2021

# Estimating and Testing Treatment Effects with Misclassified Multivariate Data

Zi Ye
*University of Kentucky*, ye.leave.2016@gmail.com
Digital Object Identifier: https://doi.org/10.13023/etd.2021.324

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Zi Ye, Student

Dr. Solomon Harrar, Major Professor

Dr. Katherine Thompson, Director of Graduate Studies

# Estimating and Testing Treatment Effects with Misclassified Multivariate Data

---

DISSERTATION

---

A dissertation submitted in partial
fulfillment of the requirements for the
degree of Doctor of Philosophy in the
College of Arts and Sciences at the
University of Kentucky

By
Zi Ye

Lexington, Kentucky

Director: Dr. Solomon Harrar, Professor of Statistics

Lexington, Kentucky

2021

ABSTRACT OF DISSERTATION

Estimating and Testing Treatment Effects with Misclassified Multivariate Data

Clinical trials are often used to assess drug efficacy and safety. Participants are sometimes pre-stratified into different groups by diagnostic tools. However, these diagnostic tools are fallible. The traditional method ignores this problem and assumes the diagnostic devices are perfect. This assumption will lead to inefficient and biased estimators. In this era of personalized medicine and measurement-based care, the issues of bias and efficiency are of paramount importance. Despite the prominence, only a few researchers evaluated the treatment effect in the presence of misclassifications in some special cases and most others focus on assessing the accuracy of the diagnostic devices. In this dissertation, we aim to fill in this methodological gap in the estimation of treatment effects in the multivariate and nonparametric contexts. We focus on a pre-post design and address the problem of misclassifications in three distinct situations.

In clinical trials with continuous multiple endpoints, we model the outcome variables as a mixture of multivariate normal distributions to account for the effect of misclassification errors. We propose two methods for estimating and testing treatment effects. When the misclassification errors are known from previous studies, we develop moment-based tests and confidence interval procedures that are accurate in finite samples. When the misclassification errors are unknown, we propose likelihood-based procedures for estimation and testing via the EM algorithm. In addition, methods for sample size and power calculations are developed. The moment-based methods can also be used when the misclassification rates are unknown if validation samples are available. In this case, consistent estimators of the misclassification error rates are derived using a novel distance-based criterion.

When the data are measured on a nonmetric scale or when the distribution of the data is heavy-tailed or skewed, the normality assumption is not valid. In this case, we develop a fully nonparametric method to assess the treatment effect. We model the distribution of the outcomes as a nonparametric mixture of unknown distributions. To overcome identifiability problems, we assume the availability of training data from the component distributions. In the nonparametric setting, functionals of these distribution functions are used to charac-

terize treatment effects. We provide consistent estimators and asymptotic distributions of the estimators of the misclassification error rates as well as the treatment effect. We do not require any assumptions regarding the existence of moments of any order.

Typically, clinical trials involve the collection of baseline covariates which are associated with the misclassification of a patient and treatment outcomes. In this situation, we propose a nonparametric finite mixture of regression models to approximate the distribution of outcomes. We establish identifiability conditions and derive an estimation procedure using the kernel methods and the EM algorithm.

Simulation results show significant advantages of the proposed methods in terms of bias reduction, coverage probability, and power. The applications of the methods are illustrated with datasets from sleep deprivation and electroencephalogram (EEG) studies.

KEYWORDS: Nonparametric analysis, EM algorithm, Multivariate Data, Asymptotic distribution

Zi Ye

August 9, 2021

Date

Estimating and Testing Treatment Effects with Misclassified Multivariate Data

By

Zi Ye

Solomon Harrar
Director of Dissertation

Katherine Thompson
Director of Graduate Studies

August 9, 2021
Date

Dedicated to my family.

# ACKNOWLEDGMENTS

First of all, I would like to express my gratitude to my advisor, Dr. Solomon W. Harrar, for his insight, guidance, encouragement, and support on my research. I also want to thank my dissertation committee members, Dr. Arnold Stromberg, Dr. Derek Young, Dr. Richard Charnigo, Dr. Olga Vsevolozhskaya, and Dr. Xiangrong Yin, for their time and valuable suggestions in improving this dissertation. Moreover, thanks to the University of Kentucky Statistics Department for providing me with a delightful and unforgettable five years. Last but not least, I would like to thank my parents and Xuan for their love, help, and support. This journey would be impossible without them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# Chapter 1 Introduction

## 1.1 Background

In drug (therapy) development, clinical trials are commonly used to assess the efficacy and safety of a treatment. In some cases, diagnostic devices or biomarkers are used, especially in the recruitment stage, to separate the sample population into subgroups that may respond differently to the treatment. By comparing the responses from subgroups, we can analyze if the treatment has different effects on these subgroups.

In this dissertation, we focus on a pre-stratified pre-post design. The participants are first stratified into different subgroups by the results of a diagnostic tool. Then, all the participants will receive the treatment, and response variables are measured before and after receiving the treatment. We can quantify the effect on each subgroup by comparing the response variables before and after the treatment. This design is commonly used in clinical trials. For example, Eling et al. (2006) examine the reduction in the demented patient by separating participants into demented and healthy groups based on an examination result. Similar examples abound in the biomedical literature (Castro et al., 2012; Gentili et al., 2008; O'Donnell et al., 1999).

However, the diagnostic tools do not usually have perfect accuracy. The traditional method ignores this problem and assumes the diagnostic devices are perfect. This assumption will lead to inefficient and biased estimators. In this era of personalized medicine and measurement-based care, the issues of bias and efficiency are of paramount importance. Despite the prominence, only a few researchers evaluated treatment effects in the presence of misclassifications in some particular cases. Most others focus on assessing the accuracy of the diagnostic devices. This dissertation aims to fill in this methodological gap and address the estimation of treatment effects in the multivariate and nonparametric contexts.

## 1.2 Multivariate Parametric Method

In clinical trials with continuous multiple endpoints, we can use multivariate normal distributions to model the outcomes. Traditionally, we assume misclassification errors do not exist and use the Hotelling $T^2$ statistic (Anderson, 2003) to make inferences about treatment effect. More precisely, let $\boldsymbol{Y}_{ij} = (\boldsymbol{Y}_{ij}^{(1)\top}, \boldsymbol{Y}_{ij}^{(2)\top})^\top$ be the pre- and post-outcome measures on a $p$-dimensional vector for the $j$th individual in the $i$th group, where $j = 1, \ldots, n_i$, and $i = 1, 2$. Here, $\{\mathbf{Y}_{11}, \cdots, \mathbf{Y}_{1n_1}\}$ is assumed to be a random sample from $N_{2p}(\boldsymbol{\eta}_1, \Sigma)$, and $\{\mathbf{Y}_{21}, \cdots, \mathbf{Y}_{2n_2}\}$ is a random sample from $N_{2p}(\boldsymbol{\eta}_2, \Sigma)$. The two samples are assumed to be mutually independent. Here $\boldsymbol{\eta}_1 = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_1 + \boldsymbol{\tau}_1)$ and $\boldsymbol{\eta}_2 = (\boldsymbol{\mu}_2, \boldsymbol{\mu}_2 + \boldsymbol{\tau}_2)$, where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ are pre-treatment response means in groups 1 and 2, respectively, and $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ are treatment effects in groups 1 and 2, respectively. The parameter of interest is

$$\boldsymbol{\Delta} = \boldsymbol{\tau}_1 - \boldsymbol{\tau}_2,$$

where $\boldsymbol{\Delta} = (d_1, \ldots, d_p)^\top$ is the vector of differences in the treatment effect between group 1 and 2. Define

$$\begin{aligned}
C &= (-I_p, I_p)_{p\times 2p}, \quad \overline{\mathbf{Y}}_1 = n_1^{-1}\sum_{j=1}^{n_1}\mathbf{Y}_{1j}, \quad \overline{\mathbf{Y}}_2 = n_2^{-1}\sum_{j=1}^{n_2}\mathbf{Y}_{2j},\\
S_P &= (n_1 + n_2 - 2)^{-1}((n_1 - 1)S_1 + (n_2 - 1)S_2),\\
S_1 &= (n_1 - 1)^{-1}\sum_{j=1}^{n_1}(\mathbf{Y}_{1j} - \overline{\mathbf{Y}}_1)(\mathbf{Y}_{1j} - \overline{\mathbf{Y}}_1)^\top \text{ and}\\
S_2 &= (n_2 - 1)^{-1}\sum_{j=1}^{n_2}(\mathbf{Y}_{2j} - \overline{\mathbf{Y}}_2)(\mathbf{Y}_{2j} - \overline{\mathbf{Y}}_2)^\top,
\end{aligned}$$

where $\mathrm{I}_p$ is identity matrix of order $p$. Then we can estimate $\boldsymbol{\Delta}$ by

$$\widehat{\boldsymbol{\Delta}} = C(\overline{\boldsymbol{Y}}_1 - \overline{\boldsymbol{Y}}_2).$$

To test the hypothesis $H_0 : \boldsymbol{\Delta} = \boldsymbol{\Delta}_0$, we can use the Hotelling $T^2$ statistic with exact distribution given by Anderson (2003) as follows

$$\begin{aligned}
T^2 =& \frac{n_1 n_2}{n}(C(\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2) - \boldsymbol{\Delta}_0)^\top (CS_PC^\top)^{-1}(C(\overline{\mathbf{Y}}_1 - \overline{\mathbf{Y}}_2) - \boldsymbol{\Delta}_0) \qquad (1.1)\\
\overset{H_0}{\sim}& \frac{(n-2)p}{n-p-1}F_{p,n-p-1},
\end{aligned}$$

Furthermore, a $(1-\alpha)$ confidence region for $\boldsymbol{\Delta}$ is obtained by inverting the Hotellings $T^2$ test as

$$\left\{\boldsymbol{\Delta}: \frac{n_1 n_2}{n_1+n_2}(C(\overline{\mathbf{Y}}_1-\overline{\mathbf{Y}}_2)-\boldsymbol{\Delta})^\top (CS_PC^\top)^{-1}(C(\overline{\mathbf{Y}}_1-\overline{\mathbf{Y}}_2)-\boldsymbol{\Delta}) \right.$$
$$\left. \leq \frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1;1-\alpha}\right\},$$

where $F_{p,n_1+n_2-p-1;1-\alpha}$ is the lower $1-\alpha$ quantile of the $F$ distribution with degrees of freedoms $(p, n_1+n_2-p-1)$.

In the context of study design, suppose we are interested in testing the null hypothesis $H_0 : \boldsymbol{\tau}_D - \boldsymbol{\tau}_H = \boldsymbol{\Delta}_0$. The distribution of the test statistic in (2.1) at the alternative $H_1 : \boldsymbol{\tau}_D - \boldsymbol{\tau}_H = \boldsymbol{\Delta}_1$ for some fixed $\boldsymbol{\Delta}_1 \neq \boldsymbol{\Delta}_0$ is

$$\begin{aligned} T^2 =& \frac{n_1 n_2}{n_1+n_2}(C\overline{\mathbf{Y}}-\boldsymbol{\Delta}_0)^\top (CS_PC^\top)^{-1}(C\overline{\mathbf{Y}}-\boldsymbol{\Delta}_0) \\ \sim& \frac{(n_1+n_2-2)p}{n_1+n_2-p-1}F_{p,n_1+n_2-p-1}\left(\frac{n_1 n_2}{n_1+n_2}(\boldsymbol{\Delta}_1-\boldsymbol{\Delta}_0)^\top (C\Sigma C^\top)^{-1}(\boldsymbol{\Delta}_1-\boldsymbol{\Delta}_0)\right), \end{aligned}$$

where $F_{p,n_1+n_2-p-1}(\xi)$ is the $F$ distribution with degrees of freedom $(p, n_1+n_2-p-1)$ and noncentrality parameter $\xi$. To guarantee a nominal test size $\alpha$ and power $1-\beta$, the required total sample size $n = n_D + n_H$, where $n_D/n_H = \pi$ and $0 < \pi < \infty$, is the solution of

$$P\left(T^2 > \frac{(n-2)p}{n-p-1}F_{p,n-p-1;1-\alpha}\middle| H_1\right) = P(Y > F_{p,n-p-1;1-\alpha}) = 1-\beta,$$

where

$$Y \sim F_{p,n-p-1}\left(\frac{n\pi}{(1+\pi)^2}(\boldsymbol{\Delta}_1-\boldsymbol{\Delta}_0)^\top (C\Sigma C^\top)^{-1}(\boldsymbol{\Delta}_1-\boldsymbol{\Delta}_0)\right).$$

This equation has to be solved numerically.

However, when misclassification errors exist, the actually observed outcome measures are affected by the error rates. Assume the misclassification error rates are $\delta_g$ in groups $g$, $g = 1, 2$. The distribution of the outcomes become mixtures of multivariate distributions. More specifically,

$$f_g(\boldsymbol{y}_{gi}) = (1-\delta_g)\phi(\boldsymbol{y}_{gi}|\boldsymbol{\eta}_g, \Sigma) + \delta_g \phi(\boldsymbol{y}_{gi}|\boldsymbol{\eta}_{g'}, \Sigma), \text{ for } g \neq g', g, g' = 1, 2.$$

Then the traditional estimator $\widehat{\boldsymbol{\Delta}}$ for $\boldsymbol{\Delta}$ has expectation

$$E(\widehat{\boldsymbol{\Delta}}) = E(C(\overline{\boldsymbol{Y}}_1 - \overline{\boldsymbol{Y}}_2)) = (1 - \delta_1 - \delta_2)(\boldsymbol{\tau}_1 - \boldsymbol{\tau}_2) = (1 - \delta_1 - \delta_2)\boldsymbol{\Delta}.$$

This shows when the misclassification errors exist, the traditional estimator is biased and the bias is affected by misclassification errors. Moreover, the outcome variables are not distributed as multivariate normal distributions, then the test statistic in (1.1) is not distributed as $F_{p,n-p-1}$ under $H_0$. The power and sample size calculation based on this test statistic become overly optimistic and misleading.

### 1.3 Estimation of Misclassification Error

Accurate estimations of the misclassification error rates of the classifiers are required to evaluate the effect of a treatment. This problem can be framed as estimating mixing proportions in mixture models. Let $X_1, \ldots, X_n$ be i.i.d random variables from a finite mixture of $m > 1$ arbitrary distributions. Suppose the cumulative distribution function of $X_i$ is

$$F = \sum_{j=1}^{m} \lambda_j F_j,$$

where $F_j$ and $\lambda_j$ are the cumulative distribution function and mixture proportion of $j$th component, $j = 1, \ldots, m$. Hall (1981) proposed nonparametric estimators for the mixture proportions when training samples are available from each component distributions, $F_1, \ldots, F_m$. The estimators are derived by combining the contaminated (original) and validation (training) data. The mixing proportions $\lambda_j$ can by estimated by minimizing

$$\Delta(\boldsymbol{\lambda}) = \left| \int_{-\infty}^{\infty} \delta \left( \widehat{F}(x) - \sum_{j=1}^{m} \lambda_j \widehat{F}_j(x) \right) w(x) dx \right|, \tag{1.2}$$

where $\widehat{F}$ and $\widehat{F}_j$ are empirical versions of $F$ and $F_j$, respectively. The primary focus of Hall (1981) is when $\delta(x) = x^2$ and $w(x) \equiv 1$. They also assume

$$\int_{-\infty}^{\infty} |x|^{1+\epsilon} dF(x) < \infty, \tag{1.3}$$

for some $\epsilon > 0$. This requirement imposes a restriction on the tails of the component distributions. Especially, it requires that the first moment of the component distributions to exist.

In clinical trial settings, more expensive and accurate diagnostic devices can sometimes be used to know the actual group membership for some participants. In this case, we can obtain validation data to enhance the accuracy of the treatment effect estimation. Then the misclassification error rates can be estimated as the mixture proportions in the mixture models.

Inspired by Hall (1981)'s method, we propose two estimators for the misclassification error rates in the multivariate and nonparametric contexts. In the multivariate setting, we obtain estimators of $\delta_g$ by minimizing the distance between the mean of the original data and the mixture of means of the two groups in the validation data. We also relaxed the requirement of (1.3) in the nonparametric model by setting the weight function in (1.2) as the weighted average of the empirical marginal distribution functions of observations in the two groups.

### 1.4 Nonparametric Method

In some applications, data are measured on a nonmetric scale, or the distribution of the data is heavy-tailed or skewed, and the normality assumption would not be valid. In these cases, nonparametric methods can be utilized to assess the treatment effect.

Suppose we have subjects from two groups $g = 1, 2$ that are observed at two (pre and post treatment) time points $t = 1, 2$ and the paired observations are denoted as $\boldsymbol{X}_{gk} = (X_{g1k}, X_{g2k})$, $k = 1, \ldots, n_g$. Let $X_{gt1}, \ldots, X_{gtn_g}$ be identically and independently distributed according to $F_{gt}$. When the misclassification errors are assumed to be zero, the normalized distribution functions $F_{gt}$ for $X_{gti}$ can be estimated by

$$\widehat{F}_{gt}(x) = \frac{1}{n_{gt}} \sum_{k=1}^{n_g} c(x - X_{gtk}), \text{ where } c(x) = \begin{cases} 0, & x < 0, \\ \frac{1}{2}, & x = 0, \\ 1, & x > 0. \end{cases}$$

To quantify the difference between two distribution functions, Brunner and Munzel (2000) proposed to formulate hypotheses in terms of the nonparametric relative effects. These effects are defined by comparing each marginal distribution function with the average distribution function. Let $G$ be the average of the distribution functions in the two groups and

at the two time points, i.e.

$$G = \frac{1}{4}(F_{11} + F_{12} + F_{21} + F_{22}).$$

Using the average distribution function $G$, define

$$p_{gt} = \int G dF_{gt} = 1 - \int F_{gt} dG,$$

for $g, t = 1, 2$ as the *nonparametric effect* at time point $t$ in group $g$ *relative* to the average of the marginal distributions, $G$. Using the nonparametric relative effects, the treatment effect of interest in the two group pre-post design is

$$p_{\text{I}} = (p_{12} - p_{11}) - (p_{22} - p_{21}).$$

According to the calculations in Harrar et al. (2020), the treatment effect can be expressed as

$$p_{\text{I}} = \frac{1}{2} \int (F_{11} + F_{22}) d(F_{12} + F_{21}) - 1.$$

We can use plug-in method and estimate the treatment effect as

$$\widehat{p}_{\text{I}} = \frac{1}{2} \int (\widehat{F}_{11} + \widehat{F}_{22}) d(\widehat{F}_{12} + \widehat{F}_{21}) - 1.$$

By results in Brunner et al. (2018), we have

$$\sqrt{N}(\widehat{p}_{\text{I}} - p_{\text{I}}) \doteqdot \frac{\sqrt{N}}{n_1} \sum_{k=1}^{n_1} W_1(\boldsymbol{X}_{1k}) + \frac{\sqrt{N}}{n_2} \sum_{k=1}^{n_2} W_2(\boldsymbol{X}_{2k}) - 2p_I,$$

where $\doteqdot$ means asymptotic equivalent,$N = 2(n_1 + n_2)$, and

$$\begin{aligned} W_1(\boldsymbol{X}_{1k}) =& \frac{1}{2} \left( F_{11}(X_{12k}) + F_{22}(X_{12k}) - F_{12}(X_{11k}) - F_{21}(X_{11k}) \right), \text{ and} \\ W_2(\boldsymbol{X}_{2k}) =& \frac{1}{2} \left( F_{11}(X_{21k}) + F_{22}(X_{21k}) - F_{12}(X_{22k}) - F_{21}(X_{22k}) \right). \end{aligned}$$

Note that $W_1(\boldsymbol{X}_{1k})$ and $W_2(\boldsymbol{X}_{2k})$ are independent and identically distributed random variables. By Central Limit Theorem, $\sqrt{N}(\widehat{p}_{\text{I}} - p_{\text{I}})$ is asymptotically normally distributed. Based on this result, if we want to test

$$H_0 : p_{\text{I}} = 0 \text{ vs } H_a : p_{\text{I}} \neq 0,$$

we can use the test statistic

$$T^2 = \frac{\sqrt{N}(\widehat{p}_{\mathrm{I}} - p_{\mathrm{I}})}{\sqrt{S^2}} \xrightarrow{D} Z \overset{H_0}{\sim} N(0,1),$$

where

$$S^2 = \frac{N}{n_1(n_1-1)} \sum_{k=1}^{n_1} (\widehat{W}_1(\boldsymbol{X}_{1k}) - \overline{\widehat{W}}_1(\boldsymbol{X}_{1k}))^2 + \frac{N}{n_2(n_2-1)} \sum_{k=1}^{n_2} (\widehat{W}_2(\boldsymbol{X}_{2k}) - \overline{\widehat{W}}_2(\boldsymbol{X}_{2k}))^2,$$

and $\widehat{W}_g(\boldsymbol{X}_{gk})$ is the empirical version of $W_g(\boldsymbol{X}_{gk})$, i.e. $F_{gt}$ is replaced by $\widehat{F}_{gt}$. An asymptotic $(1-\alpha)100\%$ confidence interval for $p_I$ can be derived from

$$P\left(\widehat{p}_{\mathrm{I}} - \frac{z_{\alpha/2}\sqrt{S^2}}{\sqrt{N}} \leq p_{\mathrm{I}} \leq \widehat{p}_{\mathrm{I}} + \frac{z_{\alpha/2}\sqrt{S^2}}{\sqrt{N}}\right) \to 1-\alpha,$$

where $z_{\alpha/2}$ denotes the $(1-\alpha/2)$th-quantile of the standard normal distribution.

However, when the classifier is fallible, the observations we obtain from one group are contaminated by observations from the other group. Suppose the misclassification error rates for the classifier is $\delta_g$ in group $g$, $g = 1, 2$. Then the distribution function of the observations is a mixture of two distribution functions, i.e.,

$$F^*_{gt} = (1-\delta_g)F_{gt} + \delta_g F_{g't},$$

where $F^*_{gt}$ is the distribution function of observations from participants classified in group $g$, $g \neq g'$, $g, g' = 1, 2$. When $\delta_g$s are known, we can use the empirical distributions to estimate $F^*_{gt}$. Notice that

$$\begin{aligned} p_I =& \frac{1}{2(1-\delta_1-\delta_2)} \int (F^*_{11} + F^*_{22}) d(F^*_{12} + F^*_{21}) \\ &+ \frac{\delta_1 - \delta_2}{2(1-\delta_1-\delta_2)^2} \int (F^*_{11} - F^*_{21}) d(F^*_{12} - F^*_{22}) - \frac{1}{1-\delta_1-\delta_2}. \end{aligned}$$

The estimation for the treatment effect can be seriously biased if we ignore the misclassification errors. Chapter 4 investigates this problem and develops a fully nonparametric method for estimating and testing the treatment effect.

## 1.5 Covariate Adjustment

Typically, clinical trials involve baseline covariates associated with the misclassification of a patient and treatment outcomes. Linear regressions are commonly used to analyze the

effects of covariates on the response variables. When misclassification errors exist, the distribution of response variables can be modeled as a mixture of linear regression models. More specifically, assume that $\{\boldsymbol{X}_{gi}, \boldsymbol{Y}_{gi}\}$, $g = 1, 2$, are covariates and outcome variables for patient $i$ diagnosed in group $g$. Assume the misclassification error rates for the classifier is $\delta_g$ in group $g$, then the conditional distribution of $\boldsymbol{Y}_g$ given $\boldsymbol{X}_g = \boldsymbol{x}$ can be written as

$$\boldsymbol{Y}_g|\boldsymbol{X}_g = \boldsymbol{x} \sim (1-\delta_g)N(B_g^\top \boldsymbol{x}, \Sigma_g) + \delta_g N(B_{g'}^\top \boldsymbol{x}, \Sigma_{g'}) \text{ for } g \neq g', g, g' = 1, 2, \quad (1.4)$$

where $B_g$ is the regression coefficient matrix for group $g$. EM algorithm can be applied to estimate the parameters $\boldsymbol{\theta} = \{\delta_1, \delta_2, B_1, B_2, \Sigma_1, \Sigma_2\}$ in (1.4).

The linear assumptions in 1.4 are restrictive for some applications. Moreover, the misclassification error rates may be affected by the covariates and do no remain constant. To relax these restrictions, we propose a mixture of nonparametric regression models.

### 1.6 Organization of This Dissertation

This dissertation is organized as follows. For outcomes with continuous multiple endpoints, Chapter 2 uses a mixture of multivariate normal distributions to account for the effect of misclassification errors. We propose two methods for estimating and testing treatment effects. In addition, methods for sample size and power calculations are developed. In Chapter 3, we refine the methods in Chapter 2 by validation (training) data when available. We derive consistent estimators of the misclassification error rates using a novel distance-based criterion. When the normality assumptions are not appropriate but validation data is available, we develop a fully nonparametric method in Chapter 4. We model the distribution of the outcomes by a nonparametric mixture of unknown distributions. Functionals of these distributions are used to characterize the treatment effects. We provide consistent estimators and asymptotic distributions of estimators of the misclassification error rates as well as the treatment effect. In Chapter 5, we propose a nonparametric finite mixture of regression models to incorporate covariates information. We establish identifiability conditions and derive an estimation procedure using the kernel methods and EM algorithm. We conclude the dissertation with discussions and conclusions in Chapter 6.

# Chapter 2 Multivariate Treatment Effects in Contaminated Clinical Trials

## 2.1 Introduction

In the development of drugs (therapy), clinical trials are commonly used to assess the efficacy and safety of a treatment. In some cases, diagnostic devices or biomarkers are used, especially in the recruitment stage, to separate the sample population into subgroups that may respond differently to the treatment. However, such diagnostic tools usually do not have perfect accuracy. In general, the misclassification error rates (false positive and false negative rates) of these devices are unknown or assumed to be zero. They will cause contamination in separating the sample populations, resulting in biased estimation of treatment effects and overly optimistic sample size and power calculations. If we do not have a sufficient sample size in clinical studies, we may fail to detect a significant effect when it is present.

In the era of personalized medicine and measurement-based care, this issue of misclassifications in pre-stratified clinical trials has become prominent. US Food and Drug Administration published a concept paper (Hinman et al., 2006) that recommends the clinical validity (i.e., the ability of a test to classify subjects correctly) and clinical utility (i.e., the ability of a test to result in a classification that will improve the benefit/reduce the risk of a drug) of a test be established in a pre-clinical pilot feasibility study. This goal can be achieved through a pre-stratified (by diagnostic devices) randomized placebo-controlled design or a pre-stratified pre-post or matched paired design. This chapter focuses on the second type of design, but the method presented can be adapted easily to the first type of design.

Despite the prominence of the issue, only a few works evaluated clinical validity in the presence of diagnostic or screening misclassification. Most works focus on evaluating the diagnostic devices themselves. Flahault et al. (2005) provide tables for sample size determination in diagnostic tests studies. Remotely related work is that of Lin et al. (2011) which proposes sample-size adjustment in the situations where the group membership cannot be

ascertained until after the collection of sample. They proposed adjusting the sample size according to a quantity called expected power in an ad hoc manner. Liu et al. (2009) investigated the estimation of continuous outcomes in the framework of enriched randomized placebo-controlled trials where randomized treatment is only conducted on the subjects diagnosed as positive and the accuracy of diagnostic devices is not perfect. Under the same design, Liu and Lin (2008) and Chen et al. (2013) studied binary and censored outcomes, respectively. Li et al. (2015) analyzed the impact of companion diagnostic device performance on the clinical validity of personalized medicine under the assumption that the true values of the model parameters are known. However, in practice, the parameter values are rarely known and need to be estimated from the observed data. Recently, Harrar et al. (2016) tackled the estimation, sample size, and power calculation for a treatment effect in the pre-post or matched-pair design accounting for possible diagnostic inaccuracy. However, their paper only considered the univariate case. In many trials, multiple outcomes (endpoints) are assessed. The estimation and testing procedure in the context of diagnostic misclassification is a lot more involved. Furthermore, the multivariate situation requires larger sample sizes, and suitable finite-sample approximation is crucial.

This chapter aims to provide a complete set of methods for estimating and testing treatment effects with multiple (multivariate) end-points in a pre-post design, where a diagnostic device used for the screening (treatment assignment) is prone to misclassification errors. The error rates may be available for some diagnostic devices from prior studies or evaluations of the devices. We develop a moment-based test and confidence set procedures that are accurate in finite (small and moderate) samples for this situation. The moment-based method is generally easier to apply, accurate, and computationally inexpensive. However, in some applications, the misclassification error rates for the diagnostic devices may not be known in advance. For this situation, we propose a likelihood-based procedure for estimation and testing via an EM algorithm. We further develop a hybrid method that aims to benefit from the advantages of both the moment- and likelihood-based approaches. In the hybrid method, the misclassification error rates are obtained from the likelihood-based procedure followed by estimation of the treatment effects by the moment-based approach. Furthermore, the chapter provides sample-size determination and power calculation for-

mulas for designing a study for a given specification by utilizing a novel finite-sample approximation for the distribution of the moment-based statistic. The formulas produce reasonable and reliable estimates of sample sizes and powers by accounting for misclassification error rates.

To achieve the above aims, we organize the chapter into seven sections, including the present one. Section 2 presents the statistical model. In Section 3, we describe the theoretical motivation and derive moment-based and likelihood-based solutions. In Section 4, we derive formulas for sample size and power calculations. We illustrate the utility of the methods developed in Sections 3 and 4 with a simulation study and real-data analysis in Sections 5 and 6, respectively. We conclude the chapter with discussions and remarks in Section 7. All proofs and technical details are placed in the Appendix.

## 2.2 Statistical Model

### Model and Parameter of Interest

Suppose a diagnostic test is applied to recruit $n = n_D + n_H$ subjects, of which $n_D$ and $n_H$ were diagnosed as positive and negative, respectively. Two probabilities of interest in diagnostic test evaluation are positive predictive value (PPV) and negative predictive value (NPV). Note that PPV is the probability that a person with a positive result will have the clinical condition, say disease, of interest. In contrast, NPV is the probability that a person with a negative result will be free from the clinical condition of interest. Here, we denote the PPV and NPV of the tool in use for clinical diagnosis as $(1-\epsilon)$ and $(1-\delta)$, respectively.

Let $\mathbf{Y}_{ij} = (\mathbf{Y}_{ij}^{(1)\top}, \mathbf{Y}_{ij}^{(2)\top})^\top$ be the pre and post outcome measures on a $p$-dimensional vector for the $j$th individual in the $i$th group, where $j = 1, \cdots, n_i$ and $i = D, H$. Further, let $Z_{ij}$ be the true disease status and $X_{ij}$ be the predicted disease status of the $j$th subject in the $i$th group. Here, $Z_{ij}$ is not an observable random variable. The set of possible values for $Z_{ij}$ and $X_{ij}$ are $\{D, H\}$, where $D$ and $H$ stand for diseased and healthy, respectively. In this notation, $X_{Dj} = D$ and $X_{Hj} = H$. Denote the conditional distributions of $\mathbf{Y}_{ij}$ by

$f(\mathbf{y}_{ij}|\cdot)$ and of $Z_{ij}$ given $X_{ij}$ by $P(Z_{ij}|X_{ij} = x)$. Using the Total Probability Law, we have

$$\begin{aligned} f(\mathbf{y}_{ij}|X_{ij} = x) =& f(\mathbf{y}_{ij}|Z_{ij} = D, X_{ij} = x)P(Z_{ij} = D|X_{ij} = x) \\ &+ f(\mathbf{y}_{ij}|Z_{ij} = H, X_{ij} = x)P(Z_{ij} = H|X_{ij} = x), \end{aligned}$$

for $x \in \{D, H\}$.

We assume that the pre and post outcome measures are continuous and can be modeled by a multivariate normal distribution given the predicted disease condition. That is, $f(\mathbf{y}_{ij}|X_{ij})$ is the pdf of a mixture of multivariate normal distributions that have equal covariance matrices and the mixing probabilities are PPV $(1 - \epsilon)$ and NPV $(1 - \delta)$ in the diseased and healthy groups, respectively. More specifically,

$$\begin{aligned} f(\mathbf{y}_{ij}|X_{ij} = x, \boldsymbol{\theta}) =& \{(1 - \epsilon)\phi(\mathbf{y}_{ij}|\boldsymbol{\eta}_D, \Sigma) + \epsilon\phi(\mathbf{y}_{ij}|\boldsymbol{\eta}_H, \Sigma)\}I_{\{D\}}(x) \\ &+ \{\delta\phi(\mathbf{y}_{ij}|\boldsymbol{\eta}_D, \Sigma) + (1 - \delta)\phi(\mathbf{y}_{ij}|\boldsymbol{\eta}_H, \Sigma)\}I_{\{H\}}(x), \end{aligned}$$

where $\boldsymbol{\theta} = (\epsilon, \delta, \boldsymbol{\eta}_D, \boldsymbol{\eta}_H, \Sigma)$, $I_A(x)$ is indicator function of the set $A$, and $\phi(\mathbf{Y}|\boldsymbol{\eta}, \Sigma)$ is the pdf of a multivariate normal distribution with mean $\boldsymbol{\eta}$ and positive definite covariate matrix $\Sigma$. Here, $\mathbf{Y}_{ij}$ is assumed to be conditionally independent of $X_{ij}$ given $Z_{ij}$, and $X_{ij}$ is a fixed design variable. We know that finite mixture of the multivariate Gaussian family are identifiable up to label switching (Yakowitz and Spragins, 1968). To avoid the label-switching problem here, we assume $0 \leq \epsilon < 0.5$, $0 \leq \delta < 0.5$ and $\boldsymbol{\mu}_D \neq \boldsymbol{\mu}_H$. For practical applications, we need information outside the collected data to ensure these assumptions hold.

We write $\boldsymbol{\eta}_D = (\boldsymbol{\mu}_D, \boldsymbol{\mu}_D + \boldsymbol{\tau}_D)^\top$ and $\boldsymbol{\eta}_H = (\boldsymbol{\mu}_H, \boldsymbol{\mu}_H + \boldsymbol{\tau}_H)^\top$, where $\boldsymbol{\mu}_D$ and $\boldsymbol{\mu}_H$ are pre-intervention response means in the diseased and healthy groups, respectively, and $\boldsymbol{\tau}_D$ and $\boldsymbol{\tau}_H$ are the effects of the treatment in the diseased and healthy groups, respectively. The parameter of interest is

$$\boldsymbol{\Delta} = \boldsymbol{\tau}_D - \boldsymbol{\tau}_H,$$

where $\boldsymbol{\Delta} = (d_1, \ldots, d_p)^\top$ is the vector of differences in the treatment effect in the diseased and healthy populations.

**Traditional Method**

Traditionally, practitioners assume $\epsilon = \delta = 0$ and use the Hotelling $T^2$ statistic (Anderson, 2003) to make inference about $\boldsymbol{\Delta}$. More precisely, $\mathbf{Y}_{D1}, \cdots, \mathbf{Y}_{Dn_D}$ is assumed to be a random sample from $N_{2p}(\boldsymbol{\eta}_D, \Sigma)$, $\mathbf{Y}_{H1}, \cdots, \mathbf{Y}_{Hn_H}$ is assumed to be a random sample from $N_{2p}(\boldsymbol{\eta}_H, \Sigma)$ and the two samples are assumed to be mutually independent. Define

$$
\begin{aligned}
C &= (-I_p, I_p)_{p\times 2p}, \quad \overline{\mathbf{Y}}_D = n_D^{-1}\sum_{j=1}^{n_D}\mathbf{Y}_{Dj}, \quad \overline{\mathbf{Y}}_H = n_H^{-1}\sum_{j=1}^{n_H}\mathbf{Y}_{Hj}, \\
S_P &= (n_D + n_H - 2)^{-1}((n_D-1)S_D + (n_H-1)S_H), \\
S_D &= (n_D-1)^{-1}\sum_{j=1}^{n_D}(\mathbf{Y}_{Dj} - \overline{\mathbf{Y}}_D)(\mathbf{Y}_{Dj} - \overline{\mathbf{Y}}_D)^\top \text{ and} \\
S_H &= (n_H-1)^{-1}\sum_{j=1}^{n_H}(\mathbf{Y}_{Hj} - \overline{\mathbf{Y}}_H)(\mathbf{Y}_{Hj} - \overline{\mathbf{Y}}_H)^\top,
\end{aligned}
$$

where $\mathrm{I}_p$ is identity matrix of order $p$.

The Hotelling $T^2$ statistic and its exact distribution given by Anderson(Anderson, 2003)

$$
\begin{aligned}
T^2 =& \frac{n_D n_H}{n}(C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \boldsymbol{\Delta}_0)^\top (CS_PC^\top)^{-1}(C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \boldsymbol{\Delta}_0) \\
\sim& \frac{(n-2)p}{n-p-1}F_{p,n-p-1},
\end{aligned} \tag{2.1}
$$

is used to test the hypothesis $H_0 : \boldsymbol{\Delta} = \boldsymbol{\Delta}_0$. Furthermore, a $(1-\alpha)$ confidence region for $\boldsymbol{\Delta}$ is obtained by inverting the Hotellings $T^2$ test as

$$
\begin{aligned}
\Bigg\{\boldsymbol{\Delta} :& \frac{n_D n_H}{n}(C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \boldsymbol{\Delta})^\top (CS_PC^\top)^{-1}(C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \boldsymbol{\Delta}) \\
&\leq \frac{(n-2)p}{n-p-1}F_{p,n-p-1;1-\alpha}\Bigg\},
\end{aligned} \tag{2.2}
$$

where $F_{p,n-p-1;1-\alpha}$ is the lower $1-\alpha$ quantile of the $F$ distribution with degrees of freedoms $(p, n-p-1)$.

In the context of study design, suppose we interest in testing the null hypothesis $H_0 : \boldsymbol{\tau}_D - \boldsymbol{\tau}_H = \boldsymbol{\Delta}_0$. The distribution of the test statistic in (2.1) at the alternative $H_1 : \boldsymbol{\tau}_D -$

$\boldsymbol{\tau}_H = \boldsymbol{\Delta}_1$ for some fixed $\boldsymbol{\Delta}_1 \neq \boldsymbol{\Delta}_0$ is

$$\begin{aligned} T^2 =& \frac{n_D}{1+\pi}(C\overline{\mathbf{Y}} - \boldsymbol{\Delta}_0)^\top (CS_PC^\top)^{-1}(C\overline{\mathbf{Y}} - \boldsymbol{\Delta}_0) \\ \sim& \frac{(n-2)p}{n-p-1} F_{p,n-p-1}\left(\frac{n_D}{1+\pi}(\boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_0)^\top (C\Sigma C^\top)^{-1}(\boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_0)\right), \end{aligned}$$

where $F_{p,n-p-1}(\xi)$ is the $F$ distribution with degrees of freedom $(p, n-p-1)$ and non-centrality parameter $\xi$. To guarantee a nominal test size $\alpha$ and power $1-\beta$, the required total sample size $n = n_D + n_H$, where $n_D/n_H = \pi$ and $0 < \pi < \infty$, is the solution of

$$P\left(T^2 > \frac{(n-2)p}{n-p-1} F_{p,n-p-1;1-\alpha} \middle| H_1\right) = P(Y > F_{p,n-p-1;1-\alpha}) = 1-\beta, \tag{2.3}$$

where

$$Y \sim F_{p,n-p-1}\left(\frac{n\pi}{(1+\pi)^2}(\boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_0)^\top (C\Sigma C^\top)^{-1}(\boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_0)\right).$$

This equation has to be solved numerically.

The traditional estimation, test, and sample size calculation procedures above ignore the diagnostic device's inaccuracies. Therefore, as demonstrated in Section 2.5, they perform poorly in terms of bias, coverage probability, type-I error rate, and power. The sample size calculations and power analysis will be overly optimistic and misleading (see Section 2.5).

## 2.3 Estimation and Test

### The Moment-Based Approach

Let us first assume that $\epsilon$ and $\delta$ are known, but at least one of them is nonzero. In this case, the statistic $T$ will not have the Hotelling's $T^2$ distribution because the distributions of $\mathbf{Y}_{Di}$ and $\mathbf{Y}_{Hi}$ are still a mixture of multivariate normal distributions and the two distributions are not the same even under the null hypothesis unless $1-\epsilon = \delta$. Now, since

$$\begin{aligned} C\mathrm{E}(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) &= (-I_p, I_p)((1-\epsilon)\boldsymbol{\eta}_D + \epsilon\boldsymbol{\eta}_H - \delta\boldsymbol{\eta}_D - (1-\delta)\boldsymbol{\eta}_H) \\ &= (1-\epsilon-\delta)(\boldsymbol{\tau}_D - \boldsymbol{\tau}_H), \end{aligned} \tag{2.4}$$

the usual mean difference has a downward bias unless an adjustment by a factor of $(1-\epsilon-\delta)^{-1}$ is made. Thus, an unbiased estimator of $\boldsymbol{\Delta} = \boldsymbol{\tau}_D - \boldsymbol{\tau}_H$ is

$$\widetilde{\boldsymbol{\Delta}} = \frac{1}{1-\epsilon-\delta} C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H).$$

Since

$$\mathrm{V}ar(\overline{\mathbf{Y}}_D) = \frac{1}{n_D}\Sigma + \frac{1}{n_D}\epsilon(1-\epsilon)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \text{ and}$$
$$\mathrm{V}ar(\overline{\mathbf{Y}}_H) = \frac{1}{n_H}\Sigma + \frac{1}{n_H}\delta(1-\delta)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top,$$

(see calculations in Appendix 2.8), the variance of $\widetilde{\boldsymbol{\Delta}}$ is given by

$$\begin{aligned}\mathrm{V}ar(\widetilde{\boldsymbol{\Delta}}) =& \frac{1}{(1-\epsilon-\delta)^2}\left[C\Sigma C^\top\left\{\frac{1}{n_D}+\frac{1}{n_H}\right\}\right. \\ &\left. + \left\{\frac{\epsilon(1-\epsilon)}{n_D}+\frac{\delta(1-\delta)}{n_H}\right\}\boldsymbol{\Delta}\boldsymbol{\Delta}^\top\right].\end{aligned} \tag{2.5}$$

This shows that the variance of estimator $\widetilde{\boldsymbol{\Delta}}$ is affected by the covariance matrix of the data, misclassification rates and the treatment effect $\boldsymbol{\Delta}$. When misclassification errors exist, larger values of $||\boldsymbol{\Delta}||^2$ will reduce the precision of $\widetilde{\boldsymbol{\Delta}}$.

An unbiased estimator of $\mathrm{V}ar(\widetilde{\boldsymbol{\Delta}})$ is

$$S_{\widetilde{\boldsymbol{\Delta}}} = (1-\epsilon-\delta)^{-2}CSC^\top, \tag{2.6}$$

where $S = n_D^{-1}S_D + n_H^{-1}S_H$. For testing the null hypothesis $H_0 : \boldsymbol{\Delta} = \boldsymbol{\Delta_0}$, we propose to use the test statistic

$$\begin{aligned}\widetilde{T}^2 &= (\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}_0)^\top (S_{\widetilde{\boldsymbol{\Delta}}})^{-1}(\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}_0) \\ &= (C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \psi\boldsymbol{\Delta}_0)^\top (CSC^\top)^{-1}(C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \psi\boldsymbol{\Delta}_0),\end{aligned}$$

where $\psi = 1-\epsilon-\delta$. The statistic $\widetilde{T}^2$ is asymptotically distributed as $\chi^2$ distribution with $p$ degree of freedom and non-centrality parameter $n_D\psi^2(\boldsymbol{\Delta} - \boldsymbol{\Delta}_0)^\top\Phi^{-1}(\boldsymbol{\Delta} - \boldsymbol{\Delta}_0)$ as $n_D \to \infty, n_H \to \infty$ and $0 < \pi = n_D/n_H < \infty$, where $\Phi = \Sigma_D + \pi\Sigma_H$, $\Sigma_D = C\Sigma C^\top + \epsilon(1-\epsilon)\boldsymbol{\Delta}\boldsymbol{\Delta}^\top$ and $\Sigma_H = C\Sigma C^\top + \delta(1-\delta)\boldsymbol{\Delta}\boldsymbol{\Delta}^\top$. Under the null hypothesis, the non-centrality parameter is $0$ and the decision rule is to reject the null hypothesis at significance level $\alpha$ if $\widetilde{T}^2 > \chi^2_{p;1-\alpha}$, where $\chi^2_{p;1-\alpha}$ is the $1-\alpha$ quantile for the $\chi^2$ distribution with degree of freedom $p$. By inverting the test $\widetilde{T}^2$, a $(1-\alpha)$ asymptotic confidence region for $\boldsymbol{\Delta}$ is given by

$$\left\{\boldsymbol{\Delta} : (C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \psi\boldsymbol{\Delta})^\top (CSC^\top)^{-1}(C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \psi\boldsymbol{\Delta}) \le \chi^2_{p;1-\alpha}\right\}.$$

The approximation by the limiting distribution tends to be inaccurate when the sample sizes are not large. For small or moderate sample sizes, we propose approximating the distribution of $\widetilde{T}^2$ by an $F$ distribution. The rationale for this approximation is as follows. By CLT, it is easy see that $C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H)$ is asymptotically distributed as a multivariate normal distribution. We propose approximating the distribution of $CSC^\top$ by the Wishart distribution $W_p(f, \Psi)$, where $f$ and $\Psi$ are determined by matching the first moment and trace of the second moment. It would then be reasonable to approximate the distribution of $\widetilde{T}^2$ by an $F$ distribution with degree of freedom $p$ and $f$. From the calculations in Appendix 2.8, we have

$$\widetilde{T}^2 \simeq pF \sim pF_{p,f}(n_D\psi^2(\boldsymbol{\Delta} - \boldsymbol{\Delta}_0)^\top \Phi^{-1}(\boldsymbol{\Delta} - \boldsymbol{\Delta}_0)), \tag{2.7}$$

where the notation "$\simeq$" means "approximately distributed as" and $f = f_1/f_2$, where

$$\begin{aligned} f_1 =& tr^2(\Sigma_D + \pi\Sigma_H) + tr((\Sigma_D + \pi\Sigma_H)^2), \quad \text{and} \\ f_2 =& \frac{1}{n_D - 1}(tr^2(\Sigma_D) + tr(\Sigma_D^2)) + \frac{\pi^3}{n_D - \pi}(tr^2(\Sigma_H) + tr(\Sigma_H^2)) \\ &+ \frac{1}{n_D}\left(1 - 6\epsilon + 6\epsilon^2) + \pi^3(1 - 6\delta + 6\delta^2)\right) tr^2(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top). \end{aligned}$$

To check the accuracy of the approximation, we ran a small-scale simulation and plotted the empirical distribution of $\widetilde{T}$ superposed with the probability density curves of $\chi^2$ and $F$ approximations above (see Figure 2.10 in Appendix 2.8). We refer the reader to Section 2.5 for details on the settings and notations of the simulations for these figures. It is evident from the figures that the $F$ approximation is more accurate than the $\chi^2$ approximation, especially when the sample sizes are not large.

Under $H_0$, the non-centrality parameter in (2.7) is 0. The estimate of the degree of freedom, $\widehat{f}_0$, is obtained by replacing $\Sigma_D$, $\Sigma_H$ and $\boldsymbol{\Delta}$ with their estimates $S_D$, $S_H$, and $\boldsymbol{\Delta}_0$, respectively, in (2.16). Then the decision rule is to reject the null hypothesis at significance level $\alpha$ if $\widetilde{T}^2 > pF_{p,\widehat{f}_0}(1-\alpha)$. Inverting this test, a $(1-\alpha)$ confidence region for $\boldsymbol{\Delta}$ is

$$\left\{\boldsymbol{\Delta} : (C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \psi\boldsymbol{\Delta})^\top (CSC^\top)^{-1}(C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \psi\boldsymbol{\Delta}) \le pF_{p,\widehat{f};1-\alpha}\right\},$$

where $\widehat{f}$ is obtained by replacing $\Sigma_D$, $\Sigma_H$ and $\boldsymbol{\Delta}$ with their estimates $S_D$, $S_H$, and $\widehat{\boldsymbol{\Delta}}$, respectively, in (2.16). One may also use the approximation in (2.7) to construct $T^2$ simultaneous intervals(Johnson et al., 2007, pp. 275-276) for the components of $\boldsymbol{\Delta}$.

**The Likelihood-Based Approach**

When the true values of $\epsilon$ and $\delta$ are unknown, and there does not exist training data to estimate them, the methods of moments developed in Section 2.3 are not directly applicable. To overcome this limitation, we propose a likelihood-based method for making inferences about $\boldsymbol{\Delta}$. Let $Z_{ij}$ be the true disease status of the $j$th subject in the $i$th group. Define the matrix of observed values as:

$$\mathbf{Y} = (\mathbf{Y}_{D1}, \cdots, \mathbf{Y}_{Dn_D}, Y_{H1}, \cdots, \mathbf{Y}_{Hn_H})^\top.$$

Let the vector of true and predicted disease status be denoted by

$$\begin{aligned}\mathbf{Z} &= (Z_{D1}, \cdots, Z_{Dn_D}, Z_{H1}, \cdots, Z_{Hn_H})^\top \quad \text{and} \\ \mathbf{X} &= (X_{D1}, \cdots, X_{Dn_D}, X_{H1}, \cdots, X_{Hn_H})^\top,\end{aligned}$$

respectively. Further, let the corresponding realizations be

$$\begin{aligned}\mathbf{y} &= (\mathbf{y}_{D1}, \cdots, \mathbf{y}_{Dn_D}, \mathbf{y}_{H1}, \cdots, \mathbf{y}_{Hn_H})^\top, \\ \mathbf{z} &= (z_{D1}, \cdots, z_{Dn_D}, z_{H1}, \cdots, z_{Hn_H})^\top, \quad \text{and} \\ \mathbf{x} &= (x_{D1}, \cdots, x_{Dn_D}, x_{H1}, \cdots, x_{Hn_H})^\top.\end{aligned}$$

The random vector $\boldsymbol{Z}$ can never be observed. Therefore, we regard it as missing information and compute the maximum likelihood estimator of $\boldsymbol{\theta}$ using expectation-maximization (EM) algorithm (Dempster et al., 1977). The log-likelihood function for the complete data (observed and missing) is

$$\begin{aligned}l_C(\boldsymbol{\theta}) = &\sum_{j=1}^{n_D} \big[I_{\{D\}}(z_{Dj})\{\log(1-\epsilon) + \log\phi(\mathbf{y}_{Dj}|\boldsymbol{\eta}_D, \Sigma)\} \\ &+ I_{\{H\}}(z_{Dj})\{\log\epsilon + \log\phi(\mathbf{y}_{Dj}|\boldsymbol{\eta}_H, \Sigma)\}\big] \\ &+ \sum_{j=1}^{n_H} \big[I_{\{D\}}(z_{Hj})\{\log\delta + \log\phi(\mathbf{y}_{Hj}|\boldsymbol{\eta}_D, \Sigma)\} \\ &+ I_{\{H\}}(z_{Hj})\{\log(1-\delta) + \log\phi(\mathbf{y}_{Hj}|\boldsymbol{\eta}_H, \Sigma)\}\big].\end{aligned}$$

The $(t+1)^{th}$ expectation (the expectation of the log-likelihood) and the maximization steps of the EM iteration set

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E_{\theta^{(t)}}[l(\boldsymbol{\theta}|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X})] \quad \text{and} \quad \frac{\partial}{\partial\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbf{0},$$

and solve the later for $\boldsymbol{\theta}$. The detailed derivations and initial values are given in Appendix 2.8. As is well known, EM algorithm can be very slow and may not converge to an appropriate root. (Lindsay and Basak Lindsay and Basak, 1993) noted in their simulation that the MOM estimates gave higher initial likelihood value than the true values themselves.

Let the maximum likelihood estimator of $\boldsymbol{\theta}$ be denoted by $\widehat{\boldsymbol{\theta}} = \left(\widehat{\epsilon}, \widehat{\delta}, \widehat{\boldsymbol{\eta}}_D, \widehat{\boldsymbol{\eta}}_H, \widehat{\Sigma}\right)$. The maximum likelihood estimator of the parameter of interest $\boldsymbol{\Delta} = \boldsymbol{\tau}_D - \boldsymbol{\tau}_H$ is $\widehat{\boldsymbol{\Delta}} = C\widehat{\boldsymbol{\eta}}_D - C\widehat{\boldsymbol{\eta}}_H$. For estimating covariance matrix of $\widehat{\boldsymbol{\Delta}}$, one may consider the supplemented EM algorithm (SEM)(Meng and Rubin, 1991). However, the sample size requirement is too large for practical application. An alternative approach is (Louis Louis, 1982) who proposed a method for estimating the expected information matrix using results from the EM algorithm. Our numerical calculations show that the resulting coverage probabilities from this covariance estimation are mostly conservative, especially for higher error rates. We propose to use the bootstrap estimator of the covariance matrix and denote it by $S_B$. For testing the null hypothesis $H_0 : \boldsymbol{\Delta} = \mathbf{0}$, we propose comparing the statistic $\widehat{T} = \widehat{\boldsymbol{\Delta}}^\top S_B^{-1} \widehat{\boldsymbol{\Delta}}$ against the appropriate percentile of the $\chi^2$-distribution with $p$ degree of freedom.

**MOM and EM Hybrid Approach**

In the absence of the true error rates or training data, we can combine EM and moment-based methods to get a hybrid procedure. More precisely, by using the EM estimates of $\epsilon$ and $\delta$, one can derive the moment-based estimators of $\boldsymbol{\Delta}$ and $\mathrm{V}ar(\widetilde{\boldsymbol{\Delta}})$ to get an alternative test and interval estimators for $\boldsymbol{\Delta}$. Specifically, a hybrid estimator of $\boldsymbol{\Delta}$ is

$$\widetilde{\boldsymbol{\Delta}} = (1 - \widehat{\epsilon} - \widehat{\delta})^{-1} C \left(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H\right),$$

where $\widehat{\epsilon}$ and $\widehat{\delta}$ are MLE from the EM algorithm. The variance of $\widetilde{\boldsymbol{\Delta}}$ can be estimated by plugging the estimators $\widehat{\epsilon}$ and $\widehat{\delta}$ into (2.6), i.e.

$$\widehat{\mathrm{V}ar(\widetilde{\boldsymbol{\Delta}})} = \left(1 - \widehat{\epsilon} - \widehat{\delta}\right)^{-2} CSC^\top.$$

## 2.4 Sample Size and Power

Sample size determination is an essential aspect of study and trial designs. Ideally, the sample size is derived based on the test statistic planned for the subsequent hypothesis

testing. However, the required sample size cannot be calculated analytically in a close form based on the bootstrap standard error recommended in the previous sections. To work around this shortcoming, we develop sample size calculation formulas based on equation (2.4).

Suppose we are interested in testing the null hypothesis $H_0 : \boldsymbol{\tau}_D - \boldsymbol{\tau}_H = \boldsymbol{\Delta}_0$ against alternative $H_1 : \boldsymbol{\tau}_D - \boldsymbol{\tau}_H = \boldsymbol{\Delta}_1$ for $\boldsymbol{\Delta}_1 \neq \boldsymbol{\Delta}_0$. For the nominal test size $\alpha$ and power $1-\beta$, a formula for the required sample size of $n = n_D + n_H$, where $n_D/n_H = \pi$ and $0 < \pi < \infty$ can be derived based on the test statistic

$$\widetilde{T}^2 = [C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \psi\boldsymbol{\Delta}_0]^\top \left(CSC^\top\right)^{-1} [C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \psi\boldsymbol{\Delta}_0], \tag{2.8}$$

where $S = \frac{1}{n_D}(S_D + \pi S_H)$ and $S_D$ and $S_H$ are the sample covariances of $\mathbf{Y}_{D_i}$ and $\mathbf{Y}_{Hi}$, respectively, and $\psi = 1 - \epsilon - \delta$.

When $\epsilon$ and $\delta$ are not equal to 0, the statistic $\widetilde{T}^2$ in (2.8) will not have the usual Hotelling $T^2$ distribution and, thus, (2.3) will give an incorrect sample size. Since $\widetilde{T}^2$ is asymptotically distributed as $\chi^2$-distribution with $p$ degree of freedom when the sample size is large, we can determine $n_D$ using the approximation

$$P(\widetilde{T}^2 > \chi^2_{p;1-\alpha}|H_1) = 1 - \beta.$$

Under $H_1$, $\widetilde{T}^2$ is asymptotically distributed as $\chi^2$ distribution with $p$ degree of freedom and non-centrality parameter $n_D\psi^2(\boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_0)^\top\Phi^{-1}(\boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_0)$, where $\Phi = (1+\pi)C\Sigma C^\top + (\epsilon(1-\epsilon) + \pi\delta(1-\delta))\boldsymbol{\Delta}_1\boldsymbol{\Delta}_1^\top$. When the required sample size are expected not to be very large, we propose to use the $F$ approximation in (2.7). Accordingly, under $H_1$, we have the approximation

$$\widetilde{T}^2 \simeq pF \sim pF_{p,f_1}(n_D\psi^2(\boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_0)^\top\Phi^{-1}(\boldsymbol{\Delta}_1 - \boldsymbol{\Delta}_0)),$$

where $f_1$ is the degree of freedom $f$ in (2.16) but $\boldsymbol{\Delta} = \boldsymbol{\Delta}_1$.

Therefore, to find $n_D$ we solve the equation

$$P(\widetilde{T}^2 > pF_{p,f_0;1-\alpha}|H_1) \approx P(F > F_{p,f_0;1-\alpha} = 1 - \beta), \tag{2.9}$$

where $f_0$ is the degree of freedom $f$ in (2.16) but $\boldsymbol{\Delta} = \boldsymbol{\Delta}_0$. Note that when $\boldsymbol{\Delta}_0 = 0$ and $\pi = 1$, $f_0$ reduces to $f_0 = n_D + n_H - 2$ as one would expect. Again, we do not have an explicit solution for (2.9) and one has to use numerical methods to solve it.

## 2.5 Numerical Study

In this section, we evaluate the performance of the proposed estimation and sample size determination methods in terms of bias, type I error rate, power, and coverage probability. We generate data from a mixture of multivariate normal distributions with common covariance matrix as described in Section 2. For the purpose of numerical evaluation, we set the covariance matrices of each component distribution of the mixture equal across pre- and post-assessment, i.e.

$$Cov(\mathbf{Y}) = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^{\top} & \Sigma_{11} \end{pmatrix}.$$

Furthermore, we set the covariance structure to

$$\Sigma_{11} = \sigma^2[(1-\rho_1)I_p + \rho_1 J_p] \quad \text{and} \quad \Sigma_{12} = \rho_2 \Sigma_{11},$$

and investigate the effects of large and small values $\sigma^2$ while fixing $\rho_1$ and $\rho_2$ at 0.1 and 0.25, respectively. Throughout the simulation section, we set level of significance to $0.05$ and confidence level to $95\%$.

### Estimation

### Simulation Design

In Section 2.3, we introduced two estimators of the difference $\boldsymbol{\Delta} = \boldsymbol{\tau}_D - \boldsymbol{\tau}_H$. These estimators are

(1) the maximum likelihood estimator via EM algorithm (EMP),

(2) the hybrid estimator when we combine the estimations $\epsilon$ and $\delta$ from EM algorithm with the moments-based estimates (Hyb).

We evaluate the performances of the estimators and compare them with each other and the traditional estimator that does not account for the misclassification errors.

The parameter settings are planned as follows. Sample sizes, $n_D$ and $n_H$, are always $100$. We consider the effects of large and small values by setting $\sigma^2 = 10$ and $\sigma^2 = 30$. The

value of $\boldsymbol{\Delta}$ varies between $20\%$ and $60\%$ of $\sigma^2$, i.e. when $\sigma^2 = 10$, we consider $\boldsymbol{\Delta} = 2\mathbf{1}_p$, $4\mathbf{1}_p$ and $6\mathbf{1}_p$ whereas when $\sigma^2 = 30$ we consider $\boldsymbol{\Delta} = 6\mathbf{1}_p$, $12\mathbf{1}_p$ and $18\mathbf{1}_p$. For the mean of multivariate normal distributions, we fix $\boldsymbol{\mu}_D = 20\mathbf{1}_p$, $\boldsymbol{\mu}_H = 10\mathbf{1}_p$, and $\boldsymbol{\tau}_H = 4\mathbf{1}_p$ but allow $\boldsymbol{\tau}_D$ to vary according to $\boldsymbol{\Delta}$. To check the effects of dimension, we consider $p = 2, 3, 4$. We also investigate the values 0.1, 0.2, 0.3 for both $\epsilon$ and $\delta$ to observe the effects of minor to moderate misclassification rates. The number of simulations for each parameter value combination is $1000$, and the number of bootstrap samples for estimating the covariance matrix in the EM algorithm is also $1000$.

Suppose we interest in estimating $\boldsymbol{\Delta}$ by $\widehat{\boldsymbol{\Delta}}$. We use three criteria for assessing the performances of the estimators.

1. Relative bias (RB$\%$):

$$RB\% = ||E(\widehat{\boldsymbol{\Delta}}) - \boldsymbol{\Delta}||/||\boldsymbol{\Delta}|| \times 100\%,$$

   where $||\cdot||$ is the Euclidean distance.

2. Standardized bias (SB$\%$):

$$SB\% = \sqrt{(E(\widehat{\boldsymbol{\Delta}}) - \boldsymbol{\Delta})SD(\widehat{\boldsymbol{\Delta}})^{-1}(E(\widehat{\boldsymbol{\Delta}}) - \boldsymbol{\Delta})} \times 100\%.$$

3. Coverage probability (CP): the proportion of intervals that cover the true value of $\boldsymbol{\Delta}$.

**Overall Comparison of All Estimators**

To facilitate the comparisons between the competing estimators, we pull all the results from different settings into a boxplot except for the simulation factors depicted in the axes or panel labels. The boxplots are presented in Figures 2.1-2.6. Figure 2.1 summarizes all the results for the estimators in boxplots. It is evident from Figure 2.1 that the traditional estimator is the worst among the three estimators. Its coverage probabilities are much lower than $95\%$, and its SB$\%$ and RB$\%$ are much higher than the other estimators. These provide evidence that when misclassification errors exist, the results from the traditional estimator are generally misleading.

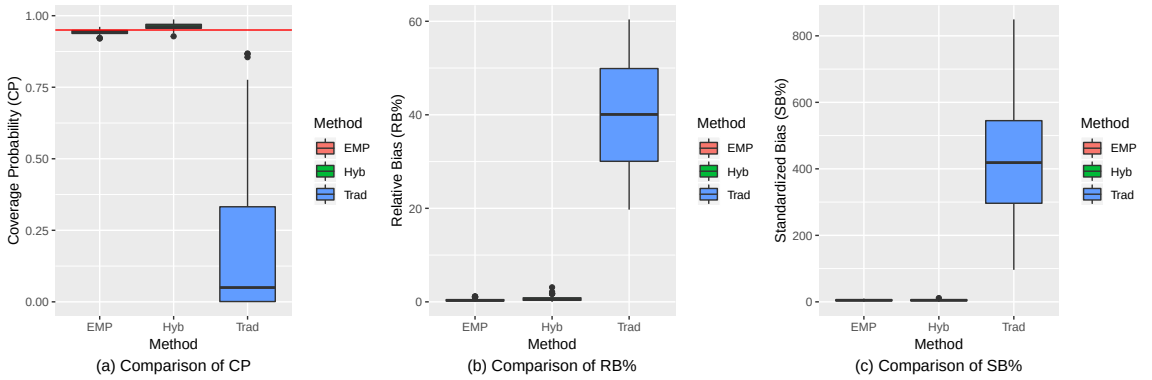



Figure 2.1: Boxplots of CP, RB%, and SB% for all estimators.Trad is for the traditional estimator; Hyb is for the hybrid estimator; EMP is for the MLE via EM algorithm.

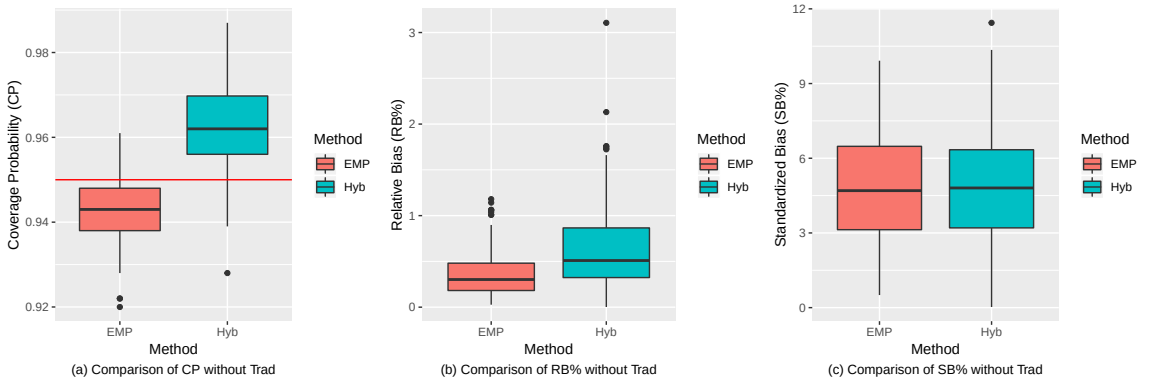



Figure 2.2: Boxplots of CP, RB%, and SB% for all methods except the traditional method. Hyb is for the hybrid estimator; EMP is for the MLE via EM algorithm.

To make the comparisons between EMP and Hyb more precise, we exclude the traditional estimator in Figures 2.2–2.6. Figure 2.2 shows that both estimators have good performances. Overall, the average SB%s are around 5%, and most RB%s are less than 1% for both estimators. The ranges and values of the RB% of EMP estimator are lower than Hyb estimator. These show EMP estimator is more accurate than Hyb estimator. The coverage probabilities of Hyb estimator are slightly higher than 95%, while the coverage probabilities of EMP estimator are slightly lower than 95%. These imply that Hyb is more conservative than EMP.

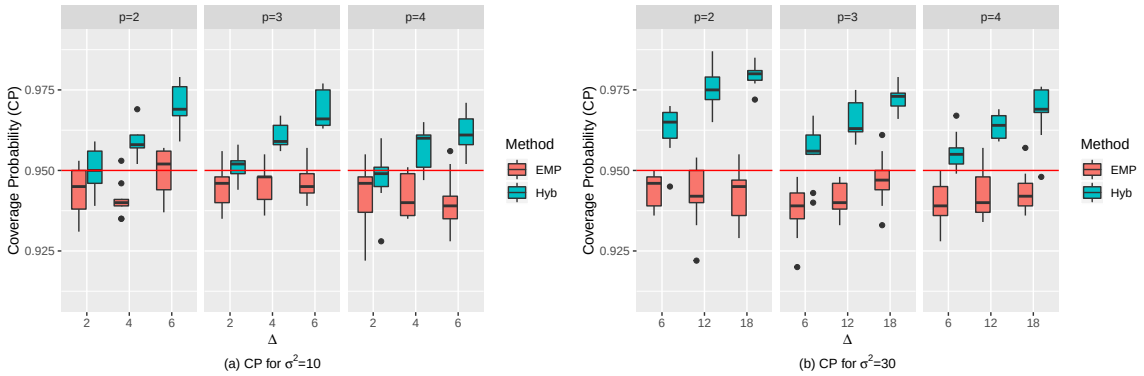



Figure 2.3: Boxplots of CP of different methods for different $p$, $\boldsymbol{\Delta}$ and $\sigma^2$. Hyb is for the hybrid estimator; EMP is for the MLE via EM algorithm.

### Effects of $p$, $\boldsymbol{\Delta}$, and $\sigma^2$

Results for different values of $p$ and $\boldsymbol{\Delta}$ are presented in Figures 2.3-2.5. Figure 2.3 shows that as the number of outcome variables increases, the coverage probabilities of Hyb become lower but get closer to the nominal level, 95%. In contrast, those of EMP remain relatively stable. The value of $\boldsymbol{\Delta}$ affects the coverage probabilities of Hyb estimator, since the coverage probabilities increases as $\boldsymbol{\Delta}$ increases. On the other hand, the coverage probabilities of EMP estimator remain stable under the changes of $\boldsymbol{\Delta}$. In Figure 2.4, we do not see much difference in RB% among the different methods when $p$ increase from 2 to 4, but we do observe a clear pattern when $\boldsymbol{\Delta}$ increases. We see that the values and ranges of RB% decrease as $\boldsymbol{\Delta}$ increases for all the methods. These may be due to division by a large number for RB% when $\boldsymbol{\Delta}$ increases. Figure 2.5 tells us that the SB% remain stable when $\boldsymbol{\Delta}$ increases, but as $p$ increase, SB%s of both estimators tend to increase.

### Effects of $\epsilon$ and $\delta$

We investigate the effects of misclassification error rates $\epsilon$ and $\delta$ in Figure 2.6. From this figure, we can see that EMP's performance is stable over different levels of misclassification error rates. However, the performance of Hyb is affected by $\epsilon$ and $\delta$. The CPs and RBs of Hyb increase as $\epsilon$ and $\delta$ get larger. When the misclassification error rates are high, Hyb becomes conservative.

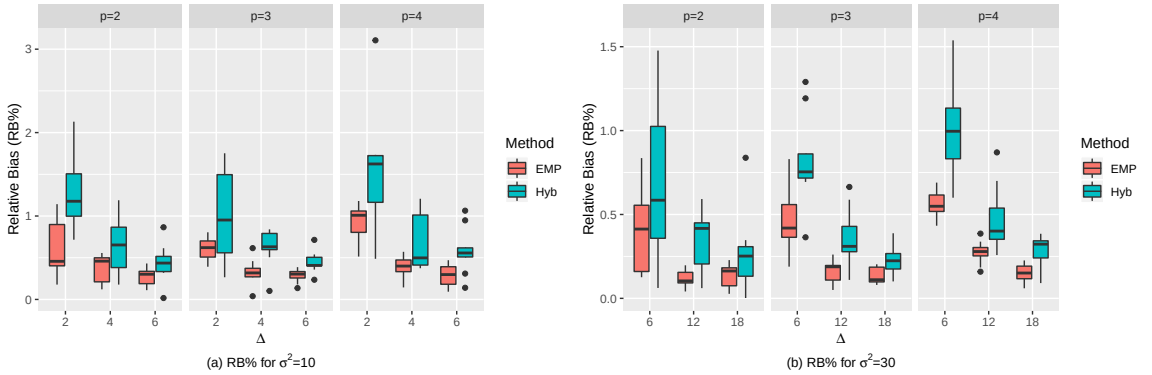



Figure 2.4: Boxplots of RB% of different methods for different $p$, $\boldsymbol{\Delta}$ and $\sigma^2$. Hyb is for the hybrid estimator; EMP is for the MLE via EM algorithm.

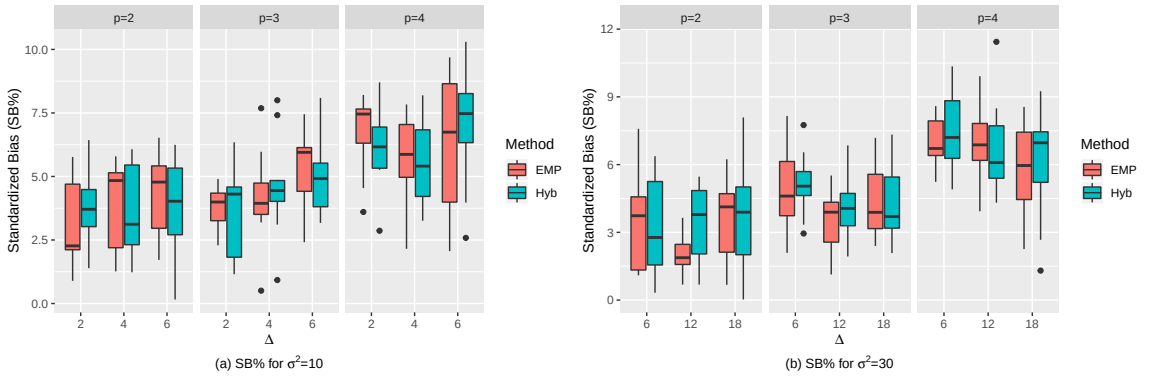



Figure 2.5: Boxplots of SB% of different methods for different $p$, $\boldsymbol{\Delta}$ and $\sigma^2$. Hyb is for the hybrid estimator; EMP is for the MLE via EM algorithm.

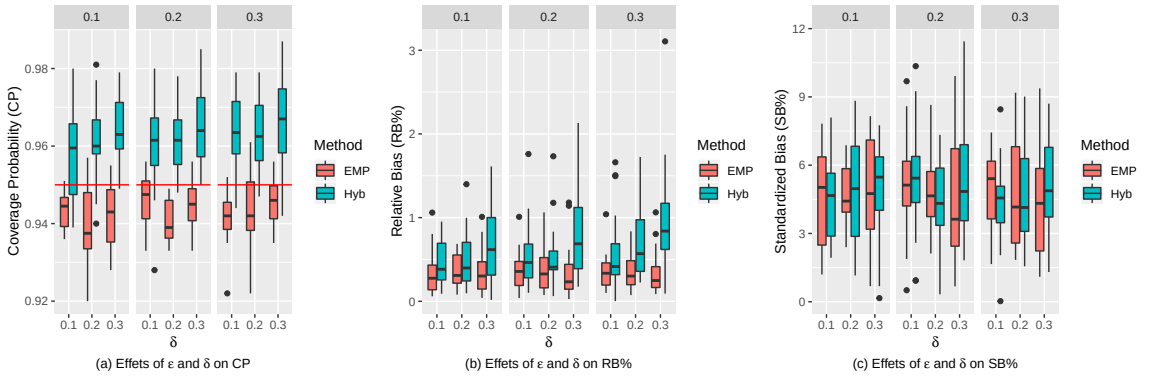



Figure 2.6: Boxplots of CP, RB%, SB% of different methods on different $\epsilon$ and $\delta$. Hyb is for the hybrid estimator; EMP is for the MLE via EM algorithm.

### Effects of Sample Size and Correlation

We also ran a simulation to check the effects of sample size, $n$, and correlations, $\rho_1$ and $\rho_2$, on the performance of estimators. The results are included in Table 2.4 and 2.5 of Appendix 2.8. We note from Table 2.4 that the traditional method's performance worsens as the sample size increases from $20$ to $200$, while both Hyb and EMP have good performances. From Table 2.5, we do not observe clear effect of correlation on the performances of either of the methods.

### Caution about EM-Based Estimators

From the comparison in Sections 2.5-2.5, we note that EMP estimator has stable performances which are not affected much by the change in the number of variables and the increase in the misclassification rates. However, by the nature of the EM algorithm, when the separation of the two-component distributions in the mixture model is poor, the estimators tend to be inaccurate and the convergence rates are low. In our case, when the distance between the mean vectors of two distributions is very small, the separation is poor. In Table 2.1, we show the relative bias (RB$\%$), standardized bias (SB$\%$), coverage probability (CP) and convergence rates (Cvg) of the EMP method. To illustrate the key ideas, we limit our investigation to $p = 1$ and the other parameters are set as $\sigma^2 = 10, \rho = 0.25, \mu_D = 10, \mu_H = 12, \tau_H = 4$ and $n_D = n_H = 100$. Also, we set $\boldsymbol{\eta}_H = (12, 16)^\top$ and vary $\boldsymbol{\eta}_D$ as $\boldsymbol{\eta_D} = (10, 16)^\top$, $\boldsymbol{\eta}_D = (10, 18)^\top$, and $\boldsymbol{\eta}_D = (10, 20)^\top$, yielding different values of $\Delta$. The separation between the two distributions is poor when $\Delta = 2$, i.e. when the euclidean distance between $\boldsymbol{\eta}_H$ and $\boldsymbol{\eta}_D$ is only 2.

From Table 2.1, we note that the results are rather unsatisfactory. The RB$\%$s and SB$\%$s are very large, especially when $\Delta = 2$ and the CPs are much lower than $95\%$. The convergence rates are about $81\%$. Therefore, the computational time is much longer than that of the other parameter settings. However, as the distance between $\boldsymbol{\eta}_H$ and $\boldsymbol{\eta}_D$ becomes larger, the EMP method's performance gets better. When $\Delta = 6$, the RB$\%$s and SB$\%$s are around $3\%$ and $20\%$, respectively, and the CPs are around $90\%$. The convergence rates are about 0.9. These show that when differences between the distribution of the two groups are larger,

we can get more accurate estimators from EMP. Also, notice that low convergence rates always accompany inaccurate estimation. In practical applications, a slow convergence rate could be indicative of poor separation of the component distributions.

Table 2.1: RB(%), SB(%), CP and Cvg (converagence rate) for EMP method when $p = 1, \sigma^2 = 10, \rho = 0.25, \mu_D = 10, \mu_H = 12, \tau_H = 4, n_D = n_H = 100$.

| | | $\epsilon$ | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 0.1 | | | | 0.3 | | | |
| $\Delta$ | $\delta$ | RB(%) | SB(%) | CP | Cvg | RB(%) | SB(%) | CP | Cvg |
| 2 | 0.1 | 82.747 | 130.793 | 0.668 | 0.813 | 84.967 | 100.368 | 0.711 | 0.811 |
| | 0.3 | 81.947 | 97.763 | 0.718 | 0.816 | 60.303 | 45.216 | 0.814 | 0.822 |
| 4 | 0.1 | 23.280 | 99.629 | 0.709 | 0.841 | 25.443 | 97.683 | 0.734 | 0.830 |
| | 0.3 | 27.384 | 105.276 | 0.704 | 0.835 | 27.962 | 80.292 | 0.853 | 0.839 |
| 6 | 0.1 | 2.345 | 17.679 | 0.913 | 0.917 | 3.288 | 21.138 | 0.892 | 0.907 |
| | 0.3 | 3.435 | 22.349 | 0.902 | 0.909 | 2.561 | 13.395 | 0.931 | 0.892 |

**Conclusion**

The conclusions from the above simulations can be summarized as follows.

(i) When misclassification errors exist, the traditional estimator for the treatment effects is severely biased.

(ii) The EM-based estimator has stable performances over a wide range of misclassification rates and the outcome's dimension. However, the accuracy of EM-based estimator is affected by the separation between the two component distributions. In practice, we need to check the EM algorithm's convergence rate to see if the separation of the two component distributions is of concern.

(iii) The moment-based estimator has small to moderate biases, but the test based on it is more conservative than the EM-based test. Our numerical investigations (not reported here to save space) revealed that the bias tends to grow with $||\mathbf{\Delta}||^2$. Looking at the variance of $\widetilde{\mathbf{\Delta}}$ in (2.5), overestimation of the variance of $\widetilde{\mathbf{\Delta}}$ is likely to happen from estimation errors in $\widehat{\epsilon}$ and $\widehat{\delta}$, and the magnitude of $\mathbf{\Delta}$. Therefore, when

$||\boldsymbol{\Delta}||^2$ gets large, so does the bias, and the tests are likely to become conservative. Nevertheless, the moment-based method is easier to use and faster to compute.

**Power and Sample Size**

In this section, we evaluate the adequacy of the sample sizes determined by (2.9) in terms of the power achieved by moment-based test in Section 2.3 and EM-based test in Section 2.3. Also, for a benchmark comparison, we compute the sample size and power of the traditional test that ignores misclassification errors. In total, we compare three methods:

1. traditional test (Tra) that ignores group classification errors with sample size from (2.3),

2. moment-based test (MMF) that uses the same test statistic as the moment-based method (MM) but the sample size is obtained through (2.9),

3. EM-based test (EMF) uses the same test statistic as EM but the sample size is obtained through (2.9).

**Criteria and Parameter Settings**

We investigate three values for $p = 2, 4, 6$. Also, we consider three settings of values for the parameters governing each component of the mixture distribution:

1. $(\sigma^2, \rho_1, \rho_2, \mu_D, \mu_D + \tau_D, \mu_H, \mu_H + \tau_H)^\top = (2, 0.25, 0.1, 4\mathbf{1}_p, 6\mathbf{1}_p, 12\mathbf{1}_p, 15.8\mathbf{1}_p)^\top$,

2. $(\sigma^2, \rho_1, \rho_2, \mu_D, \mu_D + \tau_D, \mu_H, \mu_H + \tau_H)^\top = (70, 0.3, 0.2, 40\mathbf{1}_p, 50\mathbf{1}_p, 60\mathbf{1}_p, 78\mathbf{1}_p)^\top$ and

3. $(\sigma^2, \rho_1, \rho_2, \mu_D, \mu_D + \tau_D, \mu_H, \mu_H + \tau_H)^\top = (90, 0.5, 0.1, 30\mathbf{1}_p, 50\mathbf{1}_p, 70\mathbf{1}_p, 100\mathbf{1}_p)^\top$.

The values of $\epsilon$ and $\delta$ are varied between 0.1 and 0.3 to reflect low and moderate misclassification rates. For each scenario, we generate 1000 simulated data sets. The sample sizes for each data are determined through expression (2.3) and (2.9) for a pre-specified test size $\alpha$ of $5\%$ and target power of $\beta = 80\%$. The scientific interest is to test the statistical

Table 2.2: Sample size required through traditional method (2.3) and new method (2.9) for test size $\alpha = 5\%$ and power $1 - \beta = 80\%$ when $p = 2$. Tra is the traditional method.

| | $\epsilon$ | | 0.1 | | | 0.2 | | | 0.3 | |
|---|---|---|---|---|---|---|---|---|---|---|
| Setting | $\delta$ | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.3 | 0.1 | 0.2 | 0.3 |
| 1 | Tra | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 | 17 |
| | New | 26 | 34 | 47 | 34 | 48 | 70 | 47 | 70 | 111 |
| 2 | Tra | 26 | 26 | 26 | 26 | 26 | 26 | 26 | 26 | 26 |
| | New | 39 | 52 | 71 | 52 | 71 | 104 | 71 | 104 | 164 |
| 3 | Tra | 27 | 27 | 27 | 27 | 27 | 27 | 27 | 27 | 27 |
| | New | 41 | 55 | 75 | 55 | 76 | 110 | 75 | 110 | 174 |

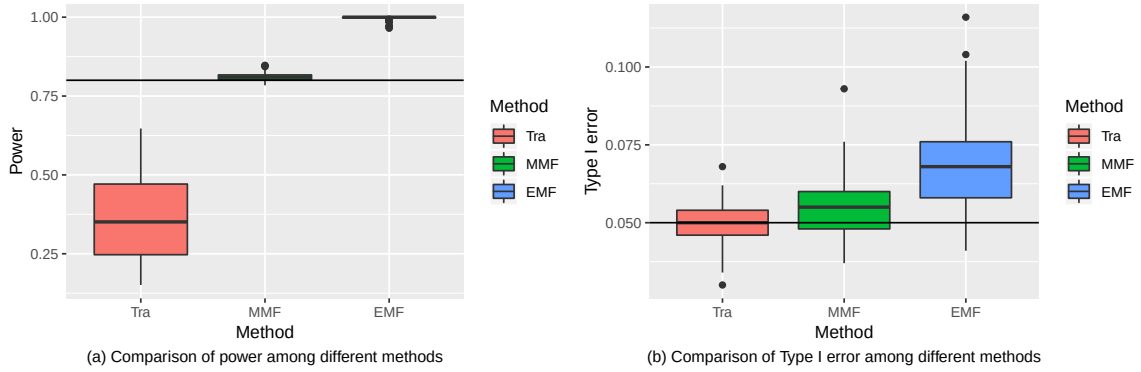
significance of null hypothesis $H_0 : \mathbf{\Delta} = 0$ vs. $H_1 : \mathbf{\Delta} \neq 0$. The empirical test size and power for each parametric combination are calculated as a proportion of data sets for which $H_0$ is rejected under the null and alternative hypothesis, respectively.

**Overall Comparisons**

From Table 2.2 we see that the sample sizes needed by the traditional method are much smaller than the other method that accounted for misclassification errors. The sample sizes needed are much higher when the misclassification errors are moderate than when they are small. Figure 2.7 provides comparisons of power and type I error rates. The power of the traditional method is too low to be reliable when classification errors exist. We also see that Type I errors of all the other methods are not far away from $5\%$. The moment-based methods have powers close to the nominal level, $80\%$, and the powers of EM-based methods are close to 1. These show EM-based methods are more powerful than moment-based methods.

**Effect of $p$**

To make the discrepancies in performance between the two new methods (MMF and EMF) clearer, we exclude the results of the traditional method in the following comparisons. Figure 2.8 shows results for the number of variables $p$ and the three different parameter settings. As the number of variables $p$ increases, the powers of all the methods increase.

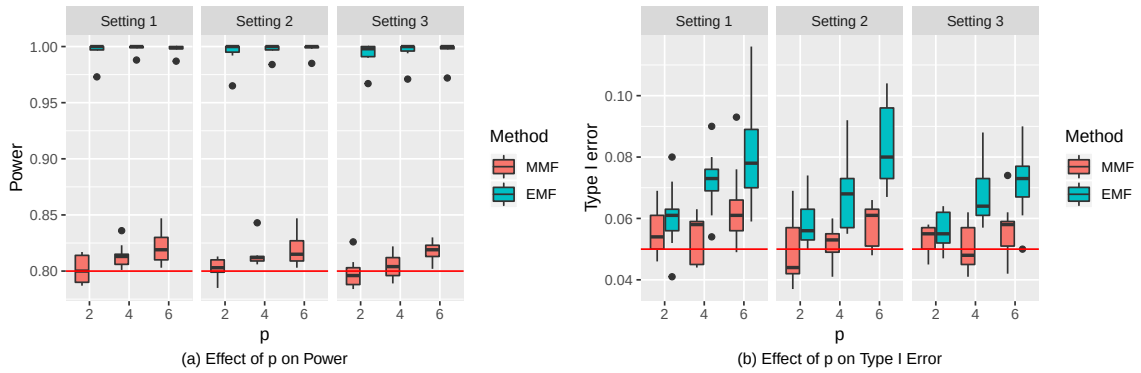



(a) Comparison of power among different methods

(b) Comparison of Type I error among different methods

Figure 2.7: Boxplots of power and Type I error for all methods. Tra, traditional test that ignores group classification errors; MMF, moment-based test; EMF, the maximum likelihood estimator and parametric bootstrapping-based test. The sample sizes for MMF and EMF are calculated using equation (2.9).

This is because we need a larger sample size when $p$ gets larger. The plot in the right panel of Figure 2.8 shows that when p increases, the type I errors of EMF increase, but MMF is less affected.




(a) Effect of p on Power

(b) Effect of p on Type I Error

Figure 2.8: Boxplots of power and Type I error for different parameter sets as $p$ increase from 2 to 6. MMF, moment-based test; EMF, the maximum likelihood estimator and parametric bootstrapping-based test. The sample sizes for MMF and EMF are calculated using equation (2.9).

### Effects of $\epsilon$ and $\delta$

To examine the roles the misclassification error rates, $\epsilon$ and $\delta$, play in the performances of the methods, we report the type I error and power results in Figure 2.9. MMF powers

remain close to 80% as $\delta$ and $\epsilon$ change, and the type I error rates decrease as the misclassification error rates increase. In this figure, it may seem that the power of EMF increases as $\epsilon$ and $\delta$ increase. Notice that the sample size needed is larger when the misclassification error rates are higher. The gain in accuracy from the increase in sample size appears to out weight the effect of the increase in misclassification error rates.

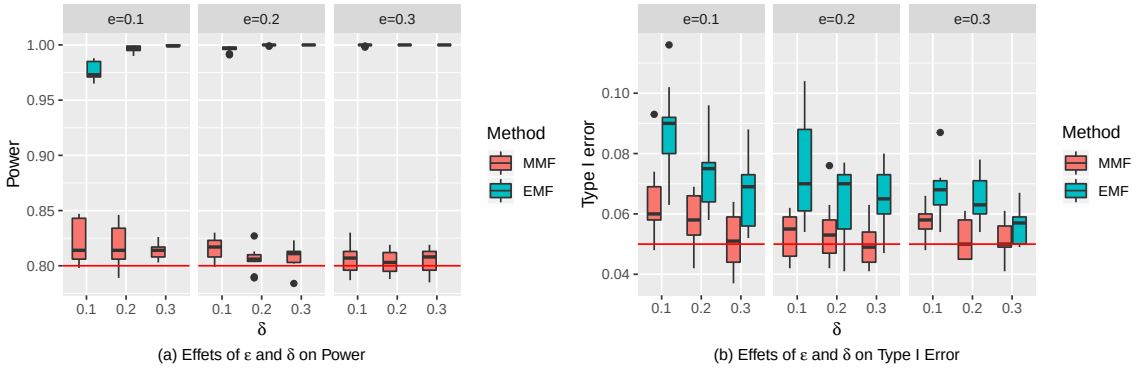



Figure 2.9: Boxplots of power and Type I error for different $\epsilon$ and $\delta$. MMF, moment-based test; EMF, the maximum likelihood estimator and parametric bootstrapping-based test. The sample sizes for MMF and EMF are calculated using equation (2.9).

## Conclusions

From the comparisons above, we can make the following conclusions. (i) Sample size calculation using the Hotelling $T^2$ tests that ignore group classification errors can severely limit the ability to detect the true treatment effect. (ii) Sample size based on F-approximation that accounts for misclassification errors achieves the desired power. Type I error of the moment-based method using the sample size based on $F$ is close to the nominal level. (iii) The test based on the maximum likelihood estimator with the parametric bootstrap covariance matrix is more powerful than the moment-based method, but it has higher Type I error rates. The advantage of the moment-based method is that it is easier to use and requires less computation time. Also, the method of moments allows the calculation of sample size for a given power and type I error rate.

## 2.6 Illustrative examples

This section analyzes a publicly available data obtained from the University of California-Irvine Machine Learning Repository [1]. The data was collected to examine Electroencephalograph (EEG) correlates of genetic predisposition to an alcohol use disorder. There are two groups of subjects, one with and the other without alcohol use disorder. There were 122 subjects. Based on the self-reported questionnaire, 77 participants were grouped as having alcohol use disorder (denoted as $D$ in this chapter), and 45 were grouped as not having alcohol use disorder (denoted as $H$ in this chapter). Their baseline brain activities were recorded using Electroencephalograph (EEG). After the baseline assessment, visual stimuli were presented, and the brain activities were measured again.

The outcome measurements are Event-Related Potentials (ERP), indicating the electrical activity level (in volts) in the region of the brain of each of the electrodes. Measurements from 64 electrodes placed on the subject's scalps were recorded for one second. Each channel (electrode) has a name identifying its location on the scalp. The names are composed of a letter and a number. The letter identifies the anatomical location of the electrode's placement (F-frontal lobe, T-temporal lobe, P-parietal lobe, and O-occipital lobe). The number identifies the brain's hemisphere (odd number - the left hemisphere, even number - the right hemisphere, and letter z (zero) - the midline). For this example, we focus on the activity recorded on EEG electrodes placed at the O1, Oz, O2, PO7, PO1, POz, PO2, and PO8. These channels correspond to the occipital lobe and parietal lobe of the brain, lobes responsible for visual processing and spatial relationships. It is hypothesized that a response to the visual stimulus would significantly differ between subjects with and without alcohol use disorder.

The classification of alcohol use disorder is based on a self-reported questionnaire. Thus, it is clear that this assessment is subject to diagnostic error, i.e., the misclassification error rates may not be 0. Therefore, we illustrate the estimation of difference in pre and post stimuli caused brain activity ($\boldsymbol{\Delta}$) between the two groups of subjects as measured by ERP.

[1] Web address: https://archive.ics.uci.edu/ml/index.php accessed on May 6, 2020.

Table 2.3: Estimates of differences in pre and post brain activity ($\boldsymbol{\Delta}$) between alcoholic and control groups and p-values for testing significance.

| | EEG electrodes | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | O1 | O2 | Oz | PO1 | PO2 | PO7 | PO8 | POz | p-value |
| Trad | 1.209 | 0.865 | 0.205 | 0.692 | 0.849 | 0.797 | 0.968 | 0.442 | 0.660 |
| Hyb | 1.514 | 1.084 | 0.256 | 0.867 | 1.063 | 0.998 | 1.212 | 0.553 | 0.750 |
| EMP | 2.143 | 1.840 | 0.637 | 1.451 | 1.549 | 1.368 | 1.778 | 0.926 | $<$0.001 |

Table 2.3 summarizes the estimates of change in activity at each of the channels and the corresponding p-values by the three methods. EM algorithm may converge at a local maximum, and we may arrive at the different maximal points from different initial values. In our case, by setting different initial values for $\epsilon$ and $\delta$, we got more than one maximal point for the likelihood function. However, there is only one maximum point that satisfy the constraints $0 \leq \epsilon \leq 1/2$ and $0 \leq \delta \leq 1/2$. The estimated value of $\epsilon$ is 0.201, and that of $\delta$ is close to 0. From these results, we observe that the estimate of pre-post mean differences for the traditional method are smaller than the estimates from both hybrid and EM estimators. The hybrid estimates are also smaller than the EM estimates. These results confirm our observations in the simulation. The hybrid estimator gives results closer to the EM estimator than the traditional method and demonstrates lower power with inflated coverage probability. According to the EM method, there are significant differences in pre and post stimuli brain activity ($\boldsymbol{\Delta}$) between the groups with and without alcohol use disorder. Therefore, alcohol use may affect brain activities related to the visual procession and spatial relationship.

For a future study where parameters are expected to be similar to the ones observed in the present study, the sample size needed according to (2.9) to detect a pre-post difference as the hybrid method estimate with $80\%$ power at $5\%$ level of significance would be $158$ patients with alcohol use disorder and $93$ patients without alcohol use disorder. On the other hand, if we need to detect a pre-post difference as estimated by the EM method estimate, according to equation (2.9), we require 96 patients with alcohol use disorder and 57 patients without alcohol use disorder. Note that the sample size calculated here is not for the EM and parametric bootstrap-based combined inference. At present, we do not have a

mechanism for sample size determination for such an approach.

## 2.7 Discussion and Conclusion

Two approaches, moment-based and likelihood-based, are proposed to estimate treatment effects in pre-post design when diagnostic devices used to classify subjects are fallible. We also derived formulas for sample size calculation based on a novel $F$ finite-sample approximation for the distribution of the moment-based test statistic. Numerical results showed that traditional methods that ignore misclassification errors lead to unacceptably-large bias and overly optimistic sample size. We should avoid the traditional methods unless there is a strong reason to believe the diagnostic device is perfect. All the methods proposed in this chapter have satisfactory performances in terms of bias, type I error rate, power, and coverage probability. The EM-based methods provide more accurate estimators and more powerful tests than the moment-based methods. However, we cannot use the EM-based test statistics to determine the required sample size because there is no closed-form expression for the covariance matrix of the estimator. Furthermore, we need to check the EM algorithm's convergence to see if the separation between the group distributions is poor. The estimators from moment-based methods are accurate, but the corresponding tests are a bit conservative. The advantage of the moment-based method is that its form is more familiar to the average practitioner and, hence, they are easier to use and faster to compute than the EM-based method.

In our model, we assumed that the covariance matrices of the two groups are the same. This assumption is not generally restrictive because it is reasonable to postulate that treatment changes only the mean of the distribution. However, in the more general setting, this assumption can be relaxed. The covariance matrices of the two groups depend on misclassification rates and yield different values when they are different. To allow different covariance matrices, we need to recalculate the covariance matrix of the moment-based estimator and the corresponding sample size determination formulas. We also need to reformulate the likelihood-based approach and recalculate the corresponding E and M steps. We will leave a detailed analysis of this problem for future research.

The problem considered in this chapter involves a continuous multivariate model in

which we assume the pre and post outcome measures are a mixture of multivariate normal distributions. It might be possible to relax this assumption and consider a fully nonparametric framework. Other outcome types such as categorical, ordinal, survival times, and functional outcomes also need further investigation. It is also essential to investigate other study designs such as clustered randomized design and cross-over design. We defer these topics for future researches.

## 2.8 Appendix

### Technical Details

In this subsection, we provide detailed calculations and technical details for the results presented in Section 2.3.

### Derivation of $\mathrm{V}ar(\overline{\mathbf{Y}}_D)$ and $\mathrm{V}ar(\overline{\mathbf{Y}}_H)$

We show the calculations only for $\mathrm{V}ar(\overline{\mathbf{Y}}_D)$, and those for $\mathrm{V}ar(\overline{\mathbf{Y}}_H)$ are the same with the obvious changes of notations. Notice that,

$$\begin{aligned}\mathrm{V}ar(\mathbf{Y}_{D1}) &= E(Y_{D1}Y_{D1}^\top) - E(\mathbf{Y}_{D1})E(\mathbf{Y}_{D1})^\top \\ &= (1-\epsilon)E(X_{D1}X_{D1}^\top) + \epsilon E(X_{H1}X_{H1}^\top) - E(\mathbf{Y}_{D1})E(\mathbf{Y}_{D1})^\top,\end{aligned}$$

where $X_{D1} \sim N(\boldsymbol{\eta}_D, \Sigma)$ and $X_{H1} \sim N(\boldsymbol{\eta}_H, \Sigma)$. Since,

$$E(X_{D1}X_{D1}^\top) = \Sigma + \boldsymbol{\eta}_D\boldsymbol{\eta}_D^\top, \quad E(X_{H1}X_{H1}^\top) = \Sigma + \boldsymbol{\eta}_H\boldsymbol{\eta}_H^\top, \quad E(\mathbf{Y}_{D1}) = (1-\epsilon)\boldsymbol{\eta}_D + \epsilon\boldsymbol{\eta}_H,$$

we have

$$\begin{aligned}\mathrm{V}ar(\mathbf{Y}_{D1}) =& (1-\epsilon)(\Sigma + \boldsymbol{\eta}_D\boldsymbol{\eta}_D^\top) + \epsilon(\Sigma + \boldsymbol{\eta}_H\boldsymbol{\eta}_H^\top) \\ &- ((1-\epsilon)\boldsymbol{\eta}_D + \epsilon\boldsymbol{\eta}_H)((1-\epsilon)\boldsymbol{\eta}_D + \epsilon\boldsymbol{\eta}_H)^\top \\ =& \Sigma + \epsilon(1-\epsilon)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top.\end{aligned}$$

Therefore,

$$\mathrm{V}ar(\overline{\mathbf{Y}}_D) = \frac{1}{n_D}\mathrm{V}ar(\mathbf{Y}_{D1}) = \frac{1}{n_D}\Sigma + \frac{1}{n_D}\epsilon(1-\epsilon)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top.$$

**Derivation of EM algorithm**

The log-likelihood function for the complete data (observed and missing) is

$$\begin{aligned} l_C(\boldsymbol{\theta}) = &\sum_{j=1}^{n_D} \big[ I_{\{D\}}(z_{Dj}) \{ \log(1-\epsilon) + \log \phi(\mathbf{y}_{Dj} | \boldsymbol{\eta}_D, \Sigma) \} \\ &+ I_{\{H\}}(z_{Dj}) \{ \log \epsilon + \log \phi(\mathbf{y}_{Dj} | \boldsymbol{\eta}_H, \Sigma) \} \big] \\ &+ \sum_{j=1}^{n_H} \big[ I_{\{D\}}(z_{Hj}) \{ \log \delta + \log \phi(\mathbf{y}_{Hj} | \boldsymbol{\eta}_D, \Sigma) \} \\ &+ I_{\{H\}}(z_{Hj}) \{ \log(1-\delta) + \log \phi(\mathbf{y}_{Hj} | \boldsymbol{\eta}_H, \Sigma) \} \big]. \end{aligned}$$

**E step:** For the $(t+1)^{th}$ expectation step of the EM algorithm,

$$\begin{aligned} Q(\boldsymbol{\theta} | \boldsymbol{\theta}^{(t)}) = &E_{\theta^{(t)}}[l(\boldsymbol{\theta} | \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{X})] \\ = &\sum_{j=1}^{n_D} K_{1j}^{(t)} \log(1-\epsilon) + \sum_{j=1}^{n_D} K_{1j}^{(t)} \log \phi(\mathbf{y}_{Dj} | \boldsymbol{\eta}_D, \Sigma) \\ &+ \sum_{j=1}^{n_D} \left(1 - K_{1j}^{(t)}\right) \log \epsilon + \sum_{j=1}^{n_D} \left(1 - K_{1j}^{(t)}\right) \log \phi(\mathbf{y}_{Dj} | \boldsymbol{\eta}_H, \Sigma) \\ &+ \sum_{j=1}^{n_H} K_{2j}^{(t)} \log \delta + \sum_{j=1}^{n_H} K_{2j}^{(t)} \log \phi(\mathbf{y}_{Hj} | \boldsymbol{\eta}_D, \Sigma) \\ &+ \sum_{j=1}^{n_H} \left(1 - K_{2j}^{(t)}\right) \log(1-\delta) + \sum_{j=1}^{n_H} \left(1 - K_{2j}^{(t)}\right) \log \phi(\mathbf{y}_{Hj} | \boldsymbol{\eta}_H, \Sigma), \end{aligned}$$

where

$$K_{1j}^{(t)} = \frac{(1-\epsilon^{(t)}) \phi(\mathbf{y}_{Dj} | \boldsymbol{\eta}_D^{(t)}, \Sigma^{(t)})}{(1-\epsilon^{(t)}) \phi(\mathbf{y}_{Dj} | \boldsymbol{\eta}_D^{(t)}, \Sigma^{(t)}) + \epsilon^{(t)} \phi(\mathbf{y}_{Dj} | \boldsymbol{\eta}_H^{(t)}, \Sigma^{(t)})} \text{ and} \tag{2.10}$$

$$K_{2j}^{(t)} = \frac{\delta^{(t)} \phi(\mathbf{y}_{Hj} | \boldsymbol{\eta}_D^{(t)}, \Sigma^{(t)})}{\delta^{(t)} \phi(\mathbf{y}_{Hj} | \boldsymbol{\eta}_D^{(t)}, \Sigma^{(t)}) + (1-\epsilon^{(t)}) \phi(\mathbf{y}_{Hj} | \boldsymbol{\eta}_H^{(t)}, \Sigma^{(t)})}. \tag{2.11}$$

Noting that

$$\begin{aligned} \log \phi(\mathbf{y}_{Dj} | \boldsymbol{\eta}_D, \Sigma) &= -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (\mathbf{y}_{Dj} - \boldsymbol{\eta}_D)^\top \Sigma^{-1} (\mathbf{y}_{Dj} - \boldsymbol{\eta}_D) + C, \\ \log \phi(\mathbf{y}_{Dj} | \boldsymbol{\eta}_H, \Sigma) &= -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (\mathbf{y}_{Dj} - \boldsymbol{\eta}_H)^\top \Sigma^{-1} (\mathbf{y}_{Dj} - \boldsymbol{\eta}_H) + C, \\ \log \phi(\mathbf{y}_{Hj} | \boldsymbol{\eta}_D, \Sigma) &= -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (\mathbf{y}_{Hj} - \boldsymbol{\eta}_D)^\top \Sigma^{-1} (\mathbf{y}_{Hj} - \boldsymbol{\eta}_D) + C \text{ and} \\ \log \phi(\mathbf{y}_{Hj} | \boldsymbol{\eta}_H, \Sigma) &= -\frac{1}{2} \log |\Sigma| - \frac{1}{2} (\mathbf{y}_{Hj} - \boldsymbol{\eta}_H)^\top \Sigma^{-1} (\mathbf{y}_{Hj} - \boldsymbol{\eta}_H) + C, \end{aligned}$$

we have

$$
\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) =& E_{\theta^{(t)}}[l(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{X})] \\
=& \sum_{j=1}^{n_D} K_{1j}^{(t)} \log(1-\epsilon) + \sum_{j=1}^{n_D} \left(1 - K_{1j}^{(t)}\right) \log \epsilon \\
&+ \sum_{j=1}^{n_H} K_{2j}^{(t)} \log \delta + \sum_{j=1}^{n_H} \left(1 - K_{2j}^{(t)}\right) \log(1-\delta) \\
&+ \sum_{j=1}^{n_D} K_{1j}^{(t)} [-\frac{1}{2} \log|\Sigma| - \frac{1}{2} (\mathbf{y}_{Dj} - \boldsymbol{\eta}_D)^\top \Sigma^{-1} (\mathbf{y}_{Dj} - \boldsymbol{\eta}_D) + C] \\
&+ \sum_{j=1}^{n_D} \left(1 - K_{1j}^{(t)}\right) [-\frac{1}{2} \log|\Sigma| - \frac{1}{2} (\mathbf{y}_{Dj} - \boldsymbol{\eta}_H)^\top \Sigma^{-1} (\mathbf{y}_{Dj} - \boldsymbol{\eta}_H) + C] \\
&+ \sum_{j=1}^{n_H} K_{2j}^{(t)} [-\frac{1}{2} \log|\Sigma| - \frac{1}{2} (\mathbf{y}_{Hj} - \boldsymbol{\eta}_D)^\top \Sigma^{-1} (\mathbf{y}_{Hj} - \boldsymbol{\eta}_D) + C] \\
&+ \sum_{j=1}^{n_H} \left(1 - K_{2j}^{(t)}\right) [-\frac{1}{2} \log|\Sigma| - \frac{1}{2} (\mathbf{y}_{Hj} - \boldsymbol{\eta}_H)^\top \Sigma^{-1} (\mathbf{y}_{Hj} - \boldsymbol{\eta}_H) + C].
\end{aligned}
$$

**M step:** For the maximization step of the EM algorithm, setting $\frac{\partial}{\partial \theta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbf{0}$,

$$
\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \epsilon} =& -\sum_{j=1}^{n_D} \frac{K_{1j}^{(t)}}{1-\epsilon} + \sum_{j=1}^{n_D} \frac{1 - K_{1j}^{(t)}}{\epsilon} = 0, \\
\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \delta} =& \sum_{j=1}^{n_H} \frac{K_{2j}^{(t)}}{\delta} + \sum_{j=1}^{n_H} \frac{1 - K_{2j}^{(t)}}{1-\delta} = 0,
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\eta}_D} =& \sum_{j=1}^{n_D} K_{1j} \Sigma^{-1} (\mathbf{y}_{Dj} - \boldsymbol{\eta}_D) + \sum_{j=1}^{n_H} K_{2j} \Sigma^{-1} (\mathbf{y}_{Hj} - \boldsymbol{\eta}_D) = \mathbf{0}, \\
\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\eta}_H} =& \sum_{j=1}^{n_D} (1 - K_{1j}) \Sigma^{-1} (\mathbf{y}_{Dj} - \boldsymbol{\eta}_H) + \sum_{j=1}^{n_H} (1 - K_{2j}) \Sigma^{-1} (\mathbf{y}_{Hj} - \boldsymbol{\eta}_H) = \mathbf{0}, \text{ and}
\end{aligned}
$$

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \Sigma} =& -\frac{1}{2}\sum_{j=1}^{n_D} K_{1j}(\Sigma^{-1} - \Sigma^{-1}(\mathbf{y}_{Dj} - \boldsymbol{\eta}_D)(\mathbf{y}_{Dj} - \boldsymbol{\eta}_D)^\top \Sigma^{-1}) \\ & -\frac{1}{2}\sum_{j=1}^{n_D}(1 - K_{1j})(\Sigma^{-1} - \Sigma^{-1}(\mathbf{y}_{Dj} - \boldsymbol{\eta}_H)(\mathbf{y}_{Dj} - \boldsymbol{\eta}_H)^\top \Sigma^{-1}) \\ & -\frac{1}{2}\sum_{j=1}^{n_H} K_{2j}(\Sigma^{-1} - \Sigma^{-1}(\mathbf{y}_{Hj} - \boldsymbol{\eta}_D)(\mathbf{y}_{Hj} - \boldsymbol{\eta}_D)^\top \Sigma^{-1}) \\ & -\frac{1}{2}\sum_{j=1}^{n_H}(1 - K_{2j})(\Sigma^{-1} - \Sigma^{-1}(\mathbf{y}_{Hj} - \boldsymbol{\eta}_H)(\mathbf{y}_{Hj} - \boldsymbol{\eta}_H)^\top \Sigma^{-1}) = 0.\end{aligned}$$

Solving for $\boldsymbol{\theta}$,

$$\epsilon^{(t+1)} = 1 - \sum_{j=1}^{n_D}\frac{K_{1j}^{(t)}}{n_D}, \quad \boldsymbol{\eta}_D^{(t+1)} = \frac{\sum_{j=1}^{n_D} K_{1j}^{(t)}\mathbf{y}_{Dj} + \sum_{j=1}^{n_H} K_{2j}^{(t)}\mathbf{y}_{Hj}}{\sum_{j=1}^{n_D} K_{1j}^{(t)} + \sum_{j=1}^{n_H} K_{2j}^{(t)}},$$

$$\delta^{(t+1)} = \sum_{j=1}^{n_H}\frac{K_{2j}^{(t)}}{n_H}, \quad \boldsymbol{\eta}_H^{(t+1)} = \frac{\sum_{j=1}^{n_D}\left(1 - K_{1j}^{(t)}\right)\mathbf{y}_{Dj} + \sum_{j=1}^{n_H}\left(1 - K_{2j}^{(t)}\right)\mathbf{y}_{Hj}}{n_D + n_H - \left(\sum_{j=1}^{n_D} K_{1j}^{(t)} + \sum_{j=1}^{n_H} K_{2j}^{(t)}\right)} \text{ and}$$

$$\begin{aligned}\Sigma^{(t+1)} =& \frac{\sum_{j=1}^{n_D}\boldsymbol{K}_{1j}^{(t)}\left(\mathbf{y}_{Dj} - \boldsymbol{\eta}_D^{(t)}\right)\left(\mathbf{y}_{Dj} - \boldsymbol{\eta}_D^{(t)}\right)^\top}{n_D + n_H} \\ & + \frac{\sum_{j=1}^{n_D}\left(1 - \boldsymbol{K}_{1j}^{(t)}\right)\left(\mathbf{y}_{Dj} - \boldsymbol{\eta}_H^{(t)}\right)\left(\mathbf{y}_{Dj} - \boldsymbol{\eta}_H^{(t)}\right)^\top}{n_D + n_H} \\ & + \frac{\sum_{j=1}^{n_D}\boldsymbol{K}_{2j}^{(t)}\left(\mathbf{y}_{Hj} - \boldsymbol{\eta}_D^{(t)}\right)\left(\mathbf{y}_{Hj} - \boldsymbol{\eta}_D^{(t)}\right)^\top}{n_D + n_H} \\ & + \frac{\sum_{j=1}^{n_D}\left(1 - \boldsymbol{K}_{2j}^{(t)}\right)\left(\mathbf{y}_{Hj} - \boldsymbol{\eta}_H^{(t)}\right)\left(\mathbf{y}_{Hj} - \boldsymbol{\eta}_H^{(t)}\right)^\top}{n_D + n_H}.\end{aligned}$$

For initial values, estimates of PPV and NPV from previous trials of effectiveness for the diagnostic tool, if available, can be used. For the other parameters, the method of moments (MOM) estimates $\widetilde{\boldsymbol{\eta}}_D$, $\widetilde{\boldsymbol{\eta}}_H$, and $\widetilde{\Sigma}$ will be used. Here,

$$\widetilde{\boldsymbol{\eta}}_D = \frac{(1-\delta)\overline{\mathbf{y}}_D - \epsilon\overline{\mathbf{y}}_H}{1 - \delta - \epsilon}, \quad \widetilde{\boldsymbol{\eta}}_H = \frac{(1-\epsilon)\overline{\mathbf{y}}_H - \delta\overline{\mathbf{y}}_D}{1 - \delta - \epsilon} \quad \text{and} \tag{2.12}$$

$$\widetilde{\Sigma} = \widetilde{S}_P - \Big(\frac{n_D}{n_D + n_H}\frac{\epsilon(1-\epsilon)}{(1-\delta-\epsilon)^2} + \frac{n_H}{n_D + n_H}\frac{\delta(1-\delta)}{(1-\delta-\epsilon)^2}\Big)(\overline{\mathbf{y}}_D - \overline{\mathbf{y}}_H)(\overline{\mathbf{y}}_D - \overline{\mathbf{y}}_H)^\top, \tag{2.13}$$

where

$$\widetilde{S}_P = (n_D + n_H)^{-1}(n_D\widetilde{S}_D + n_H\widetilde{S}_H),$$
$$\widetilde{S}_D = n_D^{-1}\sum_{j=1}^{n_D}(\mathbf{Y}_{Dj} - \overline{\mathbf{Y}}_D)(\mathbf{Y}_{Dj} - \overline{\mathbf{Y}}_D)^\top \text{ and} \tag{2.14}$$
$$\widetilde{S}_H = n_H^{-1}\sum_{j=1}^{n_H}(\mathbf{Y}_{Hj} - \overline{\mathbf{Y}}_H)(\mathbf{Y}_{Hj} - \overline{\mathbf{Y}}_H)^\top.$$

It is possible that $\widetilde{\Sigma}$ may not be positive definite. In that case, we ignore the second term in (2.13).

**Derivation of degree freedom for $F$ approximation**

Notice that

$$\widetilde{T}^2 = (C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \psi\mathbf{\Delta}_0)^\top(CSC^\top)^{-1}(C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \psi\mathbf{\Delta}_0).$$

First, we propose the approximations

$$\sqrt{\frac{n_D}{f}}\left(C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H) - \psi\mathbf{\Delta}_0\right) \simeq N\left(\sqrt{\frac{n_D}{f}}(\mathbf{\Delta} - \mathbf{\Delta}_0)\psi, \frac{\Phi}{f}\right) \quad \text{and}$$
$$n_D CSC^\top \simeq \mathrm{W}_p\left(f, \frac{\Phi}{f}\right),$$

where the notation "$\simeq$" means "approximately distributed as." The first of these approximations come from the limiting distribution, and the second one is motivated by the exact distribution of $S$ if there were no misclassification errors. It would therefore be reasonable to make the approximation

$$\widetilde{T}^2 \simeq pF \sim pF_{p,f}(n_D\psi^2(\mathbf{\Delta} - \mathbf{\Delta}_0)^\top\Phi^{-1}(\mathbf{\Delta} - \mathbf{\Delta}_0)).$$

To account for the misclassification errors in approximating the distribution of $n_D CSC^\top$, we use the method of moments to approximate the degrees of freedom $f$ and scale matrix for the Wishart distribution $\Psi$ by matching the first and trace of the second moments. For $W \sim W_p(f, \Psi)$, we know (Magnus and Neudecker, 1979) that

$$E(W) = f\Psi, \quad and \quad \mathrm{V}ar(W) = f(I_{p^2} + K_{p,p})(\Psi \otimes \Psi),$$

where $\otimes$ is the Kronecker product. On the other hand, $E[n_D CSC^\top] = \Phi$. Matching the first moment, we have $\Psi = \frac{\Phi}{f}$. Furthermore, to determine $f$ we match the trace of the second moment as

$$\begin{aligned} tr(f(I_{p^2} + K_{p,p})(\Psi \otimes \Psi)) &= tr(\mathrm{V}ar(n_D CSC^\top)) \\ &= \frac{1}{n_D^2} tr(\mathrm{V}ar(CS_D C^\top)) + \frac{1}{n_H^2} tr(\mathrm{V}ar(CS_H C^\top)). \end{aligned}$$

Solving for $f$, we get

$$f = \frac{tr((I_{p^2} + K_{p,p})(f\Psi \otimes f\Psi))}{tr(\mathrm{V}ar(n_D CSC^\top))} = \frac{tr((I_{p^2} + K_{p,p})(\Phi \otimes \Phi)}{tr(\mathrm{V}ar(n_D CSC^\top))} = \frac{tr(\Phi)^2 + tr(\Phi^2)}{n_D^2 tr(\mathrm{V}ar(CSC^\top))}.$$

The remaining task is to calculate $tr(\mathrm{V}ar(CSC^\top))$. Notice that $S = \frac{1}{n_D} S_D + \frac{1}{n_H} S_H$. Since $S_D$ and $S_H$ are independent,

$$\begin{aligned} \mathrm{V}ar(CSC^\top) &= \mathrm{V}ar\left(C\left(\frac{1}{n_D} S_D + \frac{1}{n_H} S_H\right) C^\top\right) \\ &= \frac{1}{n_D^2} \mathrm{V}ar(CS_D C^\top) + \frac{1}{n_H^2} \mathrm{V}ar(CS_H C^\top). \end{aligned}$$

The covariances $\mathrm{V}ar(CS_D C^\top)$ and $\mathrm{V}ar(CS_H C^\top)$ can now be calculated separately. Since the caculations are identical, we show the details for $\mathrm{V}ar(CS_D C^\top)$ and those for $\mathrm{V}ar(CS_D C^\top)$ are identical except the obvious changes of notations. To that end, observe that

$$\begin{aligned} CS_D C^\top &= \frac{1}{n_D - 1} \sum_{j=1}^{n_D} C(\mathbf{Y}_{Dj} - \overline{\mathbf{Y}}_D)(\mathbf{Y}_{Dj} - \overline{\mathbf{Y}}_D)^\top C^\top \\ &= \frac{1}{n_D - 1} C Y_D \left(I_{n_D} - \frac{1}{n_D} J_{n_D}\right) Y_D^\top C^\top. \end{aligned}$$

Setting $Z_D = Y_D - E(Y_D)$, $A = (a_{ij}) = \left(I_{n_D} - \frac{1}{n_D} J_{n_D}\right)$, and $B = (b_{ij}) = \left(I_{n_H} - \frac{1}{n_H} J_{n_H}\right)$, we have

$$CS_D C^\top = \frac{1}{n_D - 1} \sum_{j=1}^{n_D} C(\mathbf{Z}_{Dj} - \overline{\mathbf{Z}}_D)(\mathbf{Z}_{Dj} - \overline{\mathbf{Z}}_D)^\top C^\top = \frac{1}{n_D - 1} C Z_D A Z_D^\top C^\top.$$

Similarly,

$$\widetilde{Z}_D = C Z_D = C \begin{pmatrix} Z_D^{(1)} \\ Z_D^{(2)} \end{pmatrix} = (\mathbf{Z}_{D1}^{(2)} - \mathbf{Z}_{D1}^{(1)}, \ldots, \mathbf{Z}_{Dn_D}^{(2)} - \mathbf{Z}_{Dn_D}^{(1)}).$$

Therefore, $\widetilde{Z}_D$ is $p \times n_D$ matrix whose columns are independently distributed with mean $\mathbf{0}$ and covariance $\Sigma_D$, where

$$\Sigma_D = C\Sigma C^\top + \epsilon(1-\epsilon)\boldsymbol{\Delta}\boldsymbol{\Delta}^\top.$$

Applying Lemma 1 in Harrar and Bathke (2012), we have

$$\mathrm{V}ar\left(\widetilde{Z}_D A \widetilde{Z}_D^\top\right) = \sum_{i=1}^{n_D}\sum_{j=1}^{n_D} a_{ij}^2 (I_{p^2} + K_{p,p})(\Sigma_D \otimes \Sigma_D) + \sum_{i=1}^{n_D} a_{ii}^2 K_4(\widetilde{\mathbf{Z}}_{D_i}),$$

where

$$\begin{aligned} K_4(\widetilde{\mathbf{Z}}_{D_i}) =& K_4(\widetilde{\mathbf{Z}}_{D_1}) \\ =& E(vec(\widetilde{\boldsymbol{Z}}_{D_1}\widetilde{\boldsymbol{Z}}_{D_1}^\top)vec(\widetilde{\boldsymbol{Z}}_{D_1}\widetilde{\boldsymbol{Z}}_{D_1}^\top)^\top) \\ & - (I_{p^2} + K_{p,p})(\Sigma_D \otimes \Sigma_D) - vec(\Sigma_D)vec(\Sigma_D)^\top. \end{aligned}$$

Now it remains to calculate $E(vec(\widetilde{\boldsymbol{Z}}_{D_1}\widetilde{\boldsymbol{Z}}_{D_1}^\top)vec(\widetilde{\boldsymbol{Z}}_{D_1}\widetilde{\boldsymbol{Z}}_{D_1}^\top)^\top)$. Let $\widetilde{\boldsymbol{Z}}_{D_1} = (\widetilde{Z}_{D11}, \ldots, \widetilde{Z}_{D1p})^\top$. Then

$$\begin{aligned} & E(vec(\widetilde{\boldsymbol{Z}}_{D_1}\widetilde{\boldsymbol{Z}}_{D_1}^\top)vec(\widetilde{\boldsymbol{Z}}_{D_1}\widetilde{\boldsymbol{Z}}_{D_1}^\top)^\top) \\ =& E(\widetilde{\boldsymbol{Z}}_{D_1}\widetilde{\boldsymbol{Z}}_{D_1}^\top \otimes \widetilde{\boldsymbol{Z}}_{D_1}\widetilde{\boldsymbol{Z}}_{D_1}^\top) \\ =& E\left(\begin{pmatrix} \widetilde{Z}_{D11}\widetilde{Z}_{D11} & \cdots & \widetilde{Z}_{D11}\widetilde{Z}_{D1p} \\ \vdots & \ddots & \vdots \\ \widetilde{Z}_{D1p}\widetilde{Z}_{D11} & \cdots & \widetilde{Z}_{D1p}\widetilde{Z}_{D1p} \end{pmatrix} \otimes \begin{pmatrix} \widetilde{Z}_{D11}\widetilde{Z}_{D11} & \cdots & \widetilde{Z}_{D11}\widetilde{Z}_{D1p} \\ \vdots & \ddots & \vdots \\ \widetilde{Z}_{D1p}\widetilde{Z}_{D11} & \cdots & \widetilde{Z}_{D1p}\widetilde{Z}_{D1p} \end{pmatrix}\right) \\ =& E\begin{pmatrix} \widetilde{Z}_{D11}\widetilde{Z}_{D11}\widetilde{Z}_{D11}\widetilde{Z}_{D11} & \cdots & \widetilde{Z}_{D11}\widetilde{Z}_{D1p}\widetilde{Z}_{D11}\widetilde{Z}_{D1p} \\ \vdots & \ddots & \vdots \\ \widetilde{Z}_{D1p}\widetilde{Z}_{D11}\widetilde{Z}_{D1p}\widetilde{Z}_{D11} & \cdots & \widetilde{Z}_{D1p}\widetilde{Z}_{D1p}\widetilde{Z}_{D1p}\widetilde{Z}_{D1p} \end{pmatrix}. \end{aligned}$$

We need to find $E(\widetilde{Z}_{D1i}\widetilde{Z}_{D1j}\widetilde{Z}_{D1k}\widetilde{Z}_{D1h})$. Notice that

$$E(\widetilde{Z}_{D1i}\widetilde{Z}_{D1j}\widetilde{Z}_{D1k}\widetilde{Z}_{D1h}) = \frac{\partial^4}{\partial t_i \partial t_j \partial t_k \partial t_h}\psi_{\widetilde{\boldsymbol{Z}}_1}(t),$$

where $\psi_{\widetilde{\boldsymbol{Z}}_1}(t)$ is the characteristic function of $\widetilde{\boldsymbol{Z}}_1$. Since $\widetilde{\boldsymbol{Z}}_1 = C\boldsymbol{Z}_1 = C(\mathbf{Y}_1 - E(\mathbf{Y}_1)) = C(\mathbf{Y}_1 - [(1-\epsilon)\boldsymbol{\eta}_D + \epsilon\boldsymbol{\eta}_H])$, we have

$$\begin{aligned}
\psi_{\boldsymbol{Z}_1}(t) =& \exp(-it^\top[(1-\epsilon)\boldsymbol{\eta}_D + \epsilon\boldsymbol{\eta}_H])\psi_{\mathbf{Y}_1}(t) \\
=& \exp(-it^\top[(1-\epsilon)\boldsymbol{\eta}_D + \epsilon\boldsymbol{\eta}_H])\Big[(1-\epsilon)\exp\left(it^\top\boldsymbol{\eta}_D - \frac{1}{2}t^\top\Sigma t\right) \\
&+ \epsilon\exp\left(it^\top\boldsymbol{\eta}_H - \frac{1}{2}t^\top\Sigma t\right)\Big] \\
=&(1-\epsilon)\exp\left(\epsilon it^\top(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H) - \frac{1}{2}t^\top\Sigma t\right) \\
&+ \epsilon\exp\left((1-\epsilon)it^\top(\boldsymbol{\eta}_H - \boldsymbol{\eta}_D) - \frac{1}{2}t^\top\Sigma t\right)
\end{aligned}$$

and

$$\begin{aligned}
\psi_{\widetilde{\boldsymbol{Z}}_1}(t) =& \psi_{\boldsymbol{Z}_1}(C^\top t) \\
=&(1-\epsilon)\exp\left(\epsilon it^\top C^\top(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H) - \frac{1}{2}t^\top C\Sigma C^\top t\right) \\
&+ \epsilon\exp\left((1-\epsilon)it^\top C^\top(\boldsymbol{\eta}_H - \boldsymbol{\eta}_D) - \frac{1}{2}t^\top C\Sigma C^\top t\right) \\
=&(1-\epsilon)\exp\left(\epsilon it^\top\boldsymbol{\Delta} - \frac{1}{2}t^\top C\Sigma C^\top t\right) \\
&+ \epsilon\exp\left(-(1-\epsilon)it^\top\boldsymbol{\Delta} - \frac{1}{2}t^\top C\Sigma C^\top t\right).
\end{aligned}$$

Set $(\omega_{ij}) = C\Sigma C^\top$ and note that $\boldsymbol{\Delta} = (d_1, \ldots, d_p)$, we have

$$\begin{aligned}
&E(\widetilde{Z}_{D1i}\widetilde{Z}_{D1j}\widetilde{Z}_{D1k}\widetilde{Z}_{D1h}) \\
=&\omega_{ij}\omega_{kh} + \omega_{ik}\omega_{jh} + \omega_{jk}\omega_{ih} + \epsilon(1-\epsilon)(\epsilon^3 + (1-\epsilon^3))d_id_jd_kd_h \\
&+ \epsilon(1-\epsilon)[\omega_{ij}d_kd_h + \omega_{ik}d_jd_h + \omega_{jk}d_id_h + \omega_{ih}d_jd_k + \omega_{jh}d_id_k + \omega_{kh}d_id_j].
\end{aligned} \tag{2.15}$$

Then by Lemma 1 in Harrar and Bathke (Harrar and Bathke, 2012) and properties of

trace operation, we have

$$\begin{aligned}
tr\left(\mathrm{V}ar\left(\widetilde{Z}_D A_D \widetilde{Z}_D^\top\right)\right) =& \sum_{i=1}^{n_D}\sum_{j=1}^{n_D} a_{ij}^2 (tr(\Sigma_D)^2 + tr(\Sigma_D^2)) \\
&+ \sum_{i=1}^{n_D} a_{ii}^2 \left(tr(E(vec(\widetilde{\boldsymbol{Z}}_{D_1}\widetilde{\boldsymbol{Z}}_{D_1}^\top)vec(\widetilde{\boldsymbol{Z}}_{D_1}\widetilde{\boldsymbol{Z}}_{D_1}^\top)^\top))\right) \\
&- \sum_{i=1}^{n_D} a_{ii}^2 \left((tr(\Sigma_D)^2 + tr(\Sigma_D^2)) - tr(vec(\Sigma_D)vec(\Sigma_D)^\top\right) \\
=& \frac{n_D - 1}{n_D} tr(\Sigma_D)^2 - \frac{(n_D-2)(n_D-1)}{n_D} tr(\Sigma_D^2) \\
&+ \frac{(n_D-1)^2}{n_D} \sum_{i=1}^{p}\sum_{j=1}^{p} E(\widetilde{Z}_{D1i}^2 \widetilde{Z}_{D1j}^2).
\end{aligned}$$

From equation (2.15) and definition of $\Sigma_D$, we have

$$\begin{aligned}
\sum_{i=1}^{p}\sum_{j=1}^{p} E(\widetilde{Z}_{D1i}^2 \widetilde{Z}_{D1j}^2) =& \sum_{i=1}^{p}\sum_{j=1}^{p} (\omega_{ii}\omega_{jj} + \epsilon(1-\epsilon)\omega_{ii}d_j^2 + \epsilon(1-\epsilon)\omega_{jj}d_i^2 + \epsilon^2(1-\epsilon)^2 d_i^2 d_j^2) \\
&+ 2\sum_{i=1}^{p}\sum_{j=1}^{p} (\omega_{ij}^2 + 2\epsilon(1-\epsilon)\omega_{ij}d_i d_j + \epsilon^2(1-\epsilon)^2 d_i^2 d_j^2) \\
&+ \sum_{i=1}^{p}\sum_{j=1}^{p} \epsilon(1-\epsilon)(1-6\epsilon+6\epsilon^2) d_i^2 d_j^2, \\
tr(\Sigma_D)^2 =& \sum_{i=1}^{p}\sum_{j=1}^{p} (\omega_{ii} + \epsilon(1-\epsilon)d_i^2)(\omega_{jj} + \epsilon(1-\epsilon)d_j^2) \\
=& \sum_{i=1}^{p}\sum_{j=1}^{p} (\omega_{ii}\omega_{jj} + \epsilon(1-\epsilon)\omega_{ii}d_j^2 + \epsilon(1-\epsilon)\omega_{jj}d_i^2 \\
&+ \epsilon^2(1-\epsilon)^2 d_i^2 d_j^2) \text{ and} \\
tr(\Sigma_D^2) =& \sum_{i=1}^{p}\sum_{j=1}^{p} (\omega_{ij} + \epsilon(1-\epsilon)d_i d_j)^2 \\
=& \sum_{i=1}^{p}\sum_{j=1}^{p} (\omega_{ij}^2 + 2\epsilon(1-\epsilon)\omega_{ij}d_i d_j + \epsilon^2(1-\epsilon)^2 d_i^2 d_j^2).
\end{aligned}$$

Thus,

$$\sum_{i=1}^{p}\sum_{j=1}^{p} E(\widetilde{Z}_{D1i}^2 \widetilde{Z}_{D1j}^2) = tr^2(\Sigma_D) + 2tr(\Sigma_D^2) + (1-6\epsilon+6\epsilon^2) tr^2(\boldsymbol{\Delta}\boldsymbol{\Delta}^\top).$$

Therefore, we have

$$tr\left(\mathrm{Var}\left(\widetilde{Z}_D A\widetilde{Z}_D^\top\right)\right) =(n_D-1)(tr^2(\Sigma_D)+tr(\Sigma_D^2)) + \frac{(n_D-1)^2}{n_D}(1-6\epsilon+6\epsilon^2)tr^2(\mathbf{\Delta}\mathbf{\Delta}^\top).$$

Similarly, for $Z_H = Y_H - E(Y_H)$ and

$$\widetilde{Z}_H = CZ_H = C\begin{pmatrix} Z_H^{(1)} \\ Z_H^{(2)} \end{pmatrix} = (\mathbf{Z}_{H1}^{(2)} - \mathbf{Z}_{H1}^{(1)}, \ldots, \mathbf{Z}_{Hn_H}^{(2)} - \mathbf{Z}_{Hn_H}^{(1)}),$$

we have

$$tr\left(\mathrm{Var}\left(\widetilde{Z}_H B\widetilde{Z}_H^\top\right)\right) =(n_H-1)(tr^2(\Sigma_H)+tr(\Sigma_H^2)) + \frac{(n_H-1)^2}{n_H}(1-6\delta+6\delta^2)tr^2(\mathbf{\Delta}\mathbf{\Delta}^\top).$$

Finally,

$$\begin{aligned}
&tr(\mathrm{Var}(CSC^\top)) \\
=&tr\left(\frac{1}{n_D^2}\mathrm{Var}(CS_DC^\top)\right) + tr\left(\frac{1}{n_H^2}\mathrm{Var}(CS_HC^\top)\right) \\
=&\frac{1}{n_D^2}tr\left(\mathrm{Var}\left(\frac{1}{n_D-1}\widetilde{Z}_D A\widetilde{Z}_D^\top\right)\right) + \frac{1}{n_H^2}tr\left(\mathrm{Var}\left(\frac{1}{n_H-1}\widetilde{Z}_H B\widetilde{Z}_H^\top\right)\right) \\
=&\frac{1}{n_D^2(n_D-1)^2}tr\left(\mathrm{Var}\left(\widetilde{Z}_D A\widetilde{Z}_D^\top\right)\right) + \frac{1}{n_H^2(n_H-1)^2}tr\left(\mathrm{Var}\left(\widetilde{Z}_H B\widetilde{Z}_H^\top\right)\right) \\
=&\frac{1}{n_D^2(n_D-1)}(tr^2(\Sigma_D)+tr(\Sigma_D^2)) + \frac{1}{n_H^2(n_H-1)}(tr^2(\Sigma_H)+tr(\Sigma_H^2)) \\
&+\left(\frac{1}{n_D^3}(1-6\epsilon+6\epsilon^2) + \frac{1}{n_H^3}(1-6\delta+6\delta^2)\right)tr^2(\mathbf{\Delta}\mathbf{\Delta}^\top).
\end{aligned}$$

Recalling $\Phi = \Sigma_D + \pi\Sigma_H$ and $\frac{n_D}{n_H} = \pi$,

$$f = \frac{tr^2(\Phi)+tr(\Phi^2)}{tr(\mathrm{Var}(CSC^\top))}, \tag{2.16}$$

where

$$tr^2(\Phi)+tr(\Phi^2) = tr^2(\Sigma_D+\pi\Sigma_H) + tr((\Sigma_D+\pi\Sigma_H)^2) \quad \text{and}$$

$$\begin{aligned}
tr(\mathrm{Var}(CSC^\top)) =&\frac{1}{n_D-1}(tr^2(\Sigma_D)+tr(\Sigma_D^2)) + \frac{\pi^3}{n_D-\pi}(tr^2(\Sigma_H)+tr(\Sigma_H^2)) \\
&+\frac{1}{n_D}\left(1-6\epsilon+6\epsilon^2\right) + \pi^3(1-6\delta+6\delta^2)\big)\, tr^2(\mathbf{\Delta}\mathbf{\Delta}^\top).
\end{aligned}$$

## Supplemental Simulation Results

This subsection contains additional simulation results that are discussed in Section 2.3 and Section 2.5

## Simulation Results for $F$ and $\chi^2$ Approximations

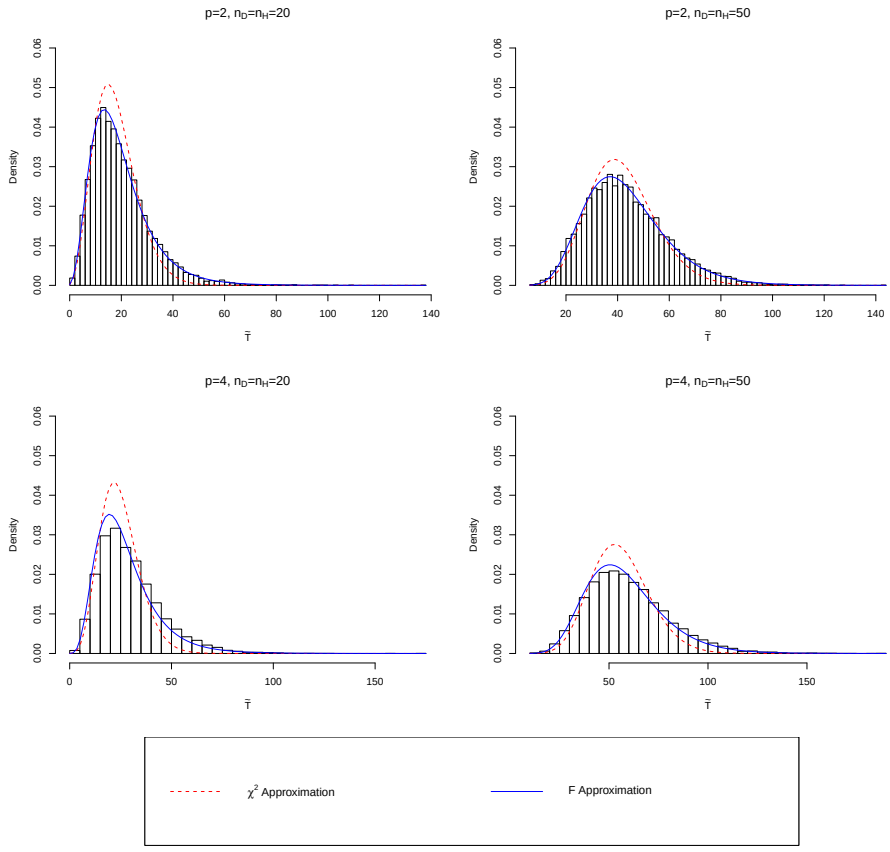



Figure 2.10: Histogram of $\widetilde{T}$ from 10000 simulations. Superposed are the density curves of $\chi^2$ (dashed line) and $F$ (solid line) approximations when $\epsilon = \delta = 0.1$, $\boldsymbol{\eta}_D = (20, 30)^\top$, $\boldsymbol{\eta}_H = (10, 14)^\top$, $\sigma^2 = 10$, $\rho_1 = 0.1$, and $\rho_2 = 0.25$.

Table 2.4: RB(%), SB(%) and CP(%) results when $p = 2, \sigma^2 = 10$.

| | | | Trad | | | Hyb | | | EMP | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $\delta$ | $\epsilon$ | $n$ | RB | SB | CP | RB | SB | CP | RB | SB | CP |
| 0.1 | 0.1 | 20 | 18.121 | 38.617 | 93.8 | 2.309 | 3.918 | 94.3 | 0.678 | 1.446 | 93.0 |
| | | 50 | 20.088 | 67.849 | 92.2 | 0.633 | 1.904 | 95.6 | 0.760 | 2.678 | 93.9 |
| | | 100 | 19.450 | 93.000 | 86.6 | 1.067 | 4.404 | 93.5 | 0.722 | 3.785 | 93.4 |
| | | 200 | 19.837 | 132.051 | 78.2 | 0.475 | 2.925 | 94.9 | 0.487 | 3.729 | 95.0 |
| | 0.3 | 20 | 41.555 | 88.950 | 88.8 | 3.059 | 3.903 | 94.3 | 2.231 | 4.284 | 92.5 |
| | | 50 | 40.017 | 136.360 | 78.9 | 2.128 | 4.716 | 95.3 | 2.066 | 7.138 | 94.4 |
| | | 100 | 40.163 | 188.878 | 61.6 | 1.777 | 5.610 | 95.3 | 1.082 | 5.695 | 95.3 |
| | | 200 | 40.715 | 272.356 | 32.2 | 1.612 | 6.838 | 95.5 | 0.828 | 5.828 | 95.0 |
| 0.3 | 0.1 | 20 | 39.692 | 84.644 | 90.5 | 4.301 | 6.102 | 94.5 | 3.187 | 7.105 | 93.7 |
| | | 50 | 39.257 | 132.102 | 81.1 | 1.434 | 3.086 | 94.1 | 0.517 | 1.853 | 95.4 |
| | | 100 | 40.181 | 195.060 | 64.2 | 0.536 | 1.649 | 95.2 | 1.283 | 6.803 | 95.8 |
| | | 200 | 40.934 | 274.266 | 30.4 | 1.496 | 6.284 | 95.6 | 0.160 | 1.110 | 93.9 |
| | 0.3 | 20 | 61.354 | 126.612 | 82.9 | 10.185 | 7.836 | 93.4 | 10.547 | 14.920 | 90.7 |
| | | 50 | 58.265 | 196.963 | 63.8 | 4.410 | 5.663 | 94.2 | 0.398 | 1.237 | 94.9 |
| | | 100 | 60.407 | 274.947 | 30.9 | 1.704 | 3.205 | 95.0 | 0.964 | 4.901 | 94.1 |
| | | 200 | 60.073 | 406.278 | 4.6 | 0.901 | 2.701 | 95.4 | 0.452 | 3.241 | 94.6 |

Table 2.5: RB(%), SB(%) and CP(%) results when $p = 2, \sigma^2 = 10$, $\mathbf{\Delta} = 2\mathbf{1}_p$.

| | | Trad | | | Hyb | | | EMP | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $(\delta, \epsilon)$ | $(\rho_1, \rho_2)$ | RB | SB | CP | RB | SB | CP | RB | SB | CP |
| (0.1,0.1) | (0.1, 0.25) | 19.450 | 93.000 | 86.6 | 1.067 | 4.404 | 93.5 | 0.722 | 3.785 | 93.4 |
| | (0.1, 0.7) | 20.044 | 147.863 | 76.7 | 0.208 | 1.336 | 96.1 | 0.121 | 0.996 | 95.7 |
| | (0.5, 0.7) | 20.775 | 133.209 | 81.0 | 1.251 | 7.082 | 95.7 | 1.168 | 9.943 | 93.5 |
| (0.1,0.3) | (0.1, 0.25) | 40.163 | 188.878 | 61.6 | 1.777 | 5.610 | 95.3 | 1.082 | 5.695 | 95.3 |
| | (0.1, 0.7) | 40.385 | 291.578 | 26.2 | 1.072 | 5.064 | 95.0 | 0.714 | 5.433 | 94.7 |
| | (0.5, 0.7) | 39.242 | 247.570 | 42.2 | 0.633 | 2.595 | 96.1 | 0.246 | 2.629 | 95.1 |
| (0.3,0.1) | (0.1, 0.25) | 40.181 | 195.060 | 64.2 | 0.536 | 1.649 | 95.2 | 1.283 | 6.803 | 95.8 |
| | (0.1, 0.7) | 40.577 | 285.623 | 26.5 | 1.327 | 6.084 | 94.2 | 0.801 | 5.906 | 93.0 |
| | (0.5, 0.7) | 39.887 | 244.200 | 40.7 | 0.433 | 2.983 | 95.7 | 0.781 | 4.933 | 94.0 |
| (0.3,0.3) | (0.1, 0.25) | 60.407 | 274.947 | 30.9 | 1.704 | 3.205 | 95.0 | 0.964 | 4.901 | 94.1 |
| | (0.1, 0.7) | 60.325 | 423.851 | 3.3 | 0.741 | 2.392 | 95.8 | 0.729 | 6.153 | 95.1 |
| | (0.5, 0.7) | 60.205 | 361.249 | 10.1 | 1.792 | 5.793 | 95.7 | 0.533 | 5.251 | 94.2 |

# Chapter 3 Estimation of Misclassification Error Rates

## 3.1 Introduction

Pre-stratified pre-post designs are commonly used in clinical trials to assess treatment effects. Diagnostic tools are used to stratify the participants into different groups, and these tools usually are imperfect. Traditional methods ignore the misclassification errors which leads to biased estimators and inaccurate tests. Only few works evaluated treatment effect when this misclassification error exist in some special situations and most works concentrate on estimating the accuracy of the diagnostic tools.

To fill in this methodological gap, Chapter 2 introduced two methods, the moment-based and the likelihood-based, for estimating and testing treatment effects when imperfect diagnostic devices are used. When available, more expensive and accurate diagnostic devices may sometimes be used to identify the actual group membership for some of the participants. In this case, we can obtain validation data to enhance the accuracy of the treatment effect estimation. Among the earliest works, Tenenbein (1970) proposed a double sampling scheme for estimating the proportion when there is misclassification on class membership. More recently, some researchers applied this strategy to assess disease prevalence. For example, Nedelman (1988) and Lie et al. (1994) used this scheme to investigate malaria prevalence in Nigeria and congenital malformations in Norway, respectively. Qiu et al. (2019) proposed test procedures for comparing disease prevalence rates in two groups when both classifiers in the double sampling scheme are fallible. However, these papers considered only count variables and focused on estimating the misclassification error rates.

This chapter extends the two methods developed in Chapter 2 to the situation where validation data is available. Section 2 presents the statistical model. We describe the theoretical motivation in Sections 3 and 4 and derive the moment-based and likelihood-based solutions. Simulation study will be conducted in Section 5 to illustrate the utility of the results derived. Section 6 concludes the chapter with discussion and remarks. All proofs and technical details are placed in Appendix.

### 3.2 Statistical Model and Parameter of Interest

Suppose we can verify the correct group membership for some of the study participants by employing a more expensive but accurate diagnostic tool. Let $\mathbf{V}_{Dj} = (\mathbf{V}_{Dj}^{(1)\top}, \mathbf{V}_{Dj}^{(2)\top})^\top$ for $j = 1, \ldots, m_D$ be the pre and post outcomes vectors for the $j$th individual whose positive disease status is validated. Further, let $\mathbf{V}_{Hj} = (\mathbf{V}_{Hj}^{(1)\top}, \mathbf{V}_{Hj}^{(2)\top})^\top$ for $j = 1, \ldots, m_H$ be outcomes measured from the $j$th individuals that is validated by the accurate classifier as healthy. Assume $\boldsymbol{V}_{D_j}$ and $\boldsymbol{V}_{H_j}$ have multivariate normal distributions, $\Phi(\boldsymbol{\eta}_D, \Sigma)$ and $\Phi(\boldsymbol{\eta}_H, \Sigma)$, respectively,with means $\boldsymbol{\eta}_D$, $\boldsymbol{\eta}_H$, respectively and common covariance $\Sigma$ are defined the same as in Section 2.2.

Let $\boldsymbol{Y}_{D_j} = \left(\boldsymbol{Y}_{D_j}^{(1)\top}, \boldsymbol{Y}_{D_j}^{(2)\top}\right)^\top$ and $\boldsymbol{Y}_{H_j} = \left(\boldsymbol{Y}_{H_j}^{(1)\top}, \boldsymbol{Y}_{H_j}^{(2)\top}\right)^\top$ be the pre and post outcomes vector for the $j$th individual classified by the fallible diagnostic tools as diseased and healthy, respectively. Denote by $(1-\epsilon)$ and $(1-\delta)$ the positvie predictive value (PPV) and negative predictive value (NPV), respectively, of the fallible diagnostic tool. We also assume that $\boldsymbol{V}_{D_i}$, $\boldsymbol{V}_{H_j}$, $\boldsymbol{Y}_{D_k}$, and $\boldsymbol{Y}_{H_l}$ are mutually independent, for $i = 1, \ldots, m_D$, $j = 1, \ldots, m_H$, $k = 1, \ldots, n_D$, and $l = 1, \ldots, n_H$. Then through the derivation in Section 2.2, the distribution $Y_{D_j}$ and $Y_{H_j}$ can be modeled as a mixture of multivariate normal distribution. That is

$$
\begin{aligned}
f_{\boldsymbol{Y}_{D_j}}(\boldsymbol{y}|\boldsymbol{\theta}) &= (1-\epsilon)\phi(\boldsymbol{y}|\boldsymbol{\eta}_D, \Sigma) + \epsilon\phi(\boldsymbol{y}|\boldsymbol{\eta}_H, \Sigma) \text{ and} \\
f_{\boldsymbol{Y}_{H_j}}(\boldsymbol{y}|\boldsymbol{\theta}) &= \delta\phi(\boldsymbol{y}|\boldsymbol{\eta}_D, \Sigma) + (1-\delta)\phi(\boldsymbol{y}|\boldsymbol{\eta}_H, \Sigma).
\end{aligned}
$$

To avoid nonidentifiability issue and switching label problems that are common in mixture models, we assume $0 \leq \epsilon < 0.5$, $0 \leq \delta < 0.5$ and $\boldsymbol{\mu}_D \neq \boldsymbol{\mu}_H$. For practical applications, we need information outside the collected data to ensure these assumptions hold. The parameter of interest is

$$
\boldsymbol{\Delta} = \boldsymbol{\tau}_D - \boldsymbol{\tau}_H,
$$

where $\boldsymbol{\Delta} = (d_1, \ldots, d_p)^\top$ is the vector of differences in the treatment effect in the diseased and healthy populations.

### 3.3 The Moment-Based Approaches

In Section 2.3, we derived an unbiased estimator for $\Delta$ using the moment-based method, i.e.,

$$\widetilde{\mathbf{\Delta}} = \frac{1}{1-\epsilon-\delta} C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H). \tag{3.1}$$

with variance

$$\begin{aligned} \mathrm{V}ar(\widetilde{\mathbf{\Delta}}) =& \frac{1}{(1-\epsilon-\delta)^2} \left[ C\Sigma C^\top \left\{ \frac{1}{n_D} + \frac{1}{n_H} \right\} \right. \\ & \left. + \left\{ \frac{\epsilon(1-\epsilon)}{n_D} + \frac{\delta(1-\delta)}{n_H} \right\} \mathbf{\Delta}\mathbf{\Delta}^\top \right]. \end{aligned} \tag{3.2}$$

Therefore, if we can estimate the misclassification errors $\epsilon$ and $\delta$, we can utilize (3.1) and (3.2) to estimate the treatment effect ($\mathbf{\Delta}$) and conduct hypotheses tests. We propose to derive consistent estimators of the misclassification error rates using novel distance-based criteria.

#### Estimation and Test on the Misclassification Error Rates ($\epsilon$ and $\delta$)

Hall (1981) proposed nonparametric estimators for the mixture proportions combining the contaminated (original) and validation data. The main idea is to estimate the mixing proportions by minimizing the distance between the empirical version of the mixture population and the linear combination of empirical versions of the component distributions. Inspired by this idea, we obtain estimates of $\epsilon$ and $\delta$ by minimizing the scaled distance between the mean of original data and the mixture of means of the two groups from the validation data. For our purpose, we choose the distance function,

$$\begin{aligned} \mathrm{D}_\Omega(\epsilon, \delta) =& ||\Omega^{-1/2} \left[ \overline{\mathbf{Y}}_D - ((1-\epsilon)\overline{\mathbf{V}}_D + \epsilon \overline{\mathbf{V}}_H) \right] ||^2 \\ & + ||\Omega^{-1/2} \left[ \overline{\mathbf{Y}}_H - (\delta \overline{\mathbf{V}}_D + (1-\delta)\overline{\mathbf{V}}_H) \right] ||^2, \end{aligned}$$

where $||\cdot||$ is the Euclidean norm and $\Omega$ is a $2p \times 2p$ symmetric positive definite matrix. The distance function will be minimized if $\epsilon$ and $\delta$ satisfy,

$$(\overline{\mathbf{V}}_D - \overline{\mathbf{V}}_H)^\top \Omega^{-1} (\overline{\mathbf{V}}_D - \overline{\mathbf{V}}_H)\epsilon = (\overline{\mathbf{V}}_D - \overline{\mathbf{V}}_H)^\top \Omega^{-1} (\overline{\mathbf{V}}_D - \overline{\mathbf{Y}}_D) \text{ and}$$

$$(\overline{\mathbf{V}}_D - \overline{\mathbf{V}}_H)^\top \Omega^{-1} (\overline{\mathbf{V}}_D - \overline{\mathbf{V}}_H)\delta = (\overline{\mathbf{V}}_H - \overline{\mathbf{V}}_D)^\top \Omega^{-1} (\overline{\mathbf{V}}_H - \overline{\mathbf{Y}}_H).$$

Thus, we have the estimators,

$$\widehat{\epsilon} = \frac{(\overline{\mathbf{V}}_D - \overline{\mathbf{V}}_H)^\top \Omega^{-1} (\overline{\mathbf{V}}_D - \overline{\mathbf{Y}}_D)}{(\overline{\mathbf{V}}_D - \overline{\mathbf{V}}_H)^\top \Omega^{-1} (\overline{\mathbf{V}}_D - \overline{\mathbf{V}}_H)} \quad \text{and} \quad \widehat{\delta} = \frac{(\overline{\mathbf{V}}_H - \overline{\mathbf{V}}_D)^\top \Omega^{-1} (\overline{\mathbf{V}}_H - \overline{\mathbf{Y}}_H)}{(\overline{\mathbf{V}}_D - \overline{\mathbf{V}}_H)^\top \Omega^{-1} (\overline{\mathbf{V}}_D - \overline{\mathbf{V}}_H)}. \tag{3.3}$$

To establish the consistency of the estimators and simplify the expression for their asymptotic variance, we need a standard proportional divergence requirements on the group sample sizes as follows.

**Assumption 3.3.1.** *There exists positive constant $C_L$ and $C_U$, such that*

$$C_L \leq \min\{\frac{N}{n_D}, \frac{N}{n_H}, \frac{N}{m_D}, \frac{N}{m_H}\} \leq \max\{\frac{N}{n_D}, \frac{N}{n_H}, \frac{N}{m_D}, \frac{N}{m_H}\} \leq C_U,$$

*where $N = n_D + n_H + m_D + m_H$. Moreover,*

$$\frac{N}{n_D} \to \kappa_{n_D} > 0, \frac{N}{n_H} \to \kappa_{n_H} > 0, \frac{N}{m_D} \to \kappa_{m_D} > 0, \frac{N}{m_H} \to \kappa_{m_H} > 0, \text{ as } N \to \infty$$

.

Proposition 3.3.1 establishes the consistency of the estimators $\widehat{\epsilon}$ and $\widehat{\delta}$ under Assumption 3.3.1.

**Proposition 3.3.1.** *Let $\widehat{\epsilon}$ and $\widehat{\delta}$ be as defined in (3.3). Under Assumption 3.3.1 , we have $\widehat{\epsilon} \xrightarrow{P} \epsilon$ and $\widehat{\delta} \xrightarrow{P} \delta$, as $N \to \infty$.*

By using the Delta method, we can derive the asymptotic distribution of $\widehat{\epsilon}$ and $\widehat{\delta}$. The results are summarized in Theorem 3.3.1. Detailed calculations are included in Appendix 3.7

**Theorem 3.3.1.** *Let $\widehat{\epsilon}$ and $\widehat{\delta}$ be defined as in 3.3. Under Assumption 3.3.1,*

$$\sqrt{N}(\widehat{\epsilon} - \epsilon) \xrightarrow{D} Z_\epsilon \sim N(0, \sigma_\epsilon^2), \tag{3.4}$$

*and*

$$\sqrt{N}(\widehat{\delta} - \delta) \xrightarrow{D} Z_\delta \sim N(0, \sigma_\delta^2), \tag{3.5}$$

*where*

$$\sigma_\epsilon^2 = \kappa_{n_D}\epsilon(1-\epsilon) + (\kappa_{n_D} + \kappa_{m_D}(1-\epsilon)^2 + \kappa_{m_H}\epsilon^2)\frac{(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1} \Sigma \Omega^{-1} (\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)}{((\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1} (\boldsymbol{\eta}_D - \boldsymbol{\eta}_H))^2},$$

*and*

$$\sigma_\delta^2 = \kappa_{n_H}\delta(1-\delta) + (\kappa_{n_H} + \kappa_{m_H}(1-\delta)^2 + \kappa_{m_D}\delta^2)\frac{(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1}\Sigma\Omega^{-1}(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)}{((\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1}(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H))^2}.$$

We can estimate $\sigma_\epsilon^2$ and $\sigma_\delta^2$ in (3.4) and (3.5) by

$$S_\epsilon = \frac{N}{n_D}\widehat{\epsilon}(1-\widehat{\epsilon}) + \left(\frac{N}{n_D} + \frac{N}{m_D}(1-\widehat{\epsilon})^2 + \frac{N}{m_H}\widehat{\epsilon}^2\right) S, \tag{3.6}$$

and

$$S_\delta = \frac{N}{n_H}\widehat{\delta}(1-\widehat{\delta}) + \left(\frac{N}{n_D} + \frac{N}{m_H}(1-\widehat{\delta})^2 + \frac{N}{m_D}\widehat{\delta}^2\right) S, \tag{3.7}$$

respectively, where

$$S = \frac{(\overline{\mathbf{V}}_D - \overline{\mathbf{V}}_H)^\top \Omega^{-1} S_p \Omega^{-1}(\overline{\mathbf{V}}_D - \overline{\mathbf{V}}_H)}{((\overline{\mathbf{V}}_D - \overline{\mathbf{V}}_H)^\top \Omega^{-1}(\overline{\mathbf{V}}_D - \overline{\mathbf{V}})^2},$$

and $S_p$ is the pooled variance calculated from the validated samples $\mathbf{V}_D$'s and $\mathbf{V}_H$'s. Based on these estimates, we can develop methods for confidence intervals and hypothesis test for $\epsilon$ and $\delta$. Suppose we are interested in testing the Hypothesis $H_0 : \epsilon = \epsilon_0$, we may use the test statistic

$$T = \frac{\sqrt{N}(\widehat{\epsilon} - \epsilon)}{\sqrt{S_\epsilon}} \xrightarrow{D} Z \sim N(0,1).$$

Further, a $(1-\alpha)100\%$ asymptotic confidence interval for $\epsilon$ can be constructed by

$$P\left(\widehat{\epsilon} - \frac{z_{\alpha/2}\sqrt{S_\epsilon}}{\sqrt{N}} \le \epsilon \le \widehat{\epsilon} + \frac{z_{\alpha/2}\sqrt{S_\epsilon}}{\sqrt{N}}\right) \to 1-\alpha,$$

where $z_{\alpha/2}$ is the $(1-\alpha/2)$th-quantile of the standard normal distribution. Test statistic and confidence interval for $\delta$ can be constructed similarly.

### Estimation and Test for $\boldsymbol{\Delta}$

Using (3.1), we can estimate $\boldsymbol{\Delta}$ by

$$\widetilde{\boldsymbol{\Delta}} = \frac{1}{1-\widehat{\epsilon}-\widehat{\delta}} C(\overline{\mathbf{Y}}_D - \overline{\mathbf{Y}}_H). \tag{3.8}$$

Since $\widehat{\epsilon}$ and $\widehat{\delta}$ are consistent for $\epsilon$ and $\delta$, respectively, by continuous mapping theorem, $\widetilde{\mathbf{\Delta}}$ is a consistent estimator of $\mathbf{\Delta}$. We may want to estimate the covariance matrix of $\tilde{\mathbf{\Delta}}$ by plugging $\widehat{\epsilon}$ and $\widehat{\delta}$ in (3.2) as

$$S_{\widetilde{\mathbf{\Delta}}} = (1-\widehat{\epsilon}-\widehat{\delta})^{-2} C S C^\top,$$

where $S = n_D^{-1} S_D + n_H^{-1} S_H$ and $S_D$ and $S_H$ are the sample covariance matrix calculated from the contaminated samples $\mathbf{Y}_D$'s and $\mathbf{Y}_H$'s, respectively. However, this estimator is inefficient, because $\widetilde{\mathbf{\Delta}}$ involves estimators of $\epsilon$ and $\delta$, and (3.2) assumes that $\epsilon$ and $\delta$ are known and it does not take into account estimation errors in $\widehat{\epsilon}$ and $\widehat{\delta}$. To obtain more accurate estimator of the covariance matrix, we use Delta method and derive the asymptotic distribution of $\widetilde{\mathbf{\Delta}}$ in Theorem 3.3.2.

**Theorem 3.3.2.** *Let $\widetilde{\mathbf{\Delta}}$ be as defined in (3.8). Under Assumption 3.3.1,*

$$\sqrt{N}(\widetilde{\mathbf{\Delta}} - \mathbf{\Delta}) \xrightarrow{D} Z \sim N(0, \Sigma_{\mathbf{\Delta}}), \tag{3.9}$$

*where*

$$\begin{aligned}
\Sigma_{\mathbf{\Delta}} =& \frac{\kappa_{n_D} + \kappa_{n_H}}{(1-\epsilon-\delta)^2} C\Sigma C^\top - \frac{\kappa_{n_D} + \kappa_{n_H}}{1-\epsilon-\delta} \cdot \frac{C\Sigma\Omega^{-1}(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top C^\top}{(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1}(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)} \\
& - \frac{\kappa_{n_D} + \kappa_{n_H}}{1-\epsilon-\delta} \cdot \frac{C(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1}\Sigma C^\top}{(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1}(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)} \\
& + \frac{(\kappa_{n_D}\epsilon(1-\epsilon) + \kappa_{n_H}\delta(1-\delta))(\epsilon+\delta)^2}{(1-\epsilon-\delta)^2} C(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top C^\top \\
& + \left[\kappa_{n_D} + \kappa_{n_H} + (\kappa_{m_D} + \kappa_{m_H})\left(2 - \frac{1}{1-\epsilon-\delta}\right)^2\right] \Sigma^*,
\end{aligned} \tag{3.10}$$

*and*

$$\Sigma^* = \frac{C(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1}\Sigma\Omega^{-1}(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top C^\top}{((\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1}(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H))^2}.$$

Similar to (3.6) and (3.7), by replacing $(\kappa_{n_D}, \kappa_{n_H}, \kappa_{m_D}, \kappa_{m_H}, \epsilon, \delta, \boldsymbol{\eta}_D, \boldsymbol{\eta}_H, \Sigma)$ with their corresponding estimators $(\frac{N}{n_D}, \frac{N}{n_H}, \frac{N}{m_D}, \frac{N}{m_H}, \widehat{\epsilon}, \widehat{\delta}, \overline{\boldsymbol{V}}_D, \overline{\boldsymbol{V}}_H, S_p)$ in (3.10), we can obtain an estimator of $\Sigma_\Delta$, which we will denote it as $\widehat{\Sigma}_\Delta$.

To test the null hypothesis $H_0 : \mathbf{\Delta} = \mathbf{\Delta}_0$, based on the result of Theorem 3.3.2, we can use the test statistic

$$\widetilde{T} = N(\widetilde{\mathbf{\Delta}} - \mathbf{\Delta}_0)^\top (\widehat{\Sigma}_{\mathbf{\Delta}})^{-1} (\widetilde{\mathbf{\Delta}} - \mathbf{\Delta}_0).$$

The statistic $\widetilde{T}$ is asymptotically distributed as $\chi^2$ distribution with $p$ degree of freedom and non-centrality parameter $N(\boldsymbol{\Delta}-\boldsymbol{\Delta}_0)^\top\Sigma_{\boldsymbol{\Delta}}^{-1}(\boldsymbol{\Delta}-\boldsymbol{\Delta}_0)$ as $N\to\infty$. Under the null hypothesis, the non-centrality parameter is zero and the decision rule is to reject the null hypothesis at significance level $\alpha$ if $\widetilde{T} > \chi_p^2(1-\alpha)$. We can also use $\widetilde{T}$ to construct a $(1-\alpha)$ asymptotic confidence region for $\boldsymbol{\Delta}$ as

$$\big\{\boldsymbol{\Delta} : N(\widetilde{\boldsymbol{\Delta}}-\boldsymbol{\Delta})^\top(\widehat{\Sigma}_{\boldsymbol{\Delta}})^{-1}(\widetilde{\boldsymbol{\Delta}}-\boldsymbol{\Delta}) \le \chi_p^2(1-\alpha)\big\}. \tag{3.11}$$

### 3.4 The Likelihood-Based Approaches

When validation data is available, we can update the EM algorithm in Section 2.3 and incorporate the information contained in the validation data to get more accurate estimators of the parameters. Let $\mathbf{V}_{Dj} = (\mathbf{V}_{Dj}^{(1)\top}, \mathbf{V}_{Dj}^{(2)\top})^\top$ be validated pre and post outcomes vector for the $j$th individual in the diseased group for $j = 1,\ldots,m_D$, and $\mathbf{V}_{Hj} = (\mathbf{V}_{Hj}^{(1)\top}, \mathbf{V}_{Hj}^{(2)\top})^\top$ be validated pre and post outcomes vector for $j$th individual in healthy group, where $j = 1,\ldots,m_H$. Using the same notation as in Section 2.3, the log-likelihood for the complete data is

$$\begin{aligned} l_C(\boldsymbol{\theta}) = &\sum_{j=1}^{n_D}\big[I_D(z_{Dj})\{\log(1-\epsilon)+\log\phi(\mathbf{y}_{Dj}|\boldsymbol{\eta}_D,\Sigma)\} \\ &+ (I_H(z_{Dj}))\{\log\epsilon+\log\phi(\mathbf{y}_{Dj}|\boldsymbol{\eta}_H,\Sigma)\}\big] \\ &+ \sum_{j=1}^{n_H}\big[I_D(z_{Hj})\{\log\delta+\log\phi(\mathbf{y}_{Hj}|\boldsymbol{\eta}_D,\Sigma)\} \\ &+ I_H(z_{Hj})\{\log(1-\delta)+\log\phi(\mathbf{y}_{Hj}|\boldsymbol{\eta}_H,\Sigma)\})\big] \\ &+ \sum_{j=1}^{m_D}\log\phi(\mathbf{v}_{D_j}|\boldsymbol{\eta}_D,\Sigma) + \sum_{j=1}^{m_H}\log\phi(\mathbf{v}_{H_j}|\boldsymbol{\eta}_H,\Sigma). \end{aligned}$$

The detailed derivations of the EM iterations are given in Appendix 3.7. For initial values, we propose using the estimates of $\epsilon$ and $\delta$ from (3.3), and weighted averages of the method of moments estimators of the original and the validated data for $\boldsymbol{\eta}_D$, $\boldsymbol{\eta}_H$, and $\Sigma$. The details are given in Appendix 3.7. Similar to Section 2.3, we do not have explicit form of the covariance matrix of $\widehat{\Delta}$. After obtaining the maximum likelihood estimator $\widehat{\boldsymbol{\theta}} = (\widehat{\epsilon},\widehat{\delta},\widehat{\boldsymbol{\eta}_D},\widehat{\boldsymbol{\eta}_H},\widehat{\Sigma})$, we again propose using the bootstrap estimator of the covariance

matrix, denoted by $S_B$. For testing the null hypothesis $H_0 : \Delta = 0$, we propose comparing the statistic $\widehat{T} = \widehat{\Delta}^\top S_B^{-1} \widehat{\Delta}$ against the appropriate percentile of the $\chi^2$-distribution with $p$ degree of freedom.

### 3.5 Numerical Study

This section evaluates the performance of the proposed estimation methods in sections 3 and 4. We check the accuracy in estimating misclassfication error rates $\epsilon$ and $\delta$ and also evaluate the two estimators for the treatment effect $\boldsymbol{\Delta} = \boldsymbol{\tau}_D - \boldsymbol{\tau}_H$. In the simulations, we set $\Omega = I_{2p}$. The rest parameter settings and the performance criteria for estimators are the same as Section 2.5.

#### Accuracy in the Estimation of $\epsilon$ and $\delta$

Generally, we expect the estimates of $\epsilon$ and $\delta$ with validated data to be more accurate. To demonstrate the accuracy, and check the effect of ratio of the contaminated and validation datasets sample sizes, $m_D/n_D$, we conducted a small-scale simulation study setting $p = 2$, $\sigma^2 = 10$, $\boldsymbol{\Delta} = 4\mathbf{1}_p$, and sample size ratios $m_D/n_D \in \{0.1, 0.3, 0.5\}$. The estimates of $\epsilon$ and $\delta$ for moment-based method are obtained from (3.3). Table 3.1 contains the bias and root means square error (RMSE) in estimating $\epsilon$ and $\delta$ from the moment-based method and EM algorithm. To make the differences more discernible, we showed both the biases and RMSEs results multiplied by $100$. From this table, we notice that both the biases and RMSEs are very small for the two methods, even when the sample size ratio is as small as $0.1$. The RMSEs for the MMV tends to get smaller when the ratio increase, while that of EMV seems to be less affected. In general, EMV is a bit more accurate than MMV.

#### Estimators for the Treatment Effect

In Sections 3 and 4, we introduced two estimators of the difference $\boldsymbol{\Delta} = \boldsymbol{\tau}_D - \boldsymbol{\tau}_H$, namely,

(1) the method of moment estimator (MMV) and

(2) the updated maximum likelihood estimator via EM algorithm (EMV).

Table 3.1: Bias $\times 100$ and root mean square error (RMSE) $\times 100$ for $\widehat{\epsilon}$ and $\widehat{\delta}$ for $p = 2, \sigma^2 = 10$ and $\mathbf{\Delta} = 4$. MMV is for the moment-based method and EMV is for the MLE via EM algorithm.

| | | | MMV | | | | EMV | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | $\widehat{\epsilon}$ | | $\widehat{\delta}$ | | $\widehat{\epsilon}$ | | $\widehat{\delta}$ | |
| $\delta$ | $\epsilon$ | ratio | Bias | RMSE | Bias | RMSE | Bias | RMSE | Bias | RMSE |
| 0.1 | 0.1 | 0.1 | -0.05 | 4.961 | 0.429 | 5.041 | -0.026 | 4.024 | 0.082 | 3.939 |
| | | 0.3 | 0.034 | 4.122 | 0.126 | 3.97 | -0.014 | 2.996 | 0.029 | 3.169 |
| | | 0.5 | 0.108 | 3.863 | 0.22 | 3.743 | 0.049 | 3.023 | 0.133 | 3.051 |
| | 0.3 | 0.1 | 0.015 | 6.105 | 0.137 | 5.088 | 0.025 | 4.721 | 0.096 | 3.013 |
| | | 0.3 | -0.196 | 5.331 | 0.291 | 4.086 | 0.009 | 4.699 | 0.04 | 3.056 |
| | | 0.5 | 0.074 | 4.879 | 0.143 | 3.718 | 0.072 | 4.549 | 0.139 | 3.001 |
| 0.3 | 0.1 | 0.1 | 0.249 | 5.045 | 0.189 | 5.723 | -0.079 | 3.001 | 0.23 | 4.443 |
| | | 0.3 | -0.226 | 4.007 | 0.11 | 5.13 | -0.125 | 3.04 | -0.073 | 4.612 |
| | | 0.5 | -0.034 | 3.814 | 0.022 | 5.082 | -0.072 | 3.009 | 0.108 | 4.647 |
| | 0.3 | 0.1 | -0.02 | 5.473 | -0.034 | 5.819 | -0.154 | 4.329 | -0.107 | 4.665 |
| | | 0.3 | 0.021 | 4.953 | -0.056 | 5.183 | -0.021 | 4.492 | 0.004 | 4.7 |
| | | 0.5 | -0.237 | 5.217 | -0.311 | 4.822 | -0.281 | 4.775 | -0.285 | 4.397 |

In the simulation, we evaluate these estimators and compare them with the traditional estimator that does not account for the classification errors. To facilitate the comparisons between the competing methods, we pull all the results from different settings into a boxplot except for the simulation factors depicted in the axes or panels labels. The boxplots are presented in Figure 3.1 and Figures 3.2-3.6 in the appendix. Figure 3.1 summarizes all the results for the three estimators. It shows that the traditional method provides misleading results when misclassification errors exist. To make the comparison between MMV and EMV more precise, we excluded the traditional method in Figure 3.2-3.6. Similar to the results in the Section 2.5, both estimators have generally good performances. EMV is more accurate and MMV test is conservative. Nevertheless, MMV does not require bootstrap method to estimate variance and, hence, is faster to compute.

### 3.6 Discussion and Conclusion

In Chapter 2, two approaches, viz, moment-based and likelihood-based, were proposed to estimate treatment effects in pre-post design when diagnostic devices used to classify subjects are fallible. In this chapter we consider the situation when validation data from an accurate diagnostic device is available. We combine the validation data with the original

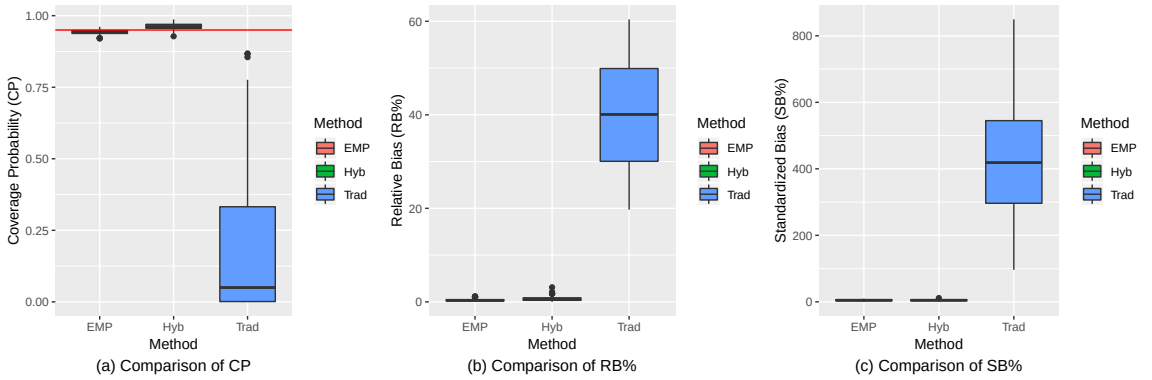



Figure 3.1: Boxplots of CP, RB%, and SB% for all methods. Trad is for the traditional method; MMV is for the moment-based method and EMV is for the MLE via EM algorithm.

data to provide updated versions of the moment- and EM-based estimators. The simulation study shows both methods have accurate estimators for the misclassification error rates, $\epsilon$ and $\delta$ and the treatment effect $\mathbf{\Delta}$. The EM-based estimators are relatively more accurate, but the moment-based methods are straight forward and computationally inexpensive. The simulation also shows that the traditional methods have unacceptably-large bias. Unless we are certain that the diagnostic tool is perfect, we should avoid using the traditional method.

The covariance matrices of the two groups are assumed to be the same in our model. Though it is reasonable to assume the treatment only changes the mean of the distribution, this assumption could be relax. To achieve that, we need to recalcuate the EM-algorithm for the likelihood-based approach. We also need to define a proper treatment effect to account for the diffierent covariance matrices in two groups. We plan to investigate these problem in future research.

## 3.7 Appendix

### Proofs

In this subsection, we give detailed proofs and technical details for the theoretical results presented in Section 3.3.

*Proof of Proposition 3.3.1.* Observe that

$$E(\mathbf{V}_{D_i}) - E(\mathbf{Y}_{D_i}) = \boldsymbol{\eta}_D - ((1-\epsilon)\boldsymbol{\eta}_D + \epsilon\boldsymbol{\eta}_H) = \epsilon(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)$$
$$= \epsilon(E(\mathbf{V}_{D_i}) - E(\mathbf{V}_{H_i})),$$

and

$$E(\mathbf{V}_{H_i}) - E(\mathbf{Y}_{H_i}) = \boldsymbol{\eta}_H - (\delta\boldsymbol{\eta}_D + (1-\delta)\boldsymbol{\eta}_H) = \delta(\boldsymbol{\eta}_H - \boldsymbol{\eta}_D)$$
$$= \epsilon(E(\mathbf{V}_{H_i}) - E(\mathbf{V}_{D_i})).$$

Thus, we have

$$\epsilon = \frac{(E(\mathbf{V}_{D_i}) - E(\mathbf{V}_{H_i}))^\top (E(\mathbf{V}_{D_i}) - E(\mathbf{Y}_{D_i}))}{(E(\mathbf{V}_{D_i}) - E(\mathbf{V}_{H_i}))^\top (E(\mathbf{V}_{D_i}) - E(\mathbf{V}_{H_i}))} \text{ and}$$
$$\delta = \frac{(E(\mathbf{V}_{H_i}) - E(\mathbf{V}_{D_i}))^\top (E(\mathbf{V}_{H_i}) - E(\mathbf{Y}_{H_i}))}{(E(\mathbf{V}_{D_i}) - E(\mathbf{V}_{H_i}))^\top (E(\mathbf{V}_{D_i}) - E(\mathbf{V}_{H_i}))}.$$

Under Assumption 3.3.1, when $N \to \infty$, $n_D, n_H, m_D, m_H \to \infty$ as well. By the Weak Law of Large Numbers and the Continuous Mapping Theorem, we can easily see that $\widehat{\epsilon}$ and $\widehat{\delta}$ converge in probability to $\epsilon$ and $\delta$, respectively. □ □

*Proof of Theorem 3.3.1.* Under the independence assumption and Assumption 3.3.1, by Central Limit Theorem, we have

$$\sqrt{N}\left(\begin{pmatrix} \overline{\mathbf{Y}}_D \\ \overline{\mathbf{Y}}_H \\ \overline{\mathbf{V}}_D \\ \overline{\mathbf{V}}_H \end{pmatrix} - \begin{pmatrix} (1-\epsilon)\boldsymbol{\eta}_D + \epsilon\boldsymbol{\eta}_H \\ (1-\delta)\boldsymbol{\eta}_H + \delta\boldsymbol{\eta}_D \\ \boldsymbol{\eta}_D \\ \boldsymbol{\eta}_H \end{pmatrix}\right) \xrightarrow{D} N\left(0, \begin{pmatrix} V_1 & 0 & 0 & 0 \\ 0 & V2 & 0 & 0 \\ 0 & 0 & V3 & 0 \\ 0 & 0 & 0 & V4 \end{pmatrix}\right),$$

where

$$V1 = \kappa_{n_D}(\Sigma + \epsilon(1-\epsilon)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top),$$
$$V2 = \kappa_{n_H}(\Sigma + \delta(1-\delta)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top), \text{ and}$$
$$V3 = \kappa_{m_D}\Sigma, \quad V4 = \kappa_{m_H}\Sigma.$$

For $\boldsymbol{\theta} = (\boldsymbol{\theta}_1^\top, \boldsymbol{\theta}_2^\top, \boldsymbol{\theta}_3^\top, \boldsymbol{\theta}_4^\top)^\top$, where $\boldsymbol{\theta}_i^\top$ is a $2p * 1$ vector, $i = 1, \ldots, 4$, define

$$g(\boldsymbol{\theta}) = \frac{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \Omega^{-1} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_3)}{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \Omega^{-1} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)}.$$

Then we have $\widehat{\epsilon} = g(\overline{\mathbf{V}}_D, \overline{\mathbf{V}}_H, \overline{\mathbf{Y}}_D)$ and $\widehat{\delta} = g(\overline{\mathbf{V}}_H, \overline{\mathbf{V}}_D, \overline{\mathbf{Y}}_H)$. To apply the multivariate Delta method, we need to calculate the Jacobain of $g$, denoted by $J_g(\boldsymbol{\theta})$, given by

$$J_g(\boldsymbol{\theta}) = \left( \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}, \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2}, \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_3} \right),$$

where

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} = \frac{(2\boldsymbol{\theta}_1 - \boldsymbol{\theta}_1 - \boldsymbol{\theta}_3)^\top \Omega^{-1}}{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \Omega^{-1} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)} - \frac{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \Omega^{-1} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_3)}{((\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \Omega^{-1} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^2} \cdot 2(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \Omega^{-1},$$

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2} = -\frac{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_3)^\top \Omega^{-1}}{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \Omega^{-1} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)} + \frac{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \Omega^{-1} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_3)}{((\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \Omega^{-1} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^2} \cdot 2(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \Omega^{-1},$$

and

$$\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_3} = -\frac{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_1)^\top \Omega^{-1}}{(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)^\top \Omega^{-1} (\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)}.$$

For $\boldsymbol{\theta_0} = (E(\overline{\mathbf{V}}_D), E(\overline{\mathbf{V}}_H), E(\overline{\mathbf{Y}}_D)) = (\boldsymbol{\eta}_D, \boldsymbol{\eta}_H, (1-\epsilon)\boldsymbol{\eta}_D + \epsilon\boldsymbol{\eta}_H)$, we have

$$J_g(\boldsymbol{\theta}_0) = \left( \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_3}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right),$$

where

$$\begin{aligned}
\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &= \frac{(1-\epsilon)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1}}{(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1} (\boldsymbol{\eta}_D - \boldsymbol{\eta}_H),} \\
\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &= -\frac{\epsilon(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1}}{(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1} (\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)}, \text{ and} \\
\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_3}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} &= -\frac{(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1}}{(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1} (\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)}.
\end{aligned}$$

Applying the multivariate Delta method, we have

$$\sqrt{N}(\widehat{\epsilon} - \epsilon) = \sqrt{N}\left( g(\overline{\mathbf{V}}_D, \overline{\mathbf{V}}_H, \overline{\mathbf{Y}}_D) - g(\boldsymbol{\theta}_0) \right) \xrightarrow{D} N(0, \sigma_\epsilon^2),$$

where

$$\begin{aligned}
\sigma_\epsilon^2 =& J_g(\boldsymbol{\theta_0}) \begin{pmatrix} V3 & 0 & 0 \\ 0 & V4 & 0 \\ 0 & 0 & V1 \end{pmatrix} J_g(\boldsymbol{\theta_0})^\top \\
=& \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} V3 \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}^\top + \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} V4 \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}^\top + \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_3}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} V1 \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_3}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}^\top \\
=& \kappa_{n_D}\epsilon(1-\epsilon) + (\kappa_{n_D} + \kappa_{m_D}(1-\epsilon)^2 + \kappa_{m_H}\epsilon^2) \frac{(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1} \Sigma \Omega^{-1} (\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)}{((\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1} (\boldsymbol{\eta}_D - \boldsymbol{\eta}_H))^2}.
\end{aligned}$$

The asymptotic distribution of $\widehat{\delta}$ can be derived similarly. □

*Proof of Theorem 3.3.2.* Similar to the proof of Theorem 3.3.1, we first define a function $\boldsymbol{g}(\boldsymbol{\theta})$ for the estimator of interest $\widetilde{\boldsymbol{\Delta}}$. Let

$$\boldsymbol{g}(\boldsymbol{\theta}) = \boldsymbol{g}(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4) = \frac{(\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)^\top \Omega^{-1}((\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4))}{(\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)^\top \Omega^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)} \cdot C(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2).$$

Then we have $\widetilde{\boldsymbol{\Delta}} = \boldsymbol{g}(\overline{\mathbf{Y}}_D, \overline{\mathbf{Y}}_H, \overline{\mathbf{V}}_D, \overline{\mathbf{V}}_H)$. Notice that

$$\begin{aligned}\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1} = -\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2} =& \frac{(\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)^\top \Omega^{-1}(\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)}{(\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)^\top \Omega^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)} \cdot C \\ & - \frac{(\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)^\top \Omega^{-1}(\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)}{((\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)^\top \Omega^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2))^2} \cdot C(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)(\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)^\top \Omega^{-1},\end{aligned}$$

and

$$\begin{aligned}\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_3} = -\frac{\partial \boldsymbol{g}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_4} =& \frac{1}{(\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)^\top \Omega^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)} \cdot 2C(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)(\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)^\top \Omega^{-1} \\ & - \frac{(\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)^\top \Omega^{-1}(\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)}{((\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)^\top \Omega^{-1}(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2))^2} C(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)(\boldsymbol{\theta}_3 - \boldsymbol{\theta}_4)^\top \Omega^{-1}.\end{aligned}$$

Then for $\boldsymbol{\theta}_0 = (E(\overline{\mathbf{Y}}_D), E(\overline{\mathbf{Y}}_H), E(\overline{\mathbf{V}}_D), E(\overline{\mathbf{V}}_H)) = ((1-\epsilon)\boldsymbol{\eta}_D + \epsilon\boldsymbol{\eta}_H, (1-\delta)\boldsymbol{\eta}_H + \delta\boldsymbol{\eta})D, \boldsymbol{\eta}_D, \boldsymbol{\eta}_H)$, we have

$$J_g(\boldsymbol{\theta}_0) = \left( \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_3}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}, \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_4}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} \right),$$

where

$$\begin{aligned}&\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = -\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \frac{C}{1-\epsilon-\delta} + \frac{C(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1}}{(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1}(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)} \text{ and} \\ &\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_3}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = -\frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_4}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} = \left(2 - \frac{1}{(1-\epsilon-\delta)^2}\right) \frac{C(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)^\top \Omega^{-1}}{(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)\Omega^{-1}(\boldsymbol{\eta}_D - \boldsymbol{\eta}_H)}.\end{aligned}$$

Applying the multivariate Delta method, we have

$$\sqrt{N}(\widetilde{\boldsymbol{\Delta}} - \boldsymbol{\Delta}) = \sqrt{N}(g(\overline{\mathbf{Y}}_D, \overline{\mathbf{Y}}_H, \overline{\mathbf{V}}_D, \overline{\mathbf{V}}_D) - g(\boldsymbol{\theta}_0) \xrightarrow{D} N(0, \Sigma_{\boldsymbol{\Delta}})),$$

where

$$\begin{aligned}
\Sigma_{\boldsymbol{\Delta}} =& \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} V1 \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}^{\top} + \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} V2 \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}^{\top} \\
&+ \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_3}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} V3 \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_3}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}^{\top} + \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_4}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0} V4 \frac{\partial g(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_4}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}^{\top} \\
=& \frac{\kappa_{n_D}+\kappa_{n_H}}{(1-\epsilon-\delta)^2} C\Sigma C^{\top} - \frac{\kappa_{n_D}+\kappa_{n_H}}{1-\epsilon-\delta} \cdot \frac{C\Sigma\Omega^{-1}(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)^{\top}C^{\top}}{(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)^{\top}\Omega^{-1}(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)} \\
&- \frac{\kappa_{n_D}+\kappa_{n_H}}{1-\epsilon-\delta} \cdot \frac{C(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)^{\top}\Omega^{-1}\Sigma C^{\top}}{(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)^{\top}\Omega^{-1}(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)} \\
&+ \frac{(\kappa_{n_D}\epsilon(1-\epsilon)+\kappa_{n_H}\delta(1-\delta))(\epsilon+\delta)^2}{(1-\epsilon-\delta)^2} C(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)^{\top}C^{\top} \\
&+ \left[\kappa_{n_D}+\kappa_{n_H}+(\kappa_{m_D}+\kappa_{m_H})\left(2-\frac{1}{1-\epsilon-\delta}\right)^2\right]\Sigma^*,
\end{aligned}$$

and

$$\Sigma^* = \frac{C(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)^{\top}\Omega^{-1}\Sigma\Omega^{-1}(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)^{\top}C^{\top}}{((\boldsymbol{\eta}_D-\boldsymbol{\eta}_H)^{\top}\Omega^{-1}(\boldsymbol{\eta}_D-\boldsymbol{\eta}_H))^2}.$$

□

**Technical Details for the EM algorithm**

Below we give some technical details and intermediate steps for the EM algorithm described in Section 3.4.

The log-likelihood function for the complete data is

$$\begin{aligned}
l_C(\boldsymbol{\theta}) =& \sum_{j=1}^{n_D}\big[I_D(z_{Dj})\{\log(1-\epsilon)+\log\phi(\mathbf{y}_{Dj}|\boldsymbol{\eta}_D,\Sigma)\} \\
&+ (I_H(z_{Dj}))\{\log\epsilon+\log\phi(\mathbf{y}_{Dj}|\boldsymbol{\eta}_H,\Sigma)\}\big] \\
&+ \sum_{j=1}^{n_H}\big[I_D(z_{Hj})\{\log\delta+\log\phi(\mathbf{y}_{Hj}|\boldsymbol{\eta}_D,\Sigma)\} \\
&+ I_H(z_{Hj})\{\log(1-\delta)+\log\phi(\mathbf{y}_{Hj}|\boldsymbol{\eta}_H,\Sigma)\})\big] \\
&+ \sum_{j=1}^{m_D}\log\phi(\mathbf{v}_{D_j}|\boldsymbol{\eta}_D,\Sigma) + \sum_{j=1}^{m_H}\log\phi(\mathbf{v}_{H_j}|\boldsymbol{\eta}_H,\Sigma).
\end{aligned}$$

**E step:**

For the $(t+1)^{th}$ expectation step of the EM algorithm,

$$\begin{aligned}
Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) =& E_{\theta^{(t)}}[l(\boldsymbol{\theta}|\boldsymbol{Y},\boldsymbol{Z},\boldsymbol{X})] \\
=& \sum_{j=1}^{n_D} K_{1j}^{(t)} \log(1-\epsilon) + \sum_{j=1}^{n_D} K_{1j}^{(t)} \log \phi(\mathbf{y}_{Dj}|\boldsymbol{\eta}_D, \Sigma) \\
&+ \sum_{j=1}^{n_D} \left(1 - K_{1j}^{(t)}\right) \log \epsilon + \sum_{j=1}^{n_D} \left(1 - K_{1j}^{(t)}\right) \log \phi(\mathbf{y}_{Dj}|\boldsymbol{\eta}_H, \Sigma) \\
&+ \sum_{j=1}^{n_H} K_{2j}^{(t)} \log \delta + \sum_{j=1}^{n_H} K_{2j}^{(t)} \log \phi(\mathbf{y}_{Hj}|\boldsymbol{\eta}_D, \Sigma) \\
&+ \sum_{j=1}^{n_H} \left(1 - K_{2j}^{(t)}\right) \log(1-\delta) + \sum_{j=1}^{n_H} \left(1 - K_{2j}^{(t)}\right) \log \phi(\mathbf{y}_{Hj}|\boldsymbol{\eta}_H, \Sigma) \\
&+ \sum_{j=1}^{m_D} \log \phi(\mathbf{v}_{D_j}|\boldsymbol{\eta}_D, \Sigma) + \sum_{j=1}^{m_H} \log \phi(\mathbf{v}_{H_j}|\boldsymbol{\eta}_H, \Sigma),
\end{aligned}$$

where

$$K_{1j}^{(t)} = \frac{(1-\epsilon^{(t)})\phi(\mathbf{y}_{Dj}|\boldsymbol{\eta}_D^{(t)}, \Sigma^{(t)})}{(1-\epsilon^{(t)})\phi(\mathbf{y}_{Dj}|\boldsymbol{\eta}_D^{(t)}, \Sigma^{(t)}) + \epsilon^{(t)}\phi(\mathbf{y}_{Dj}|\boldsymbol{\eta}_H^{(t)}, \Sigma^{(t)})} \text{ and} \tag{3.12}$$

$$K_{2j}^{(t)} = \frac{\delta^{(t)}\phi(\mathbf{y}_{Hj}|\boldsymbol{\eta}_D^{(t)}, \Sigma^{(t)})}{\delta^{(t)}\phi(\mathbf{y}_{Hj}|\boldsymbol{\eta}_D^{(t)}, \Sigma^{(t)}) + (1-\epsilon^{(t)})\phi(\mathbf{y}_{Hj}|\boldsymbol{\eta}_H^{(t)}, \Sigma^{(t)})}. \tag{3.13}$$

**M step:** For the maximization step of the EM algorithm, setting $\frac{\partial}{\partial\theta}Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbf{0}$ we have

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \epsilon} =& -\sum_{j=1}^{n_D} \frac{K_{1j}^{(t)}}{1-\epsilon} + \sum_{j=1}^{n_D} \frac{1 - K_{1j}^{(t)}}{\epsilon} = 0, \\
\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \delta} =& \sum_{j=1}^{n_H} \frac{K_{2j}^{(t)}}{\delta} + \sum_{j=1}^{n_H} \frac{1 - K_{2j}^{(t)}}{1-\delta} = 0,
\end{aligned}$$

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\eta}_D} =& \sum_{j=1}^{n_D} K_{1j}\Sigma^{-1}(\mathbf{y}_{Dj} - \boldsymbol{\eta}_D) + \sum_{j=1}^{n_H} K_{2j}\Sigma^{-1}(\mathbf{y}_{Hj} - \boldsymbol{\eta}_D) \\
&+ \sum_{j=1}^{m_D} \Sigma^{-1}(\mathbf{v}_{D_j} - \boldsymbol{\eta}_D) = \mathbf{0}, \\
\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \boldsymbol{\eta}_H} =& \sum_{j=1}^{n_D} (1 - K_{1j})\Sigma^{-1}(\mathbf{y}_{Dj} - \boldsymbol{\eta}_H) + \sum_{j=1}^{n_H} (1 - K_{2j})\Sigma^{-1}(\mathbf{y}_{Hj} - \boldsymbol{\eta}_H) \\
&+ \sum_{j=1}^{m_H} \Sigma^{-1}(\mathbf{v}_{H_j} - \boldsymbol{\eta}_H) = \mathbf{0},
\end{aligned}$$

and

$$\begin{aligned}\frac{\partial Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})}{\partial \Sigma} =& -\frac{1}{2}\sum_{j=1}^{n_D} K_{1j}(\Sigma^{-1} - \Sigma^{-1}(\mathbf{y}_{Dj} - \boldsymbol{\eta}_D)(\mathbf{y}_{Dj} - \boldsymbol{\eta}_D)^\top \Sigma^{-1}) \\ & -\frac{1}{2}\sum_{j=1}^{n_D}(1 - K_{1j})(\Sigma^{-1} - \Sigma^{-1}(\mathbf{y}_{Dj} - \boldsymbol{\eta}_H)(\mathbf{y}_{Dj} - \boldsymbol{\eta}_H)^\top \Sigma^{-1}) \\ & -\frac{1}{2}\sum_{j=1}^{n_H} K_{2j}(\Sigma^{-1} - \Sigma^{-1}(\mathbf{y}_{Hj} - \boldsymbol{\eta}_D)(\mathbf{y}_{Hj} - \boldsymbol{\eta}_D)^\top \Sigma^{-1}) \\ & -\frac{1}{2}\sum_{j=1}^{n_H}(1 - K_{2j})(\Sigma^{-1} - \Sigma^{-1}(\mathbf{y}_{Hj} - \boldsymbol{\eta}_H)(\mathbf{y}_{Hj} - \boldsymbol{\eta}_H)^\top \Sigma^{-1}) \\ & -\frac{1}{2}\sum_{j=1}^{m_D}(\Sigma^{-1} - \Sigma^{-1}(\mathbf{v}_{Dj} - \boldsymbol{\eta}_D)(\mathbf{v}_{Dj} - \boldsymbol{\eta}_D)^\top \Sigma^{-1}) \\ & -\frac{1}{2}\sum_{j=1}^{m_H}(\Sigma^{-1} - \Sigma^{-1}(\mathbf{v}_{Hj} - \boldsymbol{\eta}_H)(\mathbf{v}_{Hj} - \boldsymbol{\eta}_H)^\top \Sigma^{-1}) = 0.\end{aligned}$$

Then, for the maximization step of the EM algorithm, we set $\frac{\partial}{\partial \theta} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbf{0}$ and solve for $\boldsymbol{\theta}$ to obtain

$$\epsilon^{(t+1)} = 1 - \sum_{j=1}^{n_D} \frac{K_{1j}^{(t)}}{n_D},\ \delta^{(t+1)} = \sum_{j=1}^{n_H} \frac{K_{2j}^{(t)}}{n_H},$$

$$\begin{aligned}\boldsymbol{\eta}_D^{(t+1)} =& \frac{\sum_{j=1}^{n_D} K_{1j}^{(t)} \mathbf{y}_{Dj} + \sum_{j=1}^{n_H} K_{2j}^{(t)} \mathbf{y}_{Hj} + \sum_{j=1}^{m_D} \mathbf{v}_{D_j}}{\sum_{j=1}^{n_D} K_{1j}^{(t)} + \sum_{j=1}^{n_H} K_{2j}^{(t)} + m_D}, \\ \boldsymbol{\eta}_H^{(t+1)} =& \frac{\sum_{j=1}^{n_D} \left(1 - K_{1j}^{(t)}\right) \mathbf{y}_{Dj} + \sum_{j=1}^{n_H} \left(1 - K_{2j}^{(t)}\right) \mathbf{y}_{Hj} + \sum_{j=1}^{m_H} \mathbf{v}_{H_j}}{n_D + n_H + m_H - \left(\sum_{j=1}^{n_D} K_{1j}^{(t)} + \sum_{j=1}^{n_H} K_{2j}^{(t)}\right)} \text{ and}\end{aligned}$$

$$
\begin{aligned}
\Sigma^{(t+1)} =& \frac{\sum_{j=1}^{n_D} \boldsymbol{K}_{1j}^{(t)} \left(\mathbf{y}_{Dj} - \boldsymbol{\eta}_D^{(t)}\right)\left(\mathbf{y}_{Dj} - \boldsymbol{\eta}_D^{(t)}\right)^\top}{n_D + n_H + m_D + m_H} \\
&+ \frac{\sum_{j=1}^{n_D} \left(1 - \boldsymbol{K}_{1j}^{(t)}\right) \left(\mathbf{y}_{Dj} - \boldsymbol{\eta}_H^{(t)}\right)\left(\mathbf{y}_{Dj} - \boldsymbol{\eta}_H^{(t)}\right)^\top}{n_D + n_H + m_D + m_H} \\
&+ \frac{\sum_{j=1}^{n_D} \boldsymbol{K}_{2j}^{(t)} \left(\mathbf{y}_{Hj} - \boldsymbol{\eta}_D^{(t)}\right)\left(\mathbf{y}_{Hj} - \boldsymbol{\eta}_D^{(t)}\right)^\top}{n_D + n_H + m_D + m_H} \\
&+ \frac{\sum_{j=1}^{n_D} \left(1 - \boldsymbol{K}_{2j}^{(t)}\right) \left(\mathbf{y}_{Hj} - \boldsymbol{\eta}_H^{(t)}\right)\left(\mathbf{y}_{Hj} - \boldsymbol{\eta}_H^{(t)}\right)^\top}{n_D + n_H + m_D + m_H} \\
&+ \frac{\sum_{j=1}^{m_D} \left(\mathbf{v}_{D_j} - \boldsymbol{\eta}_D^{(t)}\right)\left(\mathbf{v}_{D_j} - \boldsymbol{\eta}_D^{(t)}\right)^\top}{n_D + n_H + m_D + m_H} \\
&+ \frac{\sum_{j=1}^{m_H} \left(\mathbf{v}_{H_j} - \boldsymbol{\eta}_H^{(t)}\right)\left(\mathbf{v}_{H_j} - \boldsymbol{\eta}_H^{(t)}\right)^\top}{n_D + n_H + m_D + m_H}.
\end{aligned}
$$

We propose the initial values

$$
\widetilde{\widetilde{\boldsymbol{\eta}}}_D = \frac{n_D \widetilde{\boldsymbol{\eta}}_D + m_D \overline{\mathbf{V}}_D}{n_D + m_D}, \quad \widetilde{\widetilde{\boldsymbol{\eta}}}_H = \frac{n_H \widetilde{\boldsymbol{\eta}}_H + m_H \overline{\mathbf{V}}_H}{n_H + m_H}, \text{ and}
$$

$$
\begin{aligned}
\widetilde{\widetilde{\Sigma}} =& \widetilde{\widetilde{S}}_P - \Big(\frac{n_D}{n_D + n_H} \frac{\epsilon(1-\epsilon)}{(1-\delta-\epsilon)^2} \\
&+ \frac{n_H}{n_D + n_H} \frac{\delta(1-\delta)}{(1-\delta-\epsilon)^2}\Big)(\overline{\mathbf{y}}_D - \overline{\mathbf{y}}_H)(\overline{\mathbf{y}}_D - \overline{\mathbf{y}}_H)^\top,
\end{aligned}
$$

where,

$$
\widetilde{\boldsymbol{\eta}}_D = \frac{(1-\widehat{\delta})\overline{\mathbf{y}}_D - \widehat{\epsilon}\overline{\mathbf{y}}_H}{1 - \widehat{\delta} - \widehat{\epsilon}}, \quad \widetilde{\boldsymbol{\eta}}_H = \frac{(1-\epsilon)\overline{\mathbf{y}}_H - \delta\overline{\mathbf{y}}_D}{1 - \delta - \epsilon},
$$

$\widehat{\epsilon}$ and $\widehat{\delta}$ are as defined in (3.3),

$$
\begin{aligned}
\widetilde{\widetilde{S}}_P &= (n_D + n_H + m_D + m_H)^{-1}(n_D \widetilde{S}_D + n_H \widetilde{S}_H + m_D \widetilde{\widetilde{S}}_D + m_H \widetilde{\widetilde{S}} H), \\
\widetilde{S}_D &= n_D^{-1} \sum_{j=1}^{n_D} (\mathbf{Y}_{Dj} - \overline{\mathbf{Y}}_D)(\mathbf{Y}_{Dj} - \overline{\mathbf{Y}}_D)^\top, \\
\widetilde{S}_H &= n_H^{-1} \sum_{j=1}^{n_H} (\mathbf{Y}_{Hj} - \overline{\mathbf{Y}}_H)(\mathbf{Y}_{Hj} - \overline{\mathbf{Y}}_H)^\top, \\
\widetilde{\widetilde{S}}_D &= m_D^{-1} \sum_{j=1}^{m_D} (\mathbf{V}_{Dj} - \overline{\mathbf{V}}_D)(\mathbf{V}_{Dj} - \overline{\mathbf{V}}_D)^\top, \quad \text{and} \\
\widetilde{\widetilde{S}}_H &= m_H^{-1} \sum_{j=1}^{m_H} (\mathbf{V}_{Hj} - \overline{\mathbf{V}}_H)(\mathbf{V}_{Hj} - \overline{\mathbf{V}}_H)^\top.
\end{aligned}
$$

## Supplemental Simulations Results

This subsection contains additional simulation results that are discussed in Section 3.5.

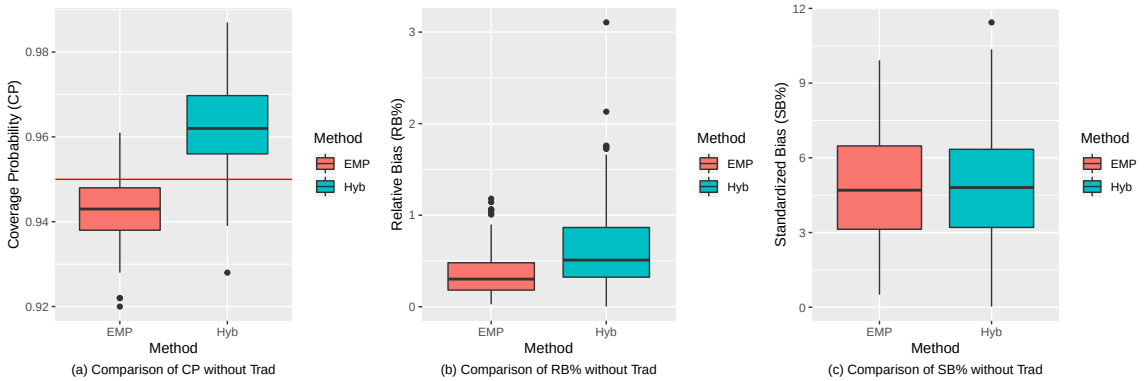



Figure 3.2: Boxplots of CP, RB%, and SB% for all methods except the traditional method. MMV is for the moment-based method and EMV is for the MLE via EM algorithm.

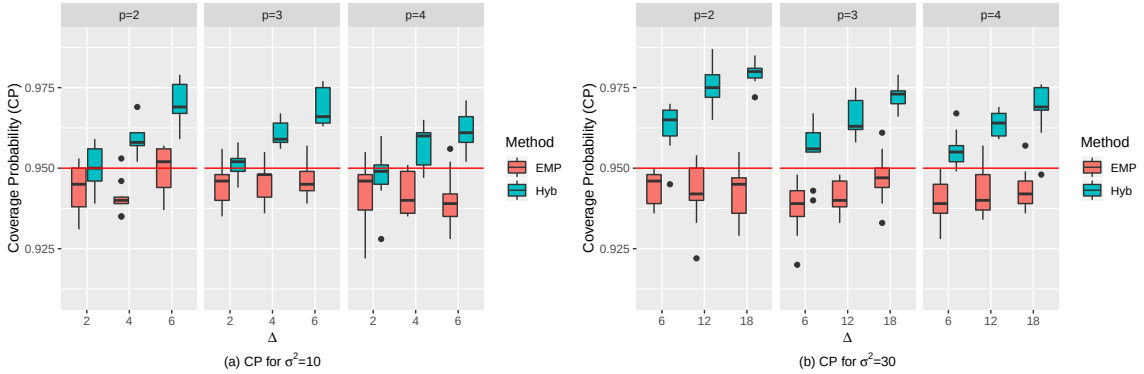



Figure 3.3: Boxplots of CP of different methods for different $p$, $\boldsymbol{\Delta}$ and $\sigma^2$. MMV is for the moment-based method and EMV is for the MLE via EM algorithm.

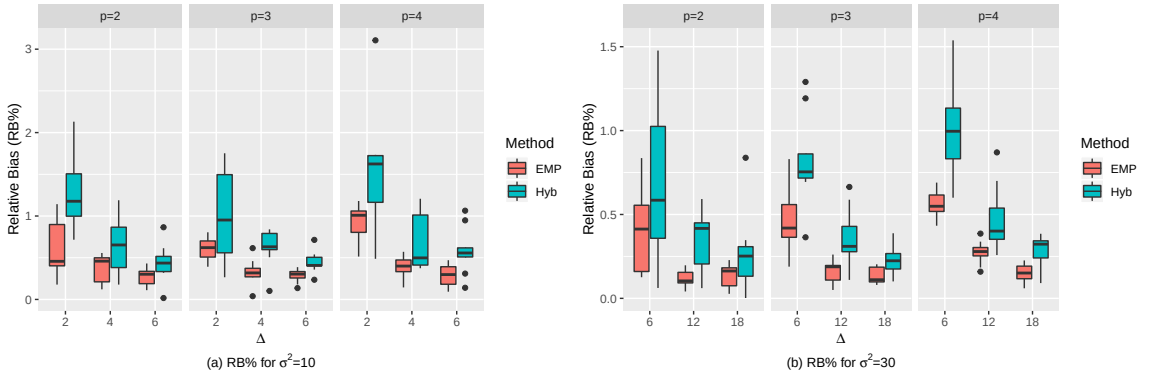



Figure 3.4: Boxplots of RB% of different methods for different $p$, $\boldsymbol{\Delta}$ and $\sigma^2$. MMV is for the moment-based method and EMV is for the MLE via EM algorithm.

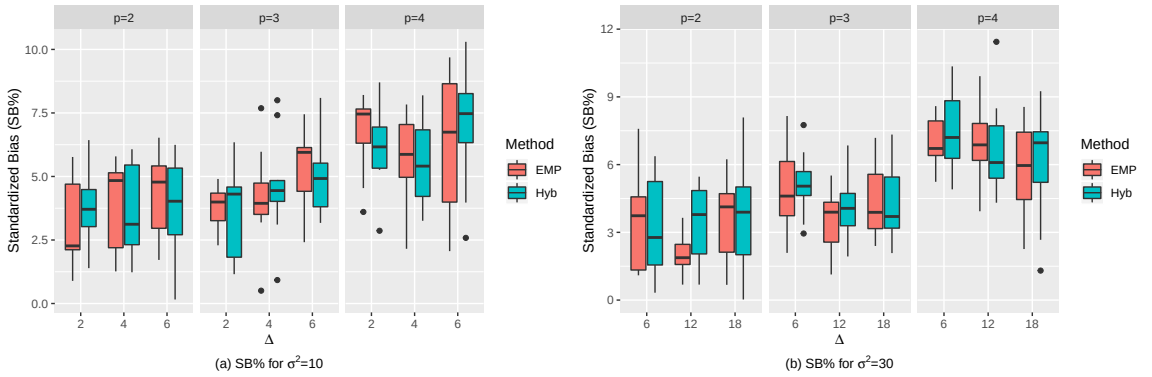



Figure 3.5: Boxplots of SB% of different methods for different $p$, $\boldsymbol{\Delta}$ and $\sigma^2$. MMV is for the moment-based method and EMV is for the MLE via EM algorithm.

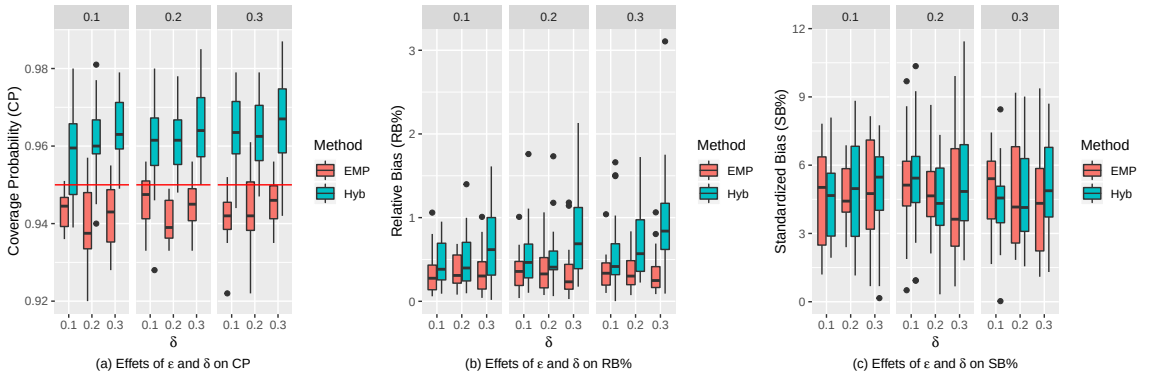



Figure 3.6: Boxplots of CP, RB%, SB% of different methods on different $\epsilon$ and $\delta$. MMV is for the moment-based method and EMV is for the MLE via EM algorithm.

# Chapter 4 Nonparametric Finite Mixture: Applications in Contaminated Trials

## 4.1 Introduction

Randomized clinical trials are commonly used to assess the efficacy and safety of a treatment. Sometimes, subjects with different conditions may respond differently to the treatment. Biomarkers, classifiers, diagnostic devices or instruments may be used at the recruitment stage to separate the sample population into different groups. Such screening tools usually do not have perfect accuracy, and their misclassification rates (false positive and false negative rates) are unknown or assumed to be zero. This leads to contamination in separating the sample populations and results in biased estimation of the treatment effect.

The issue of misclassfication in pre-stratified clinical trials has become prominent in this new era of personalized medicine and measurement-based care. US Food and Drug Administration (FDA) published a concept paper (Hinman et al., 2006) that recommends the co-development of drug and diagnostic tools. It suggests that clinical test validation (i.e. the ability of a test to classify subjects correctly) and clinical utility (i.e., the ability of a test to result in classification that will improve the benefit/reduce the risk of a drug) be established in a pre-clinical pilot feasibility study. One way to achieve this is by using a pre-stratified randomized placebo-controlled design, or in a pre-stratified pre-post or matched paired design. This chapter will focus on the second design and the methods can also be adapted for the first type of design.

Under the assumption of normality, the problem above can be solved by regarding the true status of the subjects as missing information and EM algorithm can be used to find the maximum likelihood estimators of parameters (Harrar et al., 2016). In the absence of normality, especially when the data are not measured in metric scale, or when data have heavy tails or are skewed, nonparametric methods are preferred. This chapter develops a fully nonparametric method for estimation and testing of treatment effect when the classifiers used for stratifying participants are fallible.

**Nonparametric Finite Mixture**

Accurate estimations of the misclassification rates of the classifiers are required to evaluate the effect of a treatment. This problem can be regarded as estimating mixing proportions in the nonparametric mixture models. The most general model for nonparametric multivariate mixtures is as follows: let $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ be i.i.d random variables from a finite mixture of $m > 1$ arbitrary distributions. Suppose the cumulative distribution function of $\boldsymbol{X}_i$ is

$$F = \sum_{j=1}^{m} \lambda_j F_j, \tag{4.1}$$

where $F_j$ and $\lambda_j$ are the cumulative distribution function and mixture proportion (mixing probability) of $j$th component, respectively. Obviously, model (4.1) is not identifiable and some restrictions need to be placed. Hall and Zhou (2003) introduced the conditional independence assumption. That is, the $k$ variables in $\boldsymbol{X}_i$ are independent, and each component distribution $F_j$ has a Lebesgue density $f_j$. Therefore, model (4.1) can be equivalently expressed as

$$f(\boldsymbol{x}_i) = \sum_{j=1}^{m} \lambda_j \prod_{l=1}^{k} f_{jl}(x_{il}). \tag{4.2}$$

The authors established that when $m = 2$ and $k \geq 3$, the model (4.2) is identifiable. Later on, Allman et al. (2009) establishied the identifiability of model (4.2) when $k \geq 3$, regardless of $m$. Many estimation methods have been developed under this condition. Benaglia et al. (2009) proposed estimators based on a EM-like algorithm and Levine et al. (2011) discussed some strategies of selecting the bandwidth for this algorithm. Zheng and Wu (2020) proposed an estimation method by constructing a suitable set of basis fuctions and recovering the coordinates of the component denisty functions with respect to the basis functions. Chauveau and Hoang (2016) developed an EM-like algorithms by modifying the independence assumption (4.2) to blockwise independence, but the number of blocks is still assumed to be greater than 3. In the pre-post design, measurements are taken twice for each subject, i.e. $k_2$. In this case, model (4.2) is not identifiable, and more restrictions are required. In the univariate case, Bordes et al. (2006) and Hunter et al. (2007) both proposed estimators when the component distribution belongs to location family and symmetric about zero. These assumptions are too restrictive for modeling outcomes.

Another way to overcome identifiability issue is to obtain training samples from each component distribution (Hall, 1981). More specifically, suppose we have training samples from each of the component distributions $F_1, \ldots, F_m$. Then the mixing proportions $\lambda_j$ can by estimated by minimizing the distance between empirical distribution functions, i.e.

$$\Delta(\boldsymbol{\lambda}) = \left| \int_{-\infty}^{\infty} \delta \left( \widehat{F}(x) - \sum_{j=1}^{m} \lambda_j \widehat{F}_j(x) \right) w(x) dx \right|, \tag{4.3}$$

where $\widehat{F}$ and $\widehat{F}_j$ are empirical versions of $F$ and $F_j$, respectively. The primary focus of Hall (1981) is when $\delta(x) = x^2$ and $w(x) \equiv 1$. They also assume

$$\int_{-\infty}^{\infty} |x|^{1+\epsilon} dF(x) < \infty, \tag{4.4}$$

for some $\epsilon > 0$.

This requirement imposes a restriction on the tail of the distribution. Especially, it requires that the first moment of the component distributions to exist. There are also other works that developed methods using density funtions and different distance function $\delta(x)$ to estimate mixing proportions. Titterington (1983) considered minimum distance estimators using density estimators. Qin (1999) established an empirical likelihood ratio based confidence intervals assuming the log-likelihood ratio of two component densities is linear in observations. Karunamuni and Wu (2009) proposed an estimator of mixture proportion by minimize Hellinger distance. These methods are not applicable when the distributions are not absolutely continuous (i.e. when they do not have Lebesgue density).

In the present chapter, we will assume there exists more expensive but infallible classifiers such that we can obtain a training data to estimate the mixing proportions. To make the results applicable to different types of data (discrete, continuous, binary and ordered categorical), we will use normalized cumulative functions and obtain estimators via (4.3). The normalized distribution function $F$ of a random variable $X$ is defined by $F(x) = \{P(X \leq x) + P(X < x)\}/2 = \{F^+(x) + F^-(x)\}/2$ where $F^+$ and $F^-$ are the right and left continuous, respectively, versions of the distribution functions of $X$.

### Nonparametric Relative Effects

In the absence of misclassification errors, the design we are interested in generates repeated measures (dependent) data in two groups. Nonparametric methods for dependent data have been developed in a body of literature spanning over four decades. One of the earlier approaches is the rank-based methods by Brunner and Neumann (1982), which was later generalized by Thompson (1990, 1991) for continuous data. Most of the older nonparametric methods were motivated by replacing original observations by ranks, known as rank transformation, in parametric methods to gain robustness.

In a series of papers (Akritas, 1990, 1991, 1992) the application of rank transformation is inadequate for testing some of the hypotheses in factorial desgins. This limitation motivated the development of fully nonparametric methods (Akritas and Arnold, 1994; Brunner et al., 1997; Akritas and Brunner, 1997; Brunner et al., 1999), where hypotheses are formulated in terms of marginal distributions. The (mid-)rank based procedures arise naturally as a consequence of estimating the distributions with their empirical versions. The nonparametric hypotheses have the limitation that the alternative hypotheses are generally difficult to interpret. To overcome this problem, Brunner and Munzel (2000); Konietschke et al. (2012) and Brunner et al. (2017) proposed to formulate hypotheses in terms of the so-called nonparametric relative effects. These purely nonparametric effect measures allow construction of confidence intervals. They also address the nonparametric Behrens-Fisher problem (Brunner and Munzel, 2000) in the sense that the joint and marginal distributions of the data in the various groups could still be different under the null hypothesis. In the simplest case of two independent groups, the nonparametric effects reduce to the Wilcoxon-Mann-Whitney effect (Wilcoxon, 1945; Mann and Whitney, 1947).

The asymptotic theory for the estimators and tests are generally tractable with the use of asymptotic equivalence and central limit theorem, except that the derivation of the asymptotic variance is cumbersome.

In the situation where the diagnostic tool used to stratify participants is subject to classification error, the correct nonparametric marginal model and the associated relative effect of interest require the use of nonparametric finite mixtures. The marginal distributions

of the pre and post measurements are two-component mixture distribution in both groups where the mixing probabilities are the missclassification error rates. In this design, the identifiability problem is overcome by acquisition of data from correctly classified participants, which hereinafter will be referred to as *validation* or *training* data.

Unlike existing nonparametric methods for estimation and testing of relative treatment effects, the theory for contaminated samples situation involves estimation of the mixing probabilities in addition to the relative effects, where the later depends on the former. In our approach, the mixing probabilities are estimated by minimizing the disagreement (4.3) between the estimate of the mixture distribution with the original (*contaminated*) data and that obtained by estimating the components of the mixture separately using the validation or training data. The requirement (4.4), which excludes heavy tailed distributions such as Cauchy, is rather restrictive for our application. To remove this assumption, we choose the weight function $w(x)$ to be the weighted average of the marginal empirical distribution functions.

The remainder of the chapter is organized as follows. In Section 2, we describe the statistical model and treatment effect measure. Estimation, asymptotic theory and test procedures for mixing probabilities are the subjects of Section 3. Section 4 provides estimation, asymptotic theory and testing procedures for the treatment effect. Simulation results designed to show the finite sample performance of the inferential procedures for mixing probabilities and treatment effect under various practical scenarios are presented in Section 5. In Section 6, we illustrate the application of the results using data from a sleep deprivation study. Discussions and conclusions are provided in Section 7. The Appendix contains all proofs and additional simulation results.

## 4.2 Model and Effect Size Measure

Suppose we have subjects from two groups $g = 1, 2$ that are observed at two (pre- and post-treatment) time points $t = 1, 2$. Among the subjects in each group, some of them are classified into the group by a fallible classifier, and the remaining are classified by an infallible classifier. Suppose the misclassification error rates for the fallible classifier is $\delta_g$ in group $g$. Denote the paired observations from subjects classified by the fallible classifier

as $\boldsymbol{X}_{1gk} = (X_{1g1k}, X_{1g2k})$ for $g = 1, 2$ and $k = 1, 2, \ldots, n_{1g}$, and those by the infallible classifier as $\boldsymbol{X}_{2gk} = (X_{2g1k}, X_{2g2k})$ for $g = 1, 2$ and $k = 1, 2, \ldots, n_{2g}$. We refer to these two sources of data as *contaminated* and *training (validation)* data, respectively. Let $X_{hgt1}, \ldots X_{hgtn_{hg}}$ be identically and independently distributed according to $F_{hgt}$, assumed to be nondegenerate for $h = 1, 2$, $g = 1, 2$ and $t = 1, 2$. To accommodate binary, ordered categorical, discrete and continuous data in a unified manner, $F_{hgt}$ are taken to be the normalized distribution functions defined by

$$F_{hgt}(x) := \frac{1}{2}\{F^{+}_{hgt}(x) + F^{-}_{hgt}(x)\},$$

where $F^{-}_{hgt}(x) = P(X_{hgt1} < x)$ and $F^{+}_{hgt}(x) = P(X_{hgt1} \leq x)$ are the left and right continuous, respectively, versions of the distribution function.

The distributions $F_{2gt}$s are for observations from subjects classified by infallible classifiers. Therefore, $F_{1gt}$ is a mixture of $F_{21t}$ and $F_{22t}$, mixed in proportions determined $\delta_g$, i.e.,

$$F_{1gt} = (1 - \delta_g)F_{2gt} + \delta_g F_{2g't}, \tag{4.5}$$

for $g' \neq g$, $g', g, t = 1, 2$. For the sake of convenience, we express $F_{2gt}$ in terms of $F_{11t}$ and $F_{12t}$ as,

$$F_{2gt} = F_{1gt} + \frac{\delta_g}{1 - \delta_1 - \delta_2}\left(F_{1gt} - F_{1g't}\right). \tag{4.6}$$

Equation (4.6) suggests that validation (training data) would not be needed if $\delta_g$ are known for $g = 1, 2$ as the treatment effect can be meaningfully assessed from estimation of $F_{1gt}$ and using this equation.

Nonparametric relative effects are defined by comparing each marginal distribution function with the average distribution function. Let $G$ denote the average of the distribution functions in the two groups and at the two time points defined by

$$\begin{aligned} G :&= \frac{1}{4}(F_{211} + F_{212} + F_{221} + F_{222}) \\ &= \frac{1}{4}\left(F_{111} + F_{112} + F_{121} + F_{122} + \frac{\delta_1 - \delta_2}{1 - \delta_1 - \delta_2}(F_{111} + F_{112} - F_{121} - F_{122})\right). \end{aligned}$$

Using the average distribution function $G$, define

$$p_{gt} = \int G dF_{2gt} = 1 - \int F_{2gt} dG.$$

for $g, t = 1, 2$ which is known as the *nonparametric effect* at time point $t$ and in group $g$ *relative* to the average of the marginal distributions, $G$. The magnitude of $p_{gt}$ has interpretation in terms of the corresponding marginal distribution having a tendency to generate larger or smaller values compared to the overall sample.

Using the nonparametric relative effects, the treatment effect in the two group pre-post design is defined by

$$p_I := (p_{12} - p_{11}) - (p_{22} - p_{21}). \tag{4.7}$$

To provide an intuitive meaning of $p_{\mathrm{I}}$, let $F_{2gt}$ be the distribution function of $N(\mu_{gt}, \sigma^2)$. In this case, $p_I = 0$ if and only if $(\mu_{12} - \mu_{11}) - (\mu_{22} - \mu_{21}) = 0$, which is equivalent to lack of interaction in the two group and two time point repeated measures design. Inferential methods for this design were investigated in the purely nonparametric and second-order semi-parametric contexts in Harrar et al. (2020) and Xu and Harrar (2012), respectively, but for the situation where group membership of participants can be determined without errors, i.e. $\delta_g = 0$ for $g = 1, 2$ and training data is not necessary.

Our main objective is to investigate the estimation of the nonparametric treatment effect size $p_{\mathrm{I}}$ and study its asymptotic properties that includes the asymptotic distributions. Along the way, we derive asymptotic properties for the estimators of the mixing proportions that are not only needed for the asymptotic properties of the effect size estimators but also are of interest in their own right. The asymptotic distributions of the effect size estimators will be used to develop confidence intervals and significance tests.

## 4.3 Inference on Mixing Proportions

### Estimation

To relax the requirement in (4.4), we set $\omega(x)dx = d\widehat{H}(x)$, where $\widehat{H}(x)$ is weighted average of the empirical marginal distribution functions of observations in the two groups, at

the two time points and for the two classifiers, i.e.,

$$\widehat{H} = \sum_{h=1}^{2}\sum_{g=1}^{2}\sum_{t=1}^{2} \frac{n_{hg}}{N}\widehat{F}_{hgt}, \tag{4.8}$$

where $N = 2\sum_{h=1}^{2}\sum_{g=1}^{2} n_{hg}$ and $\widehat{F}_{hgt}(x)$ is the normalized empirical counterpart of $F_{hgt}$ defined by

$$\widehat{F}_{hgt}(x) = \frac{1}{n_{hg}}\sum_{k=1}^{n_{hg}} c(x - X_{hgtk}), \text{ where } c(x) = \begin{cases} 0, & x < 0 \\ \frac{1}{2}, & x = 0 \\ 1, & x > 0 \end{cases}. \tag{4.9}$$

While this choice of the weight function is natural, it is also used to establish the consistency of the estimators of $\delta_1$ and $\delta_2$ stated in Proposition 4.3.1.

We propose to estimate $\delta_1$ and $\delta_2$ by minimizing the distance between the empirical version of left and right hand sides of (4.5). Setting $\boldsymbol{\theta} = (\delta_1, \delta_2)$, we use the distance function

$$\Delta(\boldsymbol{\theta}) = \sum_{g=1}^{2}\sum_{t=1}^{2}\int_{-\infty}^{\infty}\left(\widehat{F}_{1gt} - [(1-\delta_g)\widehat{F}_{2gt} + \delta_g\widehat{F}_{2g't}]\right)^2 d\widehat{H},$$

for $g' \neq g$ and $g' = 1, 2$.

Taking partial derivatives of $\Delta(\boldsymbol{\theta})$ with respect to $\delta_1$ and $\delta_2$ and setting them equal to 0, we have

$$\delta_g \sum_{t=1}^{2}\int (\widehat{F}_{2gt} - \widehat{F}_{2g't})^2 d\widehat{H} = \sum_{t=1}^{2}\int (\widehat{F}_{1gt} - \widehat{F}_{2gt})(\widehat{F}_{2g't} - \widehat{F}_{2gt}) d\widehat{H}.$$

In order to estimate the mixing proportions, we need to assume the component distributions are not very close to each other.

**Assumption 4.3.1.** *There exist a constant $C > 0$ such that*

$$\int \left\{(F_{221} - F_{211})^2 + (F_{222} - F_{212})^2\right\} dH > C,$$

*where $H = N^{-1}\sum_{h=1} 2\sum_{g=1}^{2}\sum_{t=1}^{2} n_{hg}F_{hgt}$.*

Under Assumption 4.3.2, the probability that the normalized empirical distribution in the two groups differ will approach 1 as the sample sizes get large. That is,

$$P\left(\sum_{t=1}^{2}\int (\widehat{F}_{21t} - \widehat{F}_{22t})^2 d\widehat{H} \neq 0\right) \to 1, \text{ as } N \to \infty.$$

Therefore, under Assumption 4.3.2, $\Delta(\boldsymbol{\theta})$ is a convex function of $\delta_g$, $g = 1, 2$, and its minimum is attained at

$$\widehat{\delta}_g = \max\left\{\frac{\sum_{t=1}^{2}\int(\widehat{F}_{1gt} - \widehat{F}_{2gt})(\widehat{F}_{2g't} - \widehat{F}_{2gt})d\widehat{H}}{\sum_{t=1}^{2}\int(\widehat{F}_{21t} - \widehat{F}_{22t})^2 d\widehat{H}}, 0\right\}, \tag{4.10}$$

for $g = 1, 2$.

To establish the consistency of $\widehat{\delta}_g$, we need a standard proportional divergence requirements on the group sample sizes.

**Assumption 4.3.2.** *There exist nonnegative numbers $M_0$ and $N_0$ such that $0 < M_0 \leq \min\{\frac{N}{n_{hg}}, h = 1, 2, g = 1, 2\} \leq \max\{\frac{N}{n_{hg}}, h = 1, 2, g = 1, 2\} \leq N_0 < \infty$.*

We also need to assume that $\delta_g$ is bounded away from 0 and $\frac{1}{2}$.

**Assumption 4.3.3.** *There exists a constant $0 < c_0 < c_1 < 1/2$, such that*

$$0 < c_0 \leq \delta_1 \leq c_1 < \frac{1}{2} \quad \textit{and} \quad 0 < c_0 \leq \delta_2 \leq c_1 < \frac{1}{2}.$$

Under Assumptions 4.3.2 and 4.3.3, $\widehat{\delta}_g$ is consistent estimators for $\delta_g$ for $g = 1, 2$. This is proved next in Proposition 4.3.1.

**Proposition 4.3.1.** *Let $X_{hgtk} \sim F_{hgt}$, $g = 1, 2$, $t = 1, 2$, $k = 1, \ldots, n_{hg}$. Further, let $\widehat{F}_{hgt}(x)$ denote the standard empirical distribution functions of $F_{hgt}(x)$. Let $\widehat{\delta}_g$ be defined as in (4.10) for $g = 1, 2$. Then, under Assumptions 4.3.1, 4.3.2, and 4.3.3, we have*

$$\widehat{\delta}_g \xrightarrow{P} \delta_g,$$

*for $g = 1, 2$.*

Although $\widehat{\delta}_g$ is consistent estimator of $\delta_g$, $g = 1, 2$, it is a biased estimator in finite samples. Proposition 4.3.2 establishes that the bias is of order $N^{-1}$.

**Proposition 4.3.2.** *Let $\widehat{\delta}_g$ be as defined in (4.10). Under Assumptions 4.3.1, 4.3.2, and 4.3.3,*

$$E(\widehat{\delta}_g) - \delta_g = O(N^{-1}),$$

*for $g = 1, 2$.*

### Asymptotic Distribution

To simplify the expression for the asymptotic variance, we assume that the ratios of the total sample size ($N$) to individual sample sizes ($n_{hg}$) are fixed as $N \to \infty$.

**Assumption 4.3.4.** $\frac{N}{n_{hg}} \to \kappa_{hg} > 0$. *as* $N \to \infty$, *for any* $h, g = 1, 2$.

We can approximate the distribution of $\widehat{\delta}_\ell$ by sums of independent random variables and derive the asymptotic distribution as follows.

**Theorem 4.3.1.** *Let* $\widehat{\delta}_\ell$, $\ell = 1, 2$, *be as defined in (4.10). Under Assumptions 4.3.1, 4.3.2, 4.3.3, and 4.3.4,*

$$\sqrt{N}(\widehat{\delta}_\ell - \delta_\ell) \xrightarrow{D} U \sim N\left(0, \kappa_{11}\sigma^2_{\ell 11} + \kappa_{12}\sigma^2_{\ell 12} + \kappa_{21}\sigma^2_{\ell 21} + \kappa_{22}\sigma^2_{\ell 22}\right), \tag{4.11}$$

*where*

$$\sigma^2_{\ell hg} = Var(V_{\ell hg}(\boldsymbol{X}_{hg1})),$$

*and* $V_{\ell hg}(\boldsymbol{X}_{hg1})$ *is defined in (4.38) for* $\ell, g, h = 1, 2$.

The quantity $\sigma^2_{\ell hg}$ can be regarded as variance of functions of observations in contaminated ($h = 1$) or validation ($h = 2$) data of group $g$ for $\widehat{\delta}_\ell$. From (4.11), the variance of $\widehat{\delta}_\ell$ is a weighted average of variances from both contaminated and validation data in the two groups. As one would expect and (4.10) reveals, estimation error of $\widehat{\delta}_\ell$ involves observations in in contaminated data of group $\ell'$, where $\ell' \neq \ell$. This is guarateed by the weight function $\widehat{H}$, which is defined as the weighted average of empirical distributions from the two data sets.

From the Weak Law of Large Numbers,

$$\widehat{\sigma}^2_{\ell hg} - \sigma^2_{\ell hg} \xrightarrow{P} 0,$$

for $\ell, g, h = 1, 2$, where

$$\begin{aligned}\widehat{\sigma}^2_{\ell hg} &= \frac{1}{n_{hg} - 1}\sum_{i=1}^{n_{hg}} \left(V_{\ell hg}(\boldsymbol{X}_{hgi}) - \overline{V}_{\ell hg}(\boldsymbol{X}_{hg\cdot})\right)^2 \text{ and } \overline{V}_{\ell hg}(\boldsymbol{X}_{hg\cdot}) \\ &= \frac{1}{n_{hg}}\sum_{i=1}^{n_{hg}} V_{\ell hg}(\boldsymbol{X}_{hgi}).\end{aligned}$$

The quantity $\widehat{\sigma}_{\ell hg}$ cannot be directly used to estimate $\sigma_{\ell hg}$ in real applications because $V_{\ell hg}(\cdot)$ is not an observable function. However, it can be shown that a consistent estimator can be obtained by replacing these functions with their analogs defined in terms of the empirical distribution functions. To that end, let $\widehat{V}_{\ell hg}(\cdot)$ be defined analogous to $V_{\ell hg}(\cdot)$, by replacing $F_{hgt}$ with $\widehat{F}_{hgt}$, for all $h, g, t = 1, 2$. Further, define

$$S^2_{\ell hg} = \frac{1}{n_{hg} - 1} \sum_{i=1}^{n_{hg}} \left( \widehat{V}_{\ell hg}(X_{hgi}) - \overline{\widehat{V}}_{\ell hg}(X_{hg\cdot}) \right)^2 \text{ and}$$

$$\overline{\widehat{V}}_{\ell hg}(X_{hg\cdot}) = \frac{1}{n_{hg}} \sum_{i=1}^{n_{hg}} \widehat{V}_{\ell hg}(X_{hgi}).$$

Obviously, the proof of $S^2_{\ell hg}$ being a consistent estimator for $\sigma_{\ell hg}$ will be complete if we can prove that $S^2_{\ell hg} - \widehat{\sigma}^2_{\ell hg} \xrightarrow{P} 0$.

**Theorem 4.3.2.** *Assume that $\sigma^2_{\ell hg} > 0$, $\ell, g, h = 1, 2$. Under Assumptions 4.3.1, 4.3.2, and 4.3.3, $S^2_{\ell hg}$ is a consistent estimator of $\sigma^2_{\ell hg}$ for $\ell, g, h = 1, 2$.*

### Test Procedure and Confidence Interval

The asymptotic theory in Section 4.3 can be used to develop methods for confidence intervals and significance test based on the estimator $\widehat{\delta}_\ell$. The hypothesis of interest is if the misclassification rates is greater than a known value $\delta_{\ell,0}$, i.e.,

$$H_0 : \delta_\ell = \delta_{\ell,0} \quad \text{vs} \quad H_a : \delta_\ell > \delta_{\ell,0}.$$

Under Assumption 4.3.1 and assuming that $\sigma^2_{\ell hg} > 0$, $\ell, h, g = 1, 2$, we have

$$T_{\mathrm{M}} = \sqrt{N} \frac{\widehat{\delta}_\ell - \delta_\ell}{\sqrt{S^2_\ell}} \xrightarrow{D} Z \sim N(0, 1),$$

where $S^2_\ell = \frac{N}{n_{11}} S^2_{\ell 11} + \frac{N}{n_{12}} S^2_{\ell 12} + \frac{N}{n_{21}} S^2_{\ell 21} + \frac{N}{n_{22}} S^2_{\ell 22}$. Taking $\delta_\ell = \delta_{\ell,0}$, we can use $T_M$ as a viable test statistic. Further, $(1 - \alpha)100\%$ asymptotic confidence interval for $\delta_\ell$ can be obtained from

$$P\left( \widehat{\delta}_\ell - \frac{z_{\alpha/2}\sqrt{S^2_\ell}}{\sqrt{N}} \leq \delta_\ell \leq \widehat{\delta}_\ell + \frac{z_{\alpha/2}\sqrt{S^2_\ell}}{\sqrt{N}} \right) \to 1 - \alpha, \tag{4.12}$$

where $z_{\alpha/2}$ denotes the $(1 - \alpha/2)$th-quantile of the standard normal distribution.

## 4.4 Estimation and Test on Effect Size

### Estimation

According to the calculations in Harrar et al. (2020),

$$p_{\mathrm{I}} = \frac{1}{2}\int (F_{211} + F_{222}) d(F_{212} + F_{221}) - 1. \tag{4.13}$$

Applying (4.6), the treatment effect $p_{\mathrm{I}}$ can be expressed in terms of $F_{1gt}$, $g, t = 1, 2$, as

$$\begin{aligned} p_I =& \frac{1}{2(1-\delta_1-\delta_2)}\int (F_{111} + F_{122}) d(F_{112} + F_{121}) \\ &+ \frac{\delta_1 - \delta_2}{2(1-\delta_1-\delta_2)^2}\int (F_{111} - F_{121}) d(F_{112} - F_{122}) - \frac{1}{1-\delta_1-\delta_2}. \end{aligned} \tag{4.14}$$

The details are shown in the Appendix. To see the effects of misclassification errors, we express (4.14) as

$$p_I = \frac{1}{1-\delta_1-\delta_2} p_I^* + \frac{\delta_1 - \delta_2}{2(1-\delta_1-\delta_2)^2}\int (F_{111} - F_{121}) d(F_{112} - F_{122}), \tag{4.15}$$

where $p_I^* = \frac{1}{2}\int (F_{111} + F_{122}) d(F_{112} + F_{121}) - 1$ is the treatment effect by traditional method ignoring the misclassification errors. Clearly, one will end up introducing bias by ignoring the errors. The estimates will still be biased even if the misclassifications are balanced in the two groups, i.e. even if $\delta_1 = \delta_2$.

Two separate estimators of the treatment effect $p_I$ can be constructed from the two sources of data. Using the contaminated data, we can plug-in the empirical versions of $F_{1gt}$, $g, t = 1, 2$ in (4.14) to estimate the treatment effect by

$$\begin{aligned} \widehat{p}_{I1} =& \frac{1}{2(1-\widehat{\delta}_1-\widehat{\delta}_2)}\int (\widehat{F}_{111} + \widehat{F}_{122}) d(\widehat{F}_{112} + \widehat{F}_{121}) \\ &+ \frac{\widehat{\delta}_1 - \widehat{\delta}_2}{2(1-\widehat{\delta}_1-\widehat{\delta}_2)^2}\int (\widehat{F}_{111} - \widehat{F}_{121}) d(\widehat{F}_{112} - \widehat{F}_{122}) - \frac{1}{1-\widehat{\delta}_1-\widehat{\delta}_2}. \end{aligned} \tag{4.16}$$

On the other hand, we can also estimate $p_I$ by using the training (validation) data to get empirical version of $F_{2gt}$, $g, t = 1, 2$, and plugging the later in (4.13) as

$$\widehat{p}_{I2} = \int \frac{1}{2}(\widehat{F}_{211} + \widehat{F}_{222}) d(\widehat{F}_{212} + \widehat{F}_{221}) - 1. \tag{4.17}$$

Combining (4.16) and (4.17), we propose to estimate $p_I$ by a weighted average of $\widehat{p}_{I1}$ and $\widehat{p}_{I2}$ as

$$\widehat{p}_I = \frac{2(n_{11}+n_{12})}{N}\widehat{p}_{I1} + \frac{2(n_{21}+n_{22})}{N}\widehat{p}_{I2}. \tag{4.18}$$

The consistency of the estimator $\widehat{p}_{\mathrm{I}}$ is proved in Proposition 4.4.1.

**Proposition 4.4.1.** *Under Assumption 4.3.1, 4.3.2, and 4.3.3, we have*

$$\widehat{p}_I \xrightarrow{P} p_I.$$

The estimator $\widehat{p}_I$ is biased for $p_I$. However, the bias vanishes at the rate of $N^{-1}$ under Assumptions 4.3.1, 4.3.2 and 4.3.3.

**Proposition 4.4.2.** *Let $p_I$ and $\widehat{p}_I$ be defined as in (4.7) and (4.18). Under Assumptions 4.3.1, 4.3.2 and 4.3.3, we have*

$$E(\widehat{p}_I) - p_I = O(N^{-1}).$$

### Asymptotic Distribution

The asymptotic theory in purely nonparametric methods generally involves the so-called Asymptotic Equivalence Theorem (Brunner and Munzel, 2000; Brunner et al., 2017; Harrar et al., 2020), where the difference between estimated and true effect size is decomposed in to a term amenable to Lindeberg's CLT and an asymptotically negligible term. Such decomposition is not possible in the nonparametric finite mixture situation. The asymptotic distribution of $\widehat{p}_{\mathrm{I}}$ in the later case is given in Theorem 4.4.1.

**Theorem 4.4.1.** *Let $\widehat{p}_I$ be defined in (4.18). Under Assumptions 4.3.1, 4.3.2, 4.3.3, and 4.3.4,*

$$\sqrt{N}(\widehat{p}_I - p_I) \xrightarrow{D} U \sim N\left(0, \kappa_{11}\sigma_{11}^2 + \kappa_{12}\sigma_{12}^2 + \kappa_{21}\sigma_{21}^2 + \kappa_{22}\sigma_{22}^2\right),$$

*where*

$$\sigma_{11}^2 = Var(A_1\left(V_1(\boldsymbol{X}_{11k}) + U_{11}(\boldsymbol{X}_{11k})\right)), \qquad \sigma_{12}^2 = Var\left(A_1\left(V_2(\boldsymbol{X}_{12k}) + U_{12}(\boldsymbol{X}_{12k})\right)\right),$$
$$\sigma_{21}^2 = Var(A_2W_1(\boldsymbol{X}_{21k}) + A_1U_{21}(\boldsymbol{X}_{21k})), \qquad \sigma_{22}^2 = Var(A_2W_2(\boldsymbol{X}_{22k}) + A_1U_{22}(\boldsymbol{X}_{22k})),$$

$A_1 = 2N^{-1}(n_{11} + n_{12})$ *and* $A_2 = 2N^{-1}(n_{21} + n_{22})$. *The functions* $V_g$, $W_g$ *and* $U_{hg}$ *are defined in (4.46), (4.44) and (4.47), respectively, for* $g, h = 1, 2$.

Similar to Theorem 4.3.1, the variance of $\widehat{p}_{\mathrm{I}}$ is a weighted average of the variances $\sigma^2_{hg}$, where $\sigma^2_{hg}$ can be regarded as variances of functions of observations from contaminated ($h = 1$) or validation ($h = 2$) data of group $g$. It is interesting to note that $\widehat{p}_{\mathrm{I}}$, defined in (4.18), is a weighted average of $\widehat{p}_{\mathrm{I}1}$ and $\widehat{p}_{\mathrm{I}2}$, terms that are not independent in finite samples. However, its variance turns out to be a linear combination of the variances of the two terms.

Since $U_{hg}(\cdot)$, $V_g(\cdot)$ and $W_g(\cdot)$ are not observable functions, we use their empirical version $\widehat{U}_{hg}(\cdot)$, $\widehat{V}_g(\cdot)$ and $\widehat{W}_g(\cdot)$, respectively, to estimate them. The components of the asymptotic variance can be estimated by

$$
\begin{aligned}
S^2_{11} &= \frac{1}{n_{11}-1}\sum_{k=1}^{n_{11}} A_1^2\left(\widehat{V}_1(\boldsymbol{X}_{11k}) + \widehat{U}_{11}(\boldsymbol{X}_{11k}) - \overline{\widehat{V}}_1(\boldsymbol{X}_{11\cdot}) - \overline{\widehat{U}}_{11}(\boldsymbol{X}_{11\cdot})\right)^2,\\
S^2_{12} &= \frac{1}{n_{12}-1}\sum_{k=1}^{n_{12}} A_1^2\left(\widehat{V}_2(\boldsymbol{X}_{12k}) + \widehat{U}_{12}(\boldsymbol{X}_{12k}) - \overline{\widehat{V}}_2(\boldsymbol{X}_{12\cdot}) - \overline{\widehat{U}}_{12}(\boldsymbol{X}_{12\cdot})\right)^2,\\
S^2_{21} &= \frac{1}{n_{21}-1}\sum_{k=1}^{n_{21}} \left(A_2\widehat{W}_1(\boldsymbol{X}_{21k}) + A_1\widehat{U}_{21}(\boldsymbol{X}_{21k}) - A_2\overline{\widehat{W}}_1(\boldsymbol{X}_{21\cdot}) - A_1\overline{\widehat{U}}_{21}(\boldsymbol{X}_{21\cdot})\right)^2,\\
&\text{and} \qquad (4.19)\\
S^2_{22} &= \frac{1}{n_{22}-1}\sum_{k=1}^{n_{22}} \left(A_2\widehat{W}_2(\boldsymbol{X}_{22k}) + A_1\widehat{U}_{22}(\boldsymbol{X}_{22k}) - A_2\overline{\widehat{W}}_2(\boldsymbol{X}_{22k}) - A_1\overline{\widehat{U}}_{22}(\boldsymbol{X}_{22\cdot})\right)^2,
\end{aligned}
$$

where $\overline{\widehat{V}}(X_{1\cdot}) = \frac{1}{n_1}\sum_{k=1}^{n_1}\widehat{V}(X_{1k})$ and $\overline{\widehat{W}}(X_{2\cdot}) = \frac{1}{n_2}\sum_{k=1}^{n_2}\widehat{W}(X_{2k})$. The consistency of the estimators in (4.19) is established in Theorem 4.4.2.

**Theorem 4.4.2.** *Under Assumptions 4.3.1, 4.3.2, and 4.3.3 and assume that* $\sigma^2_{hg} > 0$, $h, g = 1, 2$ *then* $S^2_{hg}$ *are consistent estimators of* $\sigma^2_{hg}$, *respectively.*

Note that when the mixing proportions $\delta_\ell$, $\ell = 1, 2$, are known, one can estimate the treatment effect $p_I$ using (4.14) by estimating the distribution functions from the contaminated data alone. In this case, validation data is not necessary and the estimator would

be

$$\widetilde{p}_{I1} = \frac{1}{2(1-\delta_1-\delta_2)} \int (\widehat{F}_{111} + \widehat{F}_{122}) d(\widehat{F}_{112} + \widehat{F}_{121}) \tag{4.20}$$
$$+ \frac{\delta_1 - \delta_2}{2(1-\delta_1-\delta_2)^2} \int (\widehat{F}_{111} - \widehat{F}_{121}) d(\widehat{F}_{112} - \widehat{F}_{122}) - \frac{1}{1-\delta_1-\delta_2}.$$

In proof of Theorem 4.4.1, it is established that

$$\sqrt{N}(\widetilde{p}_{I1} - p_I) \xrightarrow{D} U \sim N\left(0, \kappa_{11}\sigma_1^2 + \kappa_{12}\sigma_2^2\right), \tag{4.21}$$

where $\sigma_i^2 = Var(V_i(\boldsymbol{X}_{1i1}))$ and $V_i(\boldsymbol{X}_{1i1})$ is defined in (4.46) for $i = 1, 2$.

### Test Procedure and Confidence Interval

We can develop confidence interval and significance test procedures for the treatment effect $p_{\mathrm{I}}$ based on the estimator $\widehat{p}_I$ and its asymptotic distribution. The null hypothesis of interest is that of no treatment effect, i.e.

$$H_0 : p_I = 0 \quad \text{vs} \quad H_a : p_I \neq 0. \tag{4.22}$$

Under Assumption 4.3.1 and assuming that $\sigma_1^2 > 0$ and $\sigma_2^2 > 0$, we have

$$T_M = \frac{\sqrt{N}(\widehat{p}_I - p_I)}{\sqrt{S_I^2}} \xrightarrow{D} Z \sim N(0, 1).$$

where $S_I^2 = \frac{N}{n_{11}} S_{11}^2 + \frac{N}{n_{12}} S_{12}^2 + \frac{N}{n_{21}} S_{21}^2 + \frac{N}{n_{22}} S_{22}^2$. The quantity $T_M$ can serve as a test statistic for the hypotheses in (4.22) by taking $p_{\mathrm{I}} = 0$. An asymptotic $(1-\alpha)100\%$ confidence interval for $p_I$ can be derived from

$$P\left(\widehat{p}_I - \frac{z_{\alpha/2}\sqrt{S_I^2}}{\sqrt{N}} \leq p_I \leq \widehat{p}_I + \frac{z_{\alpha/2}\sqrt{S_I^2}}{\sqrt{N}}\right) \to 1 - \alpha,$$

where $z_{\alpha/2}$ denotes the $(1-\alpha/2)$th-quantile of the standard normal distribution.

## 4.5 Simulation Study

In this section, we use simulations to investigate the finite-sample accuracy of estimators and the asymptotic results for mixing proportions and treatment effect. We investigate the performance of the proposed methods under various settings for the distribution of the

data, sample size allocations, sample size ratio between validation and contaminated data, within-pair dependence, and the mixing proportions $\delta_1$ and $\delta_2$. In all the simulations, the run size is $10,000$.

We consider both continuous and discrete distributions to generate data. For continuous distribution, data will be generated from normal, Cauchy, and lognormal distributions which represent light-tailed, heavy-tailed and skewed distributions, respectively. The validation datasets will be generated from bivariate Normal, Cauchy or Lognormal distributions, respectively, with group mean vectors $\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12})^\top$ and $\boldsymbol{\mu}_2 = (\mu_{21}, \mu_{22})^\top$, and constant covariance matrix $\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$. The contaminated datasets are generated from mixture of these distributions such that (4.5) is satisfied. For the mean and covariance parameters, we set $\sigma^2 = 1, \mu_{11} = 1, \mu_{12} = 2, \mu_{21} = 3, \mu_{22} = 4$ and $\rho \in \{0, 0.5\}$. We will also study performance by discretizing data from the normal distributions. To investigate performance of the methods for count data, we generate bivariate Poisson data as follows. Note that, if $Z_1$, $Z_2$ and $Z_3$ are independent Poisson random variables with parameters $\lambda_1$, $\lambda_2$ and $\lambda_3$, respectively then $X_1 = Z_1 + Z_3$ and $X_2 = Z_2 + Z_3$ are marginally distributed as Poisson with $\lambda = \lambda_1 + \lambda_3$ and $\lambda_2 + \lambda_3$ and correlation $\rho = \frac{\lambda_3}{\sqrt{(\lambda_1+\lambda_3)(\lambda_2+\lambda_3)}}$. Utilizing this property, we generate bivariate Poisson data with marginal distributions Poisson($\lambda_{ht}$), $h, t = 1, 2$, where $(\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22}) = (1, 2, 3, 4)$ and correlation $\rho \in \{0, 0.5\}$. The effect of sample size is evaluated by considering the allocations $(n_{11}, n_{21}) \in \{(50, 50), (50, 100), (100, 100)\}$ and the validation-contaminated sample ratios, hereinafter referred to as *sample size ratio*, varied in $\{10\%, 30\%, 50\%\}$. To check the effects of low to moderate misclassifications, we consider $\delta_1, \delta_2 \in \{0.01, 0.1, 0.25\}$.

Suppose we are interested in estimating $\theta$ by $\widehat{\theta}$. We will use three criteria to assess the performances of estimators.

1. Bias of the estimator: Bias$= E(\widehat{\theta}) - \theta$.
2. Root mean of square error: RMSE$= \sqrt{E(\widehat{\theta} - \theta)^2}$.
3. Coverage probability (CP): the proportion of 95% confidence intervals that cover the true value of $\theta$.

### Mixing Proportions

For mixing proportions, we compare our proposed estimators with Hall (1981) estimators in terms of Bias and RMSE. We also check the asymptotic results in (4.11) by computing coverage probabilities. Thefore, the method compared are

1. the new estimator (New): the estimator defined in (4.10) with asymptotic distribution in (4.11) and

2. Hall's estimator (Hall): the estimator proposed in Hall (1981)'s where $w(x) \equiv 1$.

### Accuracy of Estimation

Table 4.1 presents the simulation results for $\widehat{\delta}_1$ and $\widehat{\delta}_2$ when $F_{2gt}$, g,t=1,2, are Normal distributions. From these results, we see that the bias and RMSE of $\widehat{\delta}_1$ and $\widehat{\delta}_2$ for both methods are small. The RMSE for the two methods are close but the bias from Hall is a bit higher than the New method. Because the weight function for Hall's method is equal to 1, the estimation of $\delta_1$ is not affected by the change of $\delta_2$, and vice versa. The new methods' weight is determined by the average of all distributions. Thus, a change in $\delta_2$ expectedly affects the estimation of $\delta_1$ but the variation is not very large.

Table 4.1: Bias$(\times 100)$ and RMSE$(\times 100)$ of $\widehat{\delta}_1$ and $\widehat{\delta}_2$ when $\sigma^2 = 1, \rho = 0, n_{11} = 100, n_{12} = 100$, ratio$= 0.5$

| Estimation | | $\widehat{\delta}_1$ | | | | $\widehat{\delta}_2$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | | New | | Hall | | New | | Hall | |
| $\delta_1$ | $\delta_2$ | Bias | RMSE | Bias | RMSE | Bias | RMSE | Bias | RMSE |
| 0.01 | 0.01 | 0.570 | 5.628 | 0.715 | 5.641 | 0.601 | 5.578 | 0.739 | 5.605 |
| | 0.10 | 0.582 | 5.693 | 0.715 | 5.641 | 0.560 | 5.928 | 0.687 | 5.994 |
| | 0.25 | 0.601 | 5.807 | 0.715 | 5.641 | 0.422 | 6.382 | 0.533 | 6.461 |
| 0.10 | 0.01 | 0.497 | 5.947 | 0.630 | 6.003 | 0.615 | 5.647 | 0.739 | 5.605 |
| | 0.10 | 0.509 | 5.996 | 0.630 | 6.003 | 0.574 | 5.978 | 0.687 | 5.994 |
| | 0.25 | 0.529 | 6.082 | 0.630 | 6.003 | 0.437 | 6.407 | 0.533 | 6.461 |
| 0.25 | 0.01 | 0.313 | 6.361 | 0.422 | 6.435 | 0.636 | 5.769 | 0.739 | 5.605 |
| | 0.10 | 0.326 | 6.387 | 0.422 | 6.435 | 0.597 | 6.07 | 0.687 | 5.994 |
| | 0.25 | 0.346 | 6.434 | 0.422 | 6.435 | 0.461 | 6.457 | 0.533 | 6.461 |

To make the comparisons clearer, we summarize the results by boxplots to illustrate the effects of various factors of the simulation design. From Figure 4.1 we see that the estimates from the new method are accurate in most scenarios, since most of the Bias and RMSE of estimators fall in the ranges (0, 0.08) and (0.05, 0.20), respectively. The biases of Hall's estimators are higher than the new method's in most scenarios. The RMSEs of the two methods are close for normal, Poisson and discretized normal distributions. But for Cauchy distribution, the bias and RMSE are the highest and the new method performs much better than Hall's estimators. This is not surprising since Hall's method requires (4.4) and Cauchy distribution does not satisfy this condition. However, for lognormal distribution, the performance of Hall's estimators of $\delta_1$ and $\delta_2$ are quite different. The bias and RMSE of $\widehat{\delta_1}$ are smaller than the New method's while the bias and RMSE of $\widehat{\delta_2}$ are much larger. The tails of lognormal distributions for group 2 are heavier than that of group 1, and the estimators seem to be affected greatly by the large values in the tail. These imply Hall's method may not be suitable for heavy tailed distribution, even if (4.4) is satisfied.

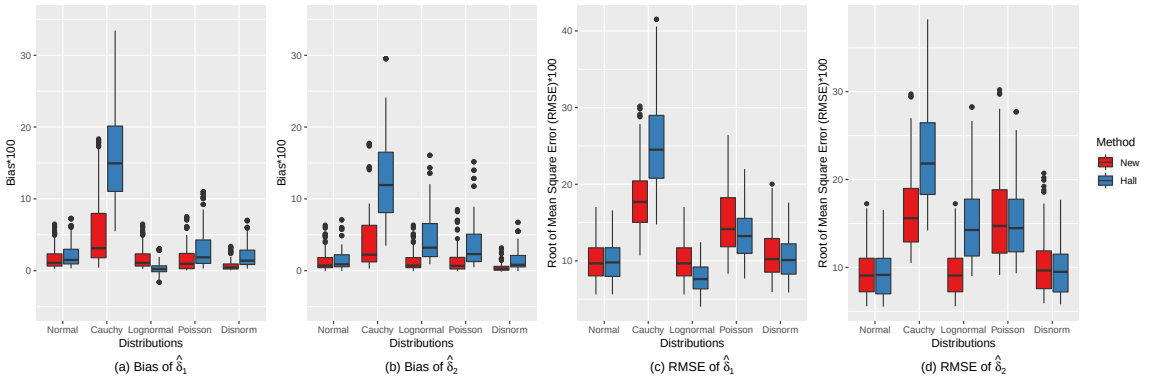



Figure 4.1: Boxplots of bias and RMSE for $\widehat{\delta_1}$ and $\widehat{\delta_2}$ by distributions. Disnorm is discretized normal distribution.

To check the effect of the other factors, we draw boxplots of Bias and RMSE by sample size allocation, mixture proportion and within-pair dependence. Since the results for $\widehat{\delta_2}$ are similar to $\widehat{\delta_1}$, we only present boxplots for $\widehat{\delta_1}$ here and the boxplots for $\widehat{\delta_2}$ are presented in the Appendix 4.8. The results for Bias of $\widehat{\delta_1}$ is presented in Figure 4.2 and Figure 4.3 contains the RMSE.

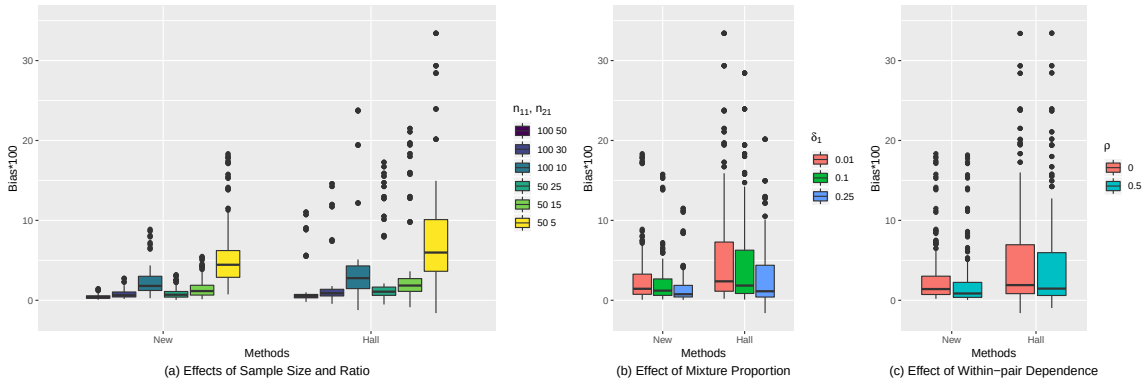



Figure 4.2: Boxplots of bias for $\widehat{\delta_1}$ by sample size, sample size ratio, $\delta_1$ and $\rho$.

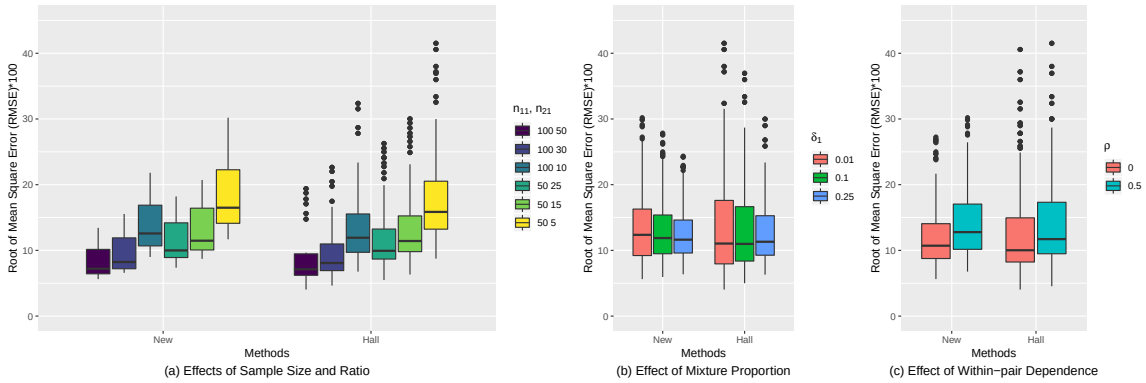



Figure 4.3: Boxplots of RMSE for $\widehat{\delta_1}$ by sample size, ratio, $\delta_1$ and $\rho$.

Overall, the comparison of the two methods are similar to what was observed in Figure 4.1. The estimators become more accurate when the sample size $n_{11}$ increase from 50 to 100. The sample size ratio between validation and contaminated data has great effects on the performance. When the ratio increase from 0.1 to 0.5, the bias and RMSE decrease rapidly. This shows that we can get accurate results if the ratio is not too low. The biases of $\widehat{\delta_1}$ decrease as the true value ($\delta_1$) gets large, but RMSEs are less affected. The biases of $\widehat{\delta_1}$ get smaller and the RMSEs get larger when $\rho$ increase from 0 to 0.5. Furthermore, the performance of the new estimator is more stable than the Hall's estimator.

**Asymptotic Distribution**

To check the asymptotic distribution in (4.11), we ran simulations and recorded the coverage probability for the 95% confidence interval in (4.12). Table 4.2 contains the results for $\widehat{\delta}_1$ and $\widehat{\delta}_2$ when $F_{2gt}$, g,t=1,2, are normal distributions. From these results, we can see that the coverage probability is close to the nominal level 95% when the sample size ratio between validation and contaminated datasets is not too small.

Table 4.2: Coverage Probability(%) of $\widehat{\delta}_1$ and $\widehat{\delta}_2$ when $\sigma^2 = 1$, $n_{11} = 100$ and $n_{12} = 100$.

| | | $\widehat{\delta}_1$ | | | | | | $\widehat{\delta}_2$ | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | | 0 | | | 0.5 | | | 0 | | | 0.5 | | |
| ratio | | 0.5 | 0.3 | 0.1 | 0.5 | 0.3 | 0.1 | 0.5 | 0.3 | 0.1 | 0.5 | 0.3 | 0.1 |
| $\delta_1$ | $\delta_2$ | | | | | | | | | | | | |
| | 0.01 | 94.0 | 93.9 | 87.8 | 94.2 | 94.2 | 88.8 | 94.2 | 93.7 | 88.1 | 94.4 | 93.8 | 88.9 |
| 0.01 | 0.10 | 94.0 | 93.9 | 87.8 | 94.2 | 94.2 | 88.8 | 94.7 | 94.2 | 90.4 | 94.7 | 94.2 | 91.2 |
| | 0.25 | 93.9 | 93.8 | 87.7 | 94.2 | 94.1 | 88.8 | 95.0 | 95.2 | 93.8 | 95.0 | 95.4 | 94.1 |
| | 0.01 | 94.5 | 94.4 | 90.4 | 94.8 | 94.8 | 90.7 | 94.0 | 93.7 | 88.0 | 94.5 | 93.8 | 89.0 |
| 0.10 | 0.10 | 94.6 | 94.3 | 90.3 | 94.8 | 94.8 | 90.6 | 94.6 | 94.2 | 90.2 | 94.7 | 94.2 | 91.1 |
| | 0.25 | 94.4 | 94.2 | 90.1 | 94.8 | 94.7 | 90.5 | 95 | 95.2 | 93.6 | 95.0 | 95.3 | 94.0 |
| | 0.01 | 95.0 | 95.2 | 94.1 | 95.0 | 95.3 | 94.1 | 94.1 | 93.6 | 88.1 | 94.5 | 93.7 | 89.1 |
| 0.25 | 0.10 | 94.9 | 95.2 | 93.9 | 95.0 | 95.3 | 94.0 | 94.6 | 94.2 | 89.9 | 94.7 | 94.1 | 90.9 |
| | 0.25 | 94.9 | 95.1 | 93.4 | 94.9 | 95.2 | 93.7 | 95.0 | 95.2 | 93.3 | 94.9 | 95.2 | 93.9 |

The effects of various simulation factors on the coverage probability are presented in Figure 4.4. The new method has reasonable performance for all the distributions considered. That is, the coverage probability is close to 95% in most scenarios. Sample size ratio between validation and contaminated data has great effects on the performance. When the ratio increase from 0.1 to 0.5, the coverage probability becomes very close the nominal level. This shows that the asymptotic distributions are good approximations if the ratio is not too low. Also, the coverage probability gets closer to the nominal level as $\delta_1$ increase.

**Treatment Effect**

In the simulation study for treatment effect, we compare our proposed methods with the traditional ones. Here also, we use bias, RMSE and coverage probability to evaluate the

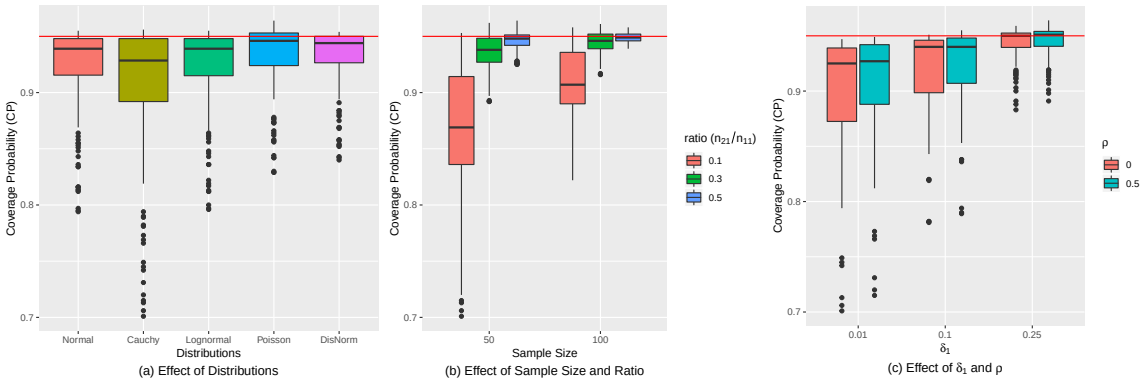



Figure 4.4: Boxplots of coverage probability for $\widehat{\delta}_1$ by distributions, sample size, ratio, $\delta_1$ and $\rho$. Disnorm is discretized normal distribution.

estimation accuracy and the asymptotic results. In addition, we also compare powers of the tests. The methods compared are

1. Tra1: the traditional method that ignores the misclassification rates for the fallible classifiers. That is, $\delta_1 = \delta_2 = 0$ is assumed,

2. Tra2: the traditional methods that only use the observations from the infallible classifiers. That is, the contaminated data is discarded.

3. Mix: the new methods that estimates the misclassification rates and treatment effects using both source of data and

In Table 4.3, we present the results for the three methods for the parameter settings as in Table 1. Since Tra2 uses only the validation datasets, its results remain unchanged as the mixing proportions change. The biases and RMSEs for Tra2 and Mix methods are small and CPs are close to the nominal level 95%. The bias of Tra1 is greatly affected by the difference between $\delta_1$ and $\delta_2$. When the difference is large, the bias of Tra1 is large and the CP falls below 95%. To facilitate further comparison, we use boxplots in Figure 4.5-4.8 to visualize the effects of the various simulation factors.

Table 4.3: Bias($\times 100$), RMSE($\times 100$) and CP of Interaction Effect when $\sigma^2 = 1, \rho = 0, n_{11} = 100, n_{12} = 100$, ratio$= 0.5$, $p_I = 0$.

| | | Method | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Tra1 | | | Tra2 | | | Mix | | |
| $\delta_1$ | $\delta_2$ | Bias | RMSE | CP | Bias | RMSE | CP | Bias | RMSE | CP |
| 0.01 | 0.01 | 0.024 | 2.797 | 0.953 | -0.009 | 4.897 | 0.948 | 0.019 | 2.986 | 0.953 |
| | 0.10 | 0.859 | 2.980 | 0.940 | -0.009 | 4.897 | 0.948 | 0.035 | 3.240 | 0.951 |
| | 0.25 | 2.002 | 3.544 | 0.891 | -0.009 | 4.897 | 0.948 | 0.067 | 3.795 | 0.944 |
| 0.10 | 0.01 | -0.808 | 2.950 | 0.942 | -0.009 | 4.897 | 0.948 | 0.008 | 3.215 | 0.954 |
| | 0.10 | 0.027 | 2.880 | 0.951 | -0.009 | 4.897 | 0.948 | 0.022 | 3.526 | 0.953 |
| | 0.25 | 1.170 | 3.157 | 0.932 | -0.009 | 4.897 | 0.948 | 0.056 | 4.238 | 0.957 |
| 0.25 | 0.01 | -1.943 | 3.505 | 0.897 | -0.009 | 4.897 | 0.948 | -0.005 | 3.768 | 0.956 |
| | 0.10 | -1.108 | 3.137 | 0.936 | -0.009 | 4.897 | 0.948 | 0.009 | 4.234 | 0.956 |
| | 0.25 | 0.036 | 2.961 | 0.951 | -0.009 | 4.897 | 0.948 | 0.054 | 5.442 | 0.962 |

**Effect of Distributions**

In Figure 4.5, we see that the bias of Tra1 is much larger than the other methods, and when the coverage probability of Tra1 is lower than 95%, especially for Poisson distribution and the true value of $p_I$ is different from $0$. Tra2 has small bias and its coverage probability is close to 95%, but the RMSE is larger than the other method. This is because Tra2 only uses the validation data and, therefore, the sample size for it is much smaller than that of the other methods. In most scenarios, the Mix method has reasonable performance for all distributions, but in some cases this method produces large biases and RMSEs. To make the comparisons clear, we omitted outliers in the boxplots for biases (16 out of 810) and RMSEs (57 out of 810).

**Effect of Sample Size**

The effects of sample size on these methods are presented in Figure 4.6. Tra1 is less affected by the sample size compared to the other methods and its coverage probabilities become lower than 95% when sample sizes increase. This shows that the performance of Tra1 will not improve as sample size increase. Tra2 and Mix have more accurate results as sample size and ratio get larger. When the sample size ratio is low, the coverage probabilities of Tra2 become lower than 95% and that of Mix tend to be more conservative. One

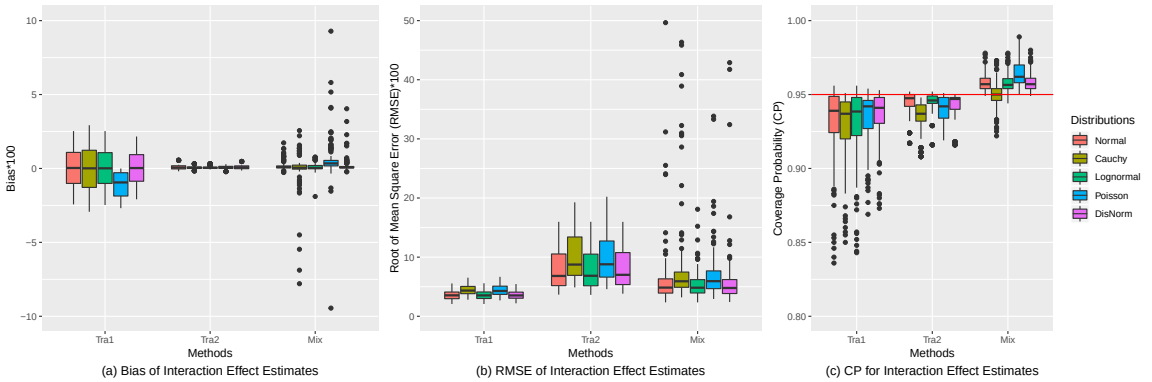



Figure 4.5: Boxplots of bias, RMSE and coverage probability for Tra1, Tra2, Mix methods' estimates of $p_I$ by distributions.

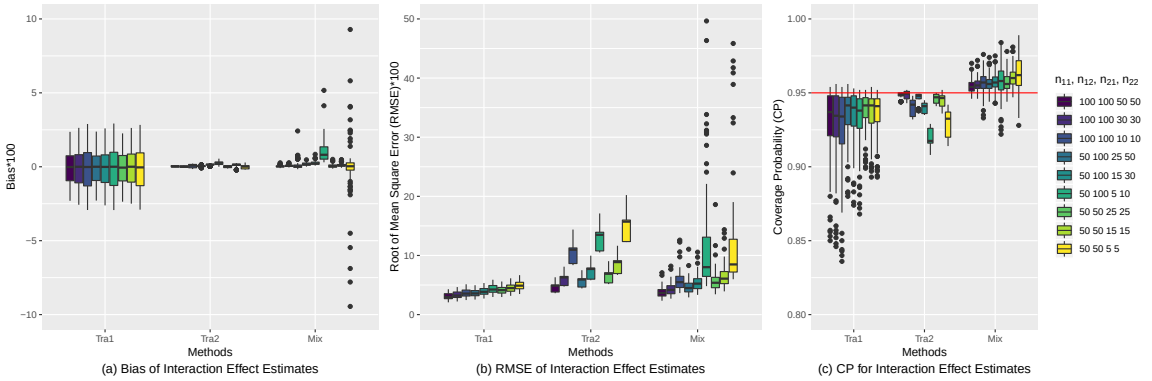



Figure 4.6: Boxplots of bias, RMSE and coverage probability for Tra1, Tra2, Mix methods' estimates of $p_I$ by sample size allocations.

thing to note is about the outliers in Mix methods. When the sample size ratio is $0.1$, Mix has many extreme value in the boxplots for biases and RMSEs. This is because when the ratio is low, the sample size of the validation data is too small that the estimates of $\delta_1$ and $\delta_2$ have large variances, and we are likely to get estimates such that $\widehat{\delta}_1 + \widehat{\delta}_2$ is close to 1. In this case, the estimator $\widehat{p}_I$ defined in (4.18) will be very large. In our simulation, the range of bias from Mix is $(-4.798, 13.132)$. But when the sample size ratio between validation and contaminated data increase to $0.3$, the range of bias for Mix is $(-0.11 to 0.008)$. This shows that to use Mix method the sample size ratio should not be too low.

### Effect of Mixing Proportions

As a result of ignoring the misclassification error, Tra1's accuracy is greatly affected by mixing proportions (see Figure 4.7). In panel (a), we can see that the bias from Tra1 is small when $\delta_1 = \delta_2$, but the bias become larger when $\delta_2 - \delta_1$ increase. From (4.15) we see that when $\delta_1 = \delta_2$, the treatment effects obtained from Tra1 will be unbiased when $p_I = 0$. When $\delta_1 \neq \delta_2$ and $p_I \neq 0$, the results from Tra1 will be misleading. Since Tra2 only uses validated dataset, its results do not change with mixing proportions. From Figure 4.7, we see that Mix has reasonable performances as mixing proportions change. The RMSE increases as sum of mixture proportion increase, and the CPs become a little bit higher than 95%.

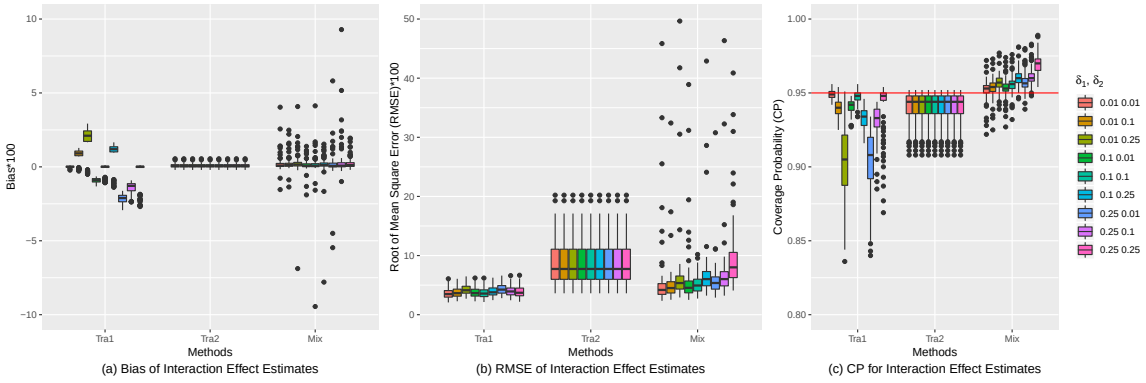



Figure 4.7: Boxplots of bias, RMSE and coverage probability for Tra1, Tra2, Mix methods' estimates of $p_I$ by mixture components.

### Effect of Within-Pair Dependence

The within-pair dependence does not show a clear effects on the performance of the methods (see Figure 4.8), except that the RMSEs get smaller when $\rho$ increases.

### Effect of Size of $p_I$

In most of the simulation scenarios above, the treatment effect was set equal to $0$, except when the distribution is Poisson. To show examples performances when $p_I$ is not equal to $0$, we set the sample size $n_{11} = n_{12} = 50$ and the sample size ratios to $0.5$. The

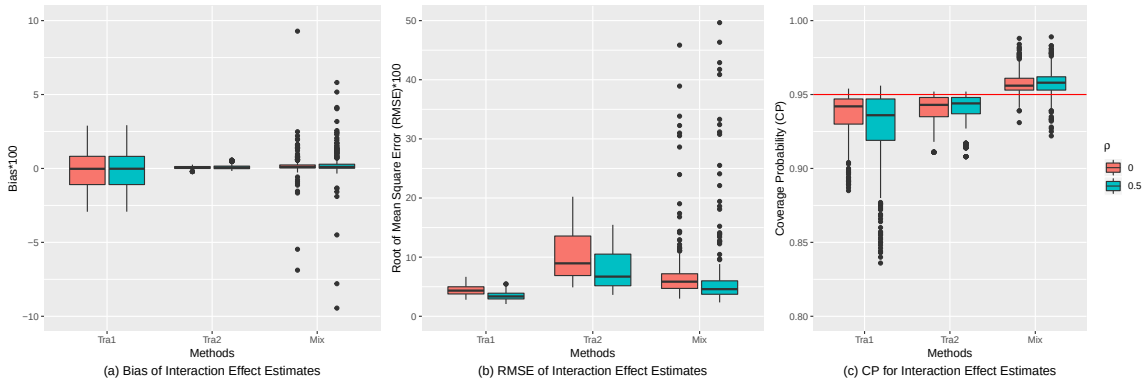



Figure 4.8: Boxplots of bias, RMSE and coverage probability for Tra1, Tra2, Mix methods' estimates of $p_I$ by within-pair dependence $\rho$.

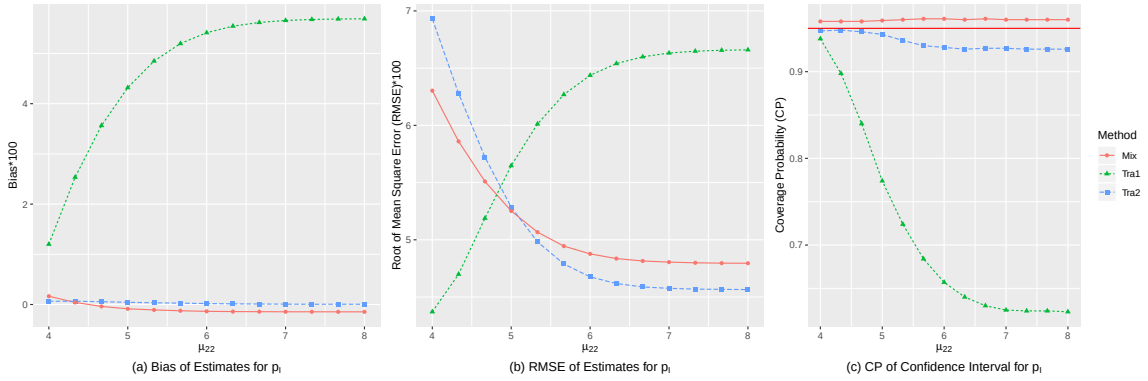



Figure 4.9: Graphs of Bias, RMSE, and CP for Tra1, Tra2, and Mix methods when $\delta_1 = 0.1$ and $\delta_2 = 0.25$.

distributions $F_{2gt}$ are set to $N(\mu_{gt}, 1)$, where $(\mu_{11}, \mu_{12}, \mu_{21}) = (1, 2, 3)$, $\mu_{22}$ increases from $4$ to $8$, and $\rho = 0$. Figure 4.9 presents graphs of Bias, RMSE and coverage probabilities (CP) of the three methods when $(\delta_1, \delta_2) = (0.1, 0.25)$. Tra2 and Mix have reasonable performances as $\mu_{22}$ increases. The bias of Tra1 become larger as $\mu_{22}$ increases and CPs drop quickly. This shows that the traditional method (ignoring misclassification errors) may produce misleading results. Interestingly, RMSE curves of Tra2 and Mix intersect and the point of intersection and its reasons need further investigation.

### Power Simulation

To show the power performances of the three methods, we fix sample size and sample size ratio same as in Section 4.5, and $\delta_1 = \delta_2 = 0.1$. The size of the test is set at $\alpha = 0.05$. We assume independence of the pre and post measurements. For the alternative hypothesis, we keep $F_{211}, F_{212}, F_{221}$ the same but add location, shape or both to $F_{222}$. We choose $F_{222}$ in such a way that when the location or shape parameter or both are zero, the treatment effect is $0$. The marginal distributions we consider are skew-normal (SN)(Azzalini, 1985) and skew-cauchy (SC)(Azzalini and Capitanio, 2003). The power curves for these distributions are shown in Figures 4.10 and 4.11.

The figures shows that powers of the Tra1 and Mix are close. Tra1 and Mix methods have clear advantage over the Tra2 when the data come from the heavy tailed distribution, skew-cauchy.

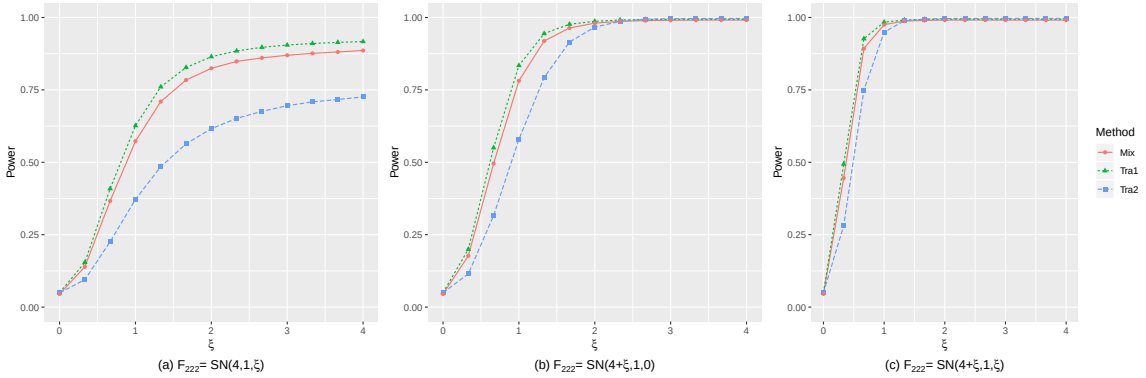



Figure 4.10: Power curves for Tra1, Tra2, and Mix methods. $F_{211}, F_{212}, F_{221}$ are distributed as N(1,1), N(2,1), and N(3,1), respectively. $F_{222}$ varies with respect to location and/or shape.

## 4.6 Real Data Example

To illustrate the application of the new method, we use a real dataset from the Total Sleep Deprivation (TSD) study (Satterfield et al., 2015). The study examined the effect of genotypes at position 308 in the TNF$\alpha$ gene on psychomotor vigilance performance impairment during total sleep deprivation. Eighty eight subjects participated in one of the five labora-

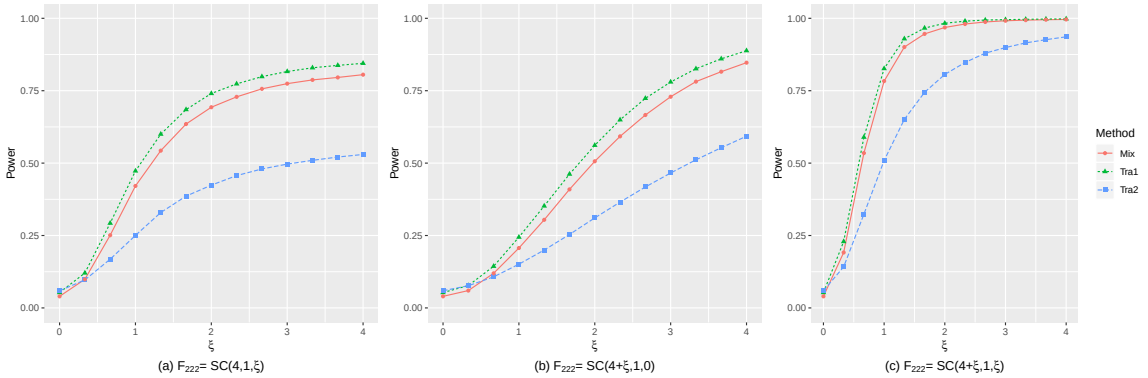



Figure 4.11: Power curves for Tra1, Tra2, and Mix methods. $F_{211}, F_{212}, F_{221}$ are distributed as Cauchy(1,1), Cauchy(2,1), and Cauchy(3,1), respectively. $F_{222}$ varies with respect to location and/or shape.

tory TSD studies. Cognitive performances of the participants were measured across 36–62 hours of sustained wakefulness. All five studies included at least 24 hours of wake extension into the night and the following day. A 10-min psychomotor vigilance test (PVT) was used as the primary performance criteria and it was administered every 2–3 hours over the course of the scheduled wakefulness. Subjects' vulnerabilities to sleep loss were quantified based on their PVT performances over the 24 hour period of sleep deprivation common to all the five studies.

To assess the effect of genotypes in the TNF$\alpha$, subjects' genetypes were determined from blood samples collected during pre-study screening sessions. The subjects were classified into three genotypes: G/G, A/G or A/A. According to Juszczynski et al. (2002), the classification method is not perfect and its positive predictive values for G/G, A/G, and A/A genotypes are $97.9\%$, $92.5\%$ and $100\%$, respectively. Since there was only one subject with A/A genotype, we will focus on comparing subjects with G/G or A/G genotypes.

There were 64 subjects with G/G genotype and 23 subjects with A/G genotypes. We regard the 24 hour period of sleep deprivation as a treatment. We use the latest PVT (20:00 or 21:00) before the 24 hour period of sleep deprivation (22:00-22:00) as pre-treatment measurements, denoted as $X_{111}$ and $X_{121}$, and the latest PVT (18:00, 20:00 or 21:00 ) in the 24 hour period as post-treatment measurement, denoted as $X_{112}$ and $X_{122}$, where $X_{11t}$ and $X_{12t}$, are measurements for subjects with G/G and A/G genotypes, respectively, for

pre and post treatment times $t = 1, 2$. Data from two subjects, one from each group, were discarded because they had missing values.

In this study, the mixing proportions are known to be $\delta_1 = 2.1\%$ and $\delta_2 = 7.5\%$. We use the estimator in (4.20) and its distribution (4.21) to estimate the treatment effect and conduct hypothesis test. The estimated treatment effect is $\widehat{p}_{\mathrm{I}} = 0.0475$. The effect is not statistically significant ($T_M = .764$ and $p$–value $= 0.445$). Therefore, we do not have evidence of difference in PVT performances by the genotypes in the TNF$\alpha$ as a result of sleep deprivation.

## 4.7 Discussion

We developed a fully nonparametric method to assess treatment effects when the classifiers used for stratifying participants are fallible. We modeled outcome distributions as mixtures of unknown distributions and developed estimators for mixing proportions when a validation (training) data exists. Consistency and the order of bias of the estimators are proved. Also, the asymptotic distributions for the estimators are derived. The estimators of the mixing proportions are used to construct inferential procedures (estimation and testing) for a purely nonparametric measure of treatment effect.

Our estimation for mixing proportions provide more accurate results compared to an existing method without requiring any distributional assumptions. Therefore, our method works for heavy tailed distributions as well as non metric data. The coverage probabilities of the proposed confidence interval for mixing proportions are close to the nominal level when the sample size ratio between contaminated and validation data is not too small. When misclassification exists, the traditional method which ignores the errors leads to a serious bias in estimation of the treatment effect. Our method have much smaller biases and stable coverage probabilities. Also, the test based on our estimator and its asymptotic theory has higher power compared to the method that only uses the validation data set.

The results derived in this chapter cover the situation where measurement is taken only once as a special case by introducing minor changes in notations. Specifically, dropping the index for time, let $F_{2g}, g = 1, 2$, be the distributions for observations from subjects classified by infallible classifiers and $F_{1g}$ be the distributions for observations from the fallible

classifier. Therefore, $F_{1g}$ is a mixture of $F_{21}$ and $F_{22}$, mixed in proportions determined by $\delta_g$. That is,

$$F_{1g} = (1 - \delta_g)F_{2g} + \delta_g F_{2g'},$$

for $g' \neq g, g', g = 1, 2$. Following the same steps as in Section 4.3, the mixing proportions are estimated by

$$\widehat{\delta}_g = \max\left\{ \frac{\int (\widehat{F}_{1g} - \widehat{F}_{2g})(\widehat{F}_{2g'} - \widehat{F}_{2g}) d\widehat{H}}{\int (\widehat{F}_{21} - \widehat{F}_{22})^2 d\widehat{H}}, 0 \right\},$$

for $g = 1, 2$, where $\widehat{H} = \sum_{h=1}^{2} \sum_{g=1}^{2} \frac{n_{hg}}{N} \widehat{F}_{hg}$. The treatment effect is measured by the nonparametric effect measure $p = \int F_{21} dF_{22}$. In the presence of misclassification errors,

$$p = \int F_{21} dF_{22} = \int F_{11} dF_{12} + \int \frac{\delta_2 F_{11} + \delta_1 F_{12}}{1 - \delta_1 - \delta_2} d(F_{12} - F_{11}).$$

Estimation and testing procedures for $p$ can be derived in a manner analogous to Section 4.4.

The accuracy of estimation of the mixing proportions will be affected by the separation of the distributions of two groups. From the simulation results, we note that when the mixing proportions $\delta_1$and $\delta_2$ are close to $0.5$ or the sample size ratio between validation and contaminated data is small, we may get estimates of $\widehat{\delta}_1$ and $\widehat{\delta}_2$ such that $\widehat{\delta}_1 + \widehat{\delta}_2$ is very close to $1$ due to sampling variation. In such cases, the estimation of $\widehat{p}_{I1}$ will be very large and it leads to large bias and RMSE for $\widehat{p}_I$, because $\widehat{p}_I$ is a weighted average of $\widehat{p}_{I1}$ and $\widehat{p}_{I2}$. To solve this problem, one idea is to choose the weight of $\widehat{p}_{I1}$ and $\widehat{p}_{I2}$ that minimizes the variance of $\widehat{p}_I$. More specifically, one can consider the estimator

$$\widehat{p}_I = \gamma\, \widehat{p}_{I1} + (1 - \gamma)\, \widehat{p}_{I2},$$

where $\gamma$ is chosen to minimize the variance of $\widehat{p}_I$. By doing so, when $\widehat{p}_{I1}$ become very large due to estimations of $\widehat{\delta}_1$ and $\widehat{\delta}_2$, $\gamma$ will be close to $0$ and, as a result, it leads to smaller bias and variation for $\widehat{p}_I$. The theory for this estimator needs futher investigation.

To avoid nonidentifiability in the finite mixtures, we assumed that the mixing proportions are known or validation (training) data exists. In some application, it may not be

possible to get either of these information. However, it may be possible to derive inferential procedures by making stronger assumption on the nature of dependence between the pre and post variables, or by using a semi-parametric dependence models. Another approach could be to use auxiliary variables or covariates that contains information about the accuracy of the classifies and use them to estimate the mixing probabilities and treatment effects simultaneously. We plan to investigate these problems in future research.

## 4.8 Appendix

### Lemmas

In this subsection, we state and prove some useful technical Lemmas before we present the proofs of the main results,

**Lemma 4.8.1.** *Let $X_{hgtk}$, $F_{hgt}$, $\widehat{F}_{hgt}$ and $N$ be defined as in Proposition 4.3.1. Then under Assumption 4.3.1, we have*

$$E\left(\int \widehat{F}_{hgt} d\widehat{F}_{slk}\right) = \int F_{hgt} dF_{slk} + O(N^{-1}), \tag{4.23}$$

$$E\left(\int \widehat{F}_{hgt}\widehat{F}_{slr} d\widehat{F}_{uvw}\right) = \int F_{hgt}F_{slr} dF_{uvw} + O(N^{-1}), \tag{4.24}$$

*for $g, t, l, r, h, s, u, v, w = 1, 2$.*

*Proof.* (i) Proof of (4.23). By definition of $\widehat{F}_{hgt}$, we have

$$\int \widehat{F}_{hgt} d\widehat{F}_{slk} = \frac{1}{n_l}\sum_{i=1}^{n_l} \widehat{F}_{hgt}(X_{slki}) = \frac{1}{n_l n_g}\sum_{i=1}^{n_l}\sum_{j=1}^{n_g} c(X_{slki} - X_{hgtj}).$$

When $X_{slki}$ and $X_{hgtj}$ are independent, by Fubini's theorem it follows that

$$E[c(X_{slki} - X_{hgtj})] = \int\int c(y-x) dF_{hgt}(x) dF_{slk}(y) = \int F_{hgt} dF_{slk}.$$

Therefore, when $(h, g) \neq (s, l)$, $X_{slki}$ and $X_{hgtj}$ are independent,

$$E\left(\int \widehat{F}_{hgt} d\widehat{F}_{slk}\right) = E\left(\frac{1}{n_l n_g}\sum_{i=1}^{n_l}\sum_{j=1}^{n_g} c(X_{slki} - X_{hgtj})\right) = \int F_{hgt} dF_{slk}.$$

On the other hand, if $(h,g)=(s,l)$, $X_{hgki}$ and $X_{hgtj}$ are independent if $i \neq j$. Thus, by Assumption 4.3.1,

$$\begin{aligned}
E\left(\int \widehat{F}_{hgt} d\widehat{F}_{hgk}\right) =& E\left(\frac{1}{n_l^2}\sum_{i=1}^{n_l}\sum_{j=1}^{n_l} c(X_{hgki}-X_{hgtj})\right) \\
=& E\left(\frac{1}{n_l^2}\sum_{i\neq j}^{n_l} c(X_{hgki}-X_{hgtj})\right) \\
&+ E\left(\frac{1}{n_l^2}\sum_{i=1}^{n_l} c(X_{hgki}-X_{hgti})\right) \\
=& \frac{n_l-1}{n_l}\int F_{hgt} dF_{hgk} + \frac{1}{n_l} E\left(c(X_{hgk1}-X_{hgt1})\right) \\
=& \int F_{hgt} dF_{hgk} + O(N^{-1}).
\end{aligned}$$

(ii) Proof of (4.24). By definition, we have

$$\begin{aligned}
\int \widehat{F}_{hgt}\widehat{F}_{slr} d\widehat{F}_{uvw} &= \frac{1}{n_{uv}}\sum_{i=1}^{n_{uv}} \widehat{F}_{hgt}(X_{uvwi})\widehat{F}_{slr}(X_{uvwi}) \\
&= \frac{1}{n_{hg}n_{sl}n_{uv}}\sum_{i=1}^{nuv}\sum_{j=1}^{n_{hg}}\sum_{k=1}^{n_{sl}} c(X_{uvwi}-X_{hgtj})c(X_{uvwi}-X_{slrk}).
\end{aligned}$$

Notice that when $X_{uvwi}$, $X_{hgtj}$, and $X_{slrk}$ are independent with each other,

$$E\left(c(X_{uvwi}-X_{hgtj})c(X_{uvwi}-X_{slrk})\right) = \int F_{hgt}F_{slr}dF_{uvw}.$$

When $i \neq j \neq k$, $Y_{uvwi}$, $Y_{hgtj}$, and $Y_{slrk}$ are independent with each other. To make the summation easier to present, suppose $n_{uv} \leq n_{hg} \leq n_{sl}$. We can change the order

of $i, j, k$ in summation if this does not hold. Therefore,

$$
\begin{aligned}
& E\left(\int \widehat{F}_{hgt}\widehat{F}_{slr}d\widehat{F}_{uvw}\right) \\
=&\frac{1}{n_{uv}n_{hg}n_{sl}}\sum_{i=1}^{n_{uv}}\sum_{j\neq i}^{n_{hg}}\sum_{k\neq i,j}^{n_{sl}} E\left(c(X_{uvwi}-X_{hgtj})c(X_{uvwi}-X_{slrk})\right) \\
&+\frac{1}{n_{uv}n_{hg}n_{sl}}\sum_{i=1}^{n_{uv}}\sum_{j\neq i}^{n_{hg}} E\left(c(X_{uvwi}-X_{hgtj})c(X_{uvwi}-X_{slrj})\right) \\
&+\frac{1}{n_{uv}n_{hg}n_{sl}}\sum_{i=1}^{n_{uv}}\sum_{j\neq i}^{n_{hg}} E\left(c(X_{uvwi}-X_{hgtj})c(X_{uvwi}-X_{slri})\right) \\
&+\frac{1}{n_{uv}n_{hg}n_{sl}}\sum_{i=1}^{n_{uv}}\sum_{k\neq i}^{n_{sl}} E\left(c(X_{uvwi}-X_{hgti})c(X_{uvwi}-X_{slrk})\right) \\
&+\frac{1}{n_{uv}n_{hg}n_{sl}}\sum_{i=1}^{n_{uv}} E\left(c(X_{uvwi}-X_{hgti})c(X_{uvwi}-X_{slri})\right) \\
=&\frac{(n_{hg}-1)(n_{sl}-2)}{n_{hg}n_{sl}}\int F_{hgt}F_{slr}dF_{uvw} \\
&+\frac{n_{hg}-1}{n_{hg}n_{sl}}E\left(c(X_{uvw1}-X_{hgt2})c(X_{uvw1}-X_{slr2})\right) \\
&+\frac{n_{hg}-1}{n_{hg}n_{sl}}E\left(c(X_{uvw1}-X_{hgt2})c(X_{uvw1}-X_{slr1})\right) \\
&+\frac{n_{sl}-1}{n_{hg}n_{sl}}E\left(c(X_{uvw1}-X_{hgt1})c(X_{uvw1}-X_{slr2})\right) \\
&+\frac{1}{n_{hg}n_{sl}}E\left(c(X_{uvw1}-X_{hgt1})c(X_{uvw1}-X_{slr1})\right) \\
=&\int F_{hgt}F_{slr}dF_{uvw}+O(N^{-1}).
\end{aligned}
$$

□

**Lemma 4.8.2.** *Let $F_{hgt}$, $\widehat{F}_{hgt}$ and $N$ be defined as in Proposition 4.3.1, and let $\widehat{H}(x)$ be defined as in (4.8). Then under Assumption 4.3.1, at any fixed point $x$ we have*

$$E[\widehat{F}_{hgt}(x)-F_{hgt}(x)]^2\leq\frac{1}{n_{hg}},\quad E[\widehat{F}_{hgt}(X_{slr})-F_{hgt}(X_{slr})]^2\leq\frac{1}{n_{hg}}, \tag{4.25}$$

$$E[\widehat{H}(x)-H(x)]^2\leq\frac{N_0}{N}=O(N^{-1}), \tag{4.26}$$

$$E\left(\int \widehat{F}_{hgt}\widehat{F}_{slr}d\widehat{F}_{uvw}-\int F_{hgt}F_{slr}dF_{uvw}\right)^2=O(N^{-1}), \tag{4.27}$$

$$E\left(\int \widehat{F}_{hgt}\widehat{F}_{slr}d\widehat{H}-\int F_{hgt}F_{slr}dH\right)^2=O(N^{-1}), \tag{4.28}$$

*for* $h, g, t, s, l, r, u, v, w = 1, 2$.

*Proof.* (i) Proof of (4.25) and (4.26). Similar to the proof of Lemma 7.4 in Brunner et al. (2018), we can get these results.

(ii) Proof of (4.27). Apply the $c_r$-inequality, we have

$$\begin{aligned}
&\left(\int \widehat{F}_{hgt}\widehat{F}_{slr}d\widehat{F}_{uvw} - \int F_{hgt}\widehat{F}_{slr}dF_{uvw}\right)^2 \\
=&\Big[\int (\widehat{F}_{hgt} - F_{hgt})F_{slr}d\widehat{F}_{uvw} + \int F_{hgt}(\widehat{F}_{slr} - F_{slk})d\widehat{F}_{uvw} \\
&+ \int F_{hgt}F_{slk}d\left(\widehat{F}_{uvw} - F_{uvw}\right)\Big]^2 \\
\leq& 3\left(\int (\widehat{F}_{hgt} - F_{hgt})\widehat{F}_{slr}d\widehat{F}_{uvw}\right)^2 + 3\left(\int F_{hgt}(\widehat{F}_{slr} - F_{slr})d\widehat{F}_{uvw}\right)^2 \\
&+ 3\left(\int F_{hgt}F_{slr}d\left(\widehat{F}_{uvw} - F_{uvw}\right)\right)^2.
\end{aligned}$$

By partial integration, we obtain that

$$\begin{aligned}
\int F_{hgt}F_{slr}d\left(\widehat{F}_{uvw} - F_{uvw}\right) = &-\int \left(\widehat{F}_{uvw} - F_{uvw}\right)F_{hgt}dF_{slr} \\
&-\int \left(\widehat{F}_{uvw} - F_{uvw}\right)F_{slr}dF_{hgt}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\left(\int \widehat{F}_{hgt}\widehat{F}_{slr}d\widehat{F}_{uvw} - \int F_{hgt}F_{slr}dF_{uvw}\right)^2 \\
\leq& 3\left(\int (\widehat{F}_{hgt} - F_{hgt})\widehat{F}_{slr}d\widehat{F}_{uvw}\right)^2 + 3\left(\int F_{hgt}(\widehat{F}_{slr} - F_{slr})d\widehat{F}_{uvw}\right)^2 \\
&+ 3\left(\int \left(\widehat{F}_{uvw} - F_{uvw}\right)F_{hgt}dF_{slr} + \int \left(\widehat{F}_{uvw} - F_{uvw}\right)F_{slr}dF_{hgt}\right)^2 \\
\leq& 3\left(\int \left(\widehat{F}_{hgt} - F_{hgt}\right)d\widehat{F}_{uvw}\right)^2 + 3\left(\int \left(\widehat{F}_{slr} - F_{slr}\right)d\widehat{F}_{uvw}\right)^2 \\
&+ 6\left(\int \left(\widehat{F}_{uvw} - F_{uvw}\right)dF_{slr}\right)^2 + 6\left(\int \left(\widehat{F}_{uvw} - F_{uvw}\right)dF_{hgt}\right)^2.
\end{aligned}$$

Using the Jensen's inequality, we have

$$\begin{aligned}
&\left(\int \widehat{F}_{hgt}\widehat{F}_{slk}d\widehat{F}_{uvw} - \int F_{hgt}F_{slk}dF_{uvw}\right)^2 \\
\leq &\frac{3}{n_{uv}}\sum_{i=1}^{n_{uv}}\left[\widehat{F}_{hgt}(X_{uvwi}) - F_{hgt}(X_{uvwi})\right]^2 + \frac{3}{n_{uv}}\sum_{i=1}^{n_{uv}}\left[\widehat{F}_{slr}(X_{uvwi}) - F_{slr}(X_{uvwi})\right]^2 \\
&+ 6\int\left(\widehat{F}_{uvw} - F_{uvw}\right)^2 dF_{slr} + 6\int\left(\widehat{F}_{uvw} - F_{uvw}\right)^2 dF_{hgt}.
\end{aligned}$$

Taking expectations on both sides and by (4.25), we have

$$E\left(\int \widehat{F}_{hgt}\widehat{F}_{slk}d\widehat{F}_{uvw} - \int F_{hgt}F_{slk}dF_{uvw}\right)^2 \leq \frac{3}{n_{hg}} + \frac{3}{n_{sl}} + \frac{12}{n_{uv}} = O\left(N^{-1}\right).$$

(iii) Proof of (4.28). By the $c_r$-inequality, we have

$$\begin{aligned}
&E\left(\int \widehat{F}_{hgt}\widehat{F}_{slr}d\widehat{H} - \int F_{hgt}F_{slr}dH\right)^2 \\
=&E\left(\sum_{u,v,w=1,2}^{2}\frac{n_{uv}}{N}\left(\int \widehat{F}_{hgt}\widehat{F}_{slr}d\widehat{F}_{uvw} - \int F_{hgt}F_{slr}dF_{uvw}\right)\right)^2 \\
\leq &\sum_{u,v,w=1,2}^{2}\frac{2n_{uv}^2}{N^2}E\left(\int \widehat{F}_{hgt}\widehat{F}_{slr}d\widehat{F}_{uvw} - \int F_{hgt}F_{slr}dF_{uvw}\right)^2 \leq O(N^{-1}).
\end{aligned}$$

□

**Lemma 4.8.3.** *Let $X_{hgtk}$, $F_{hgt}$, $\widehat{F}_{hgt}$ and $N$ be defined as in Proposition 4.3.1. Then under Assumption 4.3.1, we have*

$$B_1 = \sqrt{N}\int\left(\widehat{F}_{hgt} - F_{hgt}\right)d\left(\widehat{F}_{slr} - F_{slr}\right) \xrightarrow{P} 0, \tag{4.29}$$

$$B_2 = \sqrt{N}\int\left(\widehat{F}_{hgt}\widehat{F}_{slr} - F_{hgt}F_{slr}\right)d\left(\widehat{F}_{uvw} - F_{uvw}\right) \xrightarrow{P} 0, \tag{4.30}$$

$$B_3 = \sqrt{N}\int\left(\widehat{F}_{hgt} - F_{hgt}\right)\left(\widehat{F}_{slr} - F_{slr}\right)dF_{uvw} \xrightarrow{P} 0, \tag{4.31}$$

*for $h, g, t, s, l, r, u, v, w = 1, 2$.*

*Proof.* (i) Proof of (4.29). By definition, we have

$$\begin{aligned}
B_1 =&\frac{\sqrt{N}}{n_{sl}}\sum_{i=1}^{n_{sl}}\left(\widehat{F}_{hgt}(X_{slri}) - F_{hgt}(X_{slri}) - \int\left(\widehat{F}_{hgt}(x) - F_{hgt}(x)\right)dF_{slr}(x)\right) \\
=&\frac{\sqrt{N}}{n_{sl}}\sum_{i=1}^{n_{sl}}A_i.
\end{aligned}$$

First note that

$$E(A_i) = E\left(\widehat{F}_{hgt}(X_{slri}) - F_{hgt}(X_{slri}) - \int \left(\widehat{F}_{hgt}(x) - F_{hgt}(x)\right) dF_{slr}(x)\right) = 0,$$

for $i = 1, \ldots, n_{sl}$. Then by Fubini's theorem and independence of random variables, we have

$$E(A_i A_j) = 0 \text{ for } i \neq j, \quad i, j = 1, \ldots, n_{sl}.$$

Through the Lemma 7.4 in Brunner et al. (2018), we have

$$\begin{aligned} &E\left[\widehat{F}_{hgt}(x) - F_{hgt}(x)\right]^2 \leq \frac{1}{n_{hg}} \quad \text{and} \\ &E\left[\widehat{F}_{hgt}(X_{slri}) - F_{hgt}(X_{slri})\right]^2 \leq \frac{1}{n_{hg}}, i = 1, \ldots, n_{sl}. \end{aligned}$$

Then by $c_r$-inequality and Fubini's theorem, we have

$$\begin{aligned} E(A_i^2) \leq & 2E\left(\widehat{F}_{hgt}(X_{slri}) - F_{hgt}(X_{slri})^2\right) \\ & + 2E\left(\int \left(\widehat{F}_{hgt}(x) - F_{slri}(x)\right) dF_{slri}(x)\right)^2 \\ \leq & \frac{4}{n_{hg}}. \end{aligned}$$

Thus, we obtain that

$$E(B_1^2) = \frac{N}{n_{sl}^2} \sum_{i=1}^{n_{sl}} \sum_{j=1}^{n_{sl}} E(A_i A_j) = \frac{N}{n_{sl}^2} \sum_{i=1}^{n_{sl}} E(A_i^2) \leq \frac{4N}{n_{sl} n_{hg}}. \tag{4.32}$$

Therefore, under Assumption 4.3.1, (4.29) holds.

(ii) Proof of (4.30). Notice that

$$\begin{aligned} B_2 = & \sqrt{N} \int \widehat{F}_{hgt} \left(\widehat{F}_{slr} - F_{slr}\right) d\left(\widehat{F}_{uvw} - F_{uvw}\right) \\ & + \sqrt{N} \int F_{slr} \left(\widehat{F}_{hgt} - F_{hgt}\right) d\left(\widehat{F}_{uvw} - F_{uvw}\right) \\ = & B_{21} + B_{22}. \end{aligned}$$

By (4.32) we have

$$E(B_{2t}^2) \leq E\left(\left(\sqrt{N} \int \left(\widehat{F}_{slr} - F_{slr}\right) d\left(\widehat{F}_{uvw} - F_{uvw}\right)\right)^2\right) \leq \frac{4N}{n_{sl} n_{uv}}, \text{ for } t = 1, 2.$$

Therefore, under Assumption 4.3.1, $B_{21} \xrightarrow{P} 0$ and $B_{22} \xrightarrow{P} 0$. By the linearity of convergence in probability, (4.30) hold.

(iii) Proof of (4.31). By definition, we have

$$B_3 = \frac{\sqrt{N}}{n_{hg}n_{sl}} \sum_{i=1}^{n_{hg}} \sum_{j=1}^{n_{sl}} \int \left(c(x - X_{hgti}) - F_{hgt}(x)\right) \left(c(x - X_{slrj}) - F_{slr}(x)\right) dF_{uvw}$$
$$= \frac{\sqrt{N}}{n_{hg}n_{sl}} \sum_{i=1}^{n_{hg}} \sum_{j=1}^{n_{sl}} A_{ij}.$$

If $(h, g) \neq (s, l)$, $X_{hgti}$ and $X_{slkj}$ are independent, by Fubini's theorem, we have

$$E(A_{ij}) = 0, \text{ for } i = 1, \cdots, n_{hg} \text{ and } j = 1, \cdots, n_{sl}.$$

Also

$$E(A_{ij}A_{i'j'}) = 0, \text{if } (i, j) \neq (i'j'), \text{ and } 0 \leq E(A_{ij}^2) \leq 1,$$

for $i, i' = 1, \cdots, n_{hg}$ and $j, j' = 1, \cdots, n_{sl}$. Therefore,

$$E(B_3^2) = \frac{N}{n_{hg}^2 n_{sl}^2} \sum_{i,i'=1}^{n_{hg}} \sum_{j,j'=1}^{n_{sl}} E(A_{ij}A_{i'j'}) \tag{4.33}$$
$$= \frac{N}{n_{hg}^2 n_{sl}^2} \sum_{i=1}^{n_{hg}} \sum_{j=1}^{n_{sl}} E(A_{ij}^2) \leq \frac{N}{n_{hg}n_{sl}}.$$

On the other hand, if $(h, g) = (s, l)$, we still have

$$E(A_{ij}) = 0 \text{ and } E(A_{ii}A_{ij}) = 0 \text{ if } i \neq j, \text{ and}$$
$$-1 \leq E(A_{ii}) \leq 1, \text{ for } i, j = 1, \cdots, n_{hg}.$$

$$E(A_{ij}A_{i'j'}) = 0, \text{if } (i, j) \neq (i'j'), \text{ and}$$
$$0 \leq E(A_{ij}^2) \leq 1, \text{ for } i, i', j, j' = 1, \cdots, n_{hg}.$$

Therefore,

$$E(B_3) = \frac{N}{n_{hg}^4} \sum_{i,j,i',j'=1}^{n_{hg}} E(A_{ij}A_{i'j'}) \tag{4.34}$$
$$= \frac{N}{n_{hg}^4} \sum_{i,j=1}^{n_{hg}} E(A_{ii}A_{jj}) + \frac{N}{n_{hg}^4} \sum_{i \neq j}^{n_{hg}} E(A_{ij}^2) \leq \frac{2N}{n_{hg}^2}.$$

Combining (4.33) and (4.34), and under Assumption 4.3.1, (4.31) hold.

□

**Proofs**

In this subsection we provide detailed proofs and calculations for theoretical results in Sections 4.3 and 4.4.

*Proof of Proposition 4.3.1.* Under Assumption 4.3.3, it is suffice to consider the consistency of the first part of $\widehat{\delta}_g$. By the linearity property of convergence in probability and the results of Lemma 4.8.2, we can show that

$$\int (\widehat{F}_{1gt} - \widehat{F}_{2gt})(\widehat{F}_{2g't} - \widehat{F}_{2gt}) d\widehat{H} \xrightarrow{P} \int (F_{1gt} - F_{2gt})(F_{2g't} - F_{2gt}) dH,$$

$$\int (\widehat{F}_{2gt} - \widehat{F}_{2g't})^2 d\widehat{H} \xrightarrow{P} \int (F_{2gt} - F_{2g't})^2 dH,$$

where $g \neq g'$ and $g', g, t = 1, 2$. Then by continuous mapping theorem, we have

$$\widehat{\delta}_g \xrightarrow{P} \delta_g,$$

i.e, $\delta_g$, $g = 1, 2$, are consistent estimator of $\delta_g$. □

*Proof of Proposition 4.3.2.* By definition of $\widehat{\delta}_g$ in (4.10), we have

$$\begin{aligned}\widehat{\delta}_g - \delta_g =& \frac{\sum_{t=1}^{2} \int (\widehat{F}_{1gt} - \widehat{F}_{2gt})(\widehat{F}_{2g't} - \widehat{F}_{2gt}) d\widehat{H}}{\sum_{t=1}^{2} \int (\widehat{F}_{21t} - \widehat{F}_{22t})^2 d\widehat{H}} \\ & - \frac{\sum_{t=1}^{2} \int (F_{1gt} - F_{2gt})(F_{2g't} - F_{2gt}) dH}{\sum_{t=1}^{2} \int (F_{21t} - F_{22t})^2 dH},\end{aligned}$$

where $H = \frac{\sum_{h=1}^{2} \sum_{g=1}^{2} \sum_{t=1}^{2} n_{hg} F_{hgt}}{N}$. Set

$$\begin{aligned}\widehat{A}_{gt} &= \int (\widehat{F}_{1gt} - \widehat{F}_{2gt})(\widehat{F}_{2g't} - \widehat{F}_{2gt}) d\widehat{H}, \quad & A_{gt} &= \int (F_{1gt} - F_{2gt})(F_{2g't} - F_{2gt}) dH, \\ \widehat{S}_t &= \int (\widehat{F}_{21t} - \widehat{F}_{22t})^2 d\widehat{H}, & S_t &= \int (F_{21t} - F_{22t})^2 dH.\end{aligned}$$

where $g, t = 1, 2$. Then we have

$$\begin{aligned}E\left(\widehat{\delta}_g\right) - \delta_g =& E\left(\frac{\widehat{A}_{g1} + \widehat{A}_{g2}}{\widehat{S}_1 + \widehat{S}_2}\right) - \frac{A_{g1} + A_{g2}}{S_1 + S_2} \\ =& E\left(\left(\widehat{A}_{g1} + \widehat{A}_{g2}\right)\left(\frac{1}{\widehat{S}_1 + \widehat{S}_2} - \frac{1}{S_1 + S_2}\right)\right) \\ & + \frac{1}{S_1 + S_2} E\left(\widehat{A}_{g1} + \widehat{A}_{g2} - A_{g1} - A_{g2}\right).\end{aligned}$$

Since $\widehat{A}_{gt}$, $\widehat{S}_t$, $g,t=1,2$, can be expressed as finite summation of $\int \widehat{F}_{hgt}\widehat{F}_{slr}d\widehat{F}_{uvw}$, from Lemma 4.8.1, we have

$$E\left(\widehat{A}_{gt}\right)-A_{gt}=O(N^{-1}),\quad E\left(\widehat{S}_t\right)-S_t=O(N^{-1}).$$

Notice that $-1\leq\widehat{A}_{gt}\leq 1$, we have

$$\begin{aligned}&-2E\left(\frac{1}{\widehat{S}_1+\widehat{S}_2}-\frac{1}{S_1+S_2}\right)\\ \leq&E\left(\left(\widehat{A}_{g1}+\widehat{A}_{g2}\right)\left(\frac{1}{\widehat{S}_1+\widehat{S}_2}-\frac{1}{S_1+S_2}\right)\right)\\ \leq&2E\left(\frac{1}{\widehat{S}_1+\widehat{S}_2}-\frac{1}{S_1+S_2}\right).\end{aligned}$$

By Taylor expansion,

$$\begin{aligned}&E\left(\frac{1}{\widehat{S}_1+\widehat{S}_2}-\frac{1}{S_1+S_2}\right)\\ =&E\left(\frac{1}{S_1+S_2-(S_1+S_2-\widehat{S}_1-\widehat{S}_2)}-\frac{1}{S_1+S_2}\right)\\ =&\frac{1}{S_1+S_2}E\left(\frac{S_1-\widehat{S}_1+S_2-\widehat{S}_2}{S_1+S_2}+o\left(\frac{S_1-\widehat{S}_1+S_2-\widehat{S}_2}{S_1+S_2}\right)\right)\\ =&\frac{1}{(S_1+S_2)^2}O(N^{-1}).\end{aligned}$$

By Assumption 4.3.2, we have $S_1+S_2>C$, therefore $\frac{1}{S_1+S_2}<\frac{1}{C}$, it follows that

$$E(\widehat{\delta}_g)-\delta_g=O(N^{-1}).$$

□

*Proof of Theorem 4.3.1.* By Lemma 4.8.3, we have

$$
\begin{aligned}
&\sqrt{N}(\int \widehat{F}_{hgt}\widehat{F}_{slk}d\widehat{F}_{uvw} - \int F_{hgt}F_{slk}dF_{uvw}) \\
=&\sqrt{N}\int F_{hgt}F_{slk}d\widehat{F}_{uvw} + \sqrt{N}\int \widehat{F}_{hgt}\widehat{F}_{slk}dF_{uvw} - 2\sqrt{N}\int F_{hgt}F_{slk}dF_{uvw} + o_p(1) \\
=&\sqrt{N}\int F_{hgt}F_{slk}d\widehat{F}_{uvw} + \sqrt{N}\int F_{hgt}\widehat{F}_{slk}dF_{uvw} + \sqrt{N}\int \widehat{F}_{hgt}F_{slk}dF_{uvw} \qquad (4.35) \\
&- 3\sqrt{N}\int F_{hgt}F_{slk}dF_{uvw} + o_p(1) \\
=&\frac{\sqrt{N}}{n_{uv}}\sum_{i=1}^{n_{uv}} F_{hgt}(X_{uvwi})F_{slk}(X_{uvwi}) + \frac{\sqrt{N}}{n_{sl}}\sum_{i=1}^{n_{sl}}\int c(x - X_{slki})F_{hgt}(x)dF_{uvw}(x) \\
&+ \frac{1}{n_{hg}}\sum_{i=1}^{n_{hg}}\int c(x - X_{hgti})F_{slk}(x)dF_{uvw}(x) - 3\sqrt{N}\int F_{hgt}F_{slk}dF_{uvw} + o_p(1).
\end{aligned}
$$

By definition of $\widehat{\delta}_\ell$,

$$
\begin{aligned}
&\sqrt{N}(\widehat{\delta}_\ell - \delta_\ell) \\
=&\sqrt{N}\left(\frac{\widehat{A}_{\ell 1} + \widehat{A}_{\ell 2}}{\widehat{S}_1 + \widehat{S}_2} - \frac{A_{\ell 1} + A_{\ell 2}}{S_1 + S_2}\right) \\
=&\sqrt{N}\left(\frac{\left(\widehat{A}_{\ell 1} - A_{\ell 1}\right) + \left(\widehat{A}_{\ell 2} - A_{\ell 1}\right)}{\widehat{S}_1 + \widehat{S}_2} + (A_{\ell 1} + A_{\ell 2})\left(\frac{1}{\widehat{S}_1 + \widehat{S}_2} - \frac{1}{S_1 + S_2}\right)\right) \\
=&\sqrt{N}\left[\left(\widehat{A}_{\ell 1} - A_{\ell 1}\right) + \left(\widehat{A}_{\ell 2} - A_{\ell 1}\right)\right]\left(\frac{1}{S_1 + S_2} + \frac{(S_1 - \widehat{S}_1) + (S_2 - \widehat{S}_2)}{(S_1 + S_2)^2}\right. \\
&\left. + o\left(\frac{(S_1 - \widehat{S}_1) + (S_2 - \widehat{S}_2)}{(S_1 + S_2)^2}\right)\right) \\
&+ \sqrt{N}(A_{\ell 1} + A_{\ell 2})\left(\frac{(S_1 - \widehat{S}_1) + (S_2 - \widehat{S}_2)}{(S_1 + S_2)^2} + o\left(\frac{(S_1 - \widehat{S}_1) + (S_2 - \widehat{S}_2)}{(S_1 + S_2)^2}\right)\right) \\
=&\sqrt{N}\left[\frac{\left(\widehat{A}_{\ell 1} - A_{\ell 1}\right) + \left(\widehat{A}_{\ell 2} - A_{\ell 1}\right)}{S_1 + S_2} + \frac{A_{\ell 1} + A_{\ell 2}}{(S_1 + S_2)^2}\left((S_1 - \widehat{S}_1) + (S_2 - \widehat{S}_2)\right)\right] \\
&+ o_p(1).
\end{aligned}
$$

By (4.35), we have

$$\begin{aligned}\sqrt{N}(\widehat{A}_{\ell t} - A_{\ell t}) =& \sqrt{N}\int (F_{1\ell t} - F_{2\ell t})(F_{2\ell' t} - F_{2\ell t})d\widehat{H} \\ &+ \sqrt{N}\int (\widehat{F}_{1\ell t} - \widehat{F}_{2\ell t})(F_{2\ell' t} - F_{2\ell t})dH \\ &+ \sqrt{N}\int (F_{1\ell t} - F_{2\ell t})(\widehat{F}_{2\ell' t} - \widehat{F}_{2\ell t})dH \\ &- 3\sqrt{N}\int (F_{1\ell t} - F_{2\ell t})(F_{2\ell' t} - F_{2\ell t})dH + o_p(1),\end{aligned} \tag{4.36}$$

where $\ell' \neq \ell$, $\ell', \ell = 1, 2$.

$$\begin{aligned}\sqrt{N}(\widehat{S}_t - S_t) =& \sqrt{N}\int (F_{21t} - F_{22t})^2 d\widehat{H} + 2\sqrt{N}\int (\widehat{F}_{21t} - \widehat{F}_{22t})(F_{21t} - F_{22t})dH \\ &- 3\sqrt{N}\int (F_{21t} - F_{22t})^2 dH + o_p(1).\end{aligned} \tag{4.37}$$

Set

$$\begin{aligned}\widetilde{A}^1_{\ell t} &= \int (F_{1\ell t} - F_{2\ell t})(F_{2\ell' t} - F_{2\ell t})d\widehat{H}, \\ \widetilde{A}^2_{\ell t} &= \int (\widehat{F}_{1\ell t} - \widehat{F}_{2\ell t})(F_{2\ell' t} - F_{2\ell t})dH, \\ \widetilde{A}^3_{\ell t} &= \int (F_{1\ell t} - F_{2\ell t})(\widehat{F}_{2\ell' t} - \widehat{F}_{2\ell t})dH, \\ \widetilde{S}^1_t &= \int (F_{21t} - F_{22t})^2 d\widehat{H}, \\ \widetilde{S}^2_t &= \int (\widehat{F}_{21t} - \widehat{F}_{22t})(F_{21t} - F_{22t})dH.\end{aligned}$$

Utilizing results (4.36) and (4.37), we have

$$\begin{aligned}\sqrt{N}(\widehat{\delta}_\ell - \delta_\ell) =& \frac{1}{S_1 + S_2}\sqrt{N}\left[\widetilde{A}^1_{\ell 1} + \widetilde{A}^2_{\ell 1} + \widetilde{A}^3_{\ell 1} + \widetilde{A}^1_{\ell 2} + \widetilde{A}^2_{\ell 2} + \widetilde{A}^3_{\ell 3}\right] \\ &- \frac{A_{\ell 1} + A_{\ell 2}}{(S_1 + S_2)^2}\sqrt{N}\left[\widetilde{S}^1_1 + 2\widetilde{S}^2_1 + \widetilde{S}^1_2 + 2\widetilde{S}^2_2\right] + o_p(1).\end{aligned}$$

By definition, we have

$$\begin{aligned}
&\widetilde{A}^1_{11}+\widetilde{A}^2_{11}+\widetilde{A}^3_{11}+\widetilde{A}^1_{12}+\widetilde{A}^2_{12}+\widetilde{A}^3_{12}\\
=&\int (F_{111}-F_{211})(F_{221}-F_{211})+(F_{112}-F_{212})(F_{222}-F_{212})d\widehat{H}\\
&+\int \widehat{F}_{111}(F_{221}-F_{211})dH+\int \widehat{F}_{112}(F_{222}-F_{212})dH+\int \widehat{F}_{221}(F_{111}-F_{211})dH\\
&+\int \widehat{F}_{222}(F_{112}-F_{212})dH-\int \widehat{F}_{211}(F_{221}+F_{111}-2F_{211})dH\\
&-\int \widehat{F}_{212}(F_{222}+F_{112}-2F_{212})dH.
\end{aligned}$$

$$\begin{aligned}
&\widetilde{A}^1_{21}+\widetilde{A}^2_{21}+\widetilde{A}^3_{21}+\widetilde{A}^1_{22}+\widetilde{A}^2_{22}+\widetilde{A}^3_{22}\\
=&\int (F_{121}-F_{221})(F_{211}-F_{221})+(F_{122}-F_{222})(F_{212}-F_{222})d\widehat{H}\\
&+\int \widehat{F}_{121}(F_{211}-F_{221})dH+\int \widehat{F}_{122}(F_{212}-F_{222})dH+\int \widehat{F}_{211}(F_{121}-F_{221})dH\\
&+\int \widehat{F}_{212}(F_{122}-F_{222}dH)-\int \widehat{F}_{221}(F_{211}+F_{121}-2F_{221})dH\\
&-\int \widehat{F}_{222}(F_{212}+F_{122}-2F_{222})dH.
\end{aligned}$$

$$\begin{aligned}
&\widetilde{S}^1_1+2\widetilde{S}^2_1+\widetilde{S}^1_2+2\widetilde{S}^2_2\\
=&\int (F_{211}-F_{221})^2+(F_{212}-F_{222})^2 d\widehat{H}+2\int \widehat{F}_{211}(F_{211}-F_{221})dH\\
&+2\int \widehat{F}_{212}(F_{212}-F_{222})dH-2\int \widehat{F}_{221}(F_{211}-F_{221})dH-2\int \widehat{F}_{222}(F_{212}-F_{222})dH.
\end{aligned}$$

Set

$$\begin{aligned}
G_1&=(F_{111}-F_{211})(F_{221}-F_{211})+(F_{112}-F_{212})(F_{222}-F_{212}),\\
G_2&=(F_{121}-F_{221})(F_{211}-F_{221})+(F_{122}-F_{222})(F_{212}-F_{222}),\\
G_3&=(F_{211}-F_{221})^2+(F_{212}-F_{222})^2.
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\widetilde{A}^1_{11}+\widetilde{A}^2_{11}+\widetilde{A}^3_{11}+\widetilde{A}^1_{12}+\widetilde{A}^2_{12}+\widetilde{A}^3_{12}\\
=&\frac{1}{n_{11}}\sum_{k=1}^{n_{11}}U_{111}(\boldsymbol{X}_{11k})+\frac{1}{n_{12}}\sum_{k=1}^{n_{12}}U_{112}(\boldsymbol{X}_{12k})+\frac{1}{n_{21}}\sum_{k=1}^{n_{21}}U_{121}(\boldsymbol{X}_{21k})+\frac{1}{n_{22}}\sum_{k=1}^{n_{22}}U_{122}(\boldsymbol{X}_{22k}),
\end{aligned}$$

where

$$
\begin{aligned}
U_{111}(\boldsymbol{X}_{11k}) =& \frac{n_{11}}{N} G_1(X_{111k}) + \frac{n_{11}}{N} G_1(X_{112k}) \\
&+ \int c(x - X_{111k})(F_{221}(x) - F_{211}(x)) dH(x) \\
&+ \int c(x - X_{112k})(F_{222}(x) - F_{212}(x)) dH(x), \\
U_{112}(\boldsymbol{X}_{12k}) =& \frac{n_{12}}{N} G_1(X_{121k}) + \frac{n_{12}}{N} G_1(X_{122k}), \\
U_{121}(\boldsymbol{X}_{21k}) =& \frac{n_{21}}{N} G_1(X_{211k}) + \frac{n_{21}}{N} G_1(X_{212k}) \\
&- \int c(x - X_{211k})(F_{221}(x) + F_{111}(x) - 2F_{211}(x)) dH \\
&- \int c(x - X_{212k})(F_{222}(x) + F_{112}(x) - 2F_{212}(x)) dH, \\
U_{122}(\boldsymbol{X}_{22k}) =& \frac{n_{22}}{N} G_1(X_{221k}) + \frac{n_{22}}{N} G_1(X_{222k}) \\
&+ \int c(x - X_{221k})(F_{111}(x) - F_{211}(x)) dH(x) \\
&+ \int c(x - X_{222k})(F_{112}(x) - F_{212}(x)) dH(x).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&\widetilde{A}_{21}^1 + \widetilde{A}_{21}^2 + \widetilde{A}_{21}^3 + \widetilde{A}_{22}^1 + \widetilde{A}_{22}^2 + \widetilde{A}_{22}^3 \\
=& \frac{1}{n_{11}} \sum_{k=1}^{n_{11}} U_{211}(\boldsymbol{X}_{11k}) + \frac{1}{n_{12}} \sum_{k=1}^{n_{12}} U_{212}(\boldsymbol{X}_{12k}) + \frac{1}{n_{21}} \sum_{k=1}^{n_{21}} U_{221}(\boldsymbol{X}_{21k}) + \frac{1}{n_{22}} \sum_{k=1}^{n_{22}} U_{222}(\boldsymbol{X}_{22k}),
\end{aligned}
$$

where

$$
\begin{aligned}
U_{211}(\boldsymbol{X}_{11k}) =& \frac{n_{11}}{N} G_2(X_{111k}) + \frac{n_{11}}{N} G_2(X_{112k}), \\
U_{212}(\boldsymbol{X}_{12k}) =& \frac{n_{12}}{N} G_2(X_{121k}) + \frac{n_{12}}{N} G_2(X_{122k}) \\
&+ \int c(x - X_{121k})(F_{211}(x) - F_{221}(x)) dH(x) \\
&+ \int c(x - X_{122k})(F_{212}(x) - F_{222}(x)) dH(x), \\
U_{221}(\boldsymbol{X}_{21k}) =& \frac{n_{21}}{N} G_2(X_{211k}) + \frac{n_{21}}{N} G_2(X_{212k}) \\
&+ \int c(x - X_{211k})(F_{121}(x) - F_{221}(x)) dH(x) \\
&+ \int c(x - X_{212k})(F_{122}(x) - F_{222}(x)) dH(x), \\
U_{222}(\boldsymbol{X}_{22k}) =& \frac{n_{22}}{N} G_2(X_{221k}) + \frac{n_{22}}{N} G_2(X_{222k}) \\
&- \int c(x - X_{221k})(F_{211}(x) + F_{121}(x) - 2F_{221}(x)) dH \\
&- \int c(x - X_{222k})(F_{212}(x) + F_{122}(x) - 2F_{222}(x)) dH.
\end{aligned}
$$

Also,

$$
\begin{aligned}
\widetilde{S}_1^1 + 2\widetilde{S}_1^2 + \widetilde{S}_2^1 + 2\widetilde{S}_2^2 =& \frac{1}{n_{11}} \sum_{k=1}^{n_{11}} U_{311}(\boldsymbol{X}_{11k}) + \frac{1}{n_{12}} \sum_{k=1}^{n_{12}} U_{312}(\boldsymbol{X}_{12k}) \\
&+ \frac{1}{n_{21}} \sum_{k=1}^{n_{21}} U_{321}(\boldsymbol{X}_{21k}) + \frac{1}{n_{22}} \sum_{k=1}^{n_{22}} U_{322}(\boldsymbol{X}_{22k}),
\end{aligned}
$$

where

$$
\begin{aligned}
U_{311}(\boldsymbol{X}_{11k}) =& \frac{n_{11}}{N} G_3(X_{111k}) + \frac{n_{11}}{N} G_3(X_{112k}), \\
U_{312}(\boldsymbol{X}_{12k}) =& \frac{n_{12}}{N} G_3(X_{121k}) + \frac{n_{12}}{N} G_3(X_{122k}), \\
U_{321}(\boldsymbol{X}_{21k}) =& \frac{n_{21}}{N} G_3(X_{211k}) + \frac{n_{21}}{N} G_3(X_{212k}) \\
& + 2 \int c(x - X_{211k})(F_{211}(x) - F_{221}(x)) dH(x) \\
& + 2 \int c(x - X_{212k})(F_{212}(x) - F_{222}(x)) dH(x), \\
U_{322}(\boldsymbol{X}_{22k}) =& \frac{n_{22}}{N} G_3(X_{221k}) + \frac{n_{22}}{N} G_3(X_{222k}) \\
& - 2 \int c(x - X_{221k})(F_{211}(x) - F_{222}(x)) dH \\
& - 2 \int c(x - X_{222k})(F_{212}(x) - F_{222}(x)) dH.
\end{aligned}
$$

Therefore, we have

$$
\begin{aligned}
\sqrt{N}\left(\widehat{\delta}_\ell - \delta_\ell\right) =& \frac{\sqrt{N}}{n_{11}} \sum_{k=1}^{n_{11}} V_{\ell 11}(\boldsymbol{X}_{11k}) + \frac{\sqrt{N}}{n_{12}} \sum_{k=1}^{n_{12}} V_{\ell 12}(\boldsymbol{X}_{12k}) \\
& + \frac{\sqrt{N}}{n_{21}} \sum_{k=1}^{n_{21}} V_{\ell 21}(\boldsymbol{X}_{21k}) + \frac{\sqrt{N}}{n_{22}} \sum_{k=1}^{n_{22}} V_{\ell 22}(\boldsymbol{X}_{22k}) + o_p(1),
\end{aligned}
$$

where

$$
V_{\ell hg}(\boldsymbol{X}_{hgk}) = \frac{1}{S_1 + S_2} U_{\ell hg}(\boldsymbol{X}_{hgk}) - \frac{A_{\ell 1} + A_{\ell 2}}{(S_1 + S_2)^2} U_{3hg}(\boldsymbol{X}_{hgk}). \tag{4.38}
$$

Utilizing the Central Limit Theorem and independence between random variables, by Assumption 4.3.4, we can obtain that

$$
\sqrt{N}(\widehat{\delta}_\ell - \delta_\ell) \xrightarrow{D} U \sim N\left(0, \kappa_{11}^{-1}\sigma_{\ell 11}^2 + \kappa_{12}^{-1}\sigma_{\ell 12}^2 + \kappa_{21}^{-1}\sigma_{\ell 21}^2 + \kappa_{22}^{-1}\sigma_{\ell 22}^2\right),
$$

where

$$
\sigma_{\ell hg}^2 = Var(V_{\ell hg}(\boldsymbol{X}_{hg1})), \quad \ell, h, g = 1, 2.
$$

□

*Proof of Theorem 4.3.2.* Since $\widehat{\sigma}_{\ell hg} \xrightarrow{P} \sigma_{\ell hg}$ for $g, h, t = 1, 2$, we only need to show that $S_{\ell hg}^2 - \widehat{\sigma}_{\ell hg} \xrightarrow{P} 0$. It suffices to show that

$$
E[S_{\ell hg}^2 - \widehat{\sigma}_{\ell hg}]^2 \to 0 \text{ as } N \to \infty \text{ for } g, h, t = 1, 2.
$$

By definition of variance, we have

$$
\begin{aligned}
&E[S^2_{\ell hg} - \widehat{\sigma}_{\ell hg}]^2 \\
=&E\Bigg[\frac{1}{n_{hg}-1}\sum_{i=1}^{n_{hg}}\left(\widehat{V}_{\ell hg}(\boldsymbol{X}_{hgi}) - \overline{\widehat{V}}_{\ell hg}(\boldsymbol{X}_{hg\cdot})\right)^2 \\
&-\frac{1}{n_{hg}-1}\sum_{i=1}^{n_{hg}}\left(V_{\ell hg}(\boldsymbol{X}_{hgi}) - \overline{V}_{\ell hg}(\boldsymbol{X}_{hg\cdot})\right)^2\Bigg]^2 \\
=&\frac{1}{(n_{hg}-1)^2}E\Bigg[\sum_{i=1}^{n_{hg}}\left(\widehat{V}_{\ell hg}(\boldsymbol{X}_{hgi}) - \overline{\widehat{V}}_{\ell hg}(\boldsymbol{X}_{hg\cdot}) + V_{\ell hg}(\boldsymbol{X}_{hgi}) - \overline{V}_{\ell hg}(\boldsymbol{X}_{hg\cdot})\right) \\
&\times\left(\widehat{V}_{\ell hg}(\boldsymbol{X}_{hgi}) - V_{\ell hg}(\boldsymbol{X}_{hgi}) - \overline{\widehat{V}}_{\ell hg}(\boldsymbol{X}_{hg\cdot}) + \overline{V}_{\ell hg}(\boldsymbol{X}_{hg\cdot})\right)\Bigg]^2.
\end{aligned}
$$

Suppose

$$
\widehat{S}_1 + \widehat{S}_2 = \int(\widehat{F}_{211} - \widehat{F}_{221})^2 d\widehat{H} + \int(\widehat{F}_{212} - \widehat{F}_{222})^2\widehat{H} > C', \tag{4.39}
$$

where $0 < C' < C$ and $C$ is the constant in assumption 4.3.2. Then by Assumption 4.3.2 and range of distribution functions, $V_{\ell hg}(\boldsymbol{X}_{hgi})$ and $\widehat{V}_{\ell hg}(\boldsymbol{X}_{hgi})$ are bounded and we can find a constant $M > 0$, such that

$$
\left(\widehat{V}_{\ell hg}(\boldsymbol{X}_{hgi}) - \overline{\widehat{V}}_{\ell hg}(\boldsymbol{X}_{hg\cdot}) + V_{\ell hg}(\boldsymbol{X}_{hgi}) - \overline{V}_{\ell hg}(\boldsymbol{X}_{hg\cdot})\right)^2 \leq M.
$$

Therefore,

$$
\begin{aligned}
&E[S^2_{\ell hg} - \widehat{\sigma}_{\ell hg}]^2 \\
\leq&\frac{Mn_{hg}}{(n_{hg}-1)^2}E\left[\sum_{i=1}^{n_{hg}}\left(\widehat{V}_{\ell hg}(\boldsymbol{X}_{hgi}) - V_{\ell hg}(\boldsymbol{X}_{hgi}) - \overline{\widehat{V}}_{\ell hg}(\boldsymbol{X}_{hg\cdot}) + \overline{V}_{\ell hg}(\boldsymbol{X}_{hg\cdot})\right)\right]^2 \\
=&\frac{Mn_{hg}}{(n_{hg}-1)^2}E\left[\sum_{i=1}^{n_{hg}}\left(\widehat{V}_{\ell hg}(\boldsymbol{X}_{hgi}) - V_{\ell hg}(\boldsymbol{X}_{hgi})\right)^2\right] \\
&-\frac{Mn_{hg}^2}{(n_{hg}-1)^2}E\left[\left(\overline{\widehat{V}}_{\ell hg}(\boldsymbol{X}_{hg\cdot}) - \overline{V}_{\ell hg}(\boldsymbol{X}_{hg\cdot})\right)^2\right] \\
\leq&\frac{Mn_{hg}}{(n_{hg}-1)^2}E\left[\sum_{i=1}^{n_{hg}}\left(\widehat{V}_{\ell hg}(\boldsymbol{X}_{hgi}) - V_{\ell hg}(\boldsymbol{X}_{hgi})\right)^2\right].
\end{aligned} \tag{4.40}
$$

Notice that

$$
\begin{aligned}
& E\left(\widehat{V}_{\ell hg}(\boldsymbol{X}_{hgi}) - V_{\ell hg}(\boldsymbol{X}_{hgi})\right)^2 \\
=& E\Bigg(\frac{1}{\widehat{S}_1+\widehat{S}_2}\widehat{U}_{\ell hg}(\boldsymbol{X}_{hgi}) - \frac{\widehat{A}_{\ell 1}+\widehat{A}_{\ell 2}}{(\widehat{S}_1+\widehat{S}_2)^2}\widehat{U}_{3ht}(\boldsymbol{X}_{hgi}) \\
& - \frac{1}{S_1+S_2}U_{\ell hg}(\boldsymbol{X}_{hgi}) + \frac{A_{\ell 1}+A_{\ell 2}}{(S_1+S_2)^2}U_{3ht}(\boldsymbol{X}_{hgi})\Bigg)^2 \\
\le& 2E\left(\frac{1}{\widehat{S}_1+\widehat{S}_2}\widehat{U}_{\ell hg}(\boldsymbol{X}_{hgi}) - \frac{1}{S_1+S_2}U_{\ell hg}(\boldsymbol{X}_{hgi})\right)^2 \\
& + 2E\left(\frac{\widehat{A}_{\ell 1}+\widehat{A}_{\ell 2}}{(\widehat{S}_1+\widehat{S}_2)^2}\widehat{U}_{3ht}(\boldsymbol{X}_{hgi}) - \frac{A_{\ell 1}+A_{\ell 2}}{(S_1+S_2)^2}U_{3ht}(\boldsymbol{X}_{hgi})\right)^2 \\
\le& 4E\left(\frac{1}{\widehat{S}_1+\widehat{S}_2}\left(\widehat{U}_{\ell hg}(\boldsymbol{X}_{hgi}) - U_{\ell hg}(\boldsymbol{X}_{hgi})\right)\right)^2 \\
& + 4E\left(\left(\frac{1}{\widehat{S}_1+\widehat{S}_2} - \frac{1}{S_1+S_2}\right)U_{\ell hg}(\boldsymbol{X}_{hgi})\right)^2 \\
& + 4E\left(\frac{\widehat{A}_{\ell 1}+\widehat{A}_{\ell 2}}{(\widehat{S}_1+\widehat{S}_2)^2}\left(\widehat{U}_{3ht}(\boldsymbol{X}_{hgi}) - U_{3ht}(\boldsymbol{X}_{hgi})\right)\right)^2 \\
& + 4E\left(\left(\frac{\widehat{A}_{\ell 1}+\widehat{A}_{\ell 2}}{(\widehat{S}_1+\widehat{S}_2)^2} - \frac{A_{\ell 1}+A_{\ell 2}}{(S_1+S_2)^2}\right)U_{3ht}(\boldsymbol{X}_{hgi})\right)^2.
\end{aligned}
$$

By Lemma 4.8.2 and the $c_r$-inequality, we have

$$
\begin{aligned}
& E\left(\widehat{U}_{lhg}(\boldsymbol{X}_{hgi}) - U_{lhg}(\boldsymbol{X}_{hgi})\right)^2 = O(N^{-1}), \\
& E\left(\widehat{A}_{\ell t} - A_{\ell t}\right)^2 = O(N^{-1}), \quad E\left(\widehat{S}_t - S_t\right)^2 = O(N^{-1}),
\end{aligned}
$$

for $l = 1, 2, 3$, and $\ell, g, h, t = 1, 2$. Under the assumption in (4.39),

$$
0 \le \frac{1}{\widehat{S}_1+\widehat{S}_2} \le \frac{1}{C'}, \quad -\frac{2}{C^2} \le \frac{\widehat{A}_{\ell 1}+\widehat{A}_{\ell 2}}{(\widehat{S}_1+\widehat{S}_2)^2} \le \frac{2}{C'^2}.
$$

$$
\begin{aligned}
& E\left(\frac{1}{\widehat{S}_1+\widehat{S}_2} - \frac{1}{S_1+S_2}\right)^2 \\
&= \frac{1}{(S_1+S_2)^2}E\left(\frac{S_1-\widehat{S}_1+S_2-\widehat{S}_2}{S_1+S_2} + o\left(\frac{S_1-\widehat{S}_1+S_2-\widehat{S}_2}{S_1+S_2}\right)\right)^2 = O(N^{-1}).
\end{aligned}
$$

$$
\begin{aligned}
&E\left(\frac{\widehat{A}_{\ell 1}+\widehat{A}_{\ell 2}}{(\widehat{S}_1+\widehat{S}_2)^2}-\frac{A_{\ell 1}+A_{\ell 2}}{(S_1+S_2)^2}\right)^2\\
=&E\left(\left(\widehat{A}_{\ell 1}+\widehat{A}_{\ell 2}\right)\left(\frac{1}{\widehat{S}_1+\widehat{S}_2}-\frac{1}{S_1+S_2}\right)+\frac{1}{S_1+S_2}\left(\widehat{A}_{\ell 1}+\widehat{A}_{\ell 2}-A_{\ell 1}-A_{\ell 2}\right)\right)^2\\
\leq&8E\left(\frac{1}{\widehat{S}_1+\widehat{S}_2}-\frac{1}{S_1+S_2}\right)^2+\frac{1}{C^2}E\left(\widehat{A}_{\ell 1}-A_{\ell 1}\right)^2+\frac{1}{C^2}E\left(\widehat{A}_{\ell 2}-A_{\ell 2}\right)^2=O(N^{-1}).
\end{aligned}
$$

By the range of distribution function, we can find a constant $M_1$, such that

$$
-M_1\leq U_{\ell hg}(\boldsymbol{X}_{hgi})\leq M_1,\quad -M_1\leq \widehat{U}_{\ell hg}(\boldsymbol{X}_{hgi})\leq M_1,\ \text{for } g=1,2,3,\quad h,t=1,2.
$$

Combining these results, we have

$$
E\left(\widehat{V}_{\ell hg}(\boldsymbol{X}_{hgi})-V_{\ell hg}(\boldsymbol{X}_{hgi})\right)=O(N^{-1}).
$$

Thus,

$$
E[S^2_{\ell hg}-\widehat{\sigma}_{\ell hg}]^2\leq\frac{Mn_{hg}^2}{(n_{hg}-1)^2}E\left[\left(\widehat{V}_{\ell hg}(\boldsymbol{X}_{hg1})-V_{\ell hg}(\boldsymbol{X}_{hg1})\right)^2\right]=O(N^{-1}).
$$

Then when (4.39) holds, we have $\widehat{S}^2_{\ell hg}-\sigma^2_{\ell hg}\xrightarrow{P}0$. Notice that $\widehat{S}_1\xrightarrow{P}S_1$ and $\widehat{S}_2\xrightarrow{P}S_2$, then under Assumption 4.3.2, we have $\lim_{N\to\infty}P\left(\widehat{S}_1+\widehat{S}_2\leq C'\right)=0$. Thus, $\forall\epsilon$,

$$
\begin{aligned}
0\leq&\lim_{N\to\infty}P\left(|\widehat{S}^2_{\ell hg}-\sigma^2_{\ell hg}|>\epsilon\right)\\
\leq&\lim_{N\to\infty}P\left(\widehat{S}_1+\widehat{S}_2\leq C'\right)+\lim_{N\to\infty}P\left(\left|\widehat{S}^2_{\ell hg}-\sigma_{\ell hg}\right|>\epsilon,\widehat{S}_1+\widehat{S}_2>C'\right)=0.
\end{aligned}
$$

Hence, $\widehat{S}^2_{\ell hg}$ is a consistent estimator for $\sigma^2_{\ell hg}$. □

*Proof of Equation (4.15).* Using the equations (4.6), we have

$$
\begin{aligned}
&F_{211}+F_{222}=F_{111}+F_{122}+\frac{\delta_1}{1-\delta_1-\delta_2}(F_{111}-F_{121})-\frac{\delta_2}{1-\delta_1-\delta_2}(F_{112}-F_{122})\ \text{and}\\
&F_{212}+F_{221}=F_{112}+F_{121}+\frac{\delta_1}{1-\delta_1-\delta_2}(F_{112}-F_{122})-\frac{\delta_2}{1-\delta_1-\delta_2}(F_{111}-F_{121}).
\end{aligned}
\tag{4.41}
$$

Using (4.41), we have

$$
\begin{aligned}
&\int (F_{211}+F_{222})d(F_{212}+F_{221}) \\
=&\int (F_{111}+F_{122})d(F_{112}+F_{121}) + \frac{\delta_1}{1-\delta_1-\delta_2}\int (F_{111}-F_{121})d(F_{112}+F_{121}) \\
&- \frac{\delta_2}{1-\delta_2-\delta_1}\int (F_{112}-F_{122})d(F_{112}+F_{121}) \\
&+ \frac{\delta_1}{1-\delta_1-\delta_2}\int (F_{111}+F_{122})d(F_{112}-F_{122}) \\
&+ \left(\frac{\delta_1}{1-\delta_1-\delta_2}\right)^2\int (F_{111}-F_{121})d(F_{112}-F_{122}) \\
&- \frac{\delta_1\delta_2}{(1-\delta_2-\delta_1)^2}\int (F_{112}-F_{122})d(F_{112}-F_{122}) \\
&- \frac{\delta_2}{1-\delta_2-\delta_1}\int (F_{111}+F_{122})d(F_{111}-F_{121}) \\
&- \frac{\delta_1\delta_2}{(1-\delta_2-\delta_1)^2}\int (F_{111}-F_{121})d(F_{111}-F_{121}) \\
&+ \left(\frac{\delta_2}{1-\delta_1-\delta_2}\right)^2\int (F_{112}-F_{122})d(F_{111}-F_{121}).
\end{aligned}
$$

By the rule of partial integration, we have

$$
\begin{aligned}
\int (F_{1gh}+F_{1lk})d(F_{1g'h'}-F_{1l'k'}) &= -\int (F_{1g'h'}-F_{1l'k'})d(F_{1gh}+F_{1lk}), \\
\int (F_{1gh}-F_{1lk})d(F_{1g'h'}-F_{1l'k'}) &= -\int (F_{1g'h'}-F_{1l'k'})d(F_{1gh}-F_{1lk}), \\
\int (F_{1gh}+F_{1lk})d(F_{1g'h'}+F_{1l'k'}) &= 4-\int (F_{1g'h'}+F_{1l'k'})d(F_{1gh}+F_{1lk}),
\end{aligned}
$$

where $g, g', h, h', l, l', k, k' = 1, 2$.

Therefore, we have

$$
\begin{aligned}
&\int (F_{211}+F_{222})d(F_{212}+F_{221}) \\
=&\int (F_{111}+F_{122})d(F_{112}+F_{121}) \\
&+\frac{1}{1-\delta_1-\delta_2}\int(\delta_1 F_{111}+\delta_2 F_{122}-\delta_2 F_{112}-\delta_1 F_{121})d(F_{112}+F_{121}) \\
&+\frac{1}{1-\delta_1-\delta_2}\int(\delta_2 F_{111}+\delta_1 F_{122}-\delta_1 F_{112}-\delta_2 F_{121})d(F_{111}+F_{122}) \\
&+\frac{\delta_1^2-\delta_2^2}{(1-\delta_1-\delta_2)^2}\int(F_{111}-F_{121})d(F_{112}-F_{122}) \\
=&\frac{1}{1-\delta_1-\delta_2}\int(F_{111}+F_{122})d(F_{112}+F_{121})-\frac{2(\delta_1+\delta_2)}{1-\delta_1-\delta_2} \\
&+\left(\frac{\delta_1-\delta_2}{1-\delta_1-\delta_2}+\frac{\delta_1^2-\delta_2^2}{(1-\delta_1-\delta_2)^2}\right)\int(F_{111}-F_{121})d(F_{112}-F_{122}) \\
&+\frac{\delta_2-\delta_1}{1-\delta_1-\delta_2}\int(F_{112}-F_{122})d(F_{112}-F_{122}) \\
=&\frac{1}{1-\delta_1-\delta_2}\int(F_{111}+F_{122})d(F_{112}+F_{121})-\frac{2(\delta_1+\delta_2)}{1-\delta_1-\delta_2} \\
&+\frac{\delta_1-\delta_2}{(1-\delta_1-\delta_2)^2}\int(F_{111}-F_{121})d(F_{112}-F_{122}).
\end{aligned}
$$

Combining this result with (4.13) we have

$$
\begin{aligned}
p_I =&\frac{1}{2(1-\delta_1-\delta_2)}\int(F_{111}+F_{122})d(F_{112}+F_{121}) \\
&+\frac{\delta_1-\delta_2}{2(1-\delta_1-\delta_2)^2}\int(F_{111}-F_{121})d(F_{112}-F_{122})-\frac{1}{1-\delta_1-\delta_2}.
\end{aligned}
$$

□

*Proof of Proposition 4.4.1.* Applying Proposition 7.7 in Brunner et al. (2018), we have

$$\int \widehat{F}_{hgt}d\widehat{F}_{lij} \xrightarrow{P} \int F_{hgt}dF_{lij}.$$

Then by Proposition 4.3.1 and continuous mapping theorem, we obtain the result. □

*Proof of Proposition 4.4.2.* By definition of $p_I$ and $\widehat{p}_I$, we have

$$\widehat{p}_I - p_I = \frac{2(n_{11}+n_{12})}{N}\left(\widehat{p}_{I1}-p_I\right)+\frac{2(n_{21}+n_{22})}{N}\left(\widehat{p}_{I2}-p_I\right).$$

From Lemma 4.8.1, we have

$$E(\widehat{p}_{I2} - p_I) = O(N^{-1}). \tag{4.42}$$

For $\widehat{p}_{I1}$, by (4.15), we have

$$\begin{aligned}
&\widehat{p}_{I1} - p_I \\
=&\frac{1}{2(1-\widehat{\delta}_1-\widehat{\delta}_2)}\int(\widehat{F}_{111}+\widehat{F}_{122})d(\widehat{F}_{112}+\widehat{F}_{121}) \\
&-\frac{1}{2(1-\delta_1-\delta_2)}\int(F_{111}+F_{122})d(F_{112}+F_{121}) \\
&+\frac{\widehat{\delta}_1-\widehat{\delta}_2}{2(1-\widehat{\delta}_1-\widehat{\delta}_2)^2}\int(\widehat{F}_{111}-\widehat{F}_{121})d(\widehat{F}_{112}-\widehat{F}_{122}) \\
&-\frac{\delta_1-\delta_2}{2(1-\delta_1-\delta_2)^2}\int(F_{111}-F_{121})d(F_{112}-F_{122}) \\
&-\left(\frac{1}{1-\widehat{\delta}_1-\widehat{\delta}_2}-\frac{1}{1-\delta_1-\delta_2}\right) \\
=&\int(\widehat{F}_{111}+\widehat{F}_{122})d(\widehat{F}_{112}+\widehat{F}_{121})\left(\frac{1}{2(1-\widehat{\delta}_1-\widehat{\delta}_2)}-\frac{1}{2(1-\delta_1-\delta_2)}\right) \\
&+\frac{1}{2(1-\delta_1-\delta_2)}\left(\int(\widehat{F}_{111}+\widehat{F}_{122})d(\widehat{F}_{112}+\widehat{F}_{121})-\int(F_{111}+F_{122})d(F_{112}+F_{121})\right) \\
&+\int(\widehat{F}_{111}-\widehat{F}_{121})d(\widehat{F}_{112}-\widehat{F}_{122})\left(\frac{\widehat{\delta}_1-\widehat{\delta}_2}{2(1-\widehat{\delta}_1-\widehat{\delta}_2)^2}-\frac{\delta_1-\delta_2}{2(1-\delta_1-\delta_2)^2}\right) \\
&+\frac{\delta_1-\delta_2}{2(1-\delta_1-\delta_2)^2}\left(\int(\widehat{F}_{111}-\widehat{F}_{121})d(\widehat{F}_{112}-\widehat{F}_{122})-\int(F_{111}-F_{121})d(F_{112}-F_{122})\right) \\
&-\left(\frac{1}{1-\widehat{\delta}_1-\widehat{\delta}_2}-\frac{1}{1-\delta_1-\delta_2}\right) \\
=&\widehat{E}_1\left(\frac{1}{2(1-\widehat{\delta}_1-\widehat{\delta}_2)}-\frac{1}{2(1-\delta_1-\delta_2)}\right)+\frac{1}{2(1-\delta_1-\delta_2)}(\widehat{E}_1-E_1) \\
&+\widehat{E}_2\left(\frac{\widehat{\delta}_1-\widehat{\delta}_2}{2(1-\widehat{\delta}_1-\widehat{\delta}_2)^2}-\frac{\delta_1-\delta_2}{2(1-\delta_1-\delta_2)^2}\right)+\frac{\delta_1-\delta_2}{2(1-\delta_1-\delta_2)^2}(\widehat{E}_2-E_2) \\
&-\left(\frac{1}{1-\widehat{\delta}_1-\widehat{\delta}_2}-\frac{1}{1-\delta_1-\delta_2}\right),
\end{aligned}$$

where

$$\widehat{E}_1=\int(\widehat{F}_{111}+\widehat{F}_{122})d(\widehat{F}_{112}+\widehat{F}_{121}), \qquad E_1=\int(F_{111}+F_{122})d(F_{112}+F_{121}),$$
$$\widehat{E}_2=\int(\widehat{F}_{111}-\widehat{F}_{121})d(\widehat{F}_{112}-\widehat{F}_{122}), \qquad E_2=\int(F_{111}-F_{121})d(F_{112}-F_{122}).$$

By the range of empirical distribution function, we have $0 \leq \widehat{E}_1 \leq 4$ and $-2 \leq \widehat{E}_2 \leq 2$. From Lemma 4.8.1, we have

$$E(\widehat{E}_1) - E_1 = O(N^{-1}), \quad \text{and } E(\widehat{E}_2) - E_2 = O(N^{-1}).$$

From Proposition 4.3.2, we have $E(\widehat{\delta}_1) - \delta_1 = O(N^{-1})$ and $E(\widehat{\delta}_2 - \delta_2) = O(N^{-1})$, and from Assumption 4.3.3, we have $1 < \frac{1}{1-\delta_1-\delta_2} < \frac{1}{1-2c_2}$. Therefore,

$$\begin{aligned}
&E\left(\frac{1}{1-\widehat{\delta}_1-\widehat{\delta}_2} - \frac{1}{1-\delta_1-\delta_2}\right) \\
&= \frac{1}{1-\delta_1-\delta_2} E\left(\frac{\widehat{\delta}_1-\delta_1+\widehat{\delta}_2-\delta_2}{1-\delta_1-\delta_2} + o\left(\frac{\widehat{\delta}_1-\delta_1+\widehat{\delta}_2-\delta_2}{1-\delta_1-\delta_2}\right)\right) = O(N^{-1}).
\end{aligned}$$

$$\begin{aligned}
&E\left(\frac{\widehat{\delta}_1-\widehat{\delta}_2}{(1-\widehat{\delta}_1-\widehat{\delta}_2)^2} - \frac{\delta_1-\delta_2}{(1-\delta_1-\delta_2)^2}\right) \\
=&E\left((\widehat{\delta}_1-\widehat{\delta}_2)\left(\frac{1}{(1-\widehat{\delta}_1-\widehat{\delta}_2)^2} - \frac{1}{(1-\delta_1-\delta_2)^2}\right)\right) \\
&+ \frac{1}{(1-\delta_1-\delta_2)^2} E\left((\widehat{\delta}_1-\delta_1) - (\widehat{\delta}_2-\delta_2)\right) \\
=&E\left(\frac{\widehat{\delta}_1-\widehat{\delta}_2}{(1-\delta_1-\delta_2)^2}\left(\left(\frac{2(\widehat{\delta}_1-\delta_1+\widehat{\delta}_2-\delta_2)}{1-\delta_1-\delta_2}\right) + o\left(\frac{2(\widehat{\delta}_1-\delta_1+\widehat{\delta}_2-\delta_2)}{1-\delta_1-\delta_2}\right)\right)\right) \\
&+ \frac{1}{(1-\delta_1-\delta_2)^2} E\left((\widehat{\delta}_1-\delta_1) - (\widehat{\delta}_2-\delta_2)\right) = O(N^{-1}).
\end{aligned}$$

$$\begin{aligned}
E(\widehat{p}_{I1} - p_I) \leq& 2E\left(\frac{1}{1-\widehat{\delta}_1-\widehat{\delta}_2} - \frac{1}{1-\delta_1-\delta_2}\right) + \frac{1}{2(1-\delta_1-\delta_2)} E(\widehat{D}_1 - D_1) \\
&+ E\left(\frac{\widehat{\delta}_1-\widehat{\delta}_2}{(1-\widehat{\delta}_1-\widehat{\delta}_2)^2} - \frac{\delta_1-\delta_2}{(1-\delta_1-\delta_2)^2}\right) + \frac{\delta_1-\delta_2}{2(1-\delta_1-\delta_2)^2} E(\widehat{D}_2 - D_2) \\
&- E\left(\frac{1}{1-\widehat{\delta}_1-\widehat{\delta}_2} - \frac{1}{1-\delta_1-\delta_2}\right) = O(N^{-1}). \qquad (4.43)
\end{aligned}$$

Combining (4.42) and (4.43), we complete the proof. □

*Proof of Theorem 4.4.1.* By definition of $\widehat{p}_I$, we have

$$\sqrt{N}(\widehat{p}_I - p_I) = \frac{2(n_{11}+n_{12})}{N}\sqrt{N}(\widehat{p}_{I1} - p_I) + \frac{2(n_{21}+n_{22})}{N}\sqrt{N}(\widehat{p}_{I2} - p_I).$$

**Asymptotic Equivalence for $\sqrt{N}(\widehat{p}_{I2} - p_I)$** By results in Brunner et al. (2018), we have

$$
\begin{aligned}
&\sqrt{N}(\widehat{p}_{I2} - p_I) \\
=&\sqrt{N}\left(\int \frac{1}{2}(\widehat{F}_{211} + \widehat{F}_{222})d(\widehat{F}_{212} + \widehat{F}_{221}) - 1 - p_I\right) \\
=&\frac{\sqrt{N}}{2}\left(\int \left(\widehat{F}_{211} + \widehat{F}_{222}\right) d\left(\widehat{F}_{212} + \widehat{F}_{221}\right) - \int (F_{11} + F_{22})\, d\,(F_{12} + F_{21})\right) \\
\doteqdot&\frac{\sqrt{N}}{2}\Bigg(\int (F_{211} + F_{222})\, d\left(\widehat{F}_{212} + \widehat{F}_{221}\right) - \int (F_{212} + F_{221})\, d\left(\widehat{F}_{211} + \widehat{F}_{222}\right) \\
&+ 2 - 2\int (F_{11} + F_{22})\, d\,(F_{12} + F_{21})\Bigg) \\
=&\frac{\sqrt{N}}{2n_{21}}\sum_{i=1}^{n_{21}} (F_{211}(X_{212i}) + F_{222}(X_{212i}) - F_{212}(X_{211i}) - F_{221}(X_{211i})) \\
&+ \frac{\sqrt{N}}{2n_{22}}\sum_{i=1}^{n_{22}} (F_{211}(X_{221i}) + F_{222}(X_{221i}) - F_{212}(X_{222i}) - F_{221}(X_{222i})) - 2p_I \\
=&\frac{\sqrt{N}}{n_{21}}\sum_{i=1}^{n_{21}} W_1(\boldsymbol{X}_{21i}) + \frac{\sqrt{N}}{n_{22}}\sum_{i=1}^{n_{22}} W_2(\boldsymbol{X}_{22i}) - 2p_I,
\end{aligned}
$$

where

$$
\begin{aligned}
W_1(\boldsymbol{X}_{21i}) =&\frac{1}{2}\left(F_{211}(X_{212i}) + F_{222}(X_{212i}) - F_{212}(X_{211i}) - F_{221}(X_{211i})\right), \qquad (4.44)\\
W_2(\boldsymbol{X}_{22i}) =&\frac{1}{2}\left(F_{211}(X_{221i}) + F_{222}(X_{221i}) - F_{212}(X_{222i}) - F_{221}(X_{222i})\right).
\end{aligned}
$$

**Asymptotic Equivalence for $\sqrt{N}(\widehat{p}_{I1} - p_I)$** Then we return to $\sqrt{N}(\widehat{p}_{I1} - p_I)$, set

$$
\begin{aligned}
\widetilde{p}_{I1} =&\frac{1}{2(1-\delta_1-\delta_2)}\int (\widehat{F}_{111} + \widehat{F}_{122})d(\widehat{F}_{112} + \widehat{F}_{121}) \\
&+ \frac{\delta_1 - \delta_2}{2(1-\delta_1-\delta_2)^2}\int (\widehat{F}_{111} - \widehat{F}_{121})d(\widehat{F}_{112} - \widehat{F}_{122}) - \frac{1}{(1-\delta_1-\delta_2)}.
\end{aligned}
$$

Then we have

$$
\sqrt{N}(\widehat{p}_{I1} - p_I) = \sqrt{N}(\widehat{p}_{I1} - \widetilde{p}_{I1}) + \sqrt{N}(\widetilde{p}_{I1} - p_I).
$$

1. **Asymptotic equivalence for $\sqrt{N}(\widetilde{p}_{I1} - p_I)$**

By their definition, we have

$$\begin{aligned}
&\sqrt{N}(\widetilde{p}_{I1} - p_I) \qquad (4.45)\\
=&\frac{1}{2(1-\delta_1-\delta_2)}\sqrt{N}\int(\widehat{F}_{111}+\widehat{F}_{122})d(\widehat{F}_{112}+\widehat{F}_{121})\\
&-\frac{1}{2(1-\delta_1-\delta_2)}\sqrt{N}\int(F_{111}+F_{122})d(F_{112}+F_{121})\\
&+\frac{\delta_1-\delta_2}{2(1-\delta_1-\delta_2)^2}\sqrt{N}\int(\widehat{F}_{111}-\widehat{F}_{121})d(\widehat{F}_{112}-\widehat{F}_{122})\\
&-\frac{\delta_1-\delta_2}{2(1-\delta_1-\delta_2)^2}\sqrt{N}\int(F_{111}-F_{121})d(F_{112}-F_{122}).
\end{aligned}$$

To simplify the notation, we set

$$C_1 = \frac{1}{2(1-\delta_1-\delta_2)} \quad \text{and} \quad C_2 = \frac{\delta_1-\delta_2}{2(1-\delta_1-\delta_2)^2}.$$

Applying results (4.29) to the right hand side of (4.45), we have

$$\begin{aligned}
&\sqrt{N}(\widetilde{p}_{I1} - p_I)\\
\doteqdot&C_1\sqrt{N}\left[\int(\widehat{F}_{111}+\widehat{F}_{122})d(F_{112}+F_{121})+\int(F_{111}+F_{122})d(\widehat{F}_{112}+\widehat{F}_{121})\right]\\
&+C_2\sqrt{N}\left[\int(\widehat{F}_{111}-\widehat{F}_{121})d(F_{112}-F_{122})+\int(F_{111}-F_{121})d(\widehat{F}_{112}-\widehat{F}_{122})\right]\\
&-2C_1\sqrt{N}\int(F_{111}+F_{122})d(F_{112}+F_{121})\\
&-2C_2\sqrt{N}\int(F_{111}-F_{121})d(F_{112}-F_{122}).
\end{aligned}$$

Notice that

$$\begin{aligned}
&\int(\widehat{F}_{111}-\widehat{F}_{121})d(F_{112}-F_{122}) = -\int(F_{112}-F_{122})d(\widehat{F}_{111}-\widehat{F}_{121}) \quad \text{and}\\
&\int(\widehat{F}_{111}+\widehat{F}_{122})d(F_{112}+F_{121}) = 4-\int(F_{112}+F_{121})d(\widehat{F}_{111}+\widehat{F}_{122}),
\end{aligned}$$

we have

$$
\begin{aligned}
&\sqrt{N}(\widehat{p}_I - p_I) \\
\doteq& C_1\sqrt{N}\left[\int (F_{111}+F_{122})d(\widehat{F}_{112}+\widehat{F}_{121}) - \int (F_{112}+F_{121})d(\widehat{F}_{111}+\widehat{F}_{122})\right] \\
&+ C_2\sqrt{N}\left[\int (F_{111}-F_{121})d(\widehat{F}_{112}-\widehat{F}_{122}) - \int (F_{112}-F_{122})d(\widehat{F}_{111}-\widehat{F}_{121})\right] \\
&- 2p_I \\
\doteq& U_N.
\end{aligned}
$$

Observe that

$$
\begin{aligned}
U_N =& \frac{\sqrt{N}}{n_{11}}\sum_{k=1}^{n_{11}}\Big\{C_1[-F_{112}(X_{111k}) - F_{121}(X_{111k}) + F_{111}(X_{112k}) + F_{122}(X_{112k})] \\
&+ C_2[-F_{112}(X_{111k}) + F_{122}(X_{111k}) + F_{111}(X_{112k}) - F_{121}(X_{112k})]\Big\} \\
&+ \frac{\sqrt{N}}{n_{12}}\sum_{k=1}^{n_{12}}\Big\{C_1[F_{111}(X_{121k}) + F_{122}(X_{121k}) - F_{112}(X_{122k}) - F_{121}(X_{122k})] \\
&+ C_2[-F_{122}(X_{121k}) + F_{112}(X_{121k}) + F_{121}(X_{122k}) - F_{111}(X_{122k})]\Big\} - 2p_I\sqrt{N} \\
=& \frac{\sqrt{N}}{n_{11}}\sum_{k=1}^{n_{11}} V_1(\boldsymbol{X}_{11k}) + \frac{\sqrt{N}}{n_{12}}\sum_{k=1}^{n_{12}} V_2(\boldsymbol{X}_{2k}) - 2p_I\sqrt{N},
\end{aligned}
$$

where

$$
\begin{aligned}
V_1(\boldsymbol{X}_{11k}) =& C_1[-F_{112}(X_{111k}) - F_{121}(X_{111k}) + F_{111}(X_{112k}) + F_{122}(X_{112k})] \\
&+ C_2[-F_{112}(X_{111k}) + F_{122}(X_{111k}) + F_{111}(X_{112k}) - F_{121}(X_{112k})],
\end{aligned} \tag{4.46}
$$

$$
\begin{aligned}
V_2(\boldsymbol{X}_{2k}) =& C_1[F_{111}(X_{121k}) + F_{122}(X_{121k}) - F_{112}(X_{122k}) - F_{121}(X_{122k})] \\
&+ C_2[-F_{122}(X_{121k}) + F_{112}(X_{121k}) + F_{121}(X_{122k}) - F_{111}(X_{122k})].
\end{aligned}
$$

Thus, $U_N$ can be expressed as sums of functions of independent and identically distributed random variables.

2. **Asymptotic Equivalence for $\sqrt{N}(\widehat{p}_{I1} - \widetilde{p}_{I1})$**

Notice that

$$\begin{aligned}
&\sqrt{N}(\widehat{p}_{I1}-\widetilde{p}_{I1})\\
=&\sqrt{N}\left(\frac{1}{2(1-\widehat{\delta}_1-\widehat{\delta}_2)}-\frac{1}{2(1-\delta_1-\delta_2)}\right)\int(\widehat{F}_{111}+\widehat{F}_{122})d(\widehat{F}_{112}+\widehat{F}_{121})\\
&+\sqrt{N}\left(\frac{\widehat{\delta}_1-\widehat{\delta}_2}{2(1-\widehat{\delta}_1-\widehat{\delta}_2)^2}-\frac{\delta_1-\delta_2}{2(1-\delta_1-\delta_2)^2}\right)\int(\widehat{F}_{111}-\widehat{F}_{121})d(\widehat{F}_{112}-\widehat{F}_{122})\\
&-\sqrt{N}\left(\frac{1}{1-\widehat{\delta}_1-\widehat{\delta}_2}-\frac{1}{(1-\delta_1-\delta_2)}\right).
\end{aligned}$$

By Taylor expansion, we have

$$\begin{aligned}
&\frac{1}{1-\widehat{\delta}_1-\widehat{\delta}_2}-\frac{1}{1-\delta_1-\delta_2}\\
&=\frac{1}{1-\delta_1-\delta_2}\left(\frac{1}{1-\frac{(\widehat{\delta}_1-\delta_1)+(\widehat{\delta}_2-\delta_2)}{1-\delta_1-\delta_2}}-1\right)\\
&=\frac{1}{1-\delta_1-\delta_2}\left(\frac{(\widehat{\delta}_1-\delta_1)+(\widehat{\delta}_2-\delta_2)}{1-\delta_1-\delta_2}+\left(\frac{(\widehat{\delta}_1-\delta_1)+(\widehat{\delta}_2-\delta_2)}{1-\delta_1-\delta_2}\right)^2\right.\\
&\left.+o_p\left(\frac{(\widehat{\delta}_1-\delta_1)+(\widehat{\delta}_2-\delta_2)}{1-\delta_1-\delta_2}\right)^2\right).
\end{aligned}$$

$$\begin{aligned}
&\frac{1}{\left(1-\widehat{\delta}_1-\widehat{\delta}_2\right)^2}-\frac{1}{(1-\delta_1-\delta_2)^2}\\
=&\frac{1}{(1-\delta_1-\delta_2)^2}\left(\frac{1}{\left(1-\frac{(\widehat{\delta}_1-\delta_1)+(\widehat{\delta}_2-\delta_2)}{1-\delta_1-\delta_2}\right)^2}-1\right)\\
=&\frac{1}{(1-\delta_1-\delta_2)^2}\left(2\cdot\frac{(\widehat{\delta}_1-\delta_1)+(\widehat{\delta}_2-\delta_2)}{1-\delta_1-\delta_2}+3\cdot\left(\frac{(\widehat{\delta}_1-\delta_1)+(\widehat{\delta}_2-\delta_2)}{1-\delta_1-\delta_2}\right)^2\right.\\
&\left.+o_p\left(\left(\frac{(\widehat{\delta}_1-\delta_1)+(\widehat{\delta}_2-\delta_2)}{1-\delta_1-\delta_2}\right)^2\right)\right).
\end{aligned}$$

Since $\sqrt{N}(\widehat{\delta}_g-\delta_g)$, $g=1,2$ is asymptotically distributed as normal distributions and $\widehat{\delta}_g$ is consistent estimator for $\delta_g$, by Slutsky's theorem, we have

$$\sqrt{N}(\widehat{\delta}_1-\delta_1)^2\xrightarrow{P}0,\quad \sqrt{N}(\widehat{\delta}_2-\delta_2)^2\xrightarrow{P}0,\quad \sqrt{N}(\widehat{\delta}_2-\delta_2)(\widehat{\delta}_1-\delta_1)\xrightarrow{P}0.$$

Hence,

$$\frac{1}{1-\widehat{\delta}_1-\widehat{\delta}_2}-\frac{1}{1-\delta_1-\delta_2}=\frac{1}{(1-\delta_1-\delta_2)^2}\left((\widehat{\delta}_1-\delta_1)+(\widehat{\delta}_2-\delta_2)\right)+o_p(1).$$

$$\begin{aligned}
&\frac{\widehat{\delta}_1-\widehat{\delta}_2}{\left(1-\widehat{\delta}_1-\widehat{\delta}_2\right)^2}-\frac{\delta_1-\delta_2}{(1-\delta_1-\delta_2)^2}\\
&=\frac{(\widehat{\delta}_1-\delta_1)-(\widehat{\delta}_2-\delta_2)}{\left(1-\widehat{\delta}_1-\widehat{\delta}_2\right)^2}+(\delta_1-\delta_2)\left(\frac{1}{\left(1-\widehat{\delta}_1-\widehat{\delta}_2\right)^2}-\frac{1}{(1-\delta_1-\delta_2)^2}\right)\\
&=\frac{(\widehat{\delta}_1-\delta_1)-(\widehat{\delta}_2-\delta_2)}{(1-\delta_1-\delta_2)^2}+\frac{2(\delta_1-\delta_2)}{1-\delta_1-\delta_2}\frac{(\widehat{\delta}_1-\delta_1)+(\widehat{\delta}_2-\delta_2)}{(1-\delta_1-\delta_2)^2}+o_p(1)\\
&=\frac{1}{(1-\delta_1-\delta_2)^2}\left(\frac{1+\delta_1-3\delta_2}{1-\delta_1-\delta_2}\left(\widehat{\delta}_1-\delta_1\right)-\frac{1+\delta_2-3\delta_1}{1-\delta_1-\delta_2}\left(\widehat{\delta}_2-\delta_2\right)\right)+o_p(1).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\sqrt{N}(\widehat{p}_{I1}-\widetilde{p}_{I1})\\
=&\frac{\sqrt{N}\left((\widehat{\delta}_1-\delta_1)+(\widehat{\delta}_2-\delta_2)\right)}{2(1-\delta_1-\delta_2)^2}\int(\widehat{F}_{111}+\widehat{F}_{122})d(\widehat{F}_{112}+\widehat{F}_{121})\\
&+\frac{1}{2(1-\delta_1-\delta_2)^2}\left(\frac{1+\delta_1-3\delta_2}{1-\delta_1-\delta_2}\sqrt{N}\left(\widehat{\delta}_1-\delta_1\right)-\frac{1+\delta_2-3\delta_1}{1-\delta_1-\delta_2}\sqrt{N}\left(\widehat{\delta}_2-\delta_2\right)\right)\\
&\times\int(\widehat{F}_{111}-\widehat{F}_{121})d(\widehat{F}_{112}-\widehat{F}_{122})-\frac{\sqrt{N}\left((\widehat{\delta}_1-\delta_1)+(\widehat{\delta}_2-\delta_2)\right)}{(1-\delta_1-\delta_2)^2}+o_p(1)\\
=&\frac{\sqrt{N}(\widehat{\delta}_1-\delta_1)}{2(1-\delta_1-\delta_2)^2}\left[\int(\widehat{F}_{111}+\widehat{F}_{122})d(\widehat{F}_{112}+\widehat{F}_{121})\right.\\
&\left.+\frac{1+\delta_1-3\delta_2}{1-\delta_1-\delta_2}\int(\widehat{F}_{111}-\widehat{F}_{121})d(\widehat{F}_{112}-\widehat{F}_{122})-2\right]\\
&+\frac{\sqrt{N}(\widehat{\delta}_2-\delta_2)}{2(1-\delta_1-\delta_2)^2}\left[\int(\widehat{F}_{111}+\widehat{F}_{122})d(\widehat{F}_{112}+\widehat{F}_{121})\right.\\
&\left.-\frac{1+\delta_2-3\delta_1}{1-\delta_1-\delta_2}\int(\widehat{F}_{111}-\widehat{F}_{121})d(\widehat{F}_{112}-\widehat{F}_{122})-2\right]+o_p(1)\\
=&D_1\sqrt{N}(\widehat{\delta}_1-\delta_1)+D_2\sqrt{N}(\widehat{\delta}_2-\delta_2)+o_p(1),
\end{aligned}$$

where

$$
\begin{aligned}
D_1 = & \frac{1}{2(1-\delta_1-\delta_2)^2}\Big[\int (F_{111}+F_{122})d(F_{112}+F_{121}) \\
& + \frac{1+\delta_1-3\delta_2}{1-\delta_1-\delta_2}\int (F_{111}-F_{121})d(F_{112}-F_{122}) - 2\Big], \\
D_2 = & \frac{1}{2(1-\delta_1-\delta_2)^2}\Big[\int (F_{111}+F_{122})d(F_{112}+F_{121}) \\
& - \frac{1+\delta_2-3\delta_1}{1-\delta_1-\delta_2}\int (F_{111}-F_{121})d(F_{112}-F_{122}) - 2\Big].
\end{aligned}
$$

Hence,

$$
\begin{aligned}
& \sqrt{N}\,(\widehat{p}_I - \widetilde{p}_I) \\
= & \frac{1}{S_1+S_2}\sqrt{N}\Big[D_1\left(\widetilde{A}^1_{11}+\widetilde{A}^2_{11}+\widetilde{A}^3_{11}+\widetilde{A}^1_{12}+\widetilde{A}^2_{12}+\widetilde{A}^3_{12}\right) \\
& + D_2(\widetilde{A}^1_{21}+\widetilde{A}^2_{21}+\widetilde{A}^3_{21}+\widetilde{A}^1_{22}+\widetilde{A}^2_{22}+\widetilde{A}^3_{22})\Big] \\
& - \frac{D_1(A_{11}+A_{12})+D_2(A_{21}+A_{22})}{(S_1+S_2)^2}\sqrt{N}\left[\widetilde{S}^1_1+2\widetilde{S}^2_1+\widetilde{S}^1_2+2\widetilde{S}^2_2\right] + o_p(1) \\
= & \frac{\sqrt{N}}{n_{11}}\sum_{k=1}^{n_{11}} U_{11}(\boldsymbol{X}_{11k}) + \frac{\sqrt{N}}{n_{12}}\sum_{k=1}^{n_{12}} U_{12}(\boldsymbol{X}_{12k}) \\
& + \frac{\sqrt{N}}{n_{21}}\sum_{k=1}^{n_{21}} U_{21}(\boldsymbol{X}_{21k}) + \frac{\sqrt{N}}{n_{22}}\sum_{k=1}^{n_{22}} U_{22}(\boldsymbol{X}_{22k}),
\end{aligned}
$$

where

$$
\begin{aligned}
U_{11}(\boldsymbol{X}_{11k}) = & \frac{D_1}{S_1+S_2}U_{111}(\boldsymbol{X}_{11k}) + \frac{D_2}{S_1+S_2}U_{211}(\boldsymbol{X}_{11k}) \\
& - \frac{(D_1(A_{11}+A_{12})+D_2(A_{21}+A_{22}))}{(S_1+S_2)^2}U_{311}(\boldsymbol{X}_{11k}), \\
U_{12}(\boldsymbol{X}_{12k}) = & \frac{D_1}{S_1+S_2}U_{112}(\boldsymbol{X}_{12k}) + \frac{D_2}{S_1+S_2}U_{212}(\boldsymbol{X}_{12k}) \\
& - \frac{(D_1(A_{11}+A_{12})+D_2(A_{21}+A_{22}))}{(S_1+S_2)^2}U_{312}(\boldsymbol{X}_{12k}), \\
U_{21}(\boldsymbol{X}_{21k}) = & \frac{D_1}{S_1+S_2}U_{121}(\boldsymbol{X}_{21k}) + \frac{D_2}{S_1+S_2}U_{221}(\boldsymbol{X}_{21k}) \\
& - \frac{(D_1(A_{11}+A_{12})+D_2(A_{21}+A_{22}))}{(S_1+S_2)^2}U_{321}(\boldsymbol{X}_{21k}), \\
U_{22}(\boldsymbol{X}_{22k}) = & \frac{D_1}{S_1+S_2}U_{122}(\boldsymbol{X}_{22k}) + \frac{D_2}{S_1+S_2}U_{222}(\boldsymbol{X}_{22k}) \\
& - \frac{(D_1(A_{11}+A_{12})+D_2(A_{21}+A_{22}))}{(S_1+S_2)^2}U_{322}(\boldsymbol{X}_{22k}).
\end{aligned}
\tag{4.47}
$$

Combining the equivalence results above, we have

$$
\begin{aligned}
&\sqrt{N}(\widehat{p}_I - p_I) \\
=&\frac{2(n_{11}+n_{12})}{N}\sqrt{N}(\widehat{p}_{I1} - p_I) + \frac{2(n_{21}+n_{22})}{N}\sqrt{N}(\widehat{p}_{I2} - p_I) \\
=&\frac{2(n_{11}+n_{12})}{N}\left(\sqrt{N}(\widehat{p}_{I1} - \widetilde{p}_{I1}) + \sqrt{N}(\widetilde{p}_{I1} - p_I)\right) + \frac{2(n_{21}+n_{22})}{N}\sqrt{N}(\widetilde{p}_{I2} - p_I) \\
\doteq&\frac{2(n_{11}+n_{12})}{N}\left(\frac{\sqrt{N}}{n_{11}}\sum_{k=1}^{n_{11}} U_{11}(\boldsymbol{X}_{11k}) + \frac{\sqrt{N}}{n_{12}}\sum_{k=1}^{n_{12}} U_{12}(\boldsymbol{X}_{12k})\right) \\
&+ \frac{2(n_{11}+n_{12}}{N}\left(\frac{\sqrt{N}}{n_{21}}\sum_{k=1}^{n_{21}} U_{21}(\boldsymbol{X}_{21k}) + \frac{\sqrt{N}}{n_{22}}\sum_{k=1}^{n_{22}} U_{22}(\boldsymbol{X}_{22k})\right) \\
&+ \frac{2(n_{11}+n_{12})}{N}\left(\frac{\sqrt{N}}{n_{11}}\sum_{k=1}^{n_{11}} V_1(\boldsymbol{X}_{11k}) + \frac{\sqrt{N}}{n_{12}}\sum_{k=1}^{n_{12}} V_2(\boldsymbol{X}_{12k})\right) \\
&+ \frac{2(n_{21}+n_{22})}{N}\left(\frac{\sqrt{N}}{n_{21}}\sum_{i=1}^{n_{21}} W_1(\boldsymbol{X}_{21i}) + \frac{\sqrt{N}}{n_{22}}\sum_{i=1}^{n_{22}} W_2(\boldsymbol{X}_{22i})\right) \\
=&\frac{\sqrt{N}}{n_{11}}\sum_{k=1}^{n_{11}} A_1\left(V_1(\boldsymbol{X}_{11k}) + U_{11}(\boldsymbol{X}_{11k})\right) + \frac{\sqrt{N}}{n_{12}}\sum_{k=1}^{n_{12}} A_1\left(V_2(\boldsymbol{X}_{12k}) + U_{12}(\boldsymbol{X}_{12k})\right) \\
&+ \frac{\sqrt{N}}{n_{21}}\sum_{k=1}^{n_{21}}\left(A_2 W_1(\boldsymbol{X}_{21k}) + A_1 U_{21}(\boldsymbol{X}_{21k})\right) \\
&+ \frac{\sqrt{N}}{n_{22}}\sum_{k=1}^{n_{22}}\left(A_2 W_2(\boldsymbol{X}_{22k}) + A_1 U_{22}(\boldsymbol{X}_{22k})\right) + o_p(1).
\end{aligned}
$$

Utilizing the CLT and by Assumption 4.3.4, we can obtain that

$$
\sqrt{N}(\widehat{p}_I - p_I) \xrightarrow{D} U \sim N\left(0, \kappa_{11}^{-1}\sigma_{11}^2 + \kappa_{12}^{-1}\sigma_{12}^2 + \kappa_{21}^{-1}\sigma_{21}^2 + \kappa_{22}^{-1}\sigma_{22}^2\right),
$$

where

$$
\begin{aligned}
&\sigma_{11}^2 = Var(A_1\left(V_1(\boldsymbol{X}_{11k}) + U_{11}(\boldsymbol{X}_{11k})\right)), \qquad \sigma_{12}^2 = Var\left(A_1\left(V_2(\boldsymbol{X}_{12k}) + U_{12}(\boldsymbol{X}_{12k})\right)\right), \\
&\sigma_{21}^2 = Var(A_2 W_1(\boldsymbol{X}_{21k}) + A_1 U_{21}(\boldsymbol{X}_{21k})), \qquad \sigma_{22}^2 = Var(A_2 W_2(\boldsymbol{X}_{22k}) + A_1 U_{22}(\boldsymbol{X}_{22k})).
\end{aligned}
$$

□

*Proof of Theorem 4.4.2.* By the weak law of large number, we have

$$\widehat{\sigma}_{11}^2 = \frac{1}{n_{11}-1}\sum_{k=1}^{n_{11}} A_1^2\left(V_1(\boldsymbol{X}_{11k}) + U_{11}(\boldsymbol{X}_{11k}) - \overline{V}_1(\boldsymbol{X}_{11\cdot}) - \overline{U}_{11}(\boldsymbol{X}_{11\cdot})\right)^2 \xrightarrow{P} \sigma_{11}^2,$$

$$\widehat{\sigma}_{12}^2 = \frac{1}{n_{12}-1}\sum_{k=1}^{n_{12}} A_1^2\left(V_2(\boldsymbol{X}_{12k}) + U_{12}(\boldsymbol{X}_{12k}) - \overline{V}_2(\boldsymbol{X}_{12\cdot}) - \overline{U}_{12}(\boldsymbol{X}_{12\cdot})\right)^2 \xrightarrow{P} \sigma_{12}^2,$$

$$\widehat{\sigma}_{21}^2 = \frac{1}{n_{21}-1}\sum_{k=1}^{n_{21}} \left(A_2W_1(\boldsymbol{X}_{21k}) + A_1U_{21}(\boldsymbol{X}_{21k}) - A_2\overline{W}_1(\boldsymbol{X}_{21\cdot}) - A_1\overline{U}_{21}(\boldsymbol{X}_{21\cdot})\right)^2 \xrightarrow{P} \sigma_{21}^2,$$

$$\widehat{\sigma}_{22}^2 = \frac{1}{n_{22}-1}\sum_{k=1}^{n_{22}} \left(A_2W_2(\boldsymbol{X}_{22k}) + A_1U_{22}(\boldsymbol{X}_{22k}) - A_2\overline{W}_2(\boldsymbol{X}_{22k}) - A_1\overline{U}_{22}(\boldsymbol{X}_{22\cdot})\right)^2 \xrightarrow{P} \sigma_{22}^2.$$

as $n_{11}, n_{12}, n_{21}, n_{22} \to \infty$, where $\overline{V}_g(X_{1g\cdot}) = \frac{1}{n_{1g}}\sum_{k=1}^{n_{1g}} V(X_{1gk})$, $\overline{W}_g(X_{2g\cdot}) = \frac{1}{n_{2g}}\sum_{k=1}^{n_{2g}} W_g(X_{2gk})$ and $\overline{U}_{hg}(X_{hg\cdot}) = \frac{1}{n_{hg}}\sum_{k=1}^{n_{hg}} U_{hg}(X_{hgk})$, $h, g = 1, 2$.

Since $\widehat{\sigma}_{hg}^2 \xrightarrow{P} \sigma_{hg}^2$ , we are done with the proof if we can show $\widehat{S}_{hg}^2 - \widehat{\sigma}_{hg}^2 \xrightarrow{P} 0$. It suffices to show that

$$E[\widehat{S}_{hg}^2 - \widehat{\sigma}_{hg}]^2 \to 0 \text{ as } N \to \infty \text{ for } h, g = 1, 2.$$

Here we just show the proof for $h = g = 1$, others can be proved similarly. By definition, we have

$$\begin{aligned}
E[\widehat{S}_{11}^2 - \widehat{\sigma}_{11}]^2 =& E\Bigg[\frac{A_1^2}{n_{11}-1}\sum_{k=1}^{n_{11}}\left(\widehat{V}_1(\boldsymbol{X}_{11k}) + \widehat{U}_{11}(\boldsymbol{X}_{11k}) - \overline{\widehat{V}}_1(\boldsymbol{X}_{11\cdot}) - \overline{\widehat{U}}_{11}(\boldsymbol{X}_{11\cdot})\right)^2 \\
&- \frac{A_1^2}{n_{11}-1}\sum_{k=1}^{n_{11}}\left(V_1(\boldsymbol{X}_{11k}) + U_{11}(\boldsymbol{X}_{11k}) - \overline{V}_1(\boldsymbol{X}_{11\cdot}) - \overline{U}_{11}(\boldsymbol{X}_{11\cdot})\right)^2\Bigg]^2 \\
\le& \frac{A_1^4}{(n_{11}-1)^2}E\Bigg[\sum_{k=1}^{n_{11}}\Big(\widehat{V}_1(\boldsymbol{X}_{11k}) + \widehat{U}_{11}(\boldsymbol{X}_{11k}) - \overline{\widehat{V}}_1(\boldsymbol{X}_{11\cdot}) - \overline{\widehat{U}}_{11}(\boldsymbol{X}_{11\cdot}) \\
&+ V_1(\boldsymbol{X}_{11k}) + U_{11}(\boldsymbol{X}_{11k}) - \overline{V}_1(\boldsymbol{X}_{11\cdot}) - \overline{U}_{11}(\boldsymbol{X}_{11\cdot})\Big)^2 \\
&\times \sum_{k=1}^{n_{11}}\left(\widehat{V}_1(\boldsymbol{X}_{11k}) - V_1(\boldsymbol{X}_{11k}) + \widehat{U}_{11}(\boldsymbol{X}_{11k}) - U_{11}(\boldsymbol{X}_{11k})\right. \\
&\left.- \overline{\widehat{V}}_1(\boldsymbol{X}_{11\cdot}) + \overline{V}_1(\boldsymbol{X}_{11\cdot}) - \overline{\widehat{U}}_{11}(\boldsymbol{X}_{11\cdot}) + \overline{U}_{11}(\boldsymbol{X}_{11\cdot})\right)^2\Bigg].
\end{aligned}$$

Now suppose (4.39) holds and further assume

$$0 < \min\{\delta_1, \delta_2\} \le \max\{\widehat{\delta}_1, \widehat{\delta}_2\} < c' < \frac{1}{2}, \tag{4.48}$$

where $c < c' < \frac{1}{2}$, and $c$ is the constant in Assumption 4.3.3. Then by Assumption 4.3.2, 4.3.3 and range of distribution functions, $\widehat{V}_1(\boldsymbol{X}_{11k})$, $V_1(\boldsymbol{X}_{11k})$, $\widehat{U}_{11}(\boldsymbol{X}_{11k})$, and $U_{11}(\boldsymbol{X}_{11k})$ are bounded and we can find a constant $M > 0$, such that

$$\begin{aligned}&\big(\widehat{V}_1(\boldsymbol{X}_{11k}) + \widehat{U}_{11}(\boldsymbol{X}_{11k}) - \overline{\widehat{V}}_1(\boldsymbol{X}_{11\cdot}) - \overline{\widehat{U}}_{11}(\boldsymbol{X}_{11\cdot})\\ &+ V_1(\boldsymbol{X}_{11k}) + U_{11}(\boldsymbol{X}_{11k}) - \overline{V}_1(\boldsymbol{X}_{11\cdot}) - \overline{U}_{11}(\boldsymbol{X}_{11\cdot})\big)^2 \leq M.\end{aligned}$$

Therefore, similar to the procedures in (4.40), we have

$$\begin{aligned}&E[\widehat{S}_{11}^2 - \widehat{\sigma}_{11}]^2\\ \leq&\frac{Mn_{11}}{(n_{11}-1)^2} E\left[\sum_{k=1}^{n_{11}} \left(\widehat{V}_1(\boldsymbol{X}_{11k}) - V_1(\boldsymbol{X}_{11k})\right)^2 + \sum_{k=1}^{n_{11}} \left(\widehat{U}_{11}(\boldsymbol{X}_{11k}) - U_{11}(\boldsymbol{X}_{11k})\right)^2\right]\\ \leq&\frac{Mn_{11}}{(n_{11}-1)^2} \left[\sum_{k=1}^{n_{11}} E\left(\widehat{V}_1(\boldsymbol{X}_{11k}) - V_1(\boldsymbol{X}_{11k})\right)^2 + \sum_{k=1}^{n_{11}} E\left(\widehat{U}_{11}(\boldsymbol{X}_{11k}) - U_{11}(\boldsymbol{X}_{11k})\right)^2\right].\end{aligned}$$

By definitions, we have

$$\begin{aligned}&E\left[\widehat{V}_1(\boldsymbol{X}_{11k}) - V_1(\boldsymbol{X}_{11k})\right]^2 \qquad (4.49)\\ \leq&E\left[\widehat{C}_1\widehat{F}_{112}(X_{111k}) - C_1F_{112}(X_{111k})\right]^2 + E\left[\widehat{C}_1\widehat{F}_{121}(X_{111k}) - C_1F_{121}(X_{111k})\right]^2\\ &+ E\left[\widehat{C}_1\widehat{F}_{111}(X_{112k}) - C_1F_{111}(X_{112k})\right] + E\left[\widehat{C}_1\widehat{F}_{122}(X_{112k}) - C_1F_{122}(X_{112k})\right]^2\\ &+ E\left[\widehat{C}_2\widehat{F}_{112}(X_{111k}) - C_2F_{112}(X_{111k})\right] + E\left[\widehat{C}_2\widehat{F}_{122}(X_{111k}) - C_2F_{122}(X_{111k})\right]^2\\ &+ E\left[\widehat{C}_2\widehat{F}_{111}(X_{112k}) - C_2F_{111}(X_{112k})\right] + E\left[\widehat{C}_2\widehat{F}_{121}(X_{112k}) - C_2F_{121}(X_{112k})\right]^2,\end{aligned}$$

where $\widehat{C}_1 = \frac{1}{2(1-\widehat{\delta}_1-\widehat{\delta}_2)}$, $\widehat{C}_2 = \frac{\widehat{\delta}_1-\widehat{\delta}_2}{2(1-\widehat{\delta}_1-\widehat{\delta}_2)^2}$. By Assumption 4.3.3 and Lemma 4.8.2, we have

$$\begin{aligned}&E\left[\widehat{C}_1\widehat{F}_{112}(X_{111k}) - C_1F_{112}(X_{111k})\right]^2\\ \leq&E\left[\left(\widehat{C}_1 - C_1\right)^2 \widehat{F}_{112}^2(X_{111k})\right] + E\left[C_1^2\left(\widehat{F}_{112}(X_{111k}) - F_{112}(X_{111k})\right)^2\right]\\ \leq&E\left[\left(\widehat{C}_1 - C_1\right)^2\right] + \frac{1}{4n_{11}(1-2c)^2}.\end{aligned}$$

Since

$$
\begin{aligned}
&E\left[\left(\widehat{C}_1 - C_1\right)^2\right] \\
=&E\left[\left(\frac{1}{2(1-\widehat{\delta}_1-\widehat{\delta}_2)} - \frac{1}{2(1-\delta_1-\delta_2)}\right)^2\right] \\
=&E\left[\frac{1}{4(1-\delta_1-\delta_2)^2}\left(\frac{\widehat{\delta}_1-\delta_1+\widehat{\delta}_2-\delta_2}{1-\delta_1-\delta_2} + o\left(\frac{\widehat{\delta}_1-\delta_1+\widehat{\delta}_2-\delta_2}{1-\delta_1-\delta_2}\right)\right)^2\right] \\
\leq&\frac{1}{8(1-2c)^6}\left[E\left(\widehat{\delta}_1-\delta_1\right)^2 + E\left(\widehat{\delta}_2-\delta_2\right)^2 + E\left(o\left(\widehat{\delta}_1-\delta_1+\widehat{\delta}_2-\delta_2\right)\right)^2\right].
\end{aligned}
$$

Notice that

$$
\begin{aligned}
&E\left(\widehat{\delta}_g-\delta_g\right)^2 \\
=&E\left(\left(\widehat{A}_{g1}+\widehat{A}_{g2}\right)\left(\frac{1}{\widehat{S}_1+\widehat{S}_2}-\frac{1}{S_1+S_2}\right)+\frac{1}{S_1+S_2}\left(\widehat{A}_{g1}+\widehat{A}_{g2}-A_{g1}-A_{g2}\right)\right)^2 \\
=&O(N^{-1}),
\end{aligned}
$$

because by Lemma 4.8.2, we have

$$
E\left(\widehat{A}_{gt}-A_{gt}\right)=O(N^{-1}), \quad E\left(\widehat{S}_t-S_t\right)=O(N^{-1}).
$$

Therefore,

$$
E\left[\left(\widehat{C}_1-C_1\right)^2\right]=O(N^{-1}).
$$

Similarly, we can show that

$$
E\left[\left(\widehat{C}_2-C_2\right)^2\right]=O(N^{-1}).
$$

Hence,

$$
E\left[\widehat{C}_1\widehat{F}_{112}(X_{111k})-C_1F_{112}(X_{111k})\right]^2 \leq O(N^{-1}).
$$

The rest terms on the left hand side of (4.49) can be dealt similarly, therefore,

$$
E\left[\widehat{V}_1(\boldsymbol{X}_{11k})-V_1(\boldsymbol{X}_{11k})\right]^2 \leq O(N^{-1}). \tag{4.50}
$$

Similarly, we can show that

$$E\left[\widehat{U}_{11}(\boldsymbol{X}_{11k}) - U_{11}(\boldsymbol{X}_{11k})\right]^2 \leq O(N^{-1}). \tag{4.51}$$

Combining (4.50) and (4.51), we have

$$E[\widehat{S}_{11}^2 - \widehat{\sigma}_{11}] \leq O(N^{-1}).$$

Thus, when (4.39) and (4.48) hold, we have $\widehat{S}_{11} - \sigma_{11}^2 \xrightarrow{P} 0$. Notice that $\widehat{S}_1 \xrightarrow{P} S_1$, $\widehat{S}_2 \xrightarrow{P} S_2$, $\widehat{\delta}_1 \xrightarrow{\delta} _1$ and $\widehat{\delta}_2 \xrightarrow{\delta} _2$, then under Assumption 4.3.1, 4.3.2 and 4.3.3, we have $\lim_{N\to\infty} P(\widehat{S}_1 + \widehat{S}_2 \leq C') = 0$, $\lim_{N\to\infty} P(\widehat{\delta}_1 > c' \text{ or } \widehat{\delta}_1 < 0) = 0$, $\lim_{N\to\infty} P(\widehat{\delta}_2 > c' \text{ or } \widehat{\delta}_1 < 0) = 0$. Thus, $\forall \epsilon > 0$,

$$\begin{aligned}
0 &\leq \lim_{N\to\infty} P\left(|\widehat{S}_{11} - \sigma_{11}| > \epsilon\right) \\
&\leq \lim_{N\to\infty} P\left(\widehat{S}_1 + \widehat{S}_2 \leq C'\right) + \lim_{N\to\infty} P(\widehat{\delta}_1 > c' \text{ or } \widehat{\delta}_1 < 0) + \lim_{N\to\infty} P(\widehat{\delta}_2 > c' \text{ or } \widehat{\delta}_1 < 0) \\
&\quad + \lim_{N\to\infty} P\left(|\widehat{S}_{11} - \sigma_{11}| > \epsilon, \widehat{S}_1 + \widehat{S}_2 > C', 0 < \min\{\delta_1, \delta_2\} \leq \max\{\widehat{\delta}_1, \widehat{\delta}_2\} < c'\right) = 0.
\end{aligned}$$

Hence, $\widehat{S}_{11}$ is a constant estimator for $\sigma_{11}^2$. □

## Supplemental Simulation Results

In this subsection, we include additional simulation results that are discussed in Section 4.5.

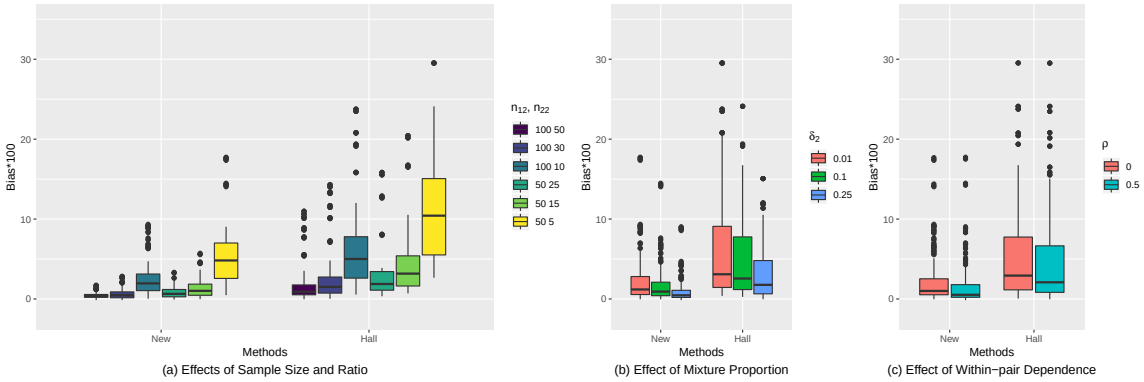



Figure 4.12: Boxplots of bias for $\widehat{\delta_2}$ by sample size, sample size ratio, $\delta_2$, and $\rho$.

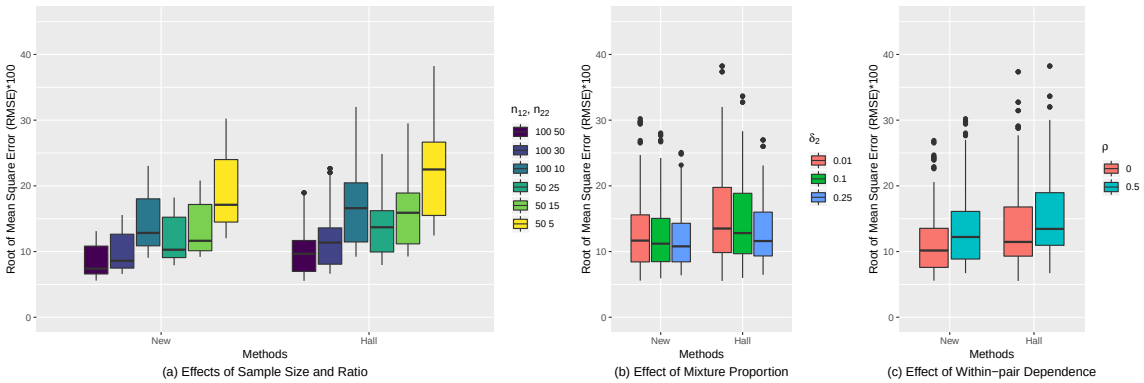



Figure 4.13: Boxplots of RMSE for $\widehat{\delta_2}$ by sample size, sample size ratio, $\delta_2$, and $\rho$.

# Chapter 5 Adjusting for Covariates in Contaminated Clinical Trials

## 5.1 Introduction

Clinical trials are often used to assess drug efficacy and safety. Sometimes, participants are classified into different groups by diagnostic tools. However, these diagnostic tools are not perfectly accurate. Their misclassification error rates are usually unknown and assumed to be zero in the traditional method. These inaccuracies lead to contamination of the group membership and, thereby, bias in estimation of treatment effect (Battistin and Sianesi, 2011). Moreover, the misclassification errors yield overly optimistic results in the sample size determinations and the power calculations. These errors may prevent the detection of significant treatment effects.

This misclassification issue are prominent in the era of personalized medicine and measurement-based care. The US Food and Drug Administration published a concept paper (Hinman et al., 2006) to support the co-development of drugs and test devices. It recommends establishing clinical validation (i.e., the accuracy of the test devices) and clinical utility (i.e., the effects of classification on the drug performances) of a test in a pre-clinical pilot feasibility study. These can be achieved using a pre-stratified randomized placebo-controlled design or a pre-stratified pre-post or matched pair design. This chapter will focus on the latter design, but the method can be easily extended to the first design.

Despite the importance of this problem, only few works evaluated treatment effect in the presence of misclassifications. Most of them focus on estimating accuracy and sample size to evaluate the devices themselves (Flahault et al., 2005). Harrar et al. (2016) recently tackled this problem and provided methods for estimating and testing treatment effect. They approximated the distribution of outcomes by a mixture of multivariate normal distribution and derived sample size determination formula. Nevertheless, their methods are only applicable for continuous outcomes and do not allow covariates.

To estimate and test treatment effects, a proper model for the distribution of outcomes is necessary. Due to misclassification errors, the distribution can be approximated by mixture

models, which are widely used in different disciplines and have well-studied theories. The finite mixture of linear regression models is a commonly used class of mixture models that can accommodate the effect of the covariates. Though powerful, their linearity assumptions are sometimes restrictive. Some efforts have been made to extend these models and relax these assumptions. Huang and Yao (2012) studied a semiparametric mixture of regression models. In their models, the regression functions are still linear functions of predictors, but the mixing proportions are smoothing functions of a covariate instead of constant. Later, Huang et al. (2013) proposed a nonparametric finite mixture of regression models. The new model relaxes the linearity assumption on the regression model and allows each component regression functions to be an unknown but smooth function of covariates. Due to the "curse of dimensionality", only a single covariate is considered.

This chapter extends Huang et al. (2013)'s model by considering multiple covariates and bivariate outcomes. We establish a more robust method for estimating and testing treatment effects in a pre-post design when a diagnostic device used for screening (treatment assignment) is prone to misclassification errors. In Section 2, we present a nonparametric finite mixture of regression models and establish its identifiability. We derive an estimation procedure using the kernel regression method and EM algorithm in Section 3. We conclude this chapter with summary and discussion for future work in Section 4.

## 5.2 Model and Identifiability

### Statistical Model

To compare treatment effects on patients with or without a specific disease in a pre-post design, it is assumed that a diagnostic device that is prone to misclassification error is used to separate the participants into two groups, diseased (group 1) and healthy (group 2). The observation on performance of each patient are measured before and after treatment. These measurements are response variables, denoted by $\boldsymbol{Y}$. We propose a mixture of nonparametric regression models to address the misclassification errors of the diagnostic device.

Let $\{\boldsymbol{X}_g, \boldsymbol{Y}_g\}, g = 1, 2$, be pair of covariates and outcome variables for each patient diagnosed in group $g$. Since the diagnostic device is not perfect, the actual group status

of each patient is unknown. We will regard it as a latent variable and denote it as $C$. Let the probability that a patient truly is classified in group $g$ be $P(C = g|\boldsymbol{X}_g = \boldsymbol{x}) = \pi_g(\boldsymbol{x})$. Conditioning on $C = g$, and $\boldsymbol{X}_g = \boldsymbol{x}$, $\boldsymbol{Y}_g$ is assumed to follow a multivariate normal distribution with mean $\boldsymbol{m}_g(\boldsymbol{x})$ and covariance matrix $\Sigma_g(\boldsymbol{x})$. We further assume that $\pi_g(\boldsymbol{x})$, $\boldsymbol{m}_g(\boldsymbol{x})$, and $\Sigma_g(\boldsymbol{x})$, $g = 1, 2$ are unknown but smooth functions. Hence, conditioning on $\boldsymbol{X}_g = \boldsymbol{x}$, $\boldsymbol{Y}_g$ follows a finite mixture of multivariate normal distributions given by

$$\boldsymbol{Y}_g|_{\boldsymbol{X}_g=\boldsymbol{x}} \sim \pi_g(\boldsymbol{x})N\{\boldsymbol{m}_g(\boldsymbol{x}), \Sigma_g(\boldsymbol{x})\} + (1 - \pi_g(\boldsymbol{x}))N\{\boldsymbol{m}_{g'}(\boldsymbol{x}), \Sigma_{g'}(\boldsymbol{x})\}, \tag{5.1}$$

where $g \neq g'$, $g, g' = 1, 2$.

### Identifiability

Huang et al. (2013) established identifiability for the nonparametric finite mixture of regression models in the univariate case. We extend their result to the multivariate case here. To establish identifiability, more restrictions for mixture models are required. The proof of Theorem 5.2.1 is included in Appendix 5.5

**Theorem 5.2.1.** *The model (5.1) is identifiable if*

1. $\pi_g(\boldsymbol{x}) > 0$, $g = 1, 2$, *are continuous functions,*

2. $\boldsymbol{m}_g(\boldsymbol{x})$ *and* $\Sigma_g(\boldsymbol{x})$, $g = 1, 2$ *are differentiable functions,*

3. *for any* $\boldsymbol{x} \in R^p$,

$$\begin{aligned}
&||\boldsymbol{m}_1(\boldsymbol{x}) - \boldsymbol{m}_2(\boldsymbol{x})||^2 + \left\| \mathrm{vec}\left(\frac{\partial \boldsymbol{m}_1(\boldsymbol{x})}{\partial \boldsymbol{x}}\right) - \mathrm{vec}\left(\frac{\partial \boldsymbol{m}_2(\boldsymbol{x})}{\partial \boldsymbol{x}}\right) \right\|^2 \\
&+ ||\mathrm{vec}(\Sigma_1(\boldsymbol{x})) - \mathrm{vec}(\Sigma_2(\boldsymbol{x}))||^2 \\
&+ \left\| \mathrm{vec}\left(\frac{\partial \mathrm{vec}(\Sigma_1(\boldsymbol{x}))}{\partial \boldsymbol{x}}\right) - \mathrm{vec}\left(\frac{\partial \mathrm{vec}(\Sigma_2(\boldsymbol{x}))}{\partial \boldsymbol{x}}\right) \right\|^2 \neq 0,
\end{aligned}$$

   *where* $\|\cdot\|$ *is the Euclidean distance and vec is the operator that transforms a matrix to a vector, and*

4. *the range of* $\boldsymbol{X}$ *is an open set in* $R^p$.

The conditions 1,2 and 4 in Theorem 5.2.1 are general conditions for model (5.1) that require the component functions $\pi_g(\boldsymbol{x}), \boldsymbol{m}_g(\boldsymbol{x}), \Sigma_g(\boldsymbol{x})$ are smooth functions on an open set in $R^p$. The conditions 3 requires $(\boldsymbol{m}_1(\boldsymbol{x}), \Sigma_1(\boldsymbol{x}))$ and $(\boldsymbol{m}_2(x), \Sigma_2(\boldsymbol{x}))$ have different derivatives on all their intersection points. If the distribution of two groups are well separated, i.e. $(\boldsymbol{m}_1(\boldsymbol{x}), \Sigma_1(\boldsymbol{x})) \neq (\boldsymbol{m}_2(x), \Sigma_2(\boldsymbol{x}))$ for all $\boldsymbol{x}$, this condition is satisfied.

### 5.3 Estimation Procedure: Nonparametric Kernel Regression

To estimate these component functions in the model (5.1), $\pi_g(x), m_g(x)$, and $\Sigma_g(x)$, $g = 1, 2$, one can use the maximum likelihood method. Suppose we have $n = n_1+n_2$ subjects in the clinical trial, of which $n_1$ and $n_2$ are diagnosed as in group 1 and group 2, respectively. Let $\boldsymbol{Y}_{gi} = (Y_{gi}^1, Y_{gi}^2)'$ be the pre and post outcomes vector and $\boldsymbol{X}_{gi}$ be the covariates for the $i$th patient in the group $g$. Denote $\phi(\boldsymbol{y}|\boldsymbol{\mu}, \Sigma)$ to be the density function of multivariate normal distribution $N(\mu, \Sigma)$. Then the log-likelihood function for the data $\{\boldsymbol{X}_{gi}, \boldsymbol{Y}_{gi}, i = 1, \ldots, n_g, g = 1, 2\}$ is

$$\mathcal{L} = \sum_{g=1}^{2}\sum_{i=1}^{n_g} \log\big[\pi_g(\boldsymbol{x}_{gi})\phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_g(\boldsymbol{x}_{gi}), \Sigma_g(\boldsymbol{x}_{gi})\} + (1-\pi_g(\boldsymbol{x}_{gi}))\phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_{g'}(\boldsymbol{x}_{gi}), \Sigma_{g'}(\boldsymbol{x}_{gi})\}\big],$$

where $g \neq g'$, $g, g' = 1, 2$. We employ the multivariate kernel regression method to estimate component functions in model (5.1). Instead of estimating the component functions directly, we use local constant $\pi_g$, $m_g$, and $\Sigma_g$ to approximate $\pi_g(x)$, $m_g(x)$, and $\Sigma_g(x)$, $g = 1, 2$. Let $K_H(x) = |H|^{-1/2}K(H^{-1/2}x)$ be a rescaled version of the kernel function $K(\cdot)$ with symmetric and positive bandwidth matrix $H$, where $K(\boldsymbol{x}) = (2\pi)^{-d/2}\exp(-1/2\boldsymbol{x}^\top\boldsymbol{x})$. Then the corresponding local log-likelihood function is

$$l_n(\boldsymbol{\theta}; x) = \sum_{g=1}^{2}\sum_{i=1}^{n_g} \log\left[\pi_g\phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_g, \Sigma_g\} + (1-\pi_g)\phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_{g'}, \Sigma_{g'}\}\right] K_H(\boldsymbol{x}_{gi} - \boldsymbol{x}), \tag{5.2}$$

where $\boldsymbol{\theta} = (\pi_1, \pi_2, m_1, m_2, \Sigma_1, \Sigma_2)$. We estimate the component functions by the maximizer of the local log-likelihood function (5.2) and denote it as $\widetilde{\boldsymbol{\theta}} = (\widetilde{\pi}_1, \widetilde{\pi}_2, \widetilde{m}_1, \widetilde{m}_2, \widetilde{\Sigma}_1, \widetilde{\Sigma}_2)$.

**EM Algorithm for Kernel Regression**

To obtain the maximizer of the function in (5.2), we formulate the problem as an incomplete-data problem and use the EM algorithm. We regard the actual group status of $i$th patient in group $g$ as the missing variable and denote it by $Z_{gi}$, i.e.,

$$Z_{gi} = \begin{cases} 1, & \text{if } (\boldsymbol{X_i}, \boldsymbol{Y_i}) \text{ classified in } g\text{th group is in the } g\text{th group,} \\ 0, & \text{otherwise.} \end{cases} \tag{5.3}$$

The complete log-likelihood function becomes

$$\sum_{g}^{2}\sum_{i=1}^{n_g}\Big\{z_{gi}[\log \pi_g(\boldsymbol{x}_{gi}) + \log \phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_g(\boldsymbol{x}_{gi}), \Sigma_g(\boldsymbol{x}_{gi})\}] \\ + (1 - z_{gi})[\log(1 - \pi_g(\boldsymbol{x}_{gi})) + \log \phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_{g'}(\boldsymbol{x}_{gi}), \Sigma_{g'}(\boldsymbol{x}_{gi})\}]\Big\}.$$

For $\boldsymbol{x} \in \{\boldsymbol{u}_1, \ldots, \boldsymbol{u}_N\}$, the set of grid points at which the unknown functions are to be evaluated, define a local complete log-likelihood as

$$\sum_{g}^{2}\sum_{i=1}^{n_g}\big\{z_{gi}[\log \pi_g + \log \phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_g, \Sigma_g\}] + (1 - z_{ig})[\log(1 - \pi_g) \\ + \log \phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_{g'}, \Sigma_{g'}\}]\big\}K_H(\boldsymbol{x}_{gi} - \boldsymbol{x}).$$

The EM algorithm involves two steps iteratively: E-step and M-step. In the E-step, we take expectations of the missing variables $Z_{gi}$. Then in the M-step, we plug in these expectations and maximize the resulting complete log-likelihood function to obtain the estimators for the component functions. Then we return to the E-step and update the expectations of $Z_{gi}$ and carry out the M-step again. We iteratively update the estimators until the algorithm converges.

For a fixed point $\boldsymbol{x}$, one can easily maximize this function using the EM algorithm. However, we are interested in evaluating the component functions at a set of grid points over an open set of $\boldsymbol{x}$. Naively implementing the EM algorithm for each point will lead to a label switching problem, a common issue in mixture models. This may lead to interchanging estimation of component distributions and result in misleading treatment effect estimation. To address this issue, we will propose a modified EM algorithm. In the E-step,

we will calculate the expectation based on the complete log-likelihood function. In the M-step, we will maximize the local log-likelihood function with the kernel function.

Suppose in the $l$th iteration, we have $\pi_g^l(\cdot)$, $\boldsymbol{m}_g^l(\cdot)$, and $\Sigma_g^{(l)}(\cdot)$, $g = 1, 2$, then in the E-step of $(l+1)$th iteration, the expectation of the latent variable $Z_{gi}$ is given by

$$r_{gi}^{(l+1)} = \frac{\pi_g(\boldsymbol{x}_{gi})\phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_g(\boldsymbol{x}_{gi}), \Sigma_g(\boldsymbol{x}_{gi})\}}{\pi_g(\boldsymbol{x}_{gi})\phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_g(\boldsymbol{x}_{gi}), \Sigma_g(\boldsymbol{x}_{gi}) + (1-\pi_g(\boldsymbol{x}_{gi}))\phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_{g'}(\boldsymbol{x}_{gi}), \Sigma_{g'}(\boldsymbol{x}_{gi})\}}. \tag{5.4}$$

In the M-step of $(l+1)$th iteration, we maximize

$$\begin{aligned}\sum_{g}^{2}\sum_{i=1}^{n_g}\Big\{ & r_{gi}^{(l+1)}[\log\pi_g + \log\phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_g, \Sigma_g\}] \\ & + (1-r_{gi}^{(l+1)})[\log(1-\pi_g) + \log\phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_{g'}, \Sigma_{g'}\}]\Big\}K_H(\boldsymbol{x}_{gi}-\boldsymbol{x}),\end{aligned} \tag{5.5}$$

for $\boldsymbol{x} = \boldsymbol{u}_j$, $j = 1, \ldots, N$. The maximization of Equation (5.5) is equivalent to maximizing

$$\sum_{i=1}^{n_g}\left[r_{gi}^{(l+1)}\log\pi_g + (1-r_{ig}^{(l+1)})\log(1-\pi_g)\right]K_H(\boldsymbol{x}_{gi}-\boldsymbol{x}) \tag{5.6}$$

for $g = 1, 2$, and

$$\sum_{g=1}^{2}\sum_{i=1}^{n_g}\left[r_{gi}^{(l+1)}\log\phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_g, \Sigma_g\} + (1-r_{ig}^{(l+1)})\log\phi\{\boldsymbol{y}_{gi}|\boldsymbol{m}_{g'}, \Sigma_{g'}\}\right]K_H(\boldsymbol{x}_{gi}-\boldsymbol{x}), \tag{5.7}$$

separately. For $\boldsymbol{x} \in \{\boldsymbol{u}_j, j = 1, \ldots, N\}$, the solution for maximization of Equation (5.6) is

$$\pi_g^{(l+1)}(x) = \frac{\sum_{i=1}^{n_g} r_{gi}^{(l+1)}K_H(\boldsymbol{x}_{gi}-\boldsymbol{x})}{\sum_{i=1}^{n_g}K_H(\boldsymbol{x}_{gi}-\boldsymbol{x})}, \tag{5.8}$$

and the closed-from solution for Equation (5.7) is

$$\boldsymbol{m}_g^{(l+1)}(\boldsymbol{x}) = \frac{\sum_{i=1}^{n_g} r_{gi}^{(l+1)}K_H(\boldsymbol{x}_{gi}-\boldsymbol{x})\boldsymbol{y}_{gi} + \sum_{i=1}^{n_{g'}}(1-r_{g'i}^{(l+1)})K_H(\boldsymbol{x}_{g'i}-\boldsymbol{x})\boldsymbol{y}_{g'i}}{\sum_{i=1}^{n_g} r_{gi}^{(l+1)}K_H(\boldsymbol{x}_{gi}-\boldsymbol{x}) + \sum_{i=1}^{n_{g'}}(1-r_{g'i}^{(l+1)})K_H(\boldsymbol{x}_{g'i}-\boldsymbol{x})}, \tag{5.9}$$

and

$$\begin{aligned}\Sigma_g^{(l+1)}(\boldsymbol{x}) = & \frac{\sum_{i=1}^{n_g} r_{gi}^{(l+1)}K_H(\boldsymbol{x}_{gi}-\boldsymbol{x})(\boldsymbol{y}_{gi}-\boldsymbol{m}_g^{(l+1)}(\boldsymbol{x}))(\boldsymbol{y}_{gi}-\boldsymbol{m}_g^{(l+1)}(\boldsymbol{x}))^\top}{\sum_{i=1}^{n_g} r_{gi}^{(l+1)}K_H(\boldsymbol{x}_{gi}-\boldsymbol{x}) + \sum_{i=1}^{n_{g'}}(1-r_{g'i}^{(l+1)})K_H(\boldsymbol{x}_{g'i}-\boldsymbol{x})} \\ & + \frac{\sum_{i=1}^{n_{g'}}(1-r_{g'i}^{(l+1)})K_H(\boldsymbol{x}_{g'i}-\boldsymbol{x})(\boldsymbol{y}_{g'i}-\boldsymbol{m}_g^{(l+1)}(\boldsymbol{x}))(\boldsymbol{y}_{g'i}-\boldsymbol{m}_g^{(l+1)}(\boldsymbol{x}))^\top}{\sum_{i=1}^{n_g} r_{gi}^{(l+1)}K_H(\boldsymbol{x}_{gi}-\boldsymbol{x}) + \sum_{i=1}^{n_{g'}}(1-r_{g'i}^{(l+1)})K_H(\boldsymbol{x}_{g'i}-\boldsymbol{x})}.\end{aligned} \tag{5.10}$$

Furthermore, we update $\pi_g^{(l+1)}(\boldsymbol{x}_{gi})$, $m_g^{(l+1)}(\boldsymbol{x}_{gi})$ and $\Sigma_g^{(l+1)}(\boldsymbol{x}_{gi})$, $g = 1, 2$, by linearly interpolating $\pi_g^{(l+1)}(\boldsymbol{u}_i)$, $m_g^{(l+1)}(\boldsymbol{u}_i)$, and $\Sigma_g^{(l+1)}(\boldsymbol{u}_i)$, $i = 1, \ldots, N$.

**Initial Value for EM algorithm**

To implement this algorithm, we need to select initial values for the parameters and the bandwidth matrix $H$ for the kernel function $K_H(\cdot)$. Huang et al. (2013) suggested using a mixture of polynomial regression to obtain the initial value first. Similar to their idea, we use a mixture of multivariate polynomial regressions as initial value for the EM algorithm. By including higher orders of the vector of the predictor variable into the covariate matrix $\boldsymbol{X}_g$, the calculations for the mixture of polynomial regression is similar to mixture of linear regressions. The **mixtools** package in R (Benaglia et al., 2009) provides code for mixture of linear regression when the response variable $y$ is univariate. In our case, the response $y$ is bivariate and we have two groups to consider simultaneously. Therefore, we need to revise the algorithm to upgrade it for mixture of multivariate linear regressions.

In a mixture of linear regressions, $\boldsymbol{m}_g(\boldsymbol{x}) = B_g^\top \boldsymbol{x}^*$, where $\boldsymbol{x}^*$ is the vector of covariates including the intercept and higher orders of the predictor variables. Here, $B_g$ is the regression coefficient matrix, whereas $\Sigma_g(\boldsymbol{x})$ and $\pi_g(\boldsymbol{x})$ are constant functions. The conditional distribution of $\boldsymbol{Y}_g$ given $\boldsymbol{X}_g^* = \boldsymbol{x}^*$ can be written as

$$\boldsymbol{Y}_g|_{\boldsymbol{X}_g^*=\boldsymbol{x}} \sim \pi_g N(B_g^\top \boldsymbol{x}^*, \Sigma_g) + (1-\pi_g) N(B_{g'}^\top \boldsymbol{x}^*, \Sigma_{g'}),$$

where $g', g = 1, 2$ and $g \neq g'$. Given the data $\{\boldsymbol{X}_{Gi}^*, \boldsymbol{Y}_{Gi}, i = 1, \ldots, n_G, G = 1, 2\}$, the log-likelihood function is

$$\mathcal{L} = \sum_{i=1}^{2} \sum_{i=1}^{n_g} \log[\pi_g \phi\{\boldsymbol{y}_{gi}|B_g^\top \boldsymbol{x}^*, \Sigma_g\} + (1-\pi_g)\phi\{\boldsymbol{y}_{gi}|B_{g'}^\top \boldsymbol{x}^*, , \Sigma_{g'}\}].$$

It is hard to derive the MLE analytically and, thus, we still utilize the EM algorithm for it. As before, we denote the true group membership by $Z_{gi}$ defined in (5.3) as missing information. The complete log-likelihood function is

$$\mathcal{L}_C = \sum_{g}^{2} \sum_{i=1}^{n_g} \Big\{ z_{gi}[\log \pi_g + \log \phi\{\boldsymbol{y}_{gi}|B_g^\top \boldsymbol{x}_{gi}^*, \Sigma_g\}] \\ + (1-z_{gi})[\log(1-\pi_g) + \log \phi\{\boldsymbol{y}_{gi}|B_{g'}^\top \boldsymbol{x}_{gi}^*, \Sigma_{g'}\}] \Big\}.$$

Suppose in the $l$th iteration, we have $\pi_g^{(l)}$, $B_g^{(l)}$, and $\Sigma_g^{(l)}$, $g = 1, 2$. In the E-step of the

$(l+1)$th iteration, the expectation of the latent variable $Z_{gi}$ is given by

$$r_{gi}^{(l+1)} = \frac{\pi_g \phi\{\boldsymbol{y}_{gi}|B_g^\top \boldsymbol{x}_{gi}^*, \Sigma_g\}}{\pi_g \phi\{\boldsymbol{y}_{gi}|B_g^\top \boldsymbol{x}_{gi}^*, \Sigma_g\} + (1-\pi_g)\phi\{\boldsymbol{y}_{gi}|B_{g'}^\top \boldsymbol{x}_{gi}^*, \Sigma_{g'}\}}. \tag{5.11}$$

In the M-step of the $(l+1)$th iteration, we maximize the following function:

$$\sum_g^2 \sum_{i=1}^{n_g} \Big\{ r_{gi}^{(l+1)} [\log \pi_g + \log \phi\{\boldsymbol{y}_{gi}|B_g^\top \boldsymbol{x}_{gi}^*, \Sigma_g\}] \tag{5.12}$$

$$+ (1 - r_{gi}^{(l+1)})[\log(1-\pi_g) + \log \phi\{\boldsymbol{y}_{gi}|B_{g'}^\top \boldsymbol{x}_{gi}^*, \Sigma_{g'}\}] \Big\}, \tag{5.13}$$

which is equivalent to maximizing

$$\mathcal{L}_1 = \sum_{i=1}^{n_g} \{ r_{gi}^{(l+1)} \log \pi_g + (1 - r_{gi}^{(l+1)}) \log(1-\pi_g) \} \tag{5.14}$$

and

$$\mathcal{L}_2 = \sum_{i=1}^{n_g} r_{gi}^{(l+1)} \log \phi(\boldsymbol{y}_{gi}|B_g^\top \boldsymbol{x}_{gi}^*, \Sigma_g) + \sum_{i=1}^{n_{g'}} (1 - r_{g'i}^{(l+1)}) \log \phi(\boldsymbol{y}_{g'i}|B_g^\top \boldsymbol{x}_{g'i}^*, \Sigma_g). \tag{5.15}$$

The maximizor for (5.14) is:

$$\pi_g^{(l+1)} = \frac{\sum_{i=1}^{n_g} r_{gi}^{(l+1)}}{n}, g = 1, 2. \tag{5.16}$$

Taking derivatives of $\mathcal{L}_2$ with respective to $B_g$ and $\Sigma_g$, we have

$$\frac{\partial \mathcal{L}_2}{\partial B_g} = \left[ \sum_{i=1}^{n_g} r_{gi}^{(l+1)} \left( -\boldsymbol{x}_{gi}^* \boldsymbol{y}_{gi}^\top + \boldsymbol{x}_{gi}^* \boldsymbol{x}_{gi}^{*\top} B_g \right) + \sum_{i=1}^{n_{g'}} \left( 1 - r_{g'i}^{(l+1)} \right) \left( -\boldsymbol{x}_{g'i}^* \boldsymbol{y}_{g'i}^\top + \boldsymbol{x}_{g'i}^* \boldsymbol{x}_{g'i}^{*\top} B_g \right) \right] \Sigma_g^{-1},$$

and

$$\begin{aligned} \frac{\partial \mathcal{L}_2}{\partial \Sigma_g} = & -\frac{1}{2} \sum_{i=1}^{n_g} r_{gi}^{(l+1)} \left( \Sigma_g^{-1} - \Sigma_g^{-1} (\boldsymbol{y}_{gi} - B_g^\top \boldsymbol{x}_{gi}^*)(\boldsymbol{y}_{gi} - B_g^\top \boldsymbol{x}_{gi}^*)^\top \Sigma_g^{-1} \right) \\ & - \frac{1}{2} \sum_{i=1}^{n_g'} \left( 1 - r_{g'i}^{(l+1)} \right) \left( \Sigma_g^{-1} - \Sigma_g^{-1} (\boldsymbol{y}_{g'i} - B_g^\top \boldsymbol{x}_{g'i}^*)(\boldsymbol{y}_{g'i} - B_g^\top \boldsymbol{x}_{g'i}^*)^\top \Sigma_g^{-1} \right). \end{aligned}$$

Thus, the maximizor for $\mathcal{L}_2$ is

$$B_g^{(l+1)} = \left[ X_g^{*\top} R_g^{(l+1)} X_g^* + X_{g'}^{*\top} (1 - R_{g'}^{(l+1)}) X_{g'}^* \right]^{-1} \left[ X_g^{*\top} R_g^{(l+1)} Y_g + X_{g'}^{*\top} (1 - R_{g'}^{(l+1)}) Y_{g'} \right], \tag{5.17}$$

and

$$\Sigma_g^{(l+1)} = \frac{\sum_{i=1}^{n_g} r_{gi}^{(l+1)} (\boldsymbol{y}_{gi} - B_g^{(l+1)\top} x_{gi}^*)(\boldsymbol{y}_{gi} - B_g^{(l+1)\top} x_{gi}^*)^\top}{\sum_{i=1}^{n_g} r_{gi}^{(l+1)} + \sum_{i=1}^{n'_g} \left(1 - r_{g'i}^{(l+1)}\right)} + \frac{\sum_{i=1}^{n'_g} \left(1 - r_{g'i}^{(l+1)}\right) (\boldsymbol{y}_{g'i} - B_g^{(l+1)\top} x_{g'i}^*)(\boldsymbol{y}_{g'i} - B_g^{(l+1)\top} x_{g'i}^*)^\top}{\sum_{i=1}^{n_g} r_{gi}^{(l+1)} + \sum_{i=1}^{n'_g} \left(1 - r_{g'i}^{(l+1)}\right)}, \tag{5.18}$$

where

$$R_g^{(l+1)} = \begin{pmatrix} r_{g1}^{(l+1)} & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & r_{gn}^{(l+1)} \end{pmatrix}, \quad X_g = \begin{pmatrix} \boldsymbol{X}_{g1}^{*\top} \\ \vdots \\ \boldsymbol{X}_{gn}^{*\top} \end{pmatrix}, \quad \text{and } Y_g = \begin{pmatrix} \boldsymbol{Y}_{g1}^{\top} \\ \vdots \\ \boldsymbol{Y}_{gn}^{\top} \end{pmatrix}.$$

For the initial values of the mixture of multivariate polynomial regression, we propose to use the previous information for mixing proportions $\pi_g$, $g = 1, 2$ and polynomial regression estimations for $B_g$ and $\Sigma_g$ disregarding the misclassification. The proposed estimation procedure is summarized as the following:

**EM Algorithm**

**Initial Value:** Utilize EM-algorithm for mixture of polynomial regressions with constant portions, regressionn coefficient matrix and variances in (5.11), (5.17) and (5.18), and obtain the estimates of coefficient matrix $\tilde{B}_g$, and $\tilde{\pi}_g$, $\tilde{\Sigma}_g$, $g = 1, 2$. Set the initial value as $\boldsymbol{m}_g^{(1)}(\boldsymbol{x}) = B_g^\top \boldsymbol{x}^*$, $\Sigma_g^{(1)}(\boldsymbol{x}) = \tilde{\Sigma}_g$, and $\pi_g^{(1)}(\boldsymbol{x}) = \tilde{\pi}_g$.

**E-step**: Use Equation (5.4) to calculate $r_{gi}^{(l)}$ for $i = 1, \dots, n$, and $g = 1, 2$.

**M-step**: For $g = 1, 2$ and $j = 1, \dots, N$, evaluate $\pi_g^{(l+1)}(u_j)$ in (5.8), $\boldsymbol{m}_g^{(l+1)}(u_j)$ in (5.9), and $\Sigma^{(l+1)}(u_j)$ in (5.10). Further obtain $\pi_g^{(l+1)}(\boldsymbol{X}_{gi})$, $\boldsymbol{m}_g^{(l+1)}(\boldsymbol{X}_{gi})$, and $\Sigma_g^{(l+1)}(\boldsymbol{X}_{gi})$ using linear interpolation.

## 5.4 Summary and Conclusion

In this chapter, we proposed a mixture of nonparametric regression models for estimating treatment effect in the presence of misclassification errors and covariate information. We provide conditions for the identifiability of the model and utilize the kernel regression method to estimate the component functions, nonparametrically. A modified EM algorithm is derived for the mixture of nonparametric regression. We propose a mixture of

multivariate polynomial regressions for obtaining initial values of which EM algorithm is also provided.

For the standard EM algorithm, it is well known that it possesses an ascent property, i.e., the likelihood function $l(\boldsymbol{\theta}^{(l)})$ increase for each subsequent iteration. The proposed EM algorithm can be viewed as a generalization of the standard EM algorithm for the nonparametric mixture of regression. Further investigation is needed to confirm if the modified algorithm still preserves the desired ascent property.

To efficiently implement the modified EM algorithm, we need to select a proper bandwidth matrix for the kernel regression. In the univariate case, Huang et al. (2013) propose a multifold cross validation (CV) method to choose the bandwidth which may not be efficient to use in the multivaraite setting. Duong (2007) introduced R package **ks** for bandwidth matrix selection in multivariate kernel smoothing. The selection is based on the Mean Integrated Squared Error (MISE) criterion,

$$\mathrm{M}ISE(H) = E \int_{R^d} \left[\hat{f}(\boldsymbol{x}, H) - f(\boldsymbol{x})\right]^2 d\boldsymbol{x},$$

where $f$ is the density function of $\boldsymbol{X}_i$ and $\hat{f}(\boldsymbol{x}, H) = n^{-1} \sum_{i=1}^{n} K_H(\boldsymbol{x} - \boldsymbol{X}_i)$. This criterion may not be suitable in our case. We will use simulation studies to check the performance of these bandwidth selectors and investigate proper bandwidth selection method for our case.

Through the iterations of the EM-algorithm, we obtain a local constant estimator $\widetilde{\boldsymbol{\theta}} = (\widetilde{\pi}_1, \widetilde{\pi}_2, \widetilde{\boldsymbol{m}}_1, \widetilde{\boldsymbol{m}}_2, \widetilde{\Sigma}_1, \widetilde{\Sigma}_2)$ as the maximizer of the local log-likelihood function (5.2). The asymptotic bias, variance, and normality of this estimator is yet to be developed.These estimators can be used to approximate the component functions. Then the treatment effects can be estimated by comparing $\widetilde{\boldsymbol{m}}_1$ and $\widetilde{\boldsymbol{m}}_2$ adjusted for the differences in $\widetilde{\Sigma}_1$ and $\widetilde{\Sigma}_2$ over the range of the covariates $\boldsymbol{X}$.

## 5.5 Appendix

In this subsection, we provide detailed proofs of for the theoretical results in Section 5.2.

*Proof of Theorem 5.2.1.* Let $S = \{\boldsymbol{x} : (\boldsymbol{m}_1(\boldsymbol{x}), \Sigma_1(\boldsymbol{x})) = (\boldsymbol{m}_2(\boldsymbol{x}), \Sigma_2(\boldsymbol{x}))\}$ be the subset of $R^p$ where the mean and covariance functions intersect. Under the assumption 3, any

point in $S$ is an isolated point. Thus, $S$ is countable and has no limit point. We can represent $S$ as a sequence $\{\boldsymbol{x}_k\}$.

Suppose that model (5.1) have another representation

$$\boldsymbol{Y}_g|_{\boldsymbol{X}_g=\boldsymbol{x}} \sim \pi_g^*(\boldsymbol{x})N\{\boldsymbol{m}_g^*(\boldsymbol{x}), \Sigma_g^*(\boldsymbol{x})\} + (1-\pi_g^*(\boldsymbol{x}))N\{\boldsymbol{m}_{g'}^*(\boldsymbol{x}), \Sigma_{g'}^*(\boldsymbol{x})\}, \tag{5.19}$$

where $g', g = 1, 2$ and $g \neq g'$. Yakowitz and Spragins (1968) established the identifiability of finite mixture of the multivariate Gaussian family up to label switching. Hence, for any given $\boldsymbol{x} \notin S$, model (5.1) is identifiable up to label switch. Therefore, there exists a permutation $\omega_{\boldsymbol{x}} = \{\omega_{\boldsymbol{x}}(1), \omega_{\boldsymbol{x}}(2)\}$ of set $\{1, 2\}$ depending on $\boldsymbol{x}$, such that

$$(\pi^*_{\omega_{\boldsymbol{x}}(g)}(\boldsymbol{x}), m^*_{\omega_{\boldsymbol{x}}(g)}(\boldsymbol{x}), \Sigma^*_{\omega_{\boldsymbol{x}}(g)}(\boldsymbol{x})) = (\pi_g(\boldsymbol{x}), m_g(\boldsymbol{x}), \Sigma_g(\boldsymbol{x})), \text{ for } g = 1, 2. \tag{5.20}$$

Because all parameter functions are continuous and they only intersect on $S$, for all open set $O$ with $O \cap S = \emptyset$, the label cannot be switched. Hence, the permutation $\omega_{\boldsymbol{x}}$ is constant on $O$. On the other hand, for $\boldsymbol{x}_k \in S$, because there is no limit point in $S$, we can find a neighborhood of $\boldsymbol{x}_k$, say $B_{\boldsymbol{x}_k}$, such that $\{B_{\boldsymbol{x}_k} \backslash \{\boldsymbol{x}_k\}\} \cap S = \emptyset$. Under assumption 3, the mean and covariance functions have different derivatives at intersection point $\boldsymbol{x}_k$ in both groups. Thus the permutation $\omega_{\boldsymbol{x}}$ must be constant on $B_{\boldsymbol{x}_k}$ since (5.20) implies the identity of parameter functions' derivatives in the neighborhood of $\boldsymbol{x}_k$. Hence, the permutation $\omega_{\boldsymbol{x}}$ is independent of $\boldsymbol{x}$ and

$$(\pi^*_{\omega(g)}(\boldsymbol{x}), m^*_{\omega(g)}(\boldsymbol{x}), \Sigma^*_{\omega(g)}(\boldsymbol{x})) = (\pi_g(\boldsymbol{x}), m_g(\boldsymbol{x}), \Sigma_g(\boldsymbol{x})),$$

for $g = 1, 2$. □

# Chapter 6 Conclusion and Future Directions

## 6.1 Conclusion

This dissertation focused on a pre-stratified pre-post design and addressed the estimation of treatment effects with misclassification problems in four distinct situations.

In Chapter 2, we addressed the problem of estimating and testing treatment effects with continuous multivariate outcomes. We proposed two methods for estimating and testing treatment effects. First, when the misclassification errors are known from previous studies, we developed moment-based test and confidence interval procedures that are accurate in finite samples. Based on this test, we also developed methods for sample size and power calculations. Second, we proposed likelihood-based procedures for estimation and testing via the EM algorithm when the misclassification errors are unknown. Chapter 3 further investigated the situation when the misclassification rates are unknown, but the validation (training) samples from infallible classifiers are available. We derived consistent estimators of the misclassification error rates using a novel distance-based criterion. Essentially, we extended the moment-based and likelihood-based procedures in Chapter 2 to the case when validation data is available.

In Chapter 4, we developed a fully nonparametric method for estimating and testing treatment effect when the normality assumption is not valid for the outcome variables, but the validation (training) data is available. We modeled the distribution of the outcomes by a nonparametric mixture of unknown distributions. We used functionals of these distribution functions to characterize treatment effects. Consistent estimators for the misclassification error rates as well as the treatment effect were provided. We also derived the asymptotic distributions of these estimators and proposed testing and estimating procedures based on these distributions.

In Chapter 5, we investigated a nonparametric finite mixture of regression models to the distributions of outcomes when some covariates associated with the misclassification error rates and treatment outcomes are collected. We established conditions for the identi-

fiability of this model. We utilized kernel methods and proposed a modified EM algorithm to estimate the component regression functions nonparametrically.

## 6.2 Future Directions

### Mixture of Multivariate Normal Model

For the finite mixture of the multivariate normal model discussed in Chapter 2 and Chapter 3, the covariance matrices of the two groups are assumed to be the same. Though it is reasonable to assume that the treatment only affects the means of the distribution, this assumption could be restrictive for some applications. The moment-based estimator does not involve covariance matrices, but its variance estimation is affected by them. The corresponding sample size determination formulas need to be recalculated when the covariance matrices are not equal. We also need to reformulate the likelihood-based approach and recalculate the corresponding E and M steps. Moreover, the treatment effect needs to adjust for the difference in covariance matrices in two groups.

### Nonparametric Finite Mixture Model

In the nonparametric finite mixture model, we assumed that the mixing proportions are known, or validation (training) data exists to avoid the nonidentifiability issue in the mixture model. This assumption can be restrictive in applications. It may be possible to establish identifiability by making some assumptions on the nature of dependence between the pre and post-measurements or using a semi-parametric dependence model. Corresponding inferential procedures need to be derived.

When the pre-and post-treatment measurements are univariate, we provided a fully nonparametric method of estimating and testing the misclassification error rates and the treatment effect. The extension of these results to the situation when the outcome measurements are multivariable variables would be useful. Proper functional of the distribution functions should be defined to assess the treatment effect. Estimators for the misclassification error rates and treatment effect need to be recalculated in the multivariate case. Their asymptotic distributions also need investigation.

**Nonparametric Finite Mixture of Regressoin Model**

The standard EM algorithm possesses an ascent property such that the likelihood function increase after each iteration. We plan to study and show that the proposed modified EM algorithm preserves this desirable property. The algorithm's performances are also affected by choice of the bandwidth matrix for the kernel regression. We plan to investigate the performance of the existing bandwidth selection method through simulations. Furthermore, we plan to develop the asymptotic results for the local log-likelihood estimators obtained from the proposed EM algorithm iterations. Finally, we will define the treatment effect in this model and establish estimating and testing procedures for this effect.

**Bibliography**


Akritas, M. G. (1990). The rank transform method in some two-factor designs. *Journal of the American Statistical Association 85*(409), 73–78.

Akritas, M. G. (1991). Limitations of the rank transform procedure: a study of repeated measures designs, Part I. *Journal of the American Statistical Association 86*(414), 457–460.

Akritas, M. G. (1992). Rank transform statistics with censored data. *Statistics & Probability Letters 13*(3), 209–221.

Akritas, M. G. and S. F. Arnold (1994). Fully nonparametric hypotheses for factorial designs I: multivariate repeated measures designs. *Journal of the American Statistical Association 89*(425), 336–343.

Akritas, M. G. and E. Brunner (1997). A unified approach to rank tests for mixed models. *Journal of Statistical Planning and Inference 61*(2), 249 – 277.

Allman, E. S., C. Matias, and J. A. Rhodes (2009). Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics 37*(6A), 3099–3132.

Anderson, T. W. (2003). *An introduction to multivariate statistical analysis*.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics 12*(2), 171–178.

Azzalini, A. and A. Capitanio (2003). Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 65*(2), 367–389.

Battistin, E. and B. Sianesi (2011). Misclassified treatment status and treatment effects: an application to returns to education in the united kingdom. *Review of Economics and Statistics 93*(2), 495–509.

Benaglia, T., D. Chauveau, D. Hunter, and D. Young (2009). mixtools: An r package for analyzing finite mixture models. *Journal of Statistical Software 32*(6), 1–29.

Benaglia, T., D. Chauveau, and D. R. Hunter (2009). An EM-like algorithm for semi- and nonparametric estimation in multivariate mixtures. *Journal of Computational and Graphical Statistics 18*(2), 505–526.

Bordes, L., S. Mottelet, and P. Vandekerkhove (2006). Semiparametric estimation of a two-component mixture model. *The Annals of Statistics 34*(3), 1204–1232.

Brunner, E., A. C. Bathke, and F. Konietschke (2018). *Rank and Pseudo-Rank Procedures for Independent Observations in Factorial Designs*. Springer.

Brunner, E., H. Dette, and A. Munk (1997). Box-type approximations in nonparametric factorial designs. *Journal of the American Statistical Association 92*(440), 1494–1502.

Brunner, E., F. Konietschke, M. Pauly, and M. L. Puri (2017). Rank-based procedures in factorial designs: hypotheses about non-parametric treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 79*(5), 1463–1485.

Brunner, E. and U. Munzel (2000). The nonparametric Behrens-Fisher problem: asymptotic theory and a small-sample approximation. *Biometrical Journal 42*(1), 17–25.

Brunner, E., U. Munzel, and M. L. Puri (1999). Rank-score tests in factorial designs with repeated measures. *Journal of Multivariate Analysis 70*(2), 286 – 317.

Brunner, E. and N. Neumann (1982). Rank tests for correlated random variables. *Biometrical Journal 24*(4), 373–389.

Castro, H., D. Pillay, C. Sabin, and D. T. Dunn (2012). Effect of misclassification of antiretroviral treatment status on the prevalence of transmitted hiv-1 drug resistance. *BMC medical research methodology 12*(1), 1–5.

Chauveau, D. and V. T. L. Hoang (2016). Nonparametric mixture models with conditionally independent multivariate component densities. *Computational Statistics & Data Analysis 103*, 1–16.

Chen, C.-F., J.-R. Lin, and J.-P. Liu (2013). Statistical inference on censored data for targeted clinical trials under enrichment design. *Pharmaceutical statistics 12*(3), 165–173.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incom-

plete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 39*(1), 1–22.

Duong, T. (2007). ks: Kernel density estimation and kernel discriminant analysis for multivariate data in r. *Journal of Statistical Software 21*(7), 1–16.

Eling, P., J. Maes, and M. Van Haaf (2006). Processing of emotionally toned pictures in dementia. *International Journal of Geriatric Psychiatry: A journal of the psychiatry of late life and allied sciences 21*(9), 831–837.

Flahault, A., M. Cadilhac, and G. Thomas (2005). Sample size calculation should be performed for design accuracy in diagnostic test studies. *Journal of clinical epidemiology 58*(8), 859–862.

Gentili, C., M. I. Gobbini, E. Ricciardi, N. Vanello, P. Pietrini, J. V. Haxby, and M. Guazzelli (2008). Differential modulation of neural activity throughout the distributed neural system for face perception in patients with social phobia and healthy subjects. *Brain research bulletin 77*(5), 286–292.

Hall, P. (1981). On the non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society: Series B (Statistical Methodological) 43*(2), 147–156.

Hall, P. and X.-H. Zhou (2003). Nonparametric estimation of component distributions in a multivariate mixture. *The Annals of Statistics 31*(1), 201–224.

Harrar, S. W., A. Amatya, and L. Kalachev (2016). Assessing treatment efficacy in the presence of diagnostic errors. *Statistics in Medicine 35*(29), 5338–5355.

Harrar, S. W. and A. C. Bathke (2012). A modified two-factor multivariate analysis of variance: asymptotics and small sample approximations. *Annals of the Institute of Statistical Mathematics 64*(1), 135–165.

Harrar, S. W., M. B. Feyasa, and E. Wencheko (2020). Nonparametric procedures for partially paired data in two groups. *Computational Statistics & Data Analysis 144*, 106903.

Hinman, L. M., S. M. Huang, J. Hackett, W. H. Koch, P. Y. Love, G. Pennello, A. Torres-Cabassa, and C. Webster (2006). The drug diagnostic co-development concept paper. *The Pharmacogenomics Journal 6*(6), 375–380.

Huang, M., R. Li, and S. Wang (2013). Nonparametric mixture of regression models. *Journal of the American Statistical Association 108*(503), 929–941.

Huang, M. and W. Yao (2012). Mixture of regression models with varying mixing proportions: a semiparametric approach. *Journal of the American Statistical Association 107*(498), 711–724.

Hunter, D. R., S. Wang, and T. P. Hettmansperger (2007). Inference for mixtures of symmetric distributions. *The Annals of Statistics 35*(1), 224–251.

Johnson, R. A., D. W. Wichern, et al. (2007). *Applied multivariate statistical analysis*, Volume 6. Prentice Hall Upper Saddle River, NJ.

Juszczynski, P., G. Woszczek, M. Borowiec, M. Kowalski, T. Robak, P. Bilinski, G. Salles, and K. Warzocha (2002). Comparison study for genotyping of a single-nucleotide polymorphism in the tumor necrosis factor promoter gene. *Diagnostic Molecular Pathology 11*(4), 228–233.

Karunamuni, R. J. and J. Wu (2009). Minimum Hellinger distance estimation in a nonparametric mixture model. *Journal of Statistical Planning and Inference 139*(3), 1118–1133.

Konietschke, F., S. W. Harrar, K. Lange, and E. Brunner (2012). Ranking procedures for matched pairs with missing data – asymptotic theory and a small sample approximation. *Computational Statistics & Data Analysis 56*(5), 1090 – 1102.

Levine, M., D. R. Hunter, and D. Chauveau (2011). Maximum smoothed likelihood for multivariate mixtures. *Biometrika 98*(2), 403–416.

Li, M., T. Yu, and Y.-F. Hu (2015). The impact of companion diagnostic device measurement performance on clinical validation of personalized medicine. *Statistics in medicine 34*(14), 2222–2234.

Lie, R. T., I. Heuch, and L. M. Irgens (1994). Maximum likelihood estimation of the proportion of congenital malformations using double registration systems. *Biometrics*, 433–444.

Lin, H., S. K. McClintock, and J. M. Williamson (2011). Correction for two-group sample size calculation with uncertain group membership. *Journal of Data Science 9*(2),

155–170.

Lindsay, B. G. and P. Basak (1993). Multivariate normal mixtures: a fast consistent method of moments. *Journal of the American Statistical Association 88*(422), 468–476.

Liu, J.-P. and J.-R. Lin (2008). Statistical methods for targeted clinical trials under enrichment design. *Journal of the Formosan Medical Association 107*(12), S35–S42.

Liu, J.-P., J.-R. Lin, and S.-C. Chow (2009). Inference on treatment effects for targeted clinical trials under enrichment design. *Pharmaceutical Statistics: The Journal of Applied Statistics in the Pharmaceutical Industry 8*(4), 356–370.

Louis, T. A. (1982). Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological) 44*(2), 226–233.

Magnus, J. R. and H. Neudecker (1979). The commutation matrix: some properties and applications. *The Annals of Statistics*, 381–394.

Mann, H. B. and D. R. Whitney (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics 18*(1), 50–60.

Meng, X.-L. and D. B. Rubin (1991). Using em to obtain asymptotic variance-covariance matrices: The sem algorithm. *Journal of the American Statistical Association 86*(416), 899–909.

Nedelman, J. (1988). The prevalence of malaria in garki, nigeria: double sampling with a fallible expert. *Biometrics*, 635–655.

O'Donnell, C. J., R. J. Glynn, T. S. Field, R. Averback, S. Satterfield, G. C. Friesenger II, J. O. Taylor, and C. H. Hennekens (1999). Misclassification and under-reporting of acute myocardial infarction by elderly persons: implications for community-based observational studies and clinical trials. *Journal of clinical epidemiology 52*(8), 745–751.

Qin, J. (1999). Empirical likelihood ratio based confidence intervals for mixture proportions. *The Annals of Statistics 27*(4), 1368–1384.

Qiu, S.-F., J. He, J.-R. Tao, M.-L. Tang, and W.-Y. Poon (2019). Comparison of disease prevalence in two populations under double-sampling scheme with two fallible clas-

sifiers. *Journal of Applied Statistics*, 1–27.

Satterfield, B. C., J. P. Wisor, S. A. Field, M. A. Schmidt, and H. P. A. Van Dongen (2015). TNF$\alpha$ G308A polymorphism is associated with resilience to sleep deprivation-induced psychomotor vigilance performance impairment in healthy young adults. *Brain, Behavior, and Immunity 47*, 66–74.

Tenenbein, A. (1970). A double sampling scheme for estimating from binomial data with misclassifications. *Journal of the American Statistical Association 65*(331), 1350–1361.

Thompson, G. L. (1990). Asymptotic distribution of rank statistics under dependencies with multivariate application. *Journal of Multivariate Analysis 33*(2), 183 – 211.

Thompson, G. L. (1991). A unified approach to rank tests for multivariate and repeated measures designs. *Journal of the American Statistical Association 86*(414), 410–419.

Titterington, D. M. (1983). Minimum distance non-parametric estimation of mixture proportions. *Journal of the Royal Statistical Society: Series B (Statistical Methodological) 45*(1), 37–46.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin 1*(6), 80–83.

Xu, J. and S. W. Harrar (2012). Accurate mean comparisons for paired samples with missing data: An application to a smoking-cessation trial. *Biometrical Journal 54*(2), 281–295.

Yakowitz, S. J. and J. D. Spragins (1968). On the identifiability of finite mixtures. *The Annals of Mathematical Statistics*, 209–214.

Zheng, C. and Y. Wu (2020). Nonparametric estimation of multivariate mixtures. *Journal of the American Statistical Association 115*(531), 1456–1471.

**Vita**

# *Zi* Ye

## EDUCATION

- **University of Kentucky** Lexington, Kentucky
  - Ph.D. in Statistics August 2018 - August 2021
  - M.S. in Statistics August 2016 - May 2018
- **Wuhan University** Wuhan, Hubei, China
  - M.S. in Probability September 2012 - June 2015
  - B.S. in Mathematics (Major) September 2008 - June 2012
  - B.A. in Finance (Minor) September 2008 - June 2012

## WORKING EXPERIENCE

- **Primary Instructor, University of Kentucky** August 2018 - May 2021
- **Teaching Assistant, University of Kentucky** August 2016 - May 2018

## PUBLICATIONS

- Ye, Z. and Harrar, S. (2021), "Multiple Treatment Effects in Contaminated Randomized Trials", *Pharmaceutical Statistics*, Major revision requested.
- Ye, Z. and Harrar, S. (2020), "Nonparametric Mixture Model: Application in Clinical Trials", *Journal of the American Statistical Association*, submitted.
- Ye, Z. (2016), "Cramer type moderate deviations for the number of renewals", *Statistics & Probability Letters*, 119: 194-199