}\beta - X^{\mathrm{T}}\beta)\right\}\partial\{\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta, X^{\mathrm{T}}\beta; l)\}/\partial\beta, \end{aligned}$$

and $\hat{d}_\beta(x^{\mathrm{T}}\beta)$ is defined in (B.8). Note that $\hat{d}_\beta(X^{\mathrm{T}}\beta)$ acts like a local linear estimator of $d_\beta(X^{\mathrm{T}}\beta)$. Similar to Lemma B.3, we have $\|\hat{d}_\beta(X_{ij}^{\mathrm{T}}\beta) - d_\beta(X_{ij}^{\mathrm{T}}\beta)\| = \mathcal{O}_r(h^2)$ for all (i, j)s, and $\sup_{X\in\mathbb{X},\beta\in\mathcal{B}} \|\hat{d}_\beta(X^{\mathrm{T}}\beta) - d_\beta(X^{\mathrm{T}}\beta)\| \overset{a.s.}{\to} 0$. Consequently,

$$\left\| \hat{q}'_\beta(X_{ij}^{\mathrm{T}}\beta)\left[X_{ij} - \hat{d}_\beta(X_{ij}^{\mathrm{T}}\beta)\right] - q'_\beta(X_{ij}^{\mathrm{T}}\beta)\left[X_{ij} - d_\beta(X_{ij}^{\mathrm{T}}\beta)\right] \right\| = \mathcal{O}_r(h).$$

Hence, it suffices to show that $\hat{R}_{Nl}(X^{\mathrm{T}}\beta, \beta) = \mathcal{O}_s(h)$ component-wisely for $s = 2r$.

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

$$\begin{aligned}\hat{R}_{Nl}(X^{\mathrm{T}}\beta,\beta) =&\hat{B}_{1l}(X^{\mathrm{T}}\beta,\beta)+\hat{B}_{2l}(X^{\mathrm{T}}\beta,\beta)+h^{-1}\left\{p_\beta(X^{\mathrm{T}}\beta)-\hat{p}_\beta(X^{\mathrm{T}}\beta)\right\}\hat{B}_{3l}(X^{\mathrm{T}}\beta,\beta)\\&+\left\{p'_\beta(X^{\mathrm{T}}\beta)-\hat{p}'_\beta(X^{\mathrm{T}}\beta)\right\}\hat{B}_{4l}(X^{\mathrm{T}}\beta,\beta),\end{aligned}$$

where

$$\begin{aligned}\hat{B}_{1l}(X^{\mathrm{T}}\beta,\beta)&=N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}b_{ij}(\hat{\mu})\left\{D_{ij}-a_{ij}(\hat{\mu})-b_{ij}(\hat{\mu})p_\beta(X_{ij}^{\mathrm{T}}\beta)\right\}\partial\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,X^{\mathrm{T}}\beta;l)/\partial\beta,\\
\hat{B}_{2l}(X^{\mathrm{T}}\beta,\beta)&=N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}b_{ij}^2(\hat{\mu})\left\{p_\beta(X_{ij}^{\mathrm{T}}\beta)-p_\beta(X^{\mathrm{T}}\beta)-p'_\beta(X^{\mathrm{T}}\beta)(X_{ij}^{\mathrm{T}}\beta-X^{\mathrm{T}}\beta)\right\}\\&\qquad\partial\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,X^{\mathrm{T}}\beta;l)/\partial\beta,\\
\hat{B}_{3l}(X^{\mathrm{T}}\beta,\beta)&=N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}b_{ij}^2(\hat{\mu})h\partial\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,X^{\mathrm{T}}\beta;l)/\partial\beta,\\
\hat{B}_{4l}(X^{\mathrm{T}}\beta,\beta)&=N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}b_{ij}^2(\hat{\mu})(X_{ij}^{\mathrm{T}}\beta-X^{\mathrm{T}}\beta)\partial\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,X^{\mathrm{T}}\beta;l)/\partial\beta.\end{aligned}$$

Define $B_{ml}(X^{\mathrm{T}}\beta,\beta)$ as the version of $\hat{B}_{ml}(X^{\mathrm{T}}\beta,\beta)$ with replacing $\hat{\mu}$ by μ_0 for $m=1,\ldots,4$. We first show that $\hat{B}_{4l}(X^{\mathrm{T}}\beta,\beta)=B_{4l}(X^{\mathrm{T}}\beta)=\mathcal{O}_s(N^{-1/2})$. By Lemma B.2, we only need show that $N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}|(X_{ij}^{\mathrm{T}}\beta-X^{\mathrm{T}}\beta)\partial\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,X^{\mathrm{T}}\beta;l)/\partial\beta|=\mathcal{O}_{2s}(1)$. Using (B.12),

$$\begin{aligned}&E\left[\left\{N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}\left|(X_{ij}^{\mathrm{T}}\beta-X^{\mathrm{T}}\beta)\partial\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,X^{\mathrm{T}}\beta;l)/\partial\beta\right|\right\}^{2s}\right]\\
\simeq&\sum_{j=1}^{J}\sum_{i=1}^{c_j}E_{X^{\mathrm{T}}\beta}\left\{\left|N^{-1}(X_{ij}^{\mathrm{T}}\beta-X^{\mathrm{T}}\beta)\partial\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,X^{\mathrm{T}}\beta;l)/\partial\beta\right|^{2s}\right\}\\
&+\left|E_{X^{\mathrm{T}}\beta}\left\{\left|(X_{ij}^{\mathrm{T}}\beta-X^{\mathrm{T}}\beta)\partial\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,X^{\mathrm{T}}\beta;l)/\partial\beta\right|\right\}\right|^{2s}\\
&+\left(\sum_{j=1}^{J}\sum_{i=1}^{c_j}E_{X^{\mathrm{T}}\beta}\left[\left\{N^{-1}(X_{ij}^{\mathrm{T}}\beta-X^{\mathrm{T}}\beta)\partial\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta,X^{\mathrm{T}}\beta;l)/\partial\beta\right\}^2\right]\right)^s.\end{aligned}$$

Letting $\psi(x) = K'(x)x^l + lK(x)x^{l-1}$,

$$E_{X^{\mathrm{T}}\beta}\left\{\left|N^{-1}(X_{ij}^{\mathrm{T}}\beta - X^{\mathrm{T}}\beta)\partial\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta, X^{\mathrm{T}}\beta; l)/\partial\beta\right|^{2s}\right\}$$
$$\simeq \int N^{-2s}h^{1-2s}|u\psi(u)|^{2s} f_{X^{\mathrm{T}}\beta}(X^{\mathrm{T}}\beta + hu)du = O(N^{-2s}h^{1-2s}).$$

Thus, $E[\{N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c_j}|(X_{ij}^{\mathrm{T}}\beta - X^{\mathrm{T}}\beta)\partial\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta, X^{\mathrm{T}}\beta; l)/\partial\beta|\}^{2s}] = O(N^{1-2s}h^{1-2s}) + O(1) + O(N^{-s}h^{-s}) = O(1)$ and $\hat{B}_{4l}(X^{\mathrm{T}}\beta, \beta) = B_{4l}(X^{\mathrm{T}}\beta) = \mathcal{O}_s(N^{-1/2})$. Similarly, one can show

$$\hat{B}_{ml}(X^{\mathrm{T}}\beta, \beta) = B_{ml}(X^{\mathrm{T}}\beta, \beta) = \mathcal{O}_s(N^{-1/2}) \text{ for } m = 2, 3.$$

Thus, we obtain that $\hat{B}_{2l}(X^{\mathrm{T}}\beta, \beta) = \mathcal{O}_s(h)$, $\hat{B}_{3l}(X^{\mathrm{T}}\beta, \beta) = \mathcal{O}_s(1)$, and $\hat{B}_{4l}(X^{\mathrm{T}}\beta, \beta) = \mathcal{O}_s(1)$. For $\hat{B}_{1l}(X^{\mathrm{T}}\beta, \beta)$, using Lemma B.2, we have

$$\hat{B}_{1l}(X^{\mathrm{T}}\beta, \beta) = B_{1l}(X^{\mathrm{T}}\beta, \beta) + \mathcal{O}_s(N^{-1/2}h^{-1}).$$

Rewrite $B_{1l}(X^{\mathrm{T}}\beta, \beta)$ as

$$B_{1l}(X^{\mathrm{T}}\beta, \beta) = \sum_{m=1}^{M} \frac{c^{(K)}J_m}{N} \cdot \frac{1}{c^{(K)}} \sum_{i=1}^{c^{(K)}} B_{1lmi}(X^{\mathrm{T}}\beta, \beta),$$

where $B_{1lmi}(X^{\mathrm{T}}\beta, \beta) = \sum_{|j|=c^{(K)}} B_{1lmij}$ and $B_{1lmij} = J_m^{-1}b_{ij}(\mu_0)\{D_{ij} - a_{ij}(\mu_0) - b_{ij}(\mu_0)$ $p_\beta(X_{ij}^{\mathrm{T}}\beta)\}\partial\mathcal{K}_h(X_{ij}^{\mathrm{T}}\beta, X^{\mathrm{T}}\beta; l)/\partial\beta$. Now, we use (B.12) to calculate the rate of $E\{B_{1lmij}\}$. We first check $E_{X^{\mathrm{T}}\beta}\{B_{1lmij}^s(X^{\mathrm{T}}\beta, \beta)\}$. Since when $\beta \neq \beta_0$, neither $E_{X_{ij}}[\{D_{ij} - a_{ij}(\mu_0) - b_{ij}(\mu_0)p_\beta(X_{ij}^{\mathrm{T}}\beta)\}X_{ij}]$ nor $E_{X_{ij}^{\mathrm{T}}\beta}[\{D_{ij} - a_{ij}(\mu_0) - b_{ij}(\mu_0)p_\beta(X_{ij}^{\mathrm{T}}\beta)\}X_{ij}]$ equals 0. We need the decomposition

$$\begin{aligned} D_{ij} - a_{ij}(\mu_0) - b_{ij}(\mu_0)p_\beta(X_{ij}^{\mathrm{T}}\beta) &= D_{ij} - a_{ij}(\mu_0) - b_{ij}(\mu_0)p_0(X_{ij}^{\mathrm{T}}\beta_0) \\ &\quad + b_{ij}(\mu_0)\left\{p_0(X_{ij}^{\mathrm{T}}\beta_0) - p_\beta(X_{ij}^{\mathrm{T}}\beta)\right\}. \end{aligned}$$

We then have $E_{X_{ij}^{\mathrm{T}}\beta}[\{D_{ij} - a_{ij}(\mu_0) - b_{ij}(\mu_0)p_0(X_{ij}^{\mathrm{T}}\beta_0)\}X_{ij}] = 0$ and $b_{ij}(\mu_0)\{p_0(X_{ij}^{\mathrm{T}}\beta_0) - p_\beta(X_{ij}^{\mathrm{T}}\beta)\} = O(N^{-1/2})$ by the smoothness of $q_\beta(X^{\mathrm{T}}\beta)$ and the condition $\|\beta - \beta_0\| = O(N^{-1/2})$. Thus $E_{X_{ij}^{\mathrm{T}}\beta}[\{D_{ij} -$

$a_{ij}(\mu_0) - b_{ij}(\mu_0)p_\beta(X_{ij}^{\mathrm{T}}\beta)\}X_{ij}] = O(N^{-1/2}h^{-1})$. Simple calculation provides that

$$\sum_{|j|=c^{(K)}} E_{X^{\mathrm{T}}\beta}\{|B_{1lmij}(X^{\mathrm{T}}\beta,\beta)|^s\} = O(N^{1-s}h^{1-2s})$$

and

$$[\sum_{|j|=c^{(K)}} E_{X^{\mathrm{T}}\beta}\{B_{1lmij}^2(X^{\mathrm{T}}\beta,\beta)\}]^{2s} = O(N^{-s/2}h^{-3s/2}).$$

Thus,

$$E_{X^{\mathrm{T}}\beta}\{B_{1lmij}^s(X^{\mathrm{T}}\beta,\beta)\} = O(N^{1-s}h^{1-2s}) + O(N^{-s/2}h^{-s}) + O(N^{-s/2}h^{-3s/2}) = O(h^s).$$

Consequently, $\hat{B}_{1l}(X^{\mathrm{T}}\beta,\beta) = \mathcal{O}_s(h) + \mathcal{O}_s(N^{-1/2}h^{-1}) = \mathcal{O}_s(h)$. Finally,

$$\hat{R}_{Nl}(X^{\mathrm{T}}\beta,\beta) = \mathcal{O}_s(h) + \mathcal{O}_s(h) + h^{-1}\mathcal{O}_s(h^2)\mathcal{O}_s(1) + \mathcal{O}_s(h)\mathcal{O}_s(1) = \mathcal{O}_r(h),$$

which completes the proof. □

Lemma B.4. *Under Conditions 3.1–3.4, we have*

$$\sup_{X\in\mathbb{X},\beta^{(1)}\in\mathcal{B}_N^{(1)}} |\hat{p}_\beta(X^{\mathrm{T}}\beta) - p_0(X^{\mathrm{T}}\beta_0)| = O_p(\{\log N/(Nh)\}^{1/2}),$$
$$\sup_{X\in\mathbb{X},\beta^{(1)}\in\mathcal{B}_N^{(1)}} \left\|\hat{p}_\beta^{(1)}(X^{\mathrm{T}}\beta) - p_0'(X^{\mathrm{T}}\beta_0)\{X - d_{\beta_0}(X^{\mathrm{T}}\beta_0)\}\right\| = O_p(\{\log N/(Nh^3)\}^{1/2}),$$

where $\mathcal{B}_N^{(1)} = \{\beta^{(1)} \in \mathcal{B}^{(1)} : \|\beta^{(1)} - \beta_0^{(1)}\| \le CN^{-1/2}\}$ *for some constant* $C > 0$.

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

Proposition B.3. *Under Conditions 3.1–3.4, we have*

$$\sup_{\beta^1\in\mathcal{B}_N^{(1)}} \left\|\hat{G}(\beta^{(1)}) - G(\beta_0^{(1)}) + N\mathcal{J}_0^{\mathrm{T}}\Omega\mathcal{J}_0(\beta - \beta_0)\right\| = o_p(N^{1/2}),$$

where $\mathcal{B}_N^{(1)} = \{\beta^{(1)} \in \mathcal{B}^{(1)} : \|\beta^{(1)} - \beta_0^{(1)}\| \le CN^{-1/2}\}$ *for some constant* $C > 0$.

Proof. We firstly denote $A_j(\beta) = \mathcal{R}^{-1}\{Z_j;\mathcal{X}_j,\beta,p_\beta(\cdot)\}$, $\hat{A}_j(\beta) = \mathcal{R}^{-1}\{Z_j;\mathcal{X}_j,\hat{\beta},\hat{p}_\beta(\cdot)\}$, $B_j(\beta) = \sum_{i=1}^{c_j}\Delta_i\{Z_j;\mathcal{X}_j,\beta,p_\beta(\cdot)\}p_\beta'(X_{ij}^{\mathrm{T}}\beta)\{X_{ij} - d_\beta(X_{ij}^{\mathrm{T}}\beta)\}$, and $\hat{B}_j(\beta) = \sum_{i=1}^{c_j}\Delta_i\{Z_j;\mathcal{X}_j,\hat{\beta},\hat{p}_\beta(\cdot)\}\hat{p}_\beta^{(1)}(X_{ij}^{\mathrm{T}}\beta)$.

Then $\hat{G}(\beta^{(1)}) = \mathcal{J}_\beta^{\mathrm{T}} \sum_{j=1}^{J} \hat{A}_j(\beta)\hat{B}_j(\beta)$ and $G(\beta) = \mathcal{J}_\beta^{\mathrm{T}} \sum_{j=1}^{J} A_j(\beta)B_j(\beta)$. Further we have the following decomposition,

$$\begin{aligned}
\hat{G}(\beta^{(1)}) - G(\beta_0^{(1)}) =& (\mathcal{J}_\beta^{\mathrm{T}} - \mathcal{J}_0^{\mathrm{T}}) \sum_{m=1}^{M} \sum_{|j|=c^{(K)}} A_j(\beta_0)B_j(\beta_0) \\
&+ \sum_{j=1}^{J} \left\{ \hat{A}_j(\beta) - \hat{A}_j(\beta_0) \right\} B_j(\beta_0) \\
&+ \mathcal{J}_\beta^{\mathrm{T}} \sum_{j=1}^{J} \left\{ \hat{A}_j(\beta_0) - A_j(\beta_0) \right\} B_j(\beta_0) \\
&+ \mathcal{J}_\beta^{\mathrm{T}} \sum_{j=1}^{J} \left\{ \hat{A}_j(\beta) - A_j(\beta_0) \right\} \times \left\{ \hat{B}_j(\beta) - B_j(\beta_0) \right\} \\
&+ \mathcal{J}_\beta^{\mathrm{T}} \sum_{j=1}^{J} A_j(\beta_0) \left\{ \hat{B}_j(\beta) - B_j(\beta_0) \right\} \\
=& I_1(\beta^{(1)}) + I_2(\beta^{(1)}) + I_3(\beta^{(1)}) + I_4(\beta^{(1)}) + I_5(\beta^{(1)}).
\end{aligned} \tag{B.17}$$

Since $\mathcal{J}_\beta - \mathcal{J}_0 = O(N^{-1/2})$ for all $\beta^{(1)} \in \mathcal{B}_N^{(1)}$, and $\sum_{|j|=c^{(K)}} A_j(\beta_0)B_j(\beta_0)$ is a sum of identical and independent random variables with mean 0 and bounded covariance matrix,

$$\sup_{\beta^{(1)} \in \mathcal{B}_N^{(1)}} \left\| I_1(\beta^{(1)}) \right\| = o_p(N^{1/2}). \tag{B.18}$$

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

$$I_2(\beta^{(1)}) = \mathcal{J}_\beta^{\mathrm{T}} \left\{ \sum_{j=1}^{J} \hat{C}_j(\bar{\beta})B_j(\beta_0)\hat{B}_j(\bar{\beta})^{\mathrm{T}} \right\} \mathcal{J}_{\bar{\beta}}(\beta - \beta_0),$$

where $\hat{C}_j(\beta) = -\mathcal{R}^{-2}\{Z_j; \mathcal{X}_j, \beta, \hat{p}_\beta(\cdot)\}$. Letting $C_j(\beta) = -\mathcal{R}^{-2}\{Z_j; \mathcal{X}_j, \beta, p_\beta(\cdot)\}$, by $\bar{\beta}^{(1)} \in \mathcal{B}_N^{(1)}$ and Lemma B.4, we have that $\sup_j \sup_{\beta^{(1)} \in \mathcal{B}_N^{(1)}} |\hat{C}_j(\beta) - C_j(\beta_0)| = o_p(1)$ and $\sup_j \sup_{\beta^{(1)} \in \mathcal{B}_N^{(1)}} \|\hat{B}_j(\beta) -$

$B_j(\beta_0)\| = o_p(1)$. Then

$$\begin{aligned}\frac{1}{N}\sum_{j=1}^{J}\hat{C}_j(\bar{\beta})B_j(\beta_0)\hat{B}_j(\bar{\beta})^{\mathrm{T}} &= \sum_{m=1}^{M}\frac{J_m}{N}\times\frac{1}{J_m}\sum_{|j|=c^{(K)}}C_j(\beta_0)B_j(\beta_0)B_j(\beta_0)^{\mathrm{T}} + o_p(1)\\ &= \sum_{m=1}^{M}\frac{\gamma_m}{c^{(K)}}E\{C_j(\beta_0)B_j(\beta_0)B_j(\beta_0)^{\mathrm{T}}\} + o_p(1)\\ &= -\,\Omega + o_p(1).\end{aligned}$$

Noticing that $\mathcal{J}_\beta = \mathcal{J}_0 + O(N^{-1/2})$, $\mathcal{J}_{\bar{\beta}} = \mathcal{J}_0 + O(N^{-1/2})$, and $\beta - \beta_0 = O(N^{-1/2})$, we obtain

$$\sup_{\beta^{(1)}\in\mathcal{B}_N^{(1)}}\left\|I_2(\beta^{(1)}) + N\mathcal{J}_0^{\mathrm{T}}\Omega\mathcal{J}_0(\beta-\beta_0)\right\| = o_p(N^{1/2}). \tag{B.19}$$

Further, by Lemma B.5 and the fact $\mathcal{J}_\beta = O(1)$ for all $\beta^{(1)} \in \mathcal{B}_N^{(1)}$, we have

$$\sup_{\beta^{(1)}\in\mathcal{B}_N^{(1)}}\left\|I_3(\beta^{(1)})\right\| = o_p(N^{1/2}). \tag{B.20}$$

The bound for $I_4(\beta^{(1)})$ follows Lemma B.4 as

$$\begin{aligned}\sup_{\beta^{(1)}\in\mathcal{B}_N^{(1)}}\left\|I_4(\beta^{(1)})\right\| &\le J\sup_j\sup_{\beta^{(1)}\in\mathcal{B}_N^{(1)}}\left|\hat{A}_j(\beta) - A_j(\beta_0)\right|\\ &\quad\times p\times\sup_j\sup_{\beta^{(1)}\in\mathcal{B}_N^{(1)}}\left\|\hat{\mathcal{J}}_\beta\{B_j(\beta) - B_j(\beta_0)\}\right\|\\ &= J\times O_p[\{\log N/(Nh)\}^{1/2}]\times O_p[\{\log N/(Nh^3)\}^{1/2}]\\ &= o_p(N^{1/2}).\end{aligned} \tag{B.21}$$

Again, by Lemma B.4,

$$\begin{aligned}\sup_{\beta^{(1)}\in\mathcal{B}_N^{(1)}}\left\|I_5(\beta^{(1)})\right\| &\le N^{1/2}\left\{N^{-1}\sum_{j=1}^{J}\sup_{z_j\in\mathcal{Z}_j}A_j^2(\beta_0)\right\}^{-1/2}\\ &\quad\times\left\{JpN^{-1}\sup_j\sup_{\beta^{(1)}\in\mathcal{B}_N^{(1)}}\left\|\hat{B}_j(\beta) - B_j(\beta_0)\right\|\right\}\\ &= o_p(N^{1/2}).\end{aligned} \tag{B.22}$$

Combining (B.17)-(B.22) completes the proof of Porposition 3. □

Lemma B.5. *Under Conditions 3.1–3.4, we have*

$$\left\| \sum_{j=1}^{J} \left\{ \hat{A}_j(\beta_0) - A_j(\beta_0) \right\} B_j(\beta_0) \right\| = o_p(N^{1/2}).$$

Proof. For ease of presentation, we assume here that the group sizes are equal. The general case follows along the same lines but notation becomes tedious. Define ν_j to be the first component of $B_j(\beta_0)$; i.e., $\nu_j = \sum_{i=1}^{c_j} \Delta_i \{Z_j; \mathcal{X}_j, \beta_0, p_0(\cdot)\} p_0'(X_{ij}^{\mathrm{T}}\beta_0)\{X_{ij1} - d_{01}(X_{ij}^{\mathrm{T}}\beta_0)\}$, where $d_{\beta_0}(u) = (d_{01}(u), \ldots, d_{op}(u))^{\mathrm{T}}$. Then, we have ν_j being bounded, identical, and independent random variable with mean 0. Further, we denote $\hat{\varsigma}_j = \mathcal{R}^{-1}\{Z_j; \mathcal{X}_j, \beta_0, \hat{p}_{\beta_0}(\cdot)\}$ and $\varsigma_j = \mathcal{R}^{-1}\{Z_j; \mathcal{X}_j, \beta_0, p_0(\cdot)\}$. To show $\sum_{j=1}^{J}(\varsigma_j - \varsigma_j)\nu_j = o_p(N^{1/2})$. By Chebyshev's inequality, it suffices to show that $E|\sum_{j=1}^{J} \nu_j(\varsigma_j - \varsigma_j)|^2 = o(N)$. To this end, we define $\hat{\varsigma}_{(-k,-l),j} = \mathcal{R}^{-1}\{Z_j; \mathcal{X}_j, \beta_0, \hat{p}_{-k,-l}(\cdot)\}$, where $\hat{p}_{-k,-l}(u)$ is the kernel estimator of $p_0(u)$ based on the data $\{Z_j, X_{ij}^{\mathrm{T}}\beta_0, i = 1, \ldots, c_j, j = 1, \ldots J, j \neq k, j \neq l\}$s. When N is large, the difference between $\hat{p}_{-k,-l}(\cdot)$ and $\hat{p}_{\beta_0}(\cdot)$ should be very small. In fact, we have

$$E\left|\hat{p}_{-k,-l}(X_{ij}^{\mathrm{T}}\beta_0) - \hat{p}_{\beta_0}(X_{ij}^{\mathrm{T}}\beta_0)\right|^r = O(N^{-r}h^{1-r})$$

for all k, l, (i, j), and $r \geq 2$. Subsequently, we have the following decomposition,

$$\begin{aligned} E\left|\sum_{j=1}^{j} \nu_j(\hat{\varsigma}_j - \varsigma_j)\right|^2 &= \sum_{j=1}^{j} E[\nu_j^2(\hat{\varsigma}_j - \varsigma_j)^2] \\ &\quad + \sum_{k\neq l} E[\nu_k\nu_l(\hat{\varsigma}_k - \hat{\varsigma}_{(-k,-l),k})(\hat{\varsigma}_l - \varsigma_l)] \\ &\quad + \sum_{k\neq l} E[\nu_k\nu_l(\hat{\varsigma}_{(-k,-l),k} - \varsigma_k)(\hat{\varsigma}_l - \hat{\varsigma}_{(-k,-l),l})] \\ &\quad + \sum_{k\neq l} E[\nu_k\nu_l(\hat{\varsigma}_{(-k,-l),k} - \varsigma_k)(\hat{\varsigma}_{(-k,-l),l} - \varsigma_l)] \\ &= I_{N1} + I_{N2} + I_{N3} + I_{N4}. \end{aligned}$$

Given $X_{ij}^{\mathrm{T}}\beta_0$s, v_k, v_l, $\hat{\varsigma}_{(-k,-l),k} - \varsigma_k$, and $\hat{\varsigma}_{(-k,-l),l} - \varsigma_l$ in term I_{N4} are independent, and we have $E(\nu_k|X_{ij}^{\mathrm{T}}\beta_0 s) = 0$ and $E(\nu_l|X_{ij}^{\mathrm{T}}\beta_0 s) = 0$. Hence $I_{N4} = 0$. By Condition 3.4, both $\mathcal{R}\{Z_j; \mathcal{X}_j, \beta_0, \hat{p}_{\beta_0}(\cdot)\}$ and $\mathcal{R}\{Z_j; \mathcal{X}_j, \beta_0, p_0(\cdot)\}$ are bounded away form 0. Further Lemma B.3 implies that $|\mathcal{R}\{Z_j; \mathcal{X}_j, \beta_0, \hat{p}_{\beta_0}(\cdot)\} - \mathcal{R}\{Z_j; \mathcal{X}_j, \beta_0, p_0(\cdot)\}| \xrightarrow{a.s.} 0$, and Proposition B.1 implies that $\mathcal{R}\{Z_j; \mathcal{X}_j, \beta_0, \hat{p}_{\beta_0}(\cdot)\} = \mathcal{R}\{Z_j; \mathcal{X}_j, \beta_0, p_0(\cdot)\} + \mathcal{O}_r(h^2)$ for any $r \geq 2$. Thus, we have $E(\hat{\varsigma}_j - \varsigma_j)^4 = O(h^8)$. Similarly, we have $\hat{\varsigma}_{(-k,-l),k} = \varsigma_k + \mathcal{O}_r(N^{-1}h^{1/r-1})$. Then, it follows

that $I_{N1} \leq \sum_{j=1}^{J}\{E(\nu_j^4)\}^{1/2}\{E(\hat{\varsigma}_j - \varsigma_j)^4\}^{1/2} = J \cdot O(h^4) = o(N)$. By Cauchy-Schwartz inequality,

$$\begin{aligned} I_{N2} &\leq \sum_{k \neq l}[(E(\nu_k^4\nu_l^4)E[\{\hat{\varsigma}_k - \hat{\varsigma}_{(-k,-l),k}\}^4])^{1/2}E\{(\hat{\varsigma}_l - \varsigma_l)^2\}]^{1/2} \\ &= J^2 \times O\left(\frac{h^{1/4}}{Nh}\right) \times O(h^2) = O(Nh^{5/4}) = o(N). \end{aligned}$$

Similarly, one can show that $I_{N3} = o(N)$ which completes the proof. □

Lemma B.6. *Under Condition 3.5, $\mathcal{J}_0^{\mathrm{T}}\Omega\mathcal{J}_0$ is a positve definite matrix. Further if Conditions 3.1–3.4 are satisfied, we have*

$$\left\|\hat{\beta}^{(1)} - \beta_0^{(1)}\right\| = O_p(N^{-1/2}).$$

Proof. By the definition of Ω, it can be seen that $\mathcal{J}_0^{\mathrm{T}}\Omega\mathcal{J}_0$ is a positive semidefinite matrix. It suffices to show that 0 is not one of its eigenvalues. By Condition 3.5, $(\mathcal{J}_0 u)^{\mathrm{T}}\Omega(\mathcal{J}_0 u) = 0$ if and only if $\mathcal{J}_0 u = r\beta_0$ for some constant $r > 0$ where

$$\mathcal{J}_0 = \begin{pmatrix} -\frac{\beta_2}{\sqrt{1-\|\beta_0^{(1)}\|^2}} & \cdots & -\frac{\beta_p}{\sqrt{1-\|\beta_0^{(1)}\|^2}} \\ 1 & & \\ & \ddots & \\ & & 1 \end{pmatrix}.$$

Solving $\mathcal{J}_0 u = r\beta_0$ results in $u = 0$ and thus $r = 0$. It is a contradiction to $r > 0$. This indicates that $\mathcal{J}_0^{\mathrm{T}}\Omega\mathcal{J}_0$ is a positive definite matrix.

To show $\|\hat{\beta}^{(1)} - \beta_0^{(1)}\| = O_p(N^{-1/2})$, by (6.3.4) on page 163 of Ortega & Rheinboldt (1973), which is also used by Weisberg & Welsh (1994) and Wang et al. (2010), it suffices to show that for any small probability τ, we can always find a constant $C > 0$, such that

$$\liminf_N P\left(\sup_{u \in U_N} u^{\mathrm{T}}\hat{G}(\beta^{(1)}) < 0\right) = 1 - \tau, \tag{B.23}$$

where $U_N = \{u \in \mathbb{R}^{p-1} : (\beta_0^{(1)} + u) \in \mathcal{B}^{(1)}, N^{1/2}\|u\| = C\}$. Let $\lambda_{\min}$ be the smallest eigenvalue of $\mathcal{J}_0^{\mathrm{T}}\Omega\mathcal{J}_0$. Then

$$\begin{aligned} u^{\mathrm{T}}G(\beta_0^{(1)}) - Nu^{\mathrm{T}}\mathcal{J}_0^{\mathrm{T}}\Omega\mathcal{J}_0 u &\leq \|N^{1/2}u\| \times \|N^{-1/2}G(\beta_0^{(1)})\| - \lambda_{\min}\|N^{1/2}u\|^2 \\ &= C \times \|N^{-1/2}G(\beta_0^{(1)})\| - \lambda_{\min} \times C^2. \end{aligned} \tag{B.24}$$

Noting that (B.24) is a quadratic function in C with $\lambda_{\min} > 0$ and $\|N^{-1/2}G(\beta_0^{(1)})\| = O_p(1)$, for any $\tau > 0$, if C is chosen large enough, we have (B.24) being negative with probability at least $1 - \tau$. Further by Propostion 3, we have

$$\sup_{u \in U_n} \left| u^{\mathrm{T}} \hat{G}(\beta^{(1)}) - \left\{ u^{\mathrm{T}} G(\beta_0^{(1)}) - N u^{\mathrm{T}} \mathcal{J}_0^{\mathrm{T}} \Omega \mathcal{J}_0 u \right\} \right| = o_p(1).$$

This proves (B.23) and hence completes the proof. □

Appendix C Technical arguments and additional simulation results related to Chapter 4

C.1 Efficient algorithms

In what follows, we provide the derivation of the efficient algorithms, discussed in Section 4.2.2 of our manuscript, for computing the probability of the observed testing outcomes under two of the most common group testing decoding algorithms; specifically Dorfman testing and three-stage halving.

C.1.1 Dorfman testing

Dorfman testing (DT) begins by combining all of the specimens in the jth group into one master pool, which is then tested; i.e., $\mathcal{P}_{j1} = \mathcal{G}_j = \{1, ..., n_j\}$. If the master pool tests negative then the screening process ends. Alternatively, if the master pool tests positive then all contributing specimens are retested individually. Using the notation developed in Section 4.2 of our manuscript, we let $Z_{\mathcal{P}_{j1}}$ denote the testing response observed from assaying the master pool. Subsequently, the testing response vector for the jth group, $\boldsymbol{Z}_j$, takes on the form $\boldsymbol{Z}_j = Z_{\mathcal{P}_{j1}} = 0$ if the master pool tests negative, and $\boldsymbol{Z}_j = (Z_{\mathcal{P}_{j1}} = 1, Z_{\mathcal{P}_{j2}}, \ldots, Z_{\mathcal{P}_{jK_j}})^{\mathrm{T}}$ otherwise, where $Z_{\mathcal{P}_{jl}}$, for $l = 2, ..., K_j$, and $K_j = n_j + 1$. Note, in this context $Z_{\mathcal{P}_{jl}}$ denotes the testing response observed from retesting the $\mathcal{P}_{jl} = \{l - 1\}$ specimen individually.

To perform maximum likelihood estimation we need only derive the probability of observing the testing response vector $\boldsymbol{Z}_j$ under its different configurations. We first focus on the event that the master pool tests negative, in which case the probability of observing $\boldsymbol{Z}_j = 0$, given the individual level covariate information, can be expressed as

$$\begin{aligned}
\text{pr}(\boldsymbol{Z}_j = 0 \mid \boldsymbol{x}_j) &= \sum_{\boldsymbol{t}_j \in \mathcal{T}_j} \text{pr}(Z_{\mathcal{P}_{j1}} = 0 \mid \boldsymbol{T}_j = \boldsymbol{t}_j, \boldsymbol{x}_j)\text{pr}(\boldsymbol{T}_j = \boldsymbol{t}_j \mid \boldsymbol{x}_j) \\
&= \sum_{\boldsymbol{t}_j \in \mathcal{T}_j} \text{pr}(Z_{\mathcal{P}_{j1}} = 0 \mid \boldsymbol{T}_j = \boldsymbol{t}_j) \prod_{i=1}^{n_j} \text{pr}(T_{ij} = t_{ij} \mid \boldsymbol{x}_{ij}) \\
&= \sum_{\boldsymbol{t}_j \in \mathcal{T}_j} M_j(0, \boldsymbol{t}_j) \prod_{i=1}^{n_j} \text{pr}(T_{ij} = t_{ij} \mid \boldsymbol{x}_{ij}) \\
&= \sum_{k=0}^{n_j} M_j(0, \mathbf{1}_{n_j:k}) \text{pr}\left(\sum_{i=1}^{n_j} T_{ij} = k \mid \boldsymbol{x}_j \right).
\end{aligned}$$

where $\boldsymbol{x}_j = (\boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_j j})^{\mathrm{T}}$ and $\mathbf{1}_{n_j:k}$ is a n_j-dimensional binary vector with the first k components being

1 and all others being 0. The last equality holds because the biomarker distributions are assumed to be independent of the individuals' covariate information, in which case $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j) = M_j(\boldsymbol{z}_j, \boldsymbol{t}'_j)$ for all $\boldsymbol{t}_j$ and $\boldsymbol{t}'_j$ such that $\sum_{i=1}^{n_j} t_{ij} = \sum_{i=1}^{n_j} t'_{ij}$. The calculation of $M_j(0, \mathbf{1}_{n_j:k})$ can easily be accomplished using the expressions provided in our manuscript. The random variable $\sum_{i=1}^{n_j} T_{ij}$ follows a Poisson binomial distribution; i.e., it is the sum of n_j independent Bernoulli random variables that are not necessarily identically distributed. Therefore, one can easily calculate the pr$(\sum_{i=1}^{n_j} T_{ij} = k \mid \boldsymbol{x}_j)$ using the methods outlined in Wang (1993).

We now turn our attention to calculating

$$\text{pr}\{\boldsymbol{Z}_j = (1, \boldsymbol{w}^{\text{T}})^{\text{T}}\} = \text{pr}(Z_{\mathcal{P}_{j1}} = 1, Z_{\mathcal{P}_{j2}} = w_1, \ldots, Z_{\mathcal{P}_{jK_j}} = w_{n_j}),$$

where $\boldsymbol{w} = (w_1, ..., w_{n_j})^{\text{T}}$ is the vector of testing response observed from retesting each of the specimens separately. Define the sets $\mathcal{I}^+(\boldsymbol{w})$ and $\mathcal{I}^-(\boldsymbol{w})$ to be the collection of indices identifying the individuals in $\mathcal{G}_j$ that tested positive and negative, respectively, according to $\boldsymbol{w}$. For example, under DT if $\boldsymbol{w} = (1, 1, 0, 0, 1)^{\text{T}}$ then $\mathcal{I}^+(\boldsymbol{w}) = \{1, 2, 5\}$ and $\mathcal{I}^-(\boldsymbol{w}) = \{3, 4\}$. Using this set notation, the probability of observing $\boldsymbol{Z}_j = (1, \boldsymbol{w}^{\text{T}})^{\text{T}}$, given the individual level covariate information, can be expressed as

$$\begin{aligned}\text{pr}\left\{\boldsymbol{Z}_j = (1, \boldsymbol{w}^{\text{T}})^{\text{T}} \mid \boldsymbol{x}_j\right\} = \sum_{k_1=0}^{|\mathcal{I}^+(\boldsymbol{w})|} \sum_{k_2=0}^{|\mathcal{I}^-(\boldsymbol{w})|} \Big[&\text{pr}\left\{\boldsymbol{Z}_j = (1, \boldsymbol{w}^{\text{T}})^{\text{T}} \mid T_j^+ = k_1, T_j^- = k_2\right\} \\ &\times \text{pr}\left(T_j^+ = k_1 \mid \boldsymbol{x}_j\right) \text{pr}\left(T_j^- = k_2 \mid \boldsymbol{x}_j\right)\Big],\end{aligned}$$

where $T_j^+ = \sum_{i \in \mathcal{I}^+(\boldsymbol{w})} T_{ij}$ and $T_j^- = \sum_{i \in \mathcal{I}^-(\boldsymbol{w})} T_{ij}$ with the convention that $T_j^+ = 0$ or $T_j^- = 0$ if $|\mathcal{I}^+(\boldsymbol{w})| = 0$ or $|\mathcal{I}^-(\boldsymbol{w})| = 0$, respectively. Again notice that T_j^+ and T_j^- each follow a Poisson binomial distribution, and the probabilities involving these variables can easily be calculated as described above. The remaining probability statement above can be calculated as follows

$$\text{pr}\left\{\boldsymbol{Z}_j = (1, \boldsymbol{w}^{\text{T}})^{\text{T}} \mid T_j^+ = k_1, T_j^- = k_2\right\} = M_j\left(\boldsymbol{\delta}_{1:1,n_j:k}, \boldsymbol{\delta}_{k:k_1,(n_j-k):k_2}\right),$$

where $\boldsymbol{\delta}_{n_1:k_1,n_2:k_2,\ldots,n_a:k_a} = (\mathbf{1}_{n_1:k_1}^{\text{T}}, \mathbf{1}_{n_2:k_2}^{\text{T}}, ...\mathbf{1}_{n_a:k_a}^{\text{T}})^{\text{T}}$ and $k = |\mathcal{I}^+(\boldsymbol{w})|$. These expressions greatly reduce the computational burden associated with evaluating the observed data log-likelihood, when the group testing data arises from Dorfman testing.

C.1.2 Three-stage halving

Three-stage halving (TH) proceeds in a similar fashion to DT with the exception that an additional decoding stage is implemented before reverting to individual testing. Specifically, TH begins by combining all of the specimens in the jth group into one master pool, which is then tested; i.e., $\mathcal{P}_{j1} = \mathcal{G}_j = \{1, ..., n_j\}$. If the master pool tests negative then the screening process ends. On the other hand, if the master pool tests positive then all contributing specimens are randomly divided into two equally sized subgroups and these subgroups are tested. If a subgroup tests negative then testing is complete, alternatively if a subgroup tests positive then all contributing specimens are retested individually.

To allow for equally sized subgroups, we consider $n_j = 2r_j$. We denote the master pool testing response as $Z_{\mathcal{P}_{j1}}$. The probability of observing $\boldsymbol{Z}_j = Z_{\mathcal{P}_{j1}} = 0$ under TH is exactly the same as DT, which was described in Section C.1.1, so we focus on the cases that involve $Z_{\mathcal{P}_{j1}} = 1$. If the master pool tests positive (i.e., $Z_{\mathcal{P}_{j1}} = 1$) then the group is divided into two equally sized subgroups. Without loss of generality, we let $\mathcal{P}_{j2} = \{1, \ldots, r_j\}$ and $\mathcal{P}_{j3} = \{r_j + 1, \ldots, 2r_j\}$ indicate the individuals assigned to the two subgroups and $Z_{\mathcal{P}_{j2}}$ and $Z_{\mathcal{P}_{j3}}$ denote the respective testing responses. The first case we consider involves both subgroups testing negative; i.e., $\boldsymbol{Z}_j = (Z_{\mathcal{P}_{j1}}, Z_{\mathcal{P}_{j2}}, Z_{\mathcal{P}_{j3}})^{\mathrm{T}} = (1, 0, 0)^{\mathrm{T}}$. The probability of observing this event, given the individual level covariate information, can be calculated as follows

$$\text{pr}\{\boldsymbol{Z}_j = (1,0,0)^{\mathrm{T}} | \boldsymbol{x}_j\} = \sum_{k_1=1}^{r_j} \sum_{k_2=0}^{r_j} M_j\left(\mathbf{1}_{3:1}, \boldsymbol{\delta}_{r_j:k_1, r_j:k_2}\right), \text{pr}\left(T_j^{(1)} = k_1 \mid \boldsymbol{x}_j\right) \text{pr}\left(T_j^{(2)} = k_2 \mid \boldsymbol{x}_j\right)$$

where $T_j^{(1)} = \sum_{i=1}^{r_j} T_{ij}$ and $T_j^{(2)} = \sum_{i=r_j+1}^{2r_j} T_{ij}$.

The next testing outcome that we consider involves exactly one of the subgroups testing positive. Under the TH protocol, if a subgroup tests positive then all contributing specimens are then retested individually. For purposes of illustration, we assume that the the pool formed from combining the $\mathcal{P}_{j2}$ specimens tests positive, while the pool formed from the $\mathcal{P}_{j3}$ specimens tests negative. So the observed testing outcome can be expressed as $\boldsymbol{Z}_j = (1, 1, 0, \boldsymbol{w}_1^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{w}_1^{\mathrm{T}}$ denotes the vector of testing responses observed from assaying each specimens in $\mathcal{P}_{j2}$ individually. The probability of observing this event, given the individual level covariate information, can be calculated as follows

$$\begin{aligned}\text{pr}\{\boldsymbol{Z}_j = (1,1,0,\boldsymbol{w}_1^{\mathrm{T}})^{\mathrm{T}} | \boldsymbol{x}_j\} \quad = \quad & \sum_{k_{11}=0}^{|\mathcal{I}^+(\boldsymbol{w}_1)|} \sum_{k_{12}=0}^{|\mathcal{I}^-(\boldsymbol{w}_1)|} \sum_{k_2=0}^{r_j} \Big\{ M_j\left(\boldsymbol{\delta}_{3:2, r_j:k_1}, \boldsymbol{\delta}_{k_1:k_{11}, (r_j-k_1):k_{12}, r_j:k_2}\right) \\ & \times \text{pr}\left(T_j^{(2)} = k_2 \mid \boldsymbol{x}_j\right) \text{pr}\left(T_j^{(1)+} = k_{11} \mid \boldsymbol{x}_j\right) \text{pr}\left(T_j^{(1)-} = k_{12} \mid \boldsymbol{x}_j\right) \Big\},\end{aligned}$$

where $T_j^{(1)+} = \sum_{i \in \mathcal{I}^+(\boldsymbol{w}_1)} T_{ij}$, $T_j^{(1)-} = \sum_{i \in \mathcal{I}^-(\boldsymbol{w}_1)} T_{ij}$, and $k_1 = |\mathcal{I}^+(\boldsymbol{w}_1)|$.

The final possible testing outcome occurs when both subgroups test positive, in which case the observed testing response can be expressed as $\boldsymbol{Z}_j = (1, 1, 1, \boldsymbol{w}_1^{\mathrm{T}}, \boldsymbol{w}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\boldsymbol{w}_2^{\mathrm{T}}$ denotes the vector of testing outcomes observed from assaying each specimens in $\mathcal{P}_{j3}$ individually. The probability of observing this event, given the individual level covariate information, can be calculated as follows

$$
\begin{aligned}
\text{pr}\{\boldsymbol{Z}_j = (1,1,1,\boldsymbol{w}_1^{\mathrm{T}}, \boldsymbol{w}_2^{\mathrm{T}})^{\mathrm{T}} | \boldsymbol{x}_j\} = \\
\sum_{k_{11}=0}^{|\mathcal{I}^+(\boldsymbol{w}_1)|} \sum_{k_{12}=0}^{|\mathcal{I}^-(\boldsymbol{w}_1)|} \sum_{k_{21}=0}^{|\mathcal{I}^+(\boldsymbol{w}_2)|} \sum_{k_{22}=0}^{|\mathcal{I}^-(\boldsymbol{w}_2)|} \Bigg\{ & M_j\left(\boldsymbol{\delta}_{3:3,r_j:k_1,r_j:k_2}, \boldsymbol{\delta}_{k_1:k_{11},(r_j-k_1):k_{12},k_2:k_{21},(r_j-k_2):k_{22}}\right) \\
& \times \text{pr}\left(T_j^{(1)+} = k_{11} \mid \boldsymbol{x}_j\right) \text{pr}\left(T_j^{(1)-} = k_{12} \mid \boldsymbol{x}_j\right) \\
& \times \text{pr}\left(T_j^{(2)+} = k_{21} \mid \boldsymbol{x}_j\right) \text{pr}\left(T_j^{(2)-} = k_{22} \mid \boldsymbol{x}_j\right) \Bigg\},
\end{aligned}
$$

where $T_j^{(2)+} = \sum_{i \in \mathcal{I}^+(\boldsymbol{w}_2)} T_{ij}$, $T_j^{(2)-} = \sum_{i \in \mathcal{I}^-(\boldsymbol{w}_2)} T_{ij}$, and $k_2 = |\mathcal{I}^+(\boldsymbol{w}_2)|$. These expressions greatly reduce the computational burden associated with evaluating the observed data log-likelihood, when the group testing data arises from three-stage halving.

C.2 Expectation maximization algorithm

In what follows we provide the expectation maximization (EM) algorithm referenced in Section 4.2.2 of our manuscript. The development of the EM algorithm begins by treating the true statuses of the individuals as latent observations. The complete data log-likelihood can then be expressed as

$$
l_c(\boldsymbol{\beta}) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left\{ T_{ij} \log[\eta^{-1}(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})] + (1 - T_{ij}) \log[1 - \eta^{-1}(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})] \right\} + \sum_{j=1}^{J} \log\left\{ M_j(\boldsymbol{z}_j, \boldsymbol{T}_j) \right\}.
$$

The E-step of an EM algorithm involves taking the expectation of $l_c(\boldsymbol{\beta})$ with respect to all latent variables (i.e., T_{ij} for $i = 1, ..., n_j$ and $j = 1, ..., J$) conditional on the observed data and the current parameter $\boldsymbol{\beta}^{(d)}$. This yields the Q function

$$
Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(d)}) = \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left\{ \omega_{ij}^{(d)} \log[\eta^{-1}(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})] + (1 - \omega_{ij}^{(d)}) \log[1 - \eta^{-1}(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})] \right\}
$$

up to an additive term that does not involve $\boldsymbol{\beta}$, where $\omega_{ij}^{(d)} = E(T_{ij} \mid \boldsymbol{z}_j, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_jj}; \boldsymbol{\beta}^{(d)})$. The M-step then determines $\boldsymbol{\beta}^{(d+1)}$ to be the value that maximizes $Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(d)})$; i.e.,

$$\boldsymbol{\beta}^{(d+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\, Q(\boldsymbol{\beta}, \boldsymbol{\beta}^{(d)}).$$

The general form of the EM algorithm can be stated succinctly as follows:

Step 1: Initialize $\boldsymbol{\beta}^{(0)}$ and set $d = 0$.

Step 2: (E-step) Using $\boldsymbol{\beta}^{(d)}$ and the observed data calculate

$$\omega_{ij}^{(d)} = E(T_{ij} \mid \boldsymbol{z}_j, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_jj}; \boldsymbol{\beta}^{(d)}), \text{ for } i = 1, \ldots, n_j \text{ and } j = 1, \ldots, J.$$

Step 3: (M-step) Set $d = d + 1$ and obtain $\boldsymbol{\beta}^{(d)}$ as,

$$\boldsymbol{\beta}^{(d)} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \left\{ \omega_{ij}^{(d-1)} \log[\eta^{-1}(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})] + (1 - \omega_{ij}^{(d-1)}) \log[1 - \eta^{-1}(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta})] \right\}.$$

Step 4: Repeat steps 2 and 3 until convergence.

The EM algorithm above is completed with the expression for $\omega_{ij}^{(d)}$ which is given by

$$E(T_{ij} \mid \boldsymbol{z}_j, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_jj}; \boldsymbol{\beta}^{(d)}) = \frac{\text{pr}(\boldsymbol{Z}_j = \boldsymbol{z}_j \mid T_{ij} = 1, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_jj}; \boldsymbol{\beta}^{(d)}) \cdot \text{pr}(T_{ij} = 1 \mid \boldsymbol{x}_{ij}; \boldsymbol{\beta}^{(d)})}{\text{pr}(\boldsymbol{Z}_j = \boldsymbol{z}_j \mid \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_jj}; \boldsymbol{\beta}^{(d)})}.$$

The first probability statement in the numerator above can be calculated as follows

$$\begin{aligned} &P(\boldsymbol{Z}_j = \boldsymbol{z}_j \mid T_{ij} = 1, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_jj}; \boldsymbol{\beta}^{(d)}) \\ &\quad = \sum_{\boldsymbol{t}_j \in \mathcal{T}_j, t_{ij}=1} \left\{ M_j(\boldsymbol{z}_j, \boldsymbol{t}_j) \prod_{r=1, r\neq i}^{n_j} \left[t_{rj}\eta^{-1}(\boldsymbol{x}_{rj}^{\mathrm{T}}\boldsymbol{\beta}^{(d)}) + (1 - t_{rj})(1 - \eta^{-1}(\boldsymbol{x}_{rj}^{\mathrm{T}}\boldsymbol{\beta}^{(d)})) \right] \right\}, \end{aligned}$$

with the two remaining probabilities being given by $P(T_{ij} = 1 \mid \boldsymbol{x}_{ij}; \boldsymbol{\beta}^{(d)}) = \eta^{-1}(\boldsymbol{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}^{(d)})$ and $P(\boldsymbol{Z}_j = \boldsymbol{z}_j \mid \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_jj}; \boldsymbol{\beta}^{(d)}) = \mathcal{R}(\boldsymbol{z}_j, \boldsymbol{x}_{1j}, ..., \boldsymbol{x}_{n_jj}, \boldsymbol{\beta}^{(d)})$.

C.3 Monte Carlo approximation of $M_j(\cdot, \cdot)$

In what follows we provide details pertaining to the Monte Carlo techniques used to approximate the joint misclassification probabilities $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j)$ referenced in Section 4.2.1 of our manuscript. Recall, we

have that

$$M_j(\boldsymbol{z}_j, \boldsymbol{t}_j) = \int_{\boldsymbol{A}(\boldsymbol{z}_j, \boldsymbol{c}_j)} \int \prod_{l=1}^{K_j} f_{\mathcal{C}|\widetilde{\mathcal{C}}_{\mathcal{P}_{jl}} = \boldsymbol{D}_{\mathcal{P}_{jl}}^{\mathrm{T}} \boldsymbol{y}}(u_l) \prod_{i=1}^{n_j} f_{\widetilde{\mathcal{C}}|T_{ij}=t_{ij}}(y_{ij}) d\boldsymbol{y} d\boldsymbol{u},$$

which is often a multi-dimensional integral and is therefore difficult to calculate analytically. Hence, we propose the following Monte Carlo approach to approximate this integral.

Step 0: Set $d = 0$.

Step 1: Based on $\boldsymbol{t}_j$ randomly generate $\widetilde{\mathcal{C}}_{ij}$, for $i = 1, ..., n_j$, according to $f_{\widetilde{\mathcal{C}}|T_{ij}=t_{ij}}$.

Step 2: Calculate $\widetilde{\mathcal{C}}_{\mathcal{P}_{jl}} = \boldsymbol{D}_{\mathcal{P}_{jl}}^{\mathrm{T}} \widetilde{\mathcal{C}}_j$, for $l = 1, ..., K_j$.

Step 3: Randomly generate $\mathcal{C}_{\mathcal{P}_{jl}}$, for $l = 1, ..., K_j$, according to $f_{\mathcal{C}|\widetilde{\mathcal{C}}_{\mathcal{P}_{jl}}}$.

Step 4: If $\boldsymbol{\mathcal{C}}_j \in \boldsymbol{A}(\boldsymbol{z}_j, \boldsymbol{c}_j)$ set $d = d + 1$, where $\boldsymbol{\mathcal{C}}_j = (\mathcal{C}_{\mathcal{P}_{j1}}, \ldots, \mathcal{C}_{\mathcal{P}_{jK_j}})^{\mathrm{T}}$.

Repeat Steps 1–4 M times, where M is chosen to be sufficiently large. Then, $M_j(\boldsymbol{z}_j, \boldsymbol{t}_j)$ can be approximated by the ratio d/M. Notice, the above algorithm can be altered to handle the case that the biomarker levels are measured without error by setting $\mathcal{C}_{\mathcal{P}_{jl}} = \widetilde{\mathcal{C}}_{\mathcal{P}_{jl}}$ in Step 3.

C.4 Additional simulation results

Table C.1: Simulation results for Model 4.1 having regression parameters $\boldsymbol{\beta} = (-3, 2)^{\mathrm{T}}$. Presented results include the sample mean (Mean) and standard deviation (SD) of the 500 estimates of $\boldsymbol{\beta}$, when $n \in \{2, 4, 6\}$ and $\sigma_+ = 1$. The average standard error (SE) and estimated 95% Wald coverage probabilities (Cov) are also provided. Assuming a 99% confidence level for the coverage probabilities, the margin of error is 0.03. Estimates outside this margin of error are shown in bold. Note, MT, DT, and TH denote individual testing, master pool testing, Dorfman testing, and three-stage halving, respectively.

When $t(c) = t_0$:			Acknowledging the Dilution Effect			Traditional Approach		
	n	Measure	MT	DT	TH	MT	DT	TH
$\hat{\beta}_0$	2	Mean(Cov)	$-3.01(0.96)$	$-3.01(0.95)$	$--(--)$	$\mathbf{-3.52(0.06)}$	$\mathbf{-3.14(0.76)}$	$--(--)$
		SD(SE)	$0.13(0.13)$	$0.12(0.11)$	$--(--)$	$0.16(0.16)$	$0.11(0.11)$	$--(--)$
	4	Mean(Cov)	$-3.02(0.97)$	$-3.01(0.95)$	$-3.01(0.95)$	$\mathbf{-4.21(0.00)}$	$\mathbf{-3.81(0.00)}$	$\mathbf{-3.74(0.00)}$
		SD(SE)	$0.22(0.22)$	$0.16(0.15)$	$0.15(0.15)$	$0.22(0.24)$	$0.16(0.15)$	$0.15(0.14)$
	6	Mean(Cov)	$-3.05(0.94)$	$-3.02(0.95)$	$-3.02(0.95)$	$\mathbf{-5.33(0.00)}$	$\mathbf{-4.64(0.00)}$	$\mathbf{-4.47(0.00)}$
		SD(SE)	$0.37(0.32)$	$0.20(0.20)$	$0.20(0.20)$	$0.37(0.41)$	$0.24(0.24)$	$0.22(0.20)$
$\hat{\beta}_1$	2	Mean(Cov)	$2.01(0.96)$	$2.00(0.95)$	$--(--)$	$\mathbf{2.14(0.91)}$	$\mathbf{1.90(0.87)}$	$--(--)$
		SD(SE)	$0.15(0.16)$	$0.13(0.13)$	$--(--)$	$0.16(0.17)$	$0.12(0.12)$	$--(--)$
	4	Mean(Cov)	$2.02(0.97)$	$2.01(0.96)$	$2.00(0.95)$	$1.88(0.96)$	$\mathbf{1.82(0.75)}$	$\mathbf{1.74(0.53)}$
		SD(SE)	$0.28(0.29)$	$0.18(0.19)$	$0.18(0.18)$	$0.21(0.24)$	$0.14(0.15)$	$0.14(0.14)$
	6	Mean(Cov)	$2.05(0.93)$	$2.02(0.95)$	$2.02(0.96)$	$2.02(0.98)$	$1.92(0.94)$	$\mathbf{1.74(0.72)}$
		SD(SE)	$0.50(0.44)$	$0.26(0.25)$	$0.25(0.25)$	$0.33(0.36)$	$0.19(0.21)$	$0.17(0.19)$
When $t(c) = t_0/c$:			Acknowledging the Dilution Effect			Traditional Approach		
	n	Measure	MT	DT	TH	MT	DT	TH
$\hat{\beta}_0$	2	Mean(Cov)	$-3.01(0.94)$	$-3.01(0.96)$	$--(--)$	$\mathbf{-2.06(0.00)}$	$\mathbf{-2.35(0.00)}$	$--(--)$
		SD(SE)	$0.18(0.18)$	$0.12(0.12)$	$--(--)$	$0.09(0.08)$	$0.08(0.08)$	$--(--)$
	4	Mean(Cov)	$-3.04(0.95)$	$-3.01(0.95)$	$-3.00(0.96)$	$\mathbf{-1.70(0.00)}$	$\mathbf{-2.23(0.00)}$	$\mathbf{-1.99(0.00)}$
		SD(SE)	$0.28(0.26)$	$0.12(0.13)$	$0.12(0.12)$	$0.09(0.09)$	$0.08(0.08)$	$0.07(0.07)$
	6	Mean(Cov)	$-3.06(0.95)$	$-3.01(0.96)$	$-3.01(0.96)$	$\mathbf{-1.63(0.00)}$	$\mathbf{-2.31(0.00)}$	$\mathbf{-2.00(0.00)}$
		SD(SE)	$0.35(0.34)$	$0.12(0.13)$	$0.12(0.13)$	$0.12(0.11)$	$0.08(0.09)$	$0.07(0.07)$
$\hat{\beta}_1$	2	Mean(Cov)	$2.01(0.95)$	$2.01(0.96)$	$--(--)$	$\mathbf{1.32(0.00)}$	$\mathbf{1.59(0.04)}$	$--(--)$
		SD(SE)	$0.21(0.20)$	$0.14(0.14)$	$--(--)$	$0.13(0.12)$	$0.11(0.10)$	$--(--)$
	4	Mean(Cov)	$2.04(0.95)$	$2.01(0.95)$	$2.00(0.95)$	$\mathbf{1.19(0.01)}$	$\mathbf{1.63(0.11)}$	$\mathbf{1.45(0.00)}$
		SD(SE)	$0.31(0.30)$	$0.14(0.14)$	$0.13(0.14)$	$0.17(0.16)$	$0.12(0.11)$	$0.10(0.10)$
	6	Mean(Cov)	$2.06(0.96)$	$2.01(0.95)$	$2.01(0.96)$	$\mathbf{1.26(0.15)}$	$\mathbf{1.70(0.28)}$	$\mathbf{1.53(0.01)}$
		SD(SE)	$0.42(0.41)$	$0.14(0.14)$	$0.13(0.14)$	$0.25(0.24)$	$0.12(0.11)$	$0.10(0.10)$

Table C.2: Simulation results for Model 4.2 having regression parameters $\boldsymbol{\beta} = (-3, 1, 0.5)^{\mathrm{T}}$. Presented results include the sample mean (Mean) and standard deviation (SD) of the 500 estimates of $\boldsymbol{\beta}$, when $n \in \{2, 4, 6\}$ and $\sigma_+ = 1$. The average standard error (SE) and estimated 95% Wald coverage probabilities (Cov) are also provided. Assuming a 99% confidence level for the coverage probabilities, the margin of error is 0.03. Estimates outside this margin of error are shown in bold. Note, MT, DT, and TH denote individual testing, master pool testing, Dorfman testing, and three-stage halving, respectively.

When $t(c) = t_0$:			Acknowledging the Dilution Effect			Traditional Approach		
	n	Measure	MT	DT	TH	MT	DT	TH
$\widehat{\beta}_0$	2	Mean(Cov)	$-3.01(0.95)$	$-3.01(0.94)$	$--(--)$	$-3.52(\mathbf{0.05})$	$-3.15(\mathbf{0.72})$	$--(--)$
		SD(SE)	$0.12(0.12)$	$0.11(0.11)$	$--(--)$	$0.17(0.16)$	$0.10(0.10)$	$--(--)$
	4	Mean(Cov)	$-3.03(0.96)$	$-3.02(0.94)$	$-3.01(0.96)$	$-4.49(\mathbf{0.00})$	$-3.83(\mathbf{0.00})$	$-3.75(\mathbf{0.00})$
		SD(SE)	$0.20(0.19)$	$0.14(0.14)$	$0.14(0.14)$	$0.43(0.43)$	$0.16(0.15)$	$0.14(0.13)$
	6	Mean(Cov)	$-3.11(0.96)$	$-3.03(0.95)$	$-3.02(0.95)$	$-7.56(\mathbf{0.14})$	$-4.93(\mathbf{0.00})$	$-4.54(\mathbf{0.00})$
		SD(SE)	$0.34(0.33)$	$0.19(0.19)$	$0.19(0.19)$	$3.90(3.10)$	$0.45(0.36)$	$0.24(0.21)$
$\widehat{\beta}_1$	2	Mean(Cov)	$1.03(0.94)$	$1.02(0.95)$	$--(--)$	$1.38(\mathbf{0.87})$	$0.99(\mathbf{0.91})$	$--(--)$
		SD(SE)	$0.19(0.17)$	$0.12(0.12)$	$--(--)$	$0.40(0.35)$	$0.12(0.11)$	$--(--)$
	4	Mean(Cov)	$1.09(0.94)$	$1.03(0.94)$	$1.03(0.96)$	$1.97(\mathbf{0.88})$	$1.10(\mathbf{0.91})$	$0.99(\mathbf{0.92})$
		SD(SE)	$0.41(0.39)$	$0.18(0.17)$	$0.15(0.16)$	$0.93(0.89)$	$0.26(0.22)$	$0.18(0.17)$
	6	Mean(Cov)	$1.25(0.95)$	$1.05(0.97)$	$1.05(0.96)$	$5.24(0.94)$	$1.85(\mathbf{0.79})$	$1.17(0.94)$
		SD(SE)	$0.81(0.75)$	$0.26(0.24)$	$0.23(0.23)$	$5.68(4.51)$	$0.88(0.67)$	$0.40(0.33)$
$\widehat{\beta}_2$	2	Mean(Cov)	$0.48(0.95)$	$0.49(0.95)$	$--(--)$	$0.32(\mathbf{0.86})$	$0.45(0.94)$	$--(--)$
		SD(SE)	$0.14(0.13)$	$0.10(0.10)$	$--(--)$	$0.24(0.20)$	$0.09(0.09)$	$--(--)$
	4	Mean(Cov)	$0.44(0.96)$	$0.48(0.97)$	$0.49(0.95)$	$-0.08(\mathbf{0.72})$	$0.29(\mathbf{0.72})$	$0.32(\mathbf{0.69})$
		SD(SE)	$0.27(0.28)$	$0.14(0.14)$	$0.14(0.14)$	$0.42(0.40)$	$0.16(0.14)$	$0.14(0.12)$
	6	Mean(Cov)	$0.37(0.93)$	$0.47(0.95)$	$0.48(0.93)$	$-1.21(\mathbf{0.88})$	$-0.04(\mathbf{0.60})$	$0.19(\mathbf{0.66})$
		SD(SE)	$0.52(0.47)$	$0.21(0.20)$	$0.21(0.20)$	$2.05(1.64)$	$0.40(0.31)$	$0.23(0.19)$
When $t(c) = t_0/c$:			Acknowledging the Dilution Effect			Traditional Approach		
	n	Measure	MT	DT	TH	MT	DT	TH
$\widehat{\beta}_0$	2	Mean(Cov)	$-3.04(0.96)$	$-3.01(0.95)$	$--(--)$	$-2.22(\mathbf{0.00})$	$-2.46(\mathbf{0.00})$	$--(--)$
		SD(SE)	$0.17(0.17)$	$0.11(0.11)$	$--(--)$	$0.08(0.08)$	$0.08(0.08)$	$--(--)$
	4	Mean(Cov)	$-3.12(0.97)$	$-3.01(0.95)$	$-3.00(0.96)$	$-1.89(\mathbf{0.00})$	$-2.32(\mathbf{0.00})$	$-2.11(\mathbf{0.00})$
		SD(SE)	$0.29(0.29)$	$0.11(0.11)$	$0.11(0.11)$	$0.09(0.10)$	$0.07(0.08)$	$0.07(0.08)$
	6	Mean(Cov)	$-3.16(0.97)$	$-3.01(0.96)$	$-3.01(0.94)$	$-1.81(\mathbf{0.00})$	$-2.37(\mathbf{0.00})$	$-2.10(\mathbf{0.00})$
		SD(SE)	$0.39(0.42)$	$0.11(0.11)$	$0.11(0.11)$	$0.12(0.13)$	$0.07(0.08)$	$0.07(0.08)$
$\widehat{\beta}_1$	2	Mean(Cov)	$1.09(0.96)$	$1.01(0.97)$	$--(--)$	$0.67(\mathbf{0.11})$	$0.82(\mathbf{0.42})$	$--(--)$
		SD(SE)	$0.30(0.30)$	$0.12(0.13)$	$--(--)$	$0.10(0.10)$	$0.08(0.09)$	$--(--)$
	4	Mean(Cov)	$1.21(0.94)$	$1.02(0.97)$	$1.01(0.95)$	$0.59(\mathbf{0.17})$	$0.84(\mathbf{0.53})$	$0.76(\mathbf{0.20})$
		SD(SE)	$0.65(0.59)$	$0.13(0.13)$	$0.13(0.12)$	$0.13(0.14)$	$0.08(0.09)$	$0.08(0.08)$
	6	Mean(Cov)	$1.29(0.93)$	$1.01(0.96)$	$1.02(0.96)$	$0.62(\mathbf{0.47})$	$0.86(\mathbf{0.67})$	$0.80(\mathbf{0.38})$
		SD(SE)	$0.90(0.85)$	$0.13(0.13)$	$0.13(0.13)$	$0.21(0.21)$	$0.09(0.09)$	$0.09(0.09)$
$\widehat{\beta}_2$	2	Mean(Cov)	$0.46(0.95)$	$0.49(0.96)$	$--(--)$	$0.43(\mathbf{0.90})$	$0.46(0.94)$	$--(--)$
		SD(SE)	$0.18(0.19)$	$0.10(0.10)$	$--(--)$	$0.08(0.09)$	$0.07(0.08)$	$--(--)$
	4	Mean(Cov)	$0.42(0.92)$	$0.49(0.95)$	$0.49(0.95)$	$0.41(0.93)$	$0.47(0.96)$	$0.44(0.92)$
		SD(SE)	$0.35(0.32)$	$0.10(0.10)$	$0.10(0.10)$	$0.11(0.12)$	$0.07(0.08)$	$0.07(0.08)$
	6	Mean(Cov)	$0.39(0.94)$	$0.49(0.96)$	$0.49(0.96)$	$0.43(\mathbf{0.98})$	$0.49(0.97)$	$0.46(0.93)$
		SD(SE)	$0.47(0.44)$	$0.10(0.10)$	$0.10(0.10)$	$0.16(0.18)$	$0.08(0.08)$	$0.08(0.08)$

Table C.3: Simulation results for Model 4.3 having regression parameters $\boldsymbol{\beta} = (-3, 2, 1)^{\mathrm{T}}$. Presented results include the sample mean (Mean) and standard deviation (SD) of the 500 estimates of $\boldsymbol{\beta}$, when $n \in \{2, 4, 6\}$ and $\sigma_+ = 1$. The average standard error (SE) and estimated 95% Wald coverage probabilities (Cov) are also provided. Assuming a 99% confidence level for the coverage probabilities, the margin of error is 0.03. Estimates outside this margin of error are shown in bold. Note, MT, DT, and TH denote individual testing, master pool testing, Dorfman testing, and three-stage halving, respectively.

When $t(c) = t_0$:			Acknowledging the Dilution Effect			Traditional Approach		
	n	Measure	MT	DT	TH	MT	DT	TH
$\hat{\beta}_0$	2	Mean(Cov)	−3.01(0.95)	−3.01(0.95)	− − (− −)	−3.49(**0.08**)	−3.13(**0.81**)	− − (− −)
		SD(SE)	0.13(0.13)	0.11(0.12)	− − (− −)	0.15(0.16)	0.11(0.11)	− − (− −)
	4	Mean(Cov)	−3.02(0.96)	−3.00(0.96)	−3.01(0.96)	−4.16(**0.00**)	−3.76(**0.00**)	−3.69(**0.00**)
		SD(SE)	0.21(0.22)	0.14(0.15)	0.15(0.15)	0.20(0.23)	0.14(0.15)	0.14(0.14)
	6	Mean(Cov)	−3.07(0.95)	−3.01(0.94)	−3.02(0.94)	−5.23(**0.00**)	−4.54(**0.00**)	−4.37(**0.00**)
		SD(SE)	0.34(0.33)	0.19(0.19)	0.21(0.20)	0.36(0.39)	0.24(0.23)	0.20(0.20)
$\hat{\beta}_1$	2	Mean(Cov)	2.01(0.95)	2.01(0.96)	− − (− −)	2.12(**0.92**)	1.90(**0.84**)	− − (− −)
		SD(SE)	0.15(0.16)	0.13(0.13)	− − (− −)	0.16(0.17)	0.12(0.12)	− − (− −)
	4	Mean(Cov)	2.01(0.95)	2.00(0.96)	2.01(0.95)	1.84(**0.91**)	1.79(**0.69**)	1.72(**0.44**)
		SD(SE)	0.28(0.29)	0.18(0.18)	0.18(0.18)	0.21(0.23)	0.14(0.15)	0.13(0.14)
	6	Mean(Cov)	2.06(0.95)	2.00(0.95)	2.02(0.94)	1.95(0.96)	1.87(**0.89**)	1.71(**0.59**)
		SD(SE)	0.45(0.43)	0.24(0.23)	0.26(0.24)	0.32(0.34)	0.19(0.20)	0.17(0.18)
$\hat{\beta}_2$	2	Mean(Cov)	1.00(0.97)	1.00(0.96)	− − (− −)	1.04(0.95)	0.94(0.95)	− − (− −)
		SD(SE)	0.24(0.25)	0.20(0.21)	− − (− −)	0.25(0.26)	0.19(0.19)	− − (− −)
	4	Mean(Cov)	0.98(0.96)	0.99(0.96)	0.98(0.94)	0.85(0.97)	0.86(0.93)	0.83(**0.90**)
		SD(SE)	0.48(0.46)	0.28(0.29)	0.30(0.29)	0.70(13.11)	0.25(0.25)	0.26(0.24)
	6	Mean(Cov)	0.94(0.96)	0.98(0.96)	0.98(0.94)	0.61(0.97)	0.86(0.95)	0.79(**0.92**)
		SD(SE)	0.70(0.71)	0.37(0.37)	0.40(0.38)	1.79(534.74)	0.35(0.34)	0.34(0.32)
When $t(c) = t_0/c$:			Acknowledging the Dilution Effect			Traditional Approach		
	n	Measure	MT	DT	TH	MT	DT	TH
$\hat{\beta}_0$	2	Mean(Cov)	−3.02(0.97)	−3.01(0.95)	− − (− −)	−2.08(**0.00**)	−2.37(**0.00**)	− − (− −)
		SD(SE)	0.17(0.18)	0.12(0.13)	− − (− −)	0.08(0.08)	0.08(0.08)	− − (− −)
	4	Mean(Cov)	−3.06(0.97)	−3.01(0.96)	−3.00(0.94)	−1.74(**0.00**)	−2.26(**0.00**)	−2.02(**0.00**)
		SD(SE)	0.27(0.27)	0.12(0.13)	0.12(0.12)	0.10(0.10)	0.08(0.09)	0.07(0.07)
	6	Mean(Cov)	−3.08(0.95)	−3.01(0.96)	−3.01(0.95)	−1.67(**0.00**)	−2.36(**0.00**)	−2.04(**0.00**)
		SD(SE)	0.35(0.35)	0.13(0.13)	0.12(0.13)	0.14(0.13)	0.09(0.09)	0.07(0.08)
$\hat{\beta}_1$	2	Mean(Cov)	2.03(0.95)	2.02(0.96)	− − (− −)	1.36(**0.00**)	1.62(**0.05**)	− − (− −)
		SD(SE)	0.19(0.19)	0.13(0.13)	− − (− −)	0.12(0.12)	0.11(0.10)	− − (− −)
	4	Mean(Cov)	2.06(0.95)	2.01(0.95)	2.01(0.94)	1.24(**0.01**)	1.66(**0.15**)	1.49(**0.00**)
		SD(SE)	0.30(0.30)	0.13(0.14)	0.14(0.13)	0.17(0.17)	0.11(0.11)	0.10(0.10)
	6	Mean(Cov)	2.07(0.96)	2.02(0.96)	2.01(0.94)	1.33(**0.28**)	1.74(**0.41**)	1.57(**0.02**)
		SD(SE)	0.41(0.40)	0.14(0.14)	0.14(0.14)	0.27(0.26)	0.12(0.11)	0.10(0.10)
$\hat{\beta}_2$	2	Mean(Cov)	1.00(0.96)	0.99(0.95)	− − (− −)	0.69(**0.69**)	0.81(**0.82**)	− − (− −)
		SD(SE)	0.28(0.28)	0.20(0.20)	− − (− −)	0.21(0.21)	0.18(0.18)	− − (− −)
	4	Mean(Cov)	1.05(0.96)	0.99(0.95)	1.00(0.95)	0.67(**0.84**)	0.84(**0.87**)	0.77(**0.77**)
		SD(SE)	0.44(0.43)	0.21(0.20)	0.20(0.20)	0.30(0.31)	0.18(0.19)	0.17(0.18)
	6	Mean(Cov)	1.02(0.96)	0.99(0.95)	1.01(0.95)	0.69(0.95)	0.87(**0.91**)	0.82(**0.86**)
		SD(SE)	0.61(0.62)	0.21(0.20)	0.20(0.20)	0.46(0.47)	0.19(0.19)	0.18(0.19)

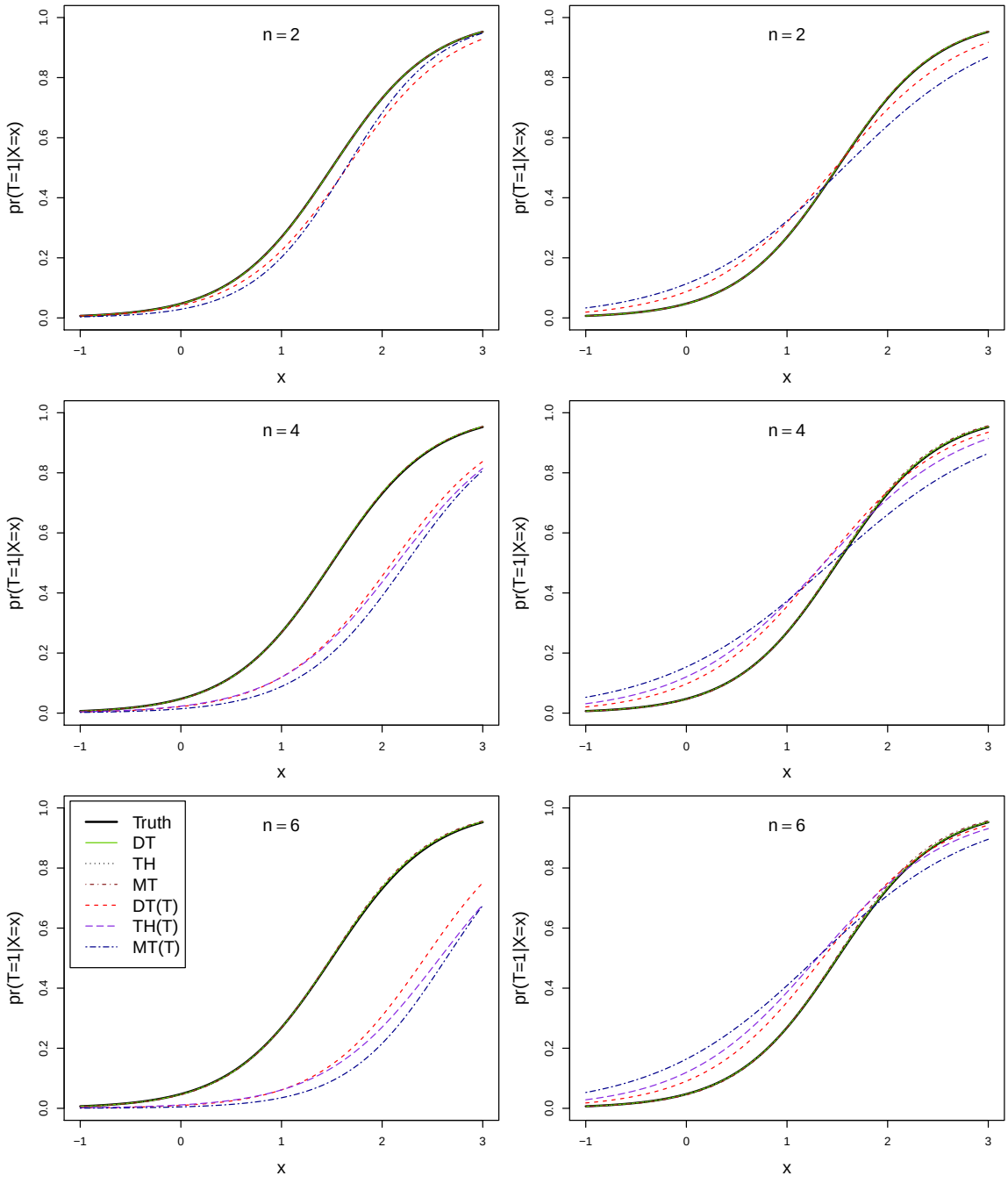


Figure C.1: Plots of the estimated regression functions averaged over 500 data sets for Model 4.1 when $\sigma_+ = 1$ and $n \in \{2, 4, 6\}$. We use DT(T), TH(T), and MT(T) to denote the results obtained under the traditional modeling assumptions for the group testing algorithms DT, TH, and MT, respectively. The panels on the left and right of the figure correspond to thresholding strategies $t(c) = t_0$ and $t(c) = t_0/c$, respectively.

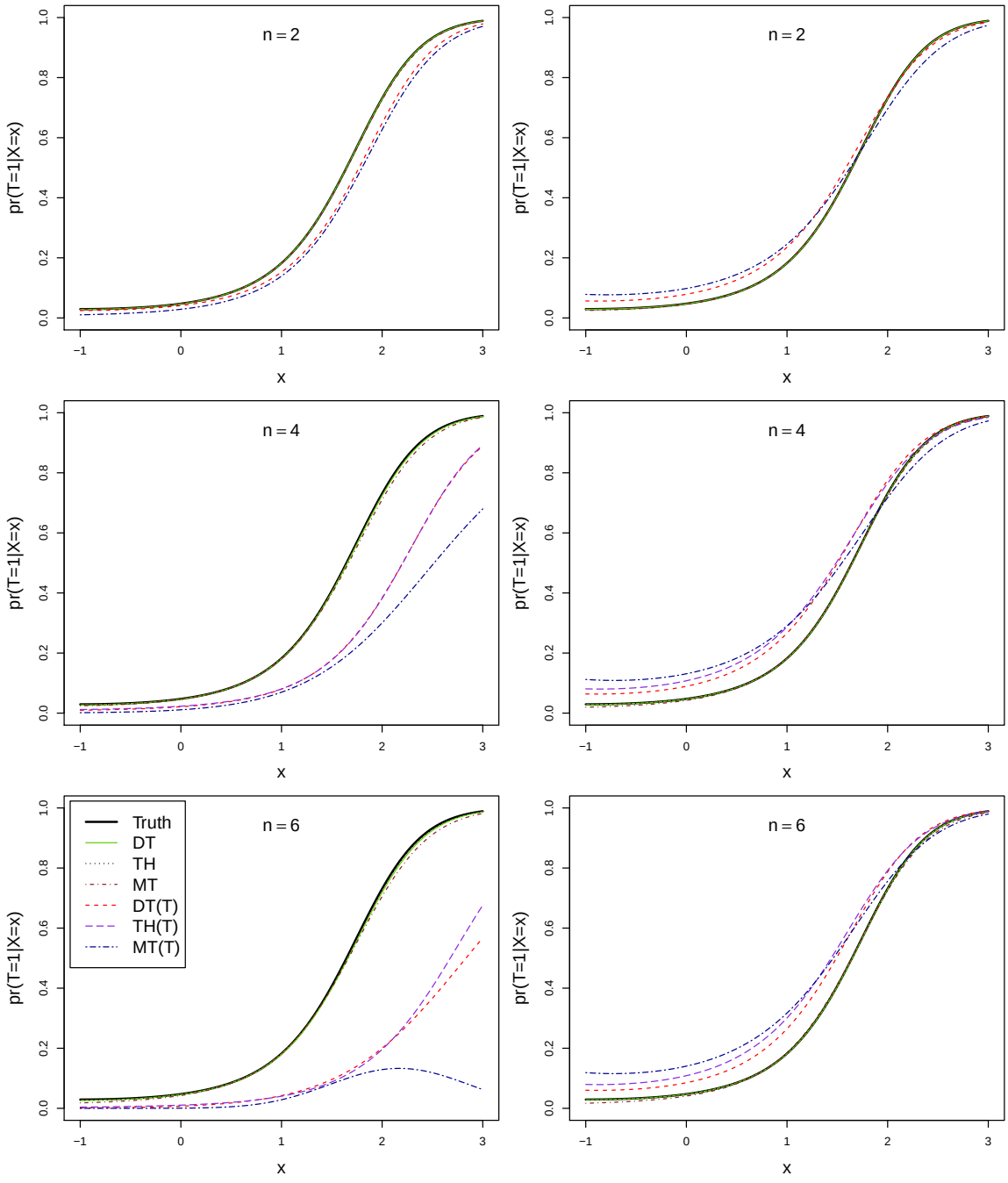


Figure C.2: Plots of the estimated regression functions averaged over 500 data sets for Model 4.2 when $\sigma_+ = 1$ and $n \in \{2, 4, 6\}$. We use DT(T), TH(T), and MT(T) to denote the results obtained under the traditional modeling assumptions for the group testing algorithms DT, TH, and MT, respectively. The panels on the left and right of the figure correspond to thresholding strategies $t(c) = t_0$ and $t(c) = t_0/c$, respectively.

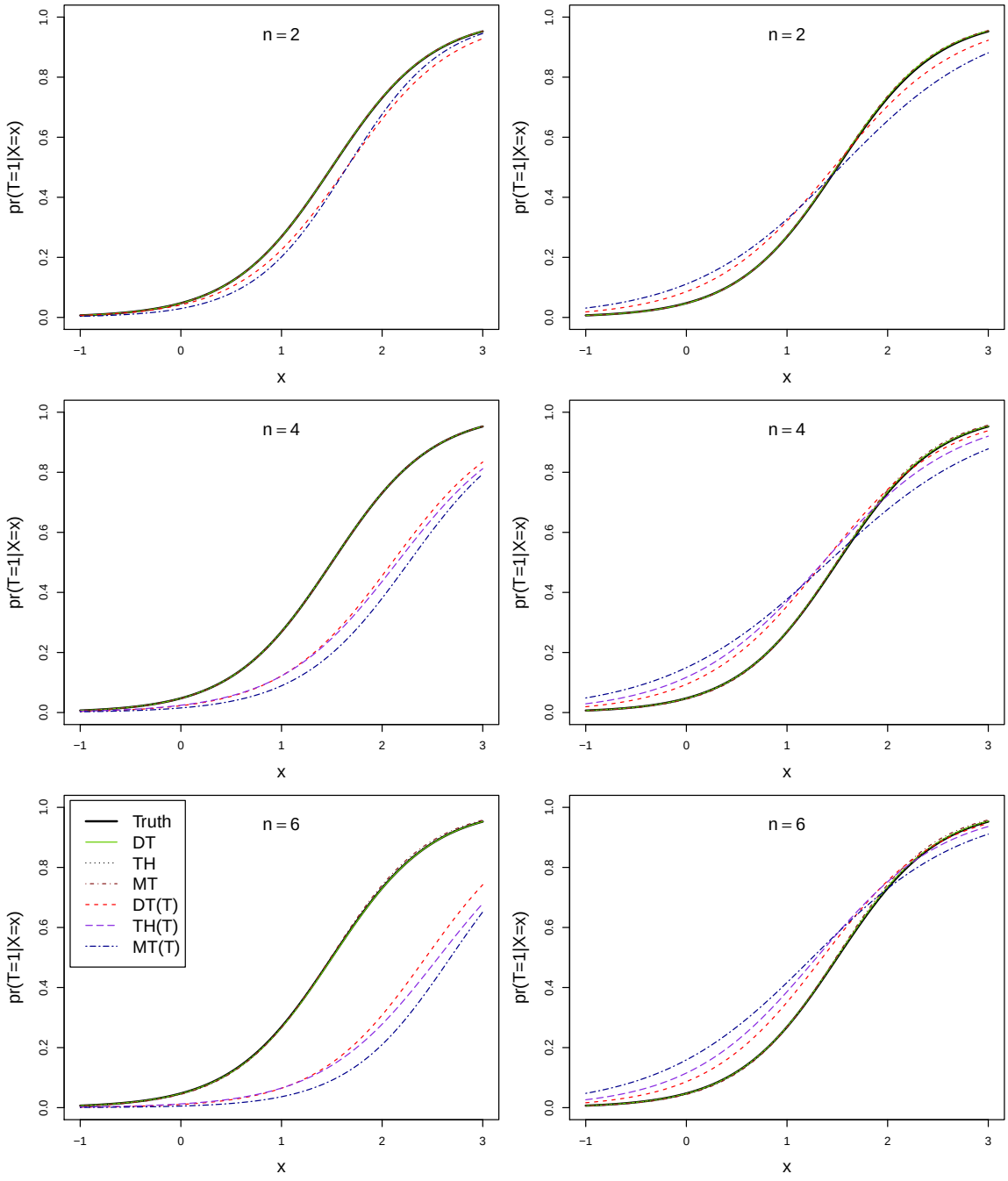


Figure C.3: Plots of the estimated regression functions averaged over 500 data sets for Model 4.3 when $\sigma_+ = 1$, $x_2 = 0$, and $n \in \{2, 4, 6\}$. We use DT(T), TH(T), and MT(T) to denote the results obtained under the traditional modeling assumptions for the group testing algorithms DT, TH, and MT, respectively. The panels on the left and right of the figure correspond to thresholding strategies $t(c) = t_0$ and $t(c) = t_0/c$, respectively.

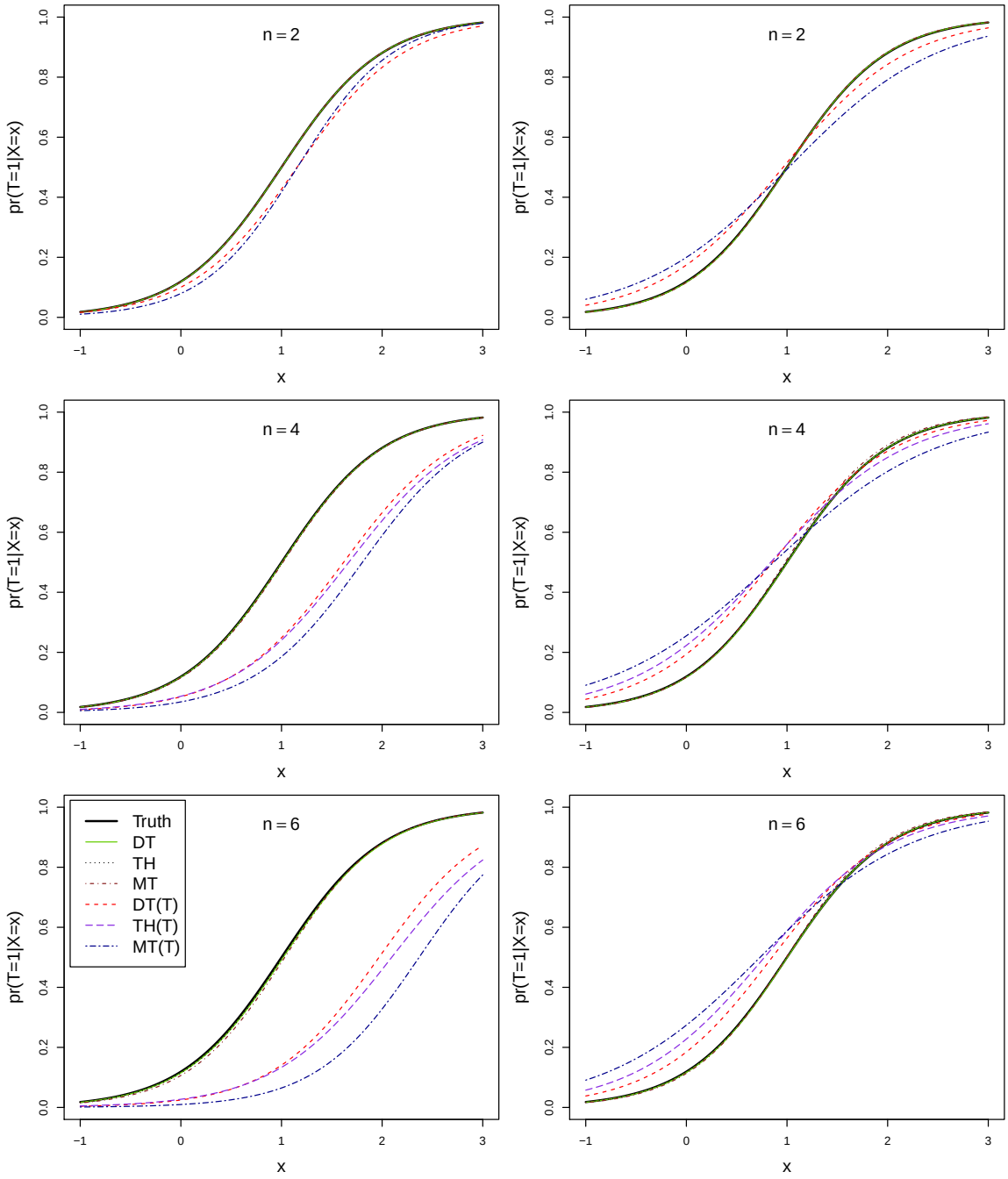


Figure C.4: Plots of the estimated regression functions averaged over 500 data sets for Model 4.3 when $\sigma_+ = 1$, $x_2 = 1$, and $n \in \{2, 4, 6\}$. We use DT(T), TH(T), and MT(T) to denote the results obtained under the traditional modeling assumptions for the group testing algorithms DT, TH, and MT, respectively. The panels on the left and right of the figure correspond to thresholding strategies $t(c) = t_0$ and $t(c) = t_0/c$, respectively.

C.5 Irish HBV Data

This appendix provides a summary of the misclassification error rates pertaining to the data analysis conducted in Section 4.4 of our manuscript. Specifically, we report the false positive rate, which is defined to be the ratio of the number of individuals diagnosed positive who are truly negative to the number of individuals who are truly negative, and the false negative rate, which is defined to be the ratio of the number of individuals diagnosed negative who are truly positive to the number of individuals who are truly positive.

Table C.4: Irish HBV data: Presented results include the mean of the false positive rates (false negative rates) of the 1000 replications under the two different thresholding strategies when $n = 2, 4, 6$. Note, DT and TH denote Dorfman testing and three-stage halving, respectively.

		Random Grouping		Homogeneous Grouping	
	n	DT	TH	DT	TH
When $t(c) = t_0$	2	$0.000(0.156)$	$--(--)$	$0.000(0.145)$	$--(--)$
	4	$0.000(0.397)$	$0.000(0.408)$	$0.000(0.398)$	$0.000(0.400)$
	6	$0.000(0.498)$	$0.000(0.525)$	$0.000(0.509)$	$0.000(0.523)$
When $t(c) = t_0/c$	2	$0.001(0.017)$	$--(--)$	$0.001(0.016)$	$--(--)$
	4	$0.001(0.017)$	$0.001(0.017)$	$0.001(0.016)$	$0.001(0.017)$
	6	$0.001(0.017)$	$0.001(0.017)$	$0.001(0.016)$	$0.001(0.017)$

Bibliography

ALLWRIGHT, S., BRADLEY, F., LONG, J., BARRY, J., THORNTON, L. & PARRY, J. (2000). Prevalence of antibodies to hepatitis B, hepatitis C, and HIV and risk factors in Irish prisoners: Results of a national cross sectional survey. *British Medical Journal* **321**, 78–82.

BILDER, C. R. & TEBBS, J. M. (2005). Empirical Bayesian estimation of the disease transmission probability in multiple-vector-transfer designs. *Biometrical Journal* **47**, 502–16.

BILDER, C. R. & TEBBS, J. M. (2009). Bias, efficiency and agreement for group-testing regression models. *J. Stat. Comput. Simul.* **79**, 67–80.

BONDELL, H., LIU, A. & SCHISTERMAN, E. (2007). Statistical inference based on pooled data: a moment-based estimating equation approach. *Journal of Applied Statistics* **34**, 129–140.

BUSCH, M., CAGLIOTI, S., ROBERTSON, E., MCAULEY, J., TOBLER, L., KAMEL, H., LINNEN, J., SHYAMALA, V., TOMASULO, P. & KLEINMAN, S. (2005). Screening the blood supply for West Nile Virus RNA by nucleic acid amplification testing. *New England Journal of Medicine* **353**, 460–467.

CARDOSO, M., KOERNER, K. & KUBANEK, B. (1998). Mini-pool screening by nucleic acid testing for hepatitis B virus, hepatitis C virus, and HIV: preliminary results. *Transfusion* **38**, 905–907.

CHEN, P., TEBBS, J. M. & BILDER, C. R. (2009). Group testing regression models with fixed and random effects. *Biometrics* **65**, 1270–78.

CUI, X., HÄRDLE, W. & ZHU, L. (2011). The EFM approach for single-index models. *Annals of Statistics* **39**, 1658–88.

CURRIE, M., MCNIVEN, M., YEE, T., SCHIEMER, U. & BOWDEN, F. (2004). Pooling of clinical specimens prior to testing for Chlamydia trachomatis by PCR is accurate and cost saving. *Journal of Clinical Microbiology* **42**, 4866–4867.

DELAIGLE, A. & HALL, P. (2012). Nonparametric regression with homogeneous group testing data. *Annals of Statistics* **40**, 131–58.

DELAIGLE, A. & MEISTER, A. (2011). Nonparametric regression analysis for group testing data. *Journal of American Statistical Association* **106**, 640–50.

DORFMAN, R. (1943). The detection of defective members of large populations. *Annals of Mathematical Statistics* **14**, 436–40.

FAN, J. & GOEBBELS, I. (1996). *Local polynomial modeling and its applications.* Chapman & Hall, London

FARAGGI, D., REISER, B. & SCHISTERMAN, E. (2003). ROC curve analysis for biomarkers based on pooled assessments. *Statistics in Medicine* **22**, 2515–2527.

GASSER, T., KNEIP, A. & KOHLER, W. (1991). A flexible and fast method for automatic smoothing. *Journal of the American Statistical Association* **86**, 643–652.

GASTWIRTH, J. (2000). The efficiency of pooling in the detection of rare mutations. *The American Journal of Human Genetics* **67**, 1036–39.

GASTWIRTH, J. & JOHNSON, W. (1994). A two-stage adaptive group-testing procedure for estimating small proportions. *Journal of the American Statistical Association* **89**, 972–981.

HAMMICK, P. & GASTWIRTH, J. (1994). Group testing for sensitive characteristics: extension to higher prevalence levels. *International Statistical Review/ Revue Internationale de Statistique* **62**, 319–331.

HÄRDLE, W., HALL, P. & ICHIMURA, H. (1993). Optimal smoothing in single-index models. *Annals of Statistics* **21**, 157–8.

HOURFAR, M., JORK, C. SCHOTTSTEDT, V., WEBER-SCHEHL, M., BRIXNER, V., BUSCH, M., GEUSENDAM, G., GUBBE, K., MAHNHARDT, C., MAYR-WOHLFART, U., PICHL, L., ROTH, W., SCHMIDT, M., SEIFRIED, E., WRIGHT, D. & GERMAN RED CROSS NAT STUDY GROUP. (2008). Experience of German Red Cross blood donor services with nucleic acid testing: results of screening more than 30 million blood donations for human immunodeficiency virus-1, hepatitis C virus, and hepatitis B virus. *Transfusion* **48**, 1558–1566.

HUANG, X. & TEBBS, J. M. (2009). On latent-variable model misspecification in structural measurement error models for binary response. *Biometrics* **65**, 710–8.

ICHIMURA, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single index models. *J. Econometrics* **58**, 71–120.

JIRSA, S. (2008). Pooling specimens: a decade of successful cost savings. National STD Prevention Conference, 2008. Chicago, IL.

JOHNSON, W. & PEARSON, L. (1999). Dual screening. *Biometrics* **55**, 867–873.

KIM, H., HUDGENS, M., DREYFUSS, J., WESTREICH, D. & PILCHER, C. (2007). Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics*, **63**, 1152–63.

KLEIN, R. W. & SPADY, R. H. (1993). An efficient semiparametric estimator for binary response models. *Econometrica* **61**, 387–421.

LENNON, J. T. (2007). Diversity and metabolism of marine bacteria cultivated on dissolved DNA. *Applied and Environmental Microbiology* **73**, 2799–2805.

LEWIS, J. L., LOCKARY, V. M. & KOBIC, S. (2012). Cost savings and increased efficiency using a stratified specimen pooling strategy for chlamydia trachomatis and neisseria gonorrhea. *Sexually Transmitted Diseases* **39**, 46–8.

LIN, W. & KULASEKERA, K. B. (2007). Identifiability of single-index models and additive-index models. *Biometrika* **94**, 496–501.

LINDAN, C., MATHUR, M., KUMTA, S., JERAJANI, H., GOGATE, A., SCHACHTER, J. & MONCADA, J. (2005). Utility of pooled urine specimens for detection of chlamydia trachomatis and neisseria gonorrhoeae in men attending public sexually transmitted infection clinics in Mumbai, India, by PCR. *Journal of Clinical Microbiology* **43**, 1674–1677.

LITVAK, E., TU, X. & PAGANO, M. (1994). Screening for presence of a disease by pooling sera samples. *Journal of the American Statistical Association* **89**, 424–434.

LIU, A. & SCHISTERMAN, E. (2003). Comparison of diagnostic accuracy of biomarkers with pooled assessments. *Biometrical Journal* **45**, 631–644.

LIU, A., SCHISTERMAN, E. & TEOH, E. (2004). Sample size and power calculation in comparing diagnostic accuracy of biomarkers with pooled assessments. *Journal of Applied Statistics* **31**, 49–59.

NAGI, M. S. & RAGGI, L. G. (1972). Importance to 'airsac' disease of water supplies contaminated with pathogenic Escherichia coli. *Avian Diseases* **16**, 718–723.

MACK, Y. P. & SILVERMAN, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Z. Wahrsch. Verw. Gebiete* **61**, 405–15.

MCMAHAN, C., TEBBS, J. & BILDER, C. (2013). Regression models for group testing data with pool dilution effects. *Biostatistics* **14**, 284–298.

MUMFORD, S., SCHISTERMAN, E., VEXLER, A. & LIU, A. (2006). Pooling biospecimens and limits of detection: effects on ROC curve analysis. *Biostatistics* **4**, 585–598.

MUÑOZ-ZANZI, C., JOHNSON, W., THURMOND, M. & HIETALA, S. (2000). Pooled-sample testing as a herd-screening tool for detection of bovine viral diarrhea virus persistently infected cattle. *J. Vet. Diagn. Invest.* **12**, 195–203.

ORTEGA, J. M. & RHEINBOLDT, W. C. (1973). *Iterative solution of nonlinear equations in several variables*. Academic Press, New York and London.

PETROV, V. V. (1995). *Limit theorems of probability theory: sequences of independent random variables*. Oxford University Press Inc., New York.

PHATARFOD, R. & SUDBURY, A. (1994). The use of a square array scheme in blood testing. *Statistics in Medicine* **13**, 2337–2343.

PILCHER, C., FISCUS, S., NGUYEN, T., FOUST, E., WOLF, L., WILLIAMS, D., ASHBY, R., O'DOWN, J., MCPHERSON, J., STALZER, B., HIGHTOW, L., MILLER, W., ERON, J., COHEN, M. & LEONE, P. (2005). Detection of acute infections during HIV testing in North Carolina. *New England Journal of Medicine* **352**, 1873–1883.

REMLINGER, K., HUGHES–OLIVER, J., YOUNG, S. & LAM, R. (2006). Statistical design of pools using optimal coverage and minimal collision. *Technometrics* **48**, 133–43.

SCHMIDT, M., ROTH, W., MEYER, H., SEIFRIED, E. & HOURFAR, M. (2005). Nucleic acid test screening of blood donors for orthopoxviruses can potentially prevent dispersion of viral agents in case of bioterrorism. *Transfusion* **45**, 399–403.

STRAMER, S., NOTARI, E., KRYSZTOF, D. & DODD, R. (2013). Hepatitis B virus testing by minipool nucleic acid testing: does it improve blood safety. *Transfusion* **53**, 2525–2537.

THOMPSON, K. (1962). Estimation of the proportion of vectors in a natural population of insects. *Biometrics* **18**, 568–578.

TU, X., LITVAK, E. & PAGANO, M. (1994). Screening tests: Can we get more by doing less? *Statistics in Medicine* **13**, 1905–1919.

VAN, T., MILLER, J., WARSHAUER, D., REISDORF, E., JERRIGAN, D., HUMES, R. & SHULT, P. (2012). Pooling nasopharyngeal/throat swab specimens to increase testing capacity for influenza viruses by PCR. *Journal of Clinical Microbiology* **50**, 891–6.

VANSTEELANDT, S., GOETGHEBEUR, E., THOMAS, I., MATHYS, E. & VAN LOOCK, F. (2005). On the viral safety of plasma pools and plasma derivatives. *Journal of the Royal Statistical Society: Series A* **168**, 345–363.

VANSTEELANDT, E., GOETGHEBEUR, E. & VERSTRAETEN, T. (2000). Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics* **56**, 1126–33.

VENETTE, R., MOON, R. & HUTCHINSON, W. (2002). Strategies and statistics of sampling for rare individuals. *Annual Review Entomology* **47**, 143–74.

VEXLER, A., SCHISTERMAN, E. & LIU, A. (2008). Estimation of ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures. *Statistics in Medicine* **27**, 280–296.

WALD, A. (1943). Tests of hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Association* **54** 426–482.

WANG, J., XUE, L., ZHU, L. & CHONG, Y. S. (2010). Estimation for a partial-linear single-index model. *Annals of Statistics* **38**, 246–74.

WANG, D., ZHOU, H. & KULASEKERA, K. B. (2013). A semi-local likelihood regression estimator of the proportion based on group testing data. *J. Nonparametr. Statist.* **25**, 209–21.

WANG, Y. H. (1993). On the number of successes in independent trials. *Statistica Sinica* **3**, 295–312.

WAHED, M. A., CHOWDHURY, D., NERMELL, B., KHAN, S. I., ILIAS, M., RAHMAN, M., PERSSON, L. A. & VAHTER, M. (2006). A modified routine analysis of arsenic content in drinking-water in Bangladesh by hydride generation-atomic absorption spectrophotometry. *Journal of Health, Population and Nutrition* **24**, 36–41.

WEIN, L. & ZENIOS, S. (1996). Pooled testing for HIV screening: capturing the dilution effect. *Operations Research* **44**, 543–569.

WEISBERG, S. & WELSH, A. H. (1994). Adapting for the missing linear link. *Annals of Statistics* **22**, 1674–1700.

XIA, C. (2006). Asymptotic distributions for two estimators of the single-index model. *Econometric Theory* **22**, 1112–37.

XIA, Y., TONG, H., LI, W. K. & ZHU, L. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B* **64**, 363–410.

XIE, M. (2001). Regression analysis of group testing samples. *Stat. Med.* **20**, 1957–69.

ZENIOS, S. & WEIN, L. (1998). Pooled testing for HIV prevalence estimation: exploiting the dilution effect. *Statistics in Medicine* **17**, 1447–1467.

ZHANG, B., BILDER, C. & TEBBS, J. (2013). Group testing regression model estimation when case identification is a goal. *Biometrical Journal* **55**, 173–89.

ZHU, L. & XUE, L. (2006). Empirical likelihood confidence regions in a partially linear single-index model. *Journal of the Royal Statistical Society: Series B* **68**, 549–70.