Summer 2019

Estimation Problems for Pooled Data

Xichen Mou

Estimation problems for pooled data

by

Xichen Mou

Bachelor of Science
Shandong University 2010

Master of Science
University of Illinois, Urbana-Champaign 2013

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Statistics

College of Arts and Sciences

University of South Carolina

2019

Accepted by:

Joshua M. Tebbs, Major Professor

Dewei Wang, Major Professor

Lianming Wang, Committee Member

Yuan Wang, Committee Member

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

Dedication

To my grandfather, Yu Mu, for his wisdom and love.

Acknowledgments

I want to express my sincere gratitude to my advisor Dr. Joshua Tebbs for his guidance in both academics and in life. He gave me considerable freedom to explore the areas in which I am interested, and he guided me in research with his erudite discussions. He is also a great friend who always listens to me and helps me whenever I meet difficulties. I want to express great thanks to my co-advisor Dr. Dewei Wang. He has encouraged me to innovate and is always ready to respond to my research questions with smart ideas. With his support, my research has made substantial progress. He is also a friend and gives excellent suggestions on how to make decisions regarding my future.

I also want to thank my committee members Dr. Lianming Wang and Dr. Yuan Wang for carefully reviewing my dissertation. Their suggestions will keep my research strong and reliable.

Finally, I want to thank my father Yuan Mu, mother Chuancui Liu, and wife Rui Qi, for their endless love and support, which help me move forward without fear.

Abstract

In epidemiological applications, individual specimens (e.g., blood, urine, etc.) are often pooled together to detect the presence of disease or to measure the concentration level of a specific biomarker. Due to the advantage of cost efficiency, pooled data are also seen in diverse areas such as genetics, animal ecology, and environmental science. With pooled data, individual observations are masked and new statistical methods are needed to estimate characteristics such as disease prevalence, the underlying density function of a biomarker, etc. We focus on three estimation problems for pooled data. Chapters 2 and 3 propose nonparametric estimators for the density function $f(Y|X)$ of a biomarker's concentration Y given a single covariate X. We consider two types of pooling strategies: random pooling in Chapter 2 and homogeneous pooling in Chapter 3. For both strategies, we derive asymptotic properties of density estimators and evaluate performance through numerical studies in a variety of settings. We further illustrate the proposed methods by applying them to a polyfluorochemical data set. In Chapter 4, we develop a method to estimate disease prevalence and diagnostic accuracy probabilities (sensitivity and specificity) simultaneously from two-stage hierarchical group testing data. Through theoretical calculation and simulation, our approach is shown to be more efficient than existing methods which utilize only pooled-level responses.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Literature Review

The idea of pooling specimens traces back to 1943, when Robert Dorfman suggested pooling blood samples to detect syphilis among American soldiers in World War II. Instead of testing the samples individually, Dorfman proposed that one could pool multiple blood samples together and then test the pooled sample. If the pool diagnosis was negative, then all individuals inside the pool would be diagnosed as negative. If the pool diagnosis was positive, then retests would be carried out for every individual inside the pool. When the disease prevalence is low, most pools end up with a negative result and no further retests are needed. This, in turn, leads to a reduction in testing costs. Since Dorfman's strategy was introduced, following the same idea, more complex pooling strategies have been proposed, and corresponding statistical methods have been developed to analyze these data. Pooling specimens, which is often referred to as group testing or pooled testing, is a common practice in numerous areas such as genetics (Gastwirth, 2000), animal ecology (Dhand, Johnson, and Toribio, 2010), and nutrition (Fahey, Ourisson, and Degnan, 2006).

One of the main advantages of group testing is cost reduction. For example, the State Hygienic Laboratory (SHL) at the University of Iowa is responsible for chlamydia and gonorrhea testing throughout the state. These screening activities are part of sexually transmitted disease assessment and prevention programs that occur in all 50 states; see also Lewis, Lockary, and Kobic (2012). During 2009-2014, the

SHL estimates that using group testing has saved $3.1 million. Besides saving costs, specimens may also be pooled to avoid undetectability below a certain concentration level. For example, Bates et al. (2005) pool sera specimens to achieve desired volume limits (approximately 50 ml) in a study of organochlorine compounds exposure among adults in New Zealand.

The group testing literature can generally be split into two areas: case identification and estimation. The case identification literature describes pooling strategies and statistical models to classify an individual as positive or negative while achieving cost efficiency and accuracy. On the other hand, the estimation literature describes methods to estimate relevant characteristics such as disease prevalence in a population, diagnostic accuracy, and regression models which connect pooled responses to covariate information.

1.1.1 Case Identification

Most case identification strategies can be classified as one of two types: hierarchical and non-hierarchical. The hierarchical method first tests master pools and then divides positive pools into non-overlapping subgroups for retesting. If positive diagnoses are still found in the subgroups, further stages of dividing and retesting are performed. Dorfman's method, which uses two stages, is the simplest version of a hierarchical algorithm. The non-hierarchical approach usually refers to array testing algorithms. For instance, two-dimensional array testing arranges specimens in a two-dimensional array. All rows and columns are pooled and are tested in the first stage. When one column and one row both show a positive result, the intersection is retested due to high suspicion of positivity. Non-hierarchical algorithms usually perform better when the disease prevalence is very low while two-stage hierarchical algorithms are easier to apply. Kim et al. (2007) offered a complete comparison of various retesting algorithms in group testing. This comparison was performed under

the assumption that all individuals have the same probability of infection.

In reality, however, one may assess the level of risk by a person's covariate information. To address this issue, "informative group testing" assumes different individual risks and was proposed by Bilder, Tebbs, and Chen (2010), Black, Bilder, and Tebbs (2012) and McMahan, Tebbs, and Bilder (2012a). Treating individual information (e.g., age, race, etc.) as covariates, these authors predict an individual's probability of disease by a regression model and then adjust group testing strategies according to the predicted probability on a per-individual basis. More recently, inspired by assays that can detect multiple diseases simultaneously, Tebbs, McMahan, and Bilder (2013) proposed a two-stage hierarchical group testing algorithm for multiple infections. Hou et al. (2017) later generalized this work to hierarchical group testing algorithms with more than two stages.

1.1.2 Estimation

There is a large literature on estimation with pooled data when the observations are viewed as concentration levels of a specific biomarker. Early works such as Faraggi, Reiser, and Schisterman (2003), Liu and Schisterman (2003) and Schisterman et al. (2005) assume the concentration level is Gaussian and utilize the estimated density function to make inference on the receiver-operating characteristic (ROC) curve. Schisterman (2003) proposed an estimator of Youden's index, Mumford et al. (2006) considered the effect of the limit of detection, and Vexler, Schisterman, and Liu (2008) proposed a nonparametric kernel type method to analyze pooled data and ROC curves. More recently, when covariate information is available, regression analysis has been utilized to analyze the relationship between the response and covariates. Ma et al. (2011) proposed a linear regression model to discover the relationship between the pooled biomarker and easily obtained covariates such as age. Malinovsky, Albert, and Schisterman (2012) extended the linear model by introducing

random effects, and Mitchell et al. (2014) considered the case when the response is right-skewed. Liu, McMahan, and Gallagher (2017) proposed a general regression framework to incorporate different parametric models, while Lin and Wang (2018) proposed a semi-parametric approach.

Regression analysis is generally limited to the mean biomarker level and thus cannot reveal the complete relationship (e.g., skewness, mode, quantiles, etc.) between the covariate and the response. Direct estimation of a biomarker's conditional density function conveys more information. In this context, the main difficulty lies in how to recover individual information from the complex convolution of pooled data. The aforementioned literature about ROC curves focuses on distributions of two distinct groups such as disease status. However, it is also meaningful to consider the case when the covariate is continuous. Based on pooled observations, Vexler, Liu, and Schisterman (2010) proposed a nonparametric estimator of a density function in the absence of covariates. Linton and Whang (2002) examined the problem of estimating a conditional density function nonparametrically when both the responses and covariates are pooled and error-prone.

When the response is binary, a common goal in group testing is to estimate the population prevalence p. Early works such as Thompson (1962) and Sobel and Elashoff (1975) derived the maximum likelihood estimator (MLE) of p in the ideal situation that no misclassification occurs. Hughes-Oliver and Swallow (1994) proposed an adaptive pooling strategy to adjust the group size based on *a priori* information from a small number of pre-tests. In practice, when diagnostic tests are imperfect, false negative and false positive results may occur. Sensitivity (S_e) and specificity (S_p) are two common criteria to measure diagnostic accuracy. Taking misclassification into account, Tu, Litvak, and Pagano (1995) provided a maximum likelihood estimator of p and derived asymptotic properties. They showed that the prevalence p can be estimated more precisely by testing pools than by using individual testing. Liu

et al. (2012) provided the upper bound on p under which group testing is more precise when fixing the number of subjects or the number of assays. Huang et al. (2017) further assumed misclassification rates are unknown and proposed a design that can estimate p, S_e, and S_p simultaneously. They proved their designs are theoretically optimal using certain criteria. Based on more modern assays that can detect multiple diseases simultaneously, Tebbs, McMahan, and Bilder (2013) estimated the prevalence of multiple infections using an expectation-maximization algorithm. Warasi et al. (2016) addressed a similar problem by using a Bayesian approach.

Similar to the continuous case, regression models have been proposed to incorporate covariate information when the response is binary. Farrington (1992) proposed a generalized linear model (GLM) to accommodate pooled data. Vansteelandt, Goetghebeur, and Verstraeten (2000) extended Farrington (1992)'s GLM by flexibly allowing imperfect tests, a broader collection of link functions, and different covariate values in the same pool. Chen, Tebbs, and Bilder (2009) developed a regression method to include random effects, and McMahan et al. (2017) proposed a Bayesian approach that can utilize historical data for any group testing protocol. The latter authors proposed a Metropolis-Hastings algorithm to fit a GLM and to estimate the covariate effects as well as assay accuracy probabilities. Warasi et al. (2017) proposed a regression model that incorporates the dilution effect and provided a hypothesis test to test for dilution.

Besides parametric models for binary responses, there has been substantial progress in developing semi-parametric and nonparametric methods in recent years. Wang et al. (2013) proposed a semi-parametric regression model that can adjust for multiple covariates. Assuming the responses are pooled randomly, Delaigle and Meister (2012) constructed a kernel-type nonparametric estimator of $p(x)$ where x is a scalar covariate measured on each individual. Delaigle and Hall (2012) considered homogeneous pooling and proposed the corresponding kernel-type nonparametric estimator. They

showed that when the pool size is large, estimation accuracy is better when pooling individuals homogeneously. Delaigle and Zhou (2015) later considered the scenario when both the binary response and a continuous covariate are pooled. Delaigle and Hall (2015) proposed a nonparametric method to accommodate the dilution effect when estimating $p(x)$ along with S_e and S_p.

1.2 Outline

This dissertation focuses on three estimation problems. Chapter 2 and Chapter 3 examine the case when the pooled outcome is continuous. In Chapter 2, we introduce a nonparametric method to estimate the conditional density function $f(Y|X)$ where Y is a continuous response; e.g., the concentration level of a biomarker, and X is a continuous covariate such as age. Instead of observing the response variable individually, the Y's are pooled randomly, and we observe only the arithmetic average $\bar{Y}$ plus a measurement error ϵ. We refer to the estimator of $f(Y|X)$ in this case as the random pooling (RP) estimator. In Chapter 3, instead of pooling the Y's randomly, we pool Y's with similar covariates. We propose a new nonparametric estimator of $f(Y|X)$ under this pooling strategy which we call the homogeneous pooling (HP) estimator. In Chapter 4, we examine the case when the pooled outcome is binary. We compare estimation efficiency of the disease prevalence, S_e, and S_p between two group testing algorithms: Dorfman testing and master pool testing. This chapter can be viewed as an extension of Huang et al. (2017) when Dorfman testing is used to resolve positive pools.

Chapter 2

Conditional density estimation with random pooling

2.1 Introduction

In epidemiological and environmental studies, it is increasingly common to pool individual specimens (e.g., blood, urine, etc.) together and then test the pools for the concentration level of a particular biomarker. There are various advantages of doing so. For example, specimens may be pooled to reduce the cost associated with assays (Caudill, 2012), to avoid undetectability below a certain concentration level (Bates et al., 2005; Schisterman and Vexler, 2008), or to conserve expensive, irreplaceable specimens (Saha-Chaudhuri and Weinberg, 2013). The idea of pooling traces back to 1943 when Dorfman pooled blood samples to detect syphilis among American soldiers in World War II. Since this seminal work, pooled data have been widely seen in diverse applications such as in genetics (Gastwirth, 2000), animal ecology (Dhand, Johnson, and Toribio, 2010), and nutrition (Fahey, Ourisson, and Degnan, 2006).

Regression is the most common method to analyze the relationship between a response variable and covariates; see, e.g., Parikh et al. (2006) and Ai et al. (2010). Considerable efforts have been made to estimate regression models with pooled data. Ma et al. (2011), Malinovsky, Albert, and Schisterman (2012), Mitchell et al. (2014), McMahan et al. (2016), and Liu, McMahan, and Gallagher (2017) focused on parametric models and Lin and Wang (2018) proposed a semi-parametric approach. However, regression models are limited to the expectation of the response; thus, they can-

not reveal the relationship between the covariate and response variable completely. Direct estimation of a biomarker's conditional density function conveys more information. One application of this is to make inference on the receiver-operating characteristic (ROC) curve, which is a well-accepted tool to analyze the efficacy of a biomarker to distinguish between two populations. For pooled data, Faraggi, Reiser, and Schisterman (2003), Liu and Schisterman (2003), and Mumford et al. (2006) studied this topic from a parametric point of view, and Vexler, Schisterman, and Liu (2008) proposed a nonparametric method.

The aforementioned literature in density estimation focuses on distributions of two distinct groups such as those bifurcated by disease status. However, it is also meaningful to estimate the density function conditional on a continuous covariate. In this chapter, our goal is to estimate $f(y|x)$, the density function of Y, which is a continuous response representing an individual's concentration level, conditional on X, which is a continuous covariate. Instead of observing the response variable directly, the Y's are pooled randomly, and we observe only the arithmetic average $\bar{Y}$ plus a measurement error ϵ, whereas the covariate is observed for every individual. This scenario is often seen when analyzing the relationship between a biomarker, which is prohibitively expensive to measure, and an easily obtained demographic covariate such as age.

This problem is closely related to the deconvolution problem which aims to recover the individual density function from the sum of independent random variables. In nonparametric statistics, deconvolution is common in the study of measurement error (Carroll and Hall, 1988; Fan, 1991b), where the observed Y is considered to be the sum of the random variable X and a measurement error ϵ. Pooled data can also be viewed as the convolution of every individual's contribution in the pool, and techniques commonly seen in measurement error are also applied in this area. Vexler, Liu, and Schisterman (2010) proposed a nonparametric density estimator without considering

continuous covariates. Delaigle and Hall (2012), Delaigle, Hall, and Wishart (2014), and Delaigle and Zhou (2015) studied deconvolution problems for pooled data with binary response.

In this chapter, we propose a nonparametric local polynomial estimator of the conditional density $f(y|x)$ when the response Y is randomly assigned to a pool. In economics, Linton and Whang (2002) considered the case where both the response and the covariate are pooled. One could apply their approach to our context by artificially pooling observed individual-level information together. However, such an aggregation leads to a loss of information and thus potentially compromises estimation. At the same time, extending their idea to our situation leads to three problems. Computationally, they directly estimate $f(y, x)$, the joint density of (Y, X), via a two-dimensional inverse Fourier transform and then construct their estimator using the formula $f(y|x) = f(y, x)/f_X(x)$, where $f_X(x)$ is the marginal density of X (which can be estimated from the data). In two dimensions, performing an inverse Fourier transform severely increases the computational burden and the difficulty in bandwidth selection. Second, their estimator involves the complex-valued root calculation of a bivariate characteristic function, which is hard to implement. Lastly, their proof implicitly assumes the characteristic function of X is real, which is only true when X is a symmetric random variable; see Delaigle and Zhou (2015). The method we introduce uses a one-dimensional inverse Fourier transform, and our theoretical results hold for skewed random variables as well.

The rest of this chapter is organized as follows. In Section 2.2, we introduce the model and notation. In Sections 2.3 and 2.4, we propose estimators of $f(y|x)$ and present their asymptotic properties. In Section 2.5, we compare our estimators with the one from Linton and Whang (2002). A practical data-driven bandwidth selection method is included. We leave a comprehensive numerical study regarding estimators of $f(y|x)$ and real data analysis to Chapter 3, where we compare our estimator with

those formulated under homogeneous pooling. Additional proofs and supplementary materials are given in Appendix A.

2.2 Model and Data

We are interested in estimating the conditional density function $f(y|x)$ of Y given $X = x$, where Y is a continuous response (e.g., the concentration level of a certain biomarker, etc.) and X is a continuous covariate (e.g., age, etc.). The ideal unobserved data consist of independent and identical distributed (iid) pairs (Y_{ij}, X_{ij}), $i = 1, \ldots, c$ and $j = 1, \ldots, J$. The index ij represents the ith individual in the jth group, and the group size is fixed to be c, an integer greater than or equal to two.

In our pooling context, the Y_{ij}'s are not available. Instead, we observe an error-laden measurement of the arithmetic average of the Y_{ij}'s in the jth pool; i.e., $\bar{Z}_j = \bar{Y}_j + \epsilon_j$, where $\bar{Y}_j = c^{-1} \sum_{i=1}^{c} Y_{ij}$ and ϵ_j denotes a measurement error. One can view $\bar{Y}_j$ as the biomarker concentration level of the jth pooled specimen; see, e.g., Weinberg and Umbach (1999), Faraggi, Reiser, and Schisterman (2003), Vexler, Schisterman, and Liu (2008), Malinovsky, Albert, and Schisterman (2012), McMahan et al. (2016), and Lin and Wang (2018). This is appropriate if the amount of specimen contributed by each individual is the same and there exists no neutralization effect during pooling. We further assume the error terms ϵ_j are iid random variables with a known density function f_ϵ and are independent of the (Y_{ij}, X_{ij})'s. We wish to estimate $f(y|x)$ nonparametrically from $\{(\bar{Z}_j, X_{1j}, \ldots, X_{cj}) : j = 1, \ldots, J\}$.

2.3 Methodology

2.3.1 Regression equation

In our problem, the starting point is to construct a response from the data for each covariate X_{ij} with respect to $f(y|x)$ such that $E(\text{response}|X_{ij} = x) = f(y|x)$. If

we had the ideal data, which contain all the Y_{ij}'s, the response for X_{ij} could be approximated by $K_h(Y_{ij} - y)$, where $K_h(\cdot) = h^{-1}K(\cdot/h)$; i.e.,

$$E\{K_h(Y_{ij} - y)|X_{ij} = x\} \approx f(y|x).$$

Then an estimator of $f(y|x)$ could be defined by the minimizer of the weighted least-squares sum

$$\sum_{j=1}^{J}\sum_{i=1}^{c}\{K_h(Y_{ij} - y) - \beta_0\}^2 \bar{K}_{\bar{h}}(X_{ij} - x) \tag{2.1}$$

with respect to β_0, where K ($\bar{K}$) is a kernel function and h ($\bar{h}$) is a bandwidth. However, when the Y_{ij}'s are aggregated to the $\bar{Z}_j$'s with potential measurement error, it is not obvious how to extract a response from $\bar{Z}_j$ to estimate $f(y|x)$. The method we propose exploits the conditional characteristic function (CF) of $Y|X = x$, which is denoted by $\phi_{Y|X=x}(t) = E\{\exp(itY)|X = x\}$ for $t \in \mathbb{R}$. Rather than extracting information from $\bar{Z}_j$, we allocate $\bar{Z}_j$ to each X_{ij} and then integrate out terms in $\bar{Z}_j$ that are not related to X_{ij}.

To be more specific, denote by $\phi_{\bar{Z}|X=x}$ the CF of $\bar{Z}_j|X_{ij} = x$. Further, let $\phi_{\bar{Z}}$, ϕ_Y, and ϕ_ϵ be the CFs of $\bar{Z}_j$, Y_{ij}, and ϵ_j, respectively. Straightforward calculation yields

$$\begin{aligned}
\phi_{\bar{Z}|X=x}(ct) &= E\{\exp(ict\bar{Z}_j)|X_{ij} = x\} \\
&= E\{\exp(ict\bar{Y}_j)|X_{ij} = x\} \times \phi_\epsilon(ct) \\
&= E\{\exp(itY_{ij})|X_{ij} = x\} \times \prod_{k=1,k\neq i}^{c} E\{\exp(itY_{kj})\} \times \phi_\epsilon(ct) \\
&= \phi_{Y|X=x}(t)\phi_Y(t)^{c-1}\phi_\epsilon(ct).
\end{aligned}$$

The second equality above is due to the independence between ϵ_j and (Y_{ij}, X_{ij}). The last two equalities are because the Y_{kj}'s, where $k \neq i$, are independent of (Y_{ij}, X_{ij}) and $E\{\exp(itY_{kj})\} = \phi_Y(t)$. Thus, $\phi_{Y|X=x}(t) = \phi_{\bar{Z}|X=x}(ct)/\{\phi_Y(t)^{c-1}\phi_\epsilon(ct)\}$ provided $\phi_Y(t) \neq 0$ and $\phi_\epsilon(t) \neq 0, \forall t \in \mathbb{R}$, which will be assumed throughout this chapter. Because Y_{ij}'s are latent and only the $\bar{Z}_j$'s are observed, we further transform ϕ_Y back

to $\phi_{\bar{Z}}$ through $\phi_{\bar{Z}}(ct) = \phi_Y(t)^c \phi_\epsilon(ct)$. Then, we have

$$\phi_{Y|X=x}(t) = \frac{\phi_{\bar{Z}|X=x}(ct)}{\{\phi_{\bar{Z}}(ct)\}^{(c-1)/c}\{\phi_\epsilon(ct)\}^{1/c}}.$$

By applying the Fourier inversion theorem, we obtain our regression equation

$$\begin{aligned} f(y|x) &= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\phi_{Y|X=x}(t)dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-ity}\phi_{\bar{Z}|X=x}(ct)}{\{\phi_{\bar{Z}}(ct)\}^{(c-1)/c}\{\phi_\epsilon(ct)\}^{1/c}}dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-ity}E(e^{ict\bar{Z}_j}|X_{ij}=x)}{\{\phi_{\bar{Z}}(ct)\}^{(c-1)/c}\{\phi_\epsilon(ct)\}^{1/c}}dt = E(Y_{ij}^*|X_{ij}=x), \end{aligned} \tag{2.2}$$

where Y_{ij}^* is the response for X_{ij} that we should extract from $\bar{Z}_j$ with respect to $f(y|x)$; this response is given by

$$Y_{ij}^* = Y_{ij}^*(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-ity}e^{ict\bar{Z}_j}}{\{\phi_{\bar{Z}}(ct)\}^{(c-1)/c}\{\phi_\epsilon(ct)\}^{1/c}}dt. \tag{2.3}$$

2.3.2 Nadaraya-Watson-type estimator

We start by constructing a basic ratio-type estimator of $f(y|x)$ via Equations (2.2) and (2.3), which is similar in spirit to the Nadaraya-Watson (NW) estimator employed in standard nonparametric regression problems. This estimator can be viewed as the minimizer of $\sum_{j=1}^{J}\sum_{i=1}^{c}(Y_{ij}^* - \beta_0)^2 \bar{K}_{\bar{h}}(X_{ij} - x)$ with respect to β_0, similar to (2.1). However, Y_{ij}^* involves $\phi_{\bar{Z}}$ which remains unknown. Thus, we use an empirical surrogate $\widehat{Y}_{ij}^*$ of Y_{ij}^*. More specifically, we estimate $\phi_{\bar{Z}}(t) = E\{\exp(it\bar{Z}_j)\}$ by $\widehat{\phi}_{\bar{Z}}(t) = J^{-1}\sum_{j=1}^{J}\exp(it\bar{Z}_j)$ and replace $\phi_{\bar{Z}}(ct)$ in Equation (2.3) by $\widehat{\phi}_{\bar{Z}}(ct)$ and define

$$\widehat{Y}_{ij}^* = \widehat{Y}_{ij}^*(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \frac{e^{-ity}e^{ict\bar{Z}_j}}{\{\widehat{\phi}_{\bar{Z}}(ct)\}^{(c-1)/c}\{\phi_\epsilon(ct)\}^{1/c}}\phi_K(ht)dt,$$

where ϕ_K denotes the Fourier transform of a symmetric kernel function K; i.e., $\phi_K(t) = \int_{-\infty}^{\infty} e^{itx}K(x)dx$, and $h = h_n > 0$ is a bandwidth.

The role of $\phi_K(t)$ is similar to that of K in (2.1). More specifically, if we define

$$\widehat{\phi}_{K^*}(ht) = \frac{\phi_K(ht)}{\{\widehat{\phi}_{\bar{Z}}(ct)\}^{(c-1)/c}\{\phi_\epsilon(ct)\}^{1/c}},$$

and let $\widehat{K^*}$ be the inverse Fourier transform of $\widehat{\phi}_{K^*}$; i.e.,

$$\widehat{K^*}(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\widehat{\phi}_{K^*}(t)dt,$$

then

$$\widehat{Y}^*_{ij} = \widehat{Y}^*_{ij}(y) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{it(c\bar{Z}_j - y)}\widehat{\phi}_{K^*}(ht)dt = \widehat{K}^*_h(c\bar{Z}_j - y), \tag{2.4}$$

where $\widehat{K}^*_h(\cdot) = h^{-1}\widehat{K^*}(\cdot/h)$. Consequently, one can view $\widehat{K}^*_h(c\bar{Z}_j - y)$ as a surrogate of $K_h(Y_{ij} - y)$ in (2.1) when the Y_{ij}'s are aggregated in the errors, and the regression Equation (2.2) can be written as $E\{\widehat{K}^*_h(c\bar{Z}_j - y)|X_{ij} = x\} \approx f(y|x)$. Consequently, our NW-type estimator of $f(y|x)$ is defined as the minimizer of

$$\sum_{j=1}^{J}\sum_{i=1}^{c}\left\{\widehat{K}^*_h(c\bar{Z}_j - y) - \beta_0\right\}^2 \bar{K}_{\bar{h}}(X_{ij} - x), \tag{2.5}$$

with respect to β_0, which is given by

$$\widehat{f}_{RN}(y|x) = \frac{\sum_{j=1}^{J}\sum_{i=1}^{c}\widehat{K}^*_h(c\bar{Z}_j - y)\bar{K}_{\bar{h}}(X_{ij} - x)}{\sum_{j=1}^{J}\sum_{i=1}^{c}\bar{K}_{\bar{h}}(X_{ij} - x)}. \tag{2.6}$$

Some comments are in order regarding potential computational issues with $\widehat{f}_{RN}$. First, plugging Equation (2.4) into Equation (2.6) yields a more computationally feasible formula

$$\widehat{f}_{RN}(y|x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\frac{\widehat{\phi}_{\bar{Z}|X=x}(ct)}{\{\widehat{\phi}_{\bar{Z}}(ct)\}^{(c-1)/c}\{\phi_\epsilon(ct)\}^{1/c}}\phi_K(ht)dt, \tag{2.7}$$

where $\widehat{\phi}_{\bar{Z}|X=x}(t) = \sum_{j=1}^{J}\sum_{i=1}^{c}\exp(it\bar{Z}_j)\bar{K}_{\bar{h}}(X_{ij} - x)/\sum_{j=1}^{J}\sum_{i=1}^{c}\bar{K}_{\bar{h}}(X_{ij} - x)$ can be viewed as an NW-type estimator of $\phi_{\bar{Z}|X=x}(t)$. Rather than computing the integral in Equation (2.4) for each $K^*_h(c\bar{Z}_j - y)$, only one integral is needed if using Equation (2.7). Second, if $\bar{Z}_j$ (ϵ_j) is asymmetric, then $\phi_{\bar{Z}}$ (ϕ_ϵ) is a complex function. This creates a problem when computing both $\{\widehat{\phi}_{\bar{Z}}(ct)\}^{(c-1)/c}$ and $\{\phi_\epsilon(ct)\}^{1/c}$ in Equation (2.7), because the roots of a complex-valued function are not unique. A careless connection among the roots would easily violate the continuity requirement of a CF on the real domain. Fortunately, Theorem 7.6.2 in Chung (2001) proved the existence

and uniqueness of a connection such that the resulting CF is valid. Building on this theorem, Meister (2007) and Delaigle and Zhou (2015) have provided numerical guidance on how to construct the ideal connection.

2.3.3 Local-polynomial-type estimator

We use the regression function in Equation (2.2) and the surrogate in Equation (2.4) to construct a local polynomial estimator (Fan, 2018); specifically, a local linear estimator of $f(y|x)$. In addition to estimating $f(y|x)$, the local polynomial approximation can estimate the derivatives $\partial^d f(y|x)/\partial x^d$, for $d \geq 1$, provided these derivatives exist. This proceeds by first approximating $f(y|X_{ij})$ in a neighborhood of x via an ℓth-order polynomial, where $\ell \geq d$, that is,

$$f(y|X_{ij}) \approx \beta_0 + \beta_1(X_{ij} - x) + \cdots + \beta_\ell(X_{ij} - x)^\ell,$$

where $\beta_d = \beta_d(y|x) = (d!)^{-1}\partial^d f(y|x)/\partial x^d$, for $d = 0, 1, \ldots, \ell$. Then, with the regression equation in (2.2) and the construction of $\widehat{Y}^*_{ij}$ in Equation (2.4), the β_d's can be estimated by minimizing the weighted least-squares objective function

$$\sum_{j=1}^{J}\sum_{i=1}^{c}\left\{\widehat{K}^*_h(c\bar{Z}_j - y) - \sum_{d=0}^{\ell}\beta_d(X_{ij} - x)^d\right\}^2 \bar{K}_{\bar{h}}(X_{ij} - x). \tag{2.8}$$

For $d \leq \ell$, the ℓth-order local polynomial estimator of $\partial^d f(y|x)/\partial x^d$ is defined by $d!\widehat{\beta}_d$, where $\widehat{\beta}_d = \boldsymbol{e}_d^T \boldsymbol{S}_N^{-1}(x)\widehat{\boldsymbol{T}}_N(y|x)$, $\boldsymbol{e}_d = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ is an $(\ell+1)$-dimensional vector with 1 as the $(d+1)$th element, $\boldsymbol{S}_N(x) = [S_{N,k_1,k_2}(x)]_{0\leq k_1,k_2\leq \ell}$ is an $(\ell+1)$-dimensional square matrix, and $\widehat{\boldsymbol{T}}_N(y|x) = (\widehat{T}_{N,0}(y|x), \ldots, \widehat{T}_{N,\ell}(y|x))^T$, where

$$S_{N,k_1,k_2} = \sum_{j=1}^{J}\sum_{i=1}^{c}\bar{K}_{\bar{h}}(X_{ij} - x)(X_{ij} - x)^{k_1+k_2}$$

and

$$\widehat{T}_{N,k}(y|x) = \sum_{j=1}^{J}\sum_{i=1}^{c}\widehat{K}^*_h(c\bar{Z}_j - y)\bar{K}_{\bar{h}}(X_{ij} - x)(X_{ij} - x)^k.$$

Consequently, we define our estimator of $f(y|x)$ by $\widehat{\beta}_0$ with $\ell = 1$; i.e.,

$$\widehat{f}_{RL}(y|x) = \frac{\widehat{T}_{N,0}(y|x)S_{N,1,1}(x) - \widehat{T}_{N,1}(y|x)S_{N,0,1}(x)}{S_{N,0,0}(x)S_{N,1,1}(x) - S_{N,0,1}(x)S_{N,1,0}(x)}. \tag{2.9}$$

Similar to Equation (2.7), one can also obtain a more computationally feasible expression of $\widehat{f}_{RL}(y|x)$.

2.4 Theoretical Properties

2.4.1 Smoothness classes

As noted earlier, our approach to estimate $f(y|x)$ falls into the general context of nonparametric deconvolution. In the statistical literature, density deconvolution has a long history that can be traced back to Carroll and Hall (1988), where the goal was to estimate the density of Y based on iid samples from the convolution $Y + \epsilon$. This problem has been thoroughly addressed by Fan (1991b) and its variations have generated much interest. One key result delivered by Fan's seminal works is that the asymptotic properties of a nonparametric deconvolution density estimator is heavily influenced by the decay rate of the error's CF. This also holds in our context; in addition, because our observed data are convolutions of the form $c^{-1}\sum_{i=1}^{c} Y_{ij} + \epsilon_j$, asymptotic properties of estimators of $f(y|x)$ not only depend on the decay rate of ϕ_ϵ but also on the one of both ϕ_Y and $\phi_{Y|X=x}$.

Following Fan (1991b), we categorize CFs into one of two classes: ordinary smooth and super smooth. For example, the Laplace, gamma, double exponential CFs (and their convolutions) are ordinary smooth, while the Cauchy and Gaussian CFs (and their convolutions) are super smooth. In this chapter, we consider the case where ϕ_Y and $\phi_{Y|X=x}$ belong to the same smoothness class. When ϕ_ϵ and ϕ_Y ($\phi_{Y|X=x}$) are ordinary smooth, we refer to this scenario as "Condition OO." When at least one of ϕ_ϵ and ϕ_Y ($\phi_{Y|X=x}$) are super smooth, we refer to this as "Condition SS." The specific definitions are stated below.

Condition OO: α_0, β_0, $\rho_0(x) > 0$ and $\rho_0(x)$ has third derivative; there exist constants A_1, A_2, and $A_3(x) > 0$ such that as $t \to \infty$,

$$|\phi_Y(t)t^{\alpha_0}| \to A_1,\ |\phi_Y'(t)t^{\alpha_0+1}| = O(1),\ |\phi_\epsilon(t)t^{\beta_0}| \to A_2,\ |\phi_\epsilon'(t)t^{\beta_0+1}| = O(1),$$
$$|\phi_{Y|X=x}(t)t^{\rho_0(x)}| \to A_3(x),\ |\partial\phi_{Y|X=x}(t)/\partial t \cdot t^{\rho_0(x)+1}| = O(1).$$

Condition SS: $\alpha_2,\ \beta_2,\ \rho_2(x) \geq 0,\ \alpha_2\beta_2 \neq 0,\ \rho_1(x)$ and $\rho_2(x)$ have third derivative; there exist constants $B_1,\ B_1',\ B_2,\ B_2',\ B_3(x) > 0$ and $\gamma,\ \zeta,\ \varrho(x) > 0$ such that as $t \to \infty$,

$$|\phi_Y(t)||t|^{-\alpha_1}\exp(|t|^{\alpha_2}/\gamma) \to B_1,\ |\phi_Y'(t)||t|^{-\alpha_1-\alpha_2+1}\exp(|t|^{\alpha_2}/\gamma) = O(1),$$
$$|\phi_\epsilon(t)||t|^{-\beta_1}\exp(|t|^{\beta_2}/\zeta) \to B_2,\ |\phi_\epsilon'(t)||t|^{-\beta_1-\beta_2+1}\exp(|t|^{\beta_2}/\zeta) = O(1),$$
$$|\phi_{Y|X=x}(t)||t|^{-\rho_1(x)}\exp\{|t|^{\rho_2(x)}/\varrho(x)\} \to B_3(x),$$
$$|\partial\phi_{Y|X=x}(t)/\partial t||t|^{-\rho_1(x)-\rho_2(x)+1}\exp\{|t|^{\rho_2(x)}/\varrho(x)\} = O(1).$$

Note that in condition SS, α_2 or $\beta_2 = 0$ means the corresponding characteristic function falls back into ordinary smoothness class.

2.4.2 Regularity conditions

Below are the regularity conditions needed for Theorems 2.1 and 2.2, which are presented in Section 2.4.3.

(C1) $\bar{K}(x)$ is a bounded symmetric density function with mean 0.

(C2) $\phi_K(t)$ is a real-valued symmetric continuous function with support $[-1, 1]$ and $\sup_{t\in\mathbb{R}}|\phi_K^{(l)}(t)| < \infty$ for $l = 0, 1$, where $\phi_K^{(l)}(t)$ is the lth of $\phi_K(t)$ derivative with respect to t.

(C3) For each x and y, $\partial^3 f(y|x)/\partial x^3 < \infty$, $\partial^2 f(y|x)/\partial y^2 < \infty$. In addition, f_ϵ is a symmetric continuously differentiable density function.

(C4) The Fourier transformation of K satisfies either $\phi_K(t) = I_{[-1,1]}(t)$ or

$$\phi_K(t) = \begin{cases} 1, & |t| < \tau \\ \phi(|t|), & \tau \leq |t| \leq 1 \\ 0, & |t| > 1, \end{cases}$$

where $0 < \tau < 1$ and $\phi : [\tau, 1] \to [0, 1]$ is a continuously differentiable, non-increasing function such that $\phi(\tau) = 1$ and $\phi^{(j)}(t) = a_j(1-t)^{b-j} + o\{1-t\}^{b-j}$ as $t \to 1$ for constants $a_0 > 0$, $a_1 < 0$, and $b > 0$, where $j = 0, 1$.

Conditions (C1) and (C2) are placed on the kernels $\bar{K}$ and K, respectively, and represent standard conditions in nonparametric regression and deconvolution problems. Common kernels that satisfy (C1) include the Gaussian and the Epanechnikov kernels, while those satisfying (C2) include the infinite order sinc kernel $\phi_K(t) = I_{[-1,1]}(t)$ and the second-order kernels $\phi_K(t) = (1-t^2)^q \cdot I_{[-1,1]}(t)$, for some positive integer q. Condition (C3) describes regular smoothness conditions of f_ϵ and $f(y|x)$. Condition (C4) defines the limiting property of $\phi_K(t)$ when $|t| \to 1$, which has been used in Delaigle and Zhou (2015). This is an extra condition for Scenario SS which leads to more accurate convergence rate estimation for Theorem 2.2.

2.4.3 Asymptotics of the RP estimator

We use the general notation $\widehat{f}_{RP}(y|x)$ to denote the NW-type estimator (local constant) and the local linear estimator for random pooling (RP) and derive the theoretical properties of $\widehat{f}_{RP}(y|x)$ under Conditions OO and SS. The proofs of our theorems are long and technical and thus are placed in Appendix A. In general, local polynomial estimators can be written as

$$\widehat{f}_{RP}(y|x) = \frac{\sum_{j=1}^{J}\sum_{i=1}^{c} w_{ij}(x)\widehat{K}_h^*(c\bar{Z}_j - y)}{\sum_{j=1}^{J}\sum_{i=1}^{c} w_{ij}(x)}, \tag{2.10}$$

where $w_{ij}(x)$ is a generalized weight function related to $\bar{K}_{\bar{h}}(\cdot)$ and X_{ij}. In the local constant (NW-type) estimator, $w_{ij}(x) = \bar{K}_{\bar{h}}(X_{ij} - x)$. In the local linear estimator,

$$w_{ij}(x) = \bar{K}_{\bar{h}}(X_{ij} - x) \sum_j \sum_i \bar{K}_{\bar{h}}(X_{ij} - x)(X_{ij} - x)^2 \\ - \bar{K}_{\bar{h}}(X_{ij} - x)(X_{ij} - x) \sum_j \sum_i \bar{K}_{\bar{h}}(X_{ij} - x)(X_{ij} - x).$$

The function $b_1(x)$ in Theorems 2.1 and 2.2 depends on whether $\widehat{f}_{RP}(y|x)$ is the local constant or the local linear estimator.

Theorem 2.1. *Assume that $f_X(x) > 0$, $c > 1$, and $(2c-1)\alpha_0 + 2\beta_0 > 1$. Under Conditions OO and (C1)–(C3), if $\bar{h} \to 0$, $h \to 0$, $Jh^{2c\alpha_0+2\beta_0} \to \infty$, and $\bar{h}^{-1}h^{2\{\alpha_0-\rho_0(x)\}} \to \infty$, then*

$$\widehat{f}_{RP}(y|x) - f(y|x) = B_{\bar{h},h}(x,y) + V^{1/2}_{\bar{h},h}(x,y),$$

where

$$B_{\bar{h},h}(x,y) = \frac{\bar{h}^2 b_1(x)}{2} \frac{\partial^2 f(y|x)}{\partial x^2} + o_p(\bar{h}^2) + \frac{1}{2\pi} \int \exp(-ity)\phi_{Y|X=x}(t)\{\phi_K(ht) - 1\}dt$$
$$V_{\bar{h},h}(x,y) = O_p\{c^{-1+2\beta_0} J^{-1} \bar{h}^{-1} h^{-2(c-1)\alpha_0 - 2\beta_0 - 1}\}.$$

Theorem 2.1 describes asymptotic properties of $\widehat{f}_{RP}(y|x)$ when ϕ_Y, $\phi_{Y|X=x}$, and ϕ_ϵ are ordinary smooth. The terms $B_{\bar{h},h}(x,y)$ and $V_{\bar{h},h}(x,y)$ describe the bias and variance, respectively. As Delaigle and Zhou (2015) discuss, if $\int uK(u)du = 0$, $\int |u|^{2+\alpha}|K(u)|du < \infty$, for $0 < \alpha \leq 1$, and $f(y|x)$ satisfies mild conditions, the last term in $B_{\bar{h},h}(x,y)$ can be expressed as $h^2\partial^2 f(y|x)/\partial y^2 \times \int u^2 K(u)du/2$. If we take $\bar{h} = J^{-\bar{d}}$, $h = J^{-d}$ when $\bar{d} = d = 1/\{6+2(c-1)\alpha_0+2\beta_0\}$, the mean-squared error (MSE) can attain the optimal rate: $J^{-2/\{3+(c-1)\alpha_0+\beta_0\}}$. When $c = 1$ (no pooling), and $\beta_0 = 0$ (no measurement error), the optimal rate is $J^{-2/3}$, which is the optimal rate in standard conditional density estimation when individual data are available.

Theorem 2.2 describes asymptotic properties of $\widehat{f}_{RP}(y|x)$ when at least one of $\phi_Y(t)$ and $\phi_\epsilon(t)$ is super smooth. Define

$$K^*(u) = \frac{1}{2\pi}\int_{-1}^{1} e^{-itu}\frac{\phi_K(t)}{\{\phi_Y(t/h)\}^{(c-1)}\phi_\epsilon(ct/h)}dt, \tag{2.11}$$

Theorem 2.2. *Assume that $f_X(x) > 0$ and $c > 1$. Under Conditions SS and (C1)–(C4), if $\bar{h} \to 0$, $h \to 0$,*

$$\bar{h}^{-1}h^{c_1}\exp\{2\varrho(x)^{-1}h^{-\rho_2(x)} - 2\gamma^{-1}h^{-\alpha_2}\} \to \infty,$$

and

$$Jh^{c_2}\exp(-2c\gamma^{-1}h^{-\alpha_2} - 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}) \to \infty,$$

for any constant c_1, c_2, then

$$\widehat{f}_{RP}(y|x) - f(y|x) = B_{\bar{h},h}(x,y) + V^{1/2}_{\bar{h},h}(x,y),$$

where

$$B_{\bar{h},h}(x,y) = \frac{\bar{h}^2 b_1(x)}{2}\frac{\partial^2 f(y|x)}{\partial x^2} + o_p(\bar{h}^2) + \frac{1}{2\pi}\int \exp(-ity)\{\phi_{Y|X=x}(t) - 1\}\phi_K(ht)dt.$$

If $\alpha_2 \vee \beta_2 < 1$,

$$V_{\bar{h},h}(x,y) = \frac{v_1(x)f_{c\bar{Z}|x}(y)}{cJ\bar{h}h}\int K^{*2}(u)du\{1 + o(1)\}.$$

If $\alpha_2 \vee \beta_2 \geq 1$,

$$V_{\bar{h},h}(x,y) = O_p[c^{-1-2\beta_1}J^{-1}\bar{h}^{-1}h^{c_3}\exp\{2(c-1)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}],$$

for some constant c_3.

Note that the value of $\alpha_2 \vee \beta_2$ determines the value of $V_{\bar{h},h}(x,y)$ in Theorem 2.2. When $\alpha_2 \vee \beta_2 < 1$, we show in the proof that $V_{\bar{h},h}(x,y)$ can be expressed by a more accurate rate; specifically,

$$c^{-1-2\beta_1}J^{-1}\bar{h}^{-1}h^{(2b^*+1)\alpha_2+(2c-2)\alpha_1+2\beta_1-1}\exp\{(2c-2)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\},$$

where $b^* = 0$ when $\phi_K(t) = I_{[-1,1]}(t)$ and $b^* = b$ when $\phi_K(t)$ is the function defined in Condition (C4). Regardless of if $\alpha_2 \vee \beta_2 \geq 1$ or $\alpha_2 \vee \beta_2 < 1$, $V_{\bar{h},h}(x,y)$ is of order

$$O_p[J^{-1}\bar{h}^{-1}h^{c_3}\exp\{2(c-1)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}].$$

2.5 Simulation Evidence

2.5.1 Comparison between Linton's estimator and the RP estimators

We evaluate the performance of three estimators: the NW-type estimator $\widehat{f}_{RN}(y|x)$, the local linear estimator $\widehat{f}_{RL}(y|x)$, and a modified version of Linton and Whang (2002)'s estimator $\widehat{f}_{LW}(y|x)$. We first compare $\widehat{f}_{RN}(y|x)$ with $\widehat{f}_{LW}(y|x)$ by Linton and Whang (2002) who considered a similar problem but with a goal to estimate the regression curve $E(Y|X=x)$. Their approach utilizes an estimator of $f(y|x)$ which can be extended to our context. The main idea of their approach comes from the fact $f(y|x) = f(y,x)/f_X(x)$, where $f(y,x)$ is the joint density function of (Y,X). Using this formula, Linton and Whang (2002) defined a three-step estimator. The first step is to invert an empirical CF of (Y,X) back to estimate $f(y,x)$ through a two-dimensional inverse Fourier transform; the second step estimates f_X; the last step takes the ratio of their estimators of $f(y,x)$ and $f_X(x)$. Extending their approach to our problem, these three steps are now described.

Step 1: Denote by $\phi_{Y,X}(t,s) = E[\exp\{i(tY+sX)\}]$, the joint CF of (Y,X). Estimate $\phi_{Y,X}(t,s)$ by $\widehat{\phi}_{Y,X}(t,s) = \{\phi_\epsilon(ct)\}^{-1/c}\{J^{-1}\sum_{j=1}^{J}\exp(ict\bar{Z}_j + ics\bar{X}_j)\}^{1/c}$, where $\bar{X}_j = c^{-1}\sum_{i=1}^{c} X_{ij}$. Obtain

$$\widehat{f}_{LW}(y,x) = \frac{1}{4\pi^2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\exp\{-i(ty+sx)\}\widehat{\phi}_{Y,X}(t,s)\phi_K(th)\phi_{\bar{K}}(s\bar{h})dsdt$$

via a two-dimensional inverse Fourier transform, where ϕ_K ($\phi_{\bar{K}}$) is the Fourier transform of the kernel K ($\bar{K}$) and h ($\bar{h}$) is a bandwidth.

Step 2: Estimate $f_X(x)$ by $\widehat{f}_{LW}(x) = N^{-1}\sum_{j=1}^{J}\sum_{i=1}^{c}\bar{K}_{\bar{h}}(X_{ij}-x)$.

Step 3: Estimate $f(y|x)$ by $\widehat{f}_{LW}(y|x) = \widehat{f}_{LW}(y,x)/\widehat{f}_{LW}(x)$.

Recall that Linton and Whang (2002) assume individual covariates X's are pooled together and the deconvolution method is used to estimate $f_X(x)$. As the individual covariates are accessible in our scenario, we utilize the ordinary kernel density

estimator of $f_X(x)$ in Step 2 above, which has a faster convergence rate than the deconvolution method. Recall also that Linton and Whang (2002) use the same bandwidth on both kernels K and $\bar{K}$. In the three steps above, we allow h and $\bar{h}$ to be different.

It is insightful to note that, mimicking Linton and Whang (2002)'s approach, $\widehat{f}_{RN}(y|x)$ could be obtained by using the following three steps:

Step 1: Denote by

$$\widehat{\phi}_{Y,X}(t,s) = \{\phi_\epsilon(ct)\}^{-1/c} \left\{ N^{-1} \sum_{j=1}^{J} \sum_{i=1}^{c} \exp(ict\bar{Z}_j + isX_{ij}) \right\} \{\widehat{\phi}_{\bar{Z}}(ct)\}^{(c-1)/c}.$$

Obtain

$$\widehat{f}_{RN}(y,x) = \frac{1}{4\pi^2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp\{-i(ty+sx)\} \widehat{\phi}_{Y,X}(t,s) \phi_K(th) \phi_{\bar{K}}(s\bar{h}) ds dt.$$

Step 2: Estimate $f_X(x)$ by $\widehat{f}_{RN}(x) = N^{-1} \sum_{j=1}^{J} \sum_{i=1}^{c} \bar{K}_{\bar{h}}(X_{ij} - x)$.

Step 3: Estimate $f(y|x)$ by $\widehat{f}_{RN}(y|x) = \widehat{f}_{RN}(y,x)/\widehat{f}_{RN}(x)$.

A comparison between the three-step procedures for both estimators provides a better understanding of the difference between them. The essential difference lies in Step 1, where different approaches are utilized to estimate the joint CF $\phi_{Y,X}(t,s)$. Note that $\bar{h}$ (h) has exactly the same effect in both the RN and the Linton and Whang (LW) estimators; $\bar{h}$ (h) is the bandwidth of kernel $\bar{K}$ (K) which controls the weight regarding X_{ij} ($\bar{Z}_j$). This enables us to make a fair comparison between $\widehat{f}_{LW}(y|x)$ and $\widehat{f}_{RN}(y|x)$ by taking the same value for $\bar{h}$ and h. However, one drawback of the LW estimator is the cth complex-valued root of a complex bivariate function in Step 1. The existence and uniqueness of this root have been discussed in Finkelstein, Tucker, and Veeh (1999); however, there is no algorithm to calculate it to the best of our knowledge.

In the example below, we perform a simulation study to evaluate the performance of our estimators. The comparison uses a model in which $\phi_{Y,X}(t,s)$ is real; more comprehensive studies are presented in Chapter 3. Along with $\widehat{f}_{RN}(y|x)$ and $\widehat{f}_{LW}(y|x)$, we include $\widehat{f}_{RL}(y|x)$ in our comparison. Specifically, we assume

Model 1: $Y_{ij} \sim Normal(X_{ij}, 0.5^2)$,

where $X_{ij} \sim Unif(-0.5, 0.5)$, $i = 1, \ldots, c$ and $j = 1, \ldots, J$. Using the same notation as in Section 2.2, Y_{ij} can not be observed directly; $\bar{Z}_j = \bar{Y}_j + \epsilon_j$ is observed instead, where $\epsilon_j \sim Normal(0, 0.05^2)$. Due to symmetry, $J^{-1}\sum_{j=1}^{J} \exp(ict\bar{Z}_j + ics\bar{X}_j)$ in Step 1 when calculating $\widehat{f}_{LW}(y|x)$ has a negligible imaginary part; thus, we do not need to worry about the cth root of a bivariate complex function. We simulated $B = 500$ data sets from this model using the group size $c = 2$.

To compare the three estimators, we calculate the mean integrated square error (MISE) defined by

$$\int_{-0.4}^{0.4}\int_{-\infty}^{\infty}\{\widehat{f}(y|x) - f(y|x)\}^2 f(x)dydx,$$

where $\widehat{f}(y|x)$ is an estimator of $f(y|x)$. We applied a truncated integration from -0.4 to 0.4 on x to rule out poor estimation near the boundary ± 0.5. This criterion has been used in the definition of MISE previously in the literature; see, e.g., Fan and Yim (2004). To keep consistent with Linton and Whang (2002), we use Gaussian kernels for both K and $\bar{K}$. Figure 2.1 displays the mean MISE of 500 simulations for different $\bar{h}$ and h. We can see that in regards to MISE, $\widehat{f}_{RN}(y|x)$ is almost uniformly better than $\widehat{f}_{LW}(y|x)$, and $\widehat{f}_{RL}(y|x)$ overall outperforms the other two estimators. When $\bar{h}$ grows larger, both $\widehat{f}_{RN}(y|x)$ and $\widehat{f}_{LW}(y|x)$ collapse to the global constant estimator and their performance tends to be similar. On the other hand, the MISE of $\widehat{f}_{RL}(y|x)$ stays low as $\bar{h}$ increases. This is not surprising because as $\bar{h}$ grows larger, $\widehat{f}_{RL}(y|x)$ collapses to global linear regression, which fits a linear relationship between X_{ij} and Y_{ij} (which is largely maintained even after pooling).

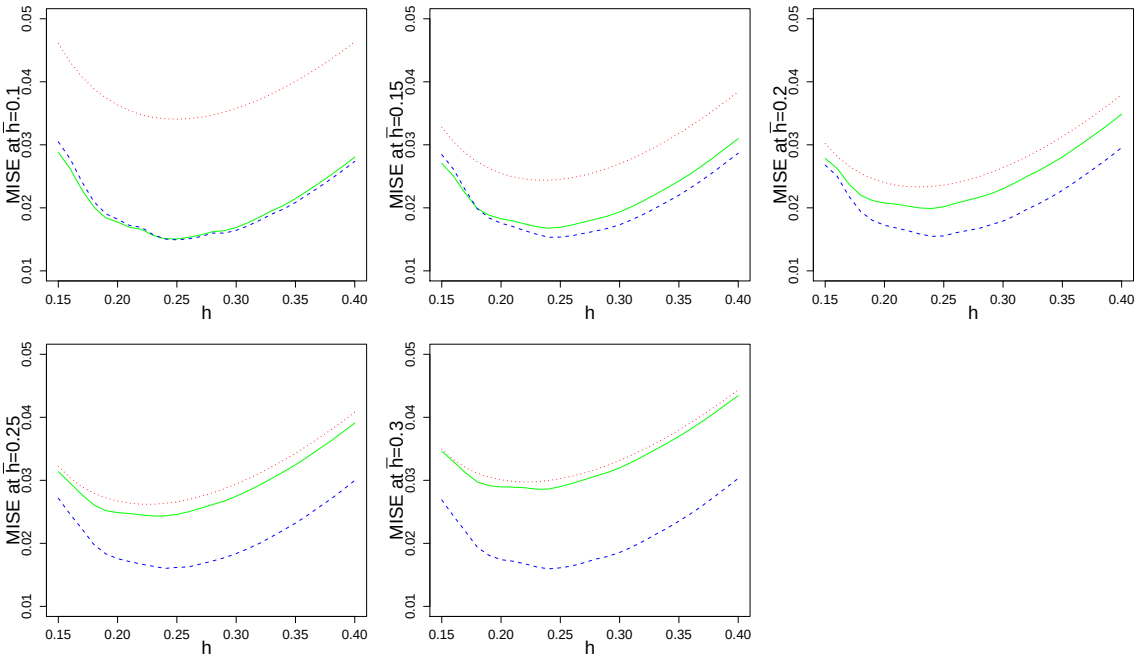

Figure 2.1: MISE of $\widehat{f}(y|x)$ for different combinations of $\bar{h}$ and h when $J = 5000$ and $c = 2$. Each plot describes the curves of mean MISE of 500 simulations at a given $\bar{h}$: Linton and Whang's estimator (dotted red), Nadaraya-Watson estimator (solid green), and local linear estimator (dotted blue). Bandwidths $\bar{h} = 0.1$, 0.15, 0.2, 0.25, 0.3 are shown.

2.5.2 Bandwidth selection

Bandwidth selection is hard in our problem. Expressions for the asymptotic bias and variance involve complicated forms of unknown parameters, which prevents us from using the plug-in method, and cross-validation methods are computationally intense. We propose a data-driven two-step bandwidth selection method which chooses $\bar{h}$ first and then h second. Simulation results show we can estimate $f(y|x)$ accurately across different values of x by using this method. Throughout this subsection, we use $\bar{K}$ as the normal kernel and K as the infinite order sinc kernel; both are commonly used kernels in the literature.

To determine $\bar{h}$, if we could observe $(Y_{1j}, X_{1j}), \ldots, (Y_{cj}, X_{cj})$ for each $j = 1, \ldots, J$, then we could estimate $f_{Y|X=x}$ by using a local constant estimator $\sum_{i,j} \bar{K}_{\bar{h}}(X_{ij} - x)K_h(Y_{ij} - y)/\sum_{i,j} \bar{K}_{\bar{h}}(X_{ij} - x)$ or a higher order local-polynomial-type estimator. In this setting, methods for selecting h and $\bar{h}$ are well developed; see, e.g., Bashtannyk and Hyndman (2001) and Hyndman and Yao (2002). However in our scenario, we do not observe the Y_{ij}'s. Instead, we extend the grouped data $(\bar{Z}_j, X_{1j}, \ldots, X_{cj})$ to c paired data $(\bar{Z}_j, X_{1j}), \ldots, (\bar{Z}_j, X_{cj})$ for each $j = 1, \ldots, J$ and construct the estimator in Equation (2.6); i.e., $\sum_{i,j} \bar{K}_{\bar{h}}(X_{ij} - x)K^*_h(\bar{Z}_j - y)/\sum_{i,j} \bar{K}_{\bar{h}}(X_{ij} - x)$ or a higher order estimator described in Equation (2.9). We use the bandwidth selection method described in Bashtannyk and Hyndman (2001) and Hyndman and Yao (2002) on the $(\bar{Z}_j, X_{ij})$'s instead of the (Y_{ij}, X_{ij})'s and take their $\bar{h}$ as our selection. Our simulation results show this approximation performs very well when the measurement error ϵ is small. Once $\bar{h}$ is determined, we can then estimate $\phi_{Y|X=x}(t)$ by

$$\widehat{\phi}_{Y|X=x}(t) = \frac{\widehat{\phi}_{\bar{Z}|X=x}(ct)}{\{\widehat{\phi}_{\bar{Z}}(ct)\}^{(c-1)/c}\{\phi_\epsilon(ct)\}^{1/c}}.$$

The bandwidth h is the tuning parameter of $\phi_K(ht)$ in the inverse Fourier transformation. We develop an adaptive bandwidth selection method for h called the Empirical Rule Selection (ERS). Here "adaptive" means h is adapted for different values of x

in the estimation process.

Our ERS method is motivated by the fact that the empirical characteristic function will become wiggly when t goes beyond a certain boundary. Diggle and Hall (1993) discussed this phenomenon and developed a selection method for nonparametric density estimation with known measurement error. However, their method involves subjective judgment and does not fit into our pooled data structure. Motivated by Diggle and Hall (1993)'s method, we develop a concise but efficient method to select h. In our case, $\phi_K(ht)$ is a truncated function between $[-1/h, 1/h]$. Letting $T = 1/h$, we apply the following algorithm to determine T:

Step 1: Locate the closest turning point T_0 in the plot of $|\widehat{\phi}_{Y|X=x}(t)|$, i.e., $T_0 = \inf\{t > 0 : \partial|\widehat{\phi}_{Y|X=x}(t)|/\partial t > 0\}$, assuming $\partial|\widehat{\phi}_{Y|X=x}(t)|/\partial t$ exists for $t \in \mathbb{R}$.

Step 2: Select T to be a multiple of T_0, i.e., $T = \lambda T_0$.

Selecting λ is based on the belief that the "best" T should be neither too conservative to be less than T_0 (which will cause an over-smooth estimate) nor too far away (which will cause a spiky estimate). Based on the observation that $\widehat{\phi}_{Y|X=x}(t)$ quickly enters a "wiggly" area, which is believed to contain little information about the data, we recommend using $\lambda = 1.2$. Even though our method involves empirical assessments, this selection turns out to be quite effective in practice. Numerical evidence demonstrating this is provided in Chapter 3, where we evaluate the performance of our estimators further.

Chapter 3

Conditional density estimation with homogeneous pooling

3.1 Introduction

In Chapter 2, we proposed nonparametric estimators of $f(y|x)$ when the response Y is randomly assigned to pools. When individual covariate information is accessible ahead of time, we can pool specimens that have similar covariates together. This is called homogeneous pooling (HP). In regression settings with continuous response, it has been shown homogeneous pooling can attain better estimation performance than random pooling in parametric settings (Mitchell et al., 2015). Delaigle and Hall (2012) proposed a nonparametric estimator for group testing binary regression with homogeneous pooled data. They also showed homogeneous pooling gains more efficiency than random pooling, especially when the group size is large.

In this chapter, we propose local polynomial estimators of $f(y|x)$ when homogeneous pooling is used. As in Chapter 2, we focus on the case where the response variable Y is continuous and show theoretically, excluding the impact of measurement error, the performance of the HP estimator essentially depends on the decay rate of $\phi_{Y|X=x}(t)$ while the random pooling (RP) estimator in Chapter 2 depends only on the decay rate of the marginal CF $\phi_Y(t)$. Furthermore, in the special case that the decay rate of $\phi_{Y|X=x}(t)$ does not depend on the covariate x, the HP estimator achieves a faster rate of convergence than the RP estimator, especially when the pool size is large. Our theoretical conclusions are reinforced through simulation,

and we apply our estimation methods to a data set regarding the bio-accumulation of perfluorooctane sulfonate (PFOS).

3.2 Data and Methodology

In homogeneous pooling, individual specimens with similar covariates are pooled together. Assuming covariates (e.g., age, etc.) are observed ahead of time, we sort the specimens in ascending order by the covariate x and pool every c specimens together. Denote $\mathcal{X} = \{X_{(1)}, \ldots, X_{(N)}\}$ as the order statistics of $\{X_1, \ldots, X_N\}$. The jth homogeneous group consists of individuals whose covariates values are

$$\{X_{(cj-c+1)}, X_{(cj-c+2)}, \ldots, X_{(cj)}\}.$$

To keep notation concise, we use the subscript (ij) to represent the ith individual in the jth homogeneous pool and relabel the ideal data as $\{Y_{(ij)}, X_{(ij)}\}$, for $i = 1, \ldots, c$, $j = 1, \ldots, J$. Note that we abuse the notation slightly by letting $X_{(ij)}$ denote the $(cj - c + i)$th order statistic of $X_1, \ldots, X_N$. Again, the $Y_{(ij)}$'s are latent, and we instead observe $\bar{Z}_{(j)} = c^{-1} \sum_{i=1}^{c} Y_{(ij)} + \epsilon_j$.

Let $\bar{X}_{(j)}$ denote the arithmetic average of the jth homogeneous pool. It is reasonable to assume the covariate values in the same pool are "close" as $N \to \infty$, i.e., $\bar{X}_{(j)} \approx X_{(ij)}$, for $i = 1, \ldots, c$. Based on this idea, Delaigle and Hall (2012) developed a nonparametric estimator of the probability of response given one covariate. With a continuous response, our approach constructs an estimator for $q(x, t) = E(e^{itY_{(ij)}} | X_{(ij)} = x) = \phi_{Y|X=x}(t)$ and applies a Fourier transformation to estimate $f(y|x)$. We start by noting

$$E(e^{ict\bar{Z}_{(j)}} | \mathcal{X}) = \prod_{i=1}^{c} E(e^{itY_{(ij)}} | X_{(ij)}) \phi_\epsilon(ct) = \prod_{i=1}^{c} q(X_{(ij)}, t) \phi_\epsilon(ct) \approx q(\bar{X}_{(j)}, t)^c \phi_\epsilon(ct)$$

and

$$E(e^{ict\bar{Z}_{(j)}} | \mathcal{X}) = E(e^{ict\bar{Z}_{(j)}} | X_{(cj-c+1)}, X_{(cj-c+2)}, \ldots, X_{(cj)}) \approx E(e^{ict\bar{Z}_{(j)}} | \bar{X}_{(j)}).$$

Combining the two equations above and letting $\bar{X}_{(j)} = x$, we obtain

$$E(e^{ict\bar{Z}_{(j)}}|\bar{X}_{(j)} = x) \approx q(x,t)^c \phi_\epsilon(ct).$$

We suggest estimating the left side using $\widehat{\phi}_{\bar{Z}|\bar{X}=x}(ct)$, which is defined by

$$\widehat{\phi}_{\bar{Z}|\bar{X}=x}(ct) = \frac{\sum_{j=1}^{J} w_j(x) \exp(ict\bar{Z}_{(j)})}{\sum_{j=1}^{J} w_j(x)}, \tag{3.1}$$

where $w_j(x)$ is a generalized weight. Analogous to that in Chapter 2, for the NW-type (local constant) estimator, $w_j(x) = K_{\bar{h}}(\bar{X}_{(j)} - x)$. For the local linear estimator,

$$\begin{aligned} w_j(x) = \bar{K}_{\bar{h}}(X_{(j)} - x) \sum_j \bar{K}_{\bar{h}}(X_{(j)} - x)(X_{(j)} - x)^2 \\ - \bar{K}_{\bar{h}}(X_{(j)} - x)(X_{(j)} - x) \sum_j \bar{K}_{\bar{h}}(X_{(j)} - x)(X_{(j)} - x). \end{aligned}$$

Then we estimate $q(x,t)$ by $\{\widehat{\phi}_{\bar{Z}|\bar{X}=x}(ct)/\phi_\epsilon(ct)\}^{1/c}$. Lastly, we use the Fourier inversion formula to estimate $f(y|x)$ by

$$\widehat{f}_{HP}(y|x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ity} \left\{ \frac{\widehat{\phi}_{\bar{Z}|\bar{X}=x}(ct)}{\phi_\epsilon(ct)} \right\}^{1/c} \phi_K(ht) dt, \tag{3.2}$$

where $\widehat{f}_{HP}(y|x)$ is the general notation for the NW-type estimator and the local linear estimator. The kernel function K ($\bar{K}$) and the bandwidth h ($\bar{h}$) are defined as in Equation (2.1).

3.3 Theoretical Properties

3.3.1 Asymptotics of the HP estimator

We derive the theoretical properties of $\widehat{f}_{HP}(y|x)$ under Conditions OO and SS which were stated in Chapter 2. Proofs of our theorems are in Appendix B. The conditions for the generalized weight function $w_j(x)$ in $\widehat{f}_{HP}(y|x)$ are generalized to Condition H, which are also listed in this appendix. The quantity $b_2(x)$, which appears in the theorems, is also included in Condition H.

Denote $\partial^l q(x,t)/\partial^l x$ as $q^{(l)}(x,t)$. The technical conditions about the decay rate of $q^{(l)}(x,t)$ are generalized to conditions TO and TS when $q(x,t)$ is ordinary smooth and super smooth, respectively. We include these in Appendix B.

Theorem 3.1. *Assume that $f_X(x) > 0$, $c > 1$, and $\rho_0(x) > 1$. Suppose Conditions OO, H, TO, (C1)–(C3) hold and $\bar{h} \to 0$, $h \to 0$, and $J\bar{h}h^{2c\rho_0(x)+2\beta_0} \to \infty$. For $\epsilon > 0$, if $J^{1-\epsilon}\bar{h} \to \infty$ and $\bar{h}\log(h) \to 0$, then*

$$\widehat{f}_{HP}(y|x) - f(y|x) = B_{\bar{h},h}(x,y) + V_{\bar{h},h}^{1/2}(x,y),$$

where

$$\begin{aligned} B_{\bar{h},h}(x,y) &= \bar{h}^2 b_2(x)\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\left\{\frac{1}{2}q''(x,t) + \frac{c-1}{2}\frac{q'(x,t)^2}{q(x,t)}\right\}\phi_K(ht)dt \\ &\quad + o_p(\bar{h}^2) + \frac{1}{2\pi}\int \exp(-ity)q(x,t)\{\phi_K(ht)-1\}dt \\ V_{\bar{h},h}(x,y) &= O_p\{c^{-2+2\beta_0}J^{-1}\bar{h}^{-1}h^{-2(c-1)\rho_0(x)-2\beta_0-1}\}. \end{aligned}$$

Theorem 3.2. *Assume that $f_X(x) > 0$ and $c > 1$. Suppose Conditions SS, H, TS, (C1)–(C3) hold and $\bar{h} \to 0$, $h \to 0$, and*

$$J\bar{h}h^{d_1}\exp\{-2c\varrho(x)^{-1}h^{-\rho_2(x)} - 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\} \to \infty,$$

for any d_1. For $\epsilon > 0$, if $J^{1-\epsilon}\bar{h} \to \infty$ and $\bar{h}h^{d_2} \to 0$ for any d_2, then

$$\widehat{f}_{HP}(y|x) - f(y|x) = B_{\bar{h},h}(x,y) + V_{\bar{h},h}^{1/2}(x,y),$$

where

$$\begin{aligned} B_{\bar{h},h}(x,y) &= \bar{h}^2 b_2(x)\frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\left\{\frac{1}{2}q''(x,t) + \frac{c-1}{2}\frac{q'(x,t)^2}{q(x,t)}\right\}\phi_K(ht)dt \\ &\quad + o_p(\bar{h}^2) + \frac{1}{2\pi}\int \exp(-ity)q(x,t)\{\phi_K(ht)-1\}dt \\ V_{\bar{h},h}(x,y) &= O_p[c^{-2-2\beta_1}J^{-1}\bar{h}^{-1}h^{d_3}\exp\{2(c-1)\varrho(x)^{-1}h^{-\rho_2(x)} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}], \end{aligned}$$

for some constant d_3.

The conclusions of Theorems 3.1 and 3.2 are similar to that of Theorems 2.1 and 2.2. We leave a comparison to next section.

3.3.2 Comparison of the asymptotic properties of the RP and HP estimators

If we exclude the impact of measurement error, the asymptotics of the HP estimator of $f(y|x)$ depends on $\rho_0(x)$ in the ordinary smooth case and on $\rho_2(x)$ in super smooth case. That is, the performance of $\widehat{f}_{HP}(y|x)$ is closely related to the tail behavior of $q(x,t) = \phi_{Y|X=x}(t)$ as $t \to \infty$. In this subsection, we discuss the scenario when the decay rate of $\phi_{Y|X=x}(t)$ does not depend on x, that is, $\rho_0(x) = \alpha_0$ and $\rho_2(x) = \alpha_2$. This means a change in x does not affect the shape of $f(y|x)$. For example, in the normal distribution, x is only related to the mean. Or, in the gamma distribution, x is only related to the rate parameter but not the shape parameter.

In the ordinary smooth case when $\alpha = \rho_0(x)$, both estimators have the same bias rate:

$$O_p(\bar{h}^2) + \frac{1}{2\pi}\int_{-\infty}^{\infty} \exp(-ity)\phi_{Y|X=x}(t)\{\phi_K(ht) - 1\}dt.$$

The variance of random pooling is $O_p\{c^{-1+2\beta_0}J^{-1}\bar{h}^{-1}h^{-2(c-1)\alpha_0-2\beta_0-1}\}$ and that of the homogeneous pooling is $O_p\{c^{-2+2\beta_0}J^{-1}\bar{h}^{-1}h^{-2(c-1)\alpha_0-2\beta_0-1}\}$. We can see the convergence rate is the same in regards to J but the HP estimator's variance is c times smaller than that of the RP estimator. The same conclusion holds for the super smooth case. This conclusion is consistent with Delaigle and Hall (2012), who compare RP and HP when the response variable Y is binary. When the decay rate of $\phi_{Y|X=x}(t)$ is related to x, the situation becomes much more complicated. We use simulation to explore this further.

3.4 Numerical Study

We illustrate a data-driven bandwidth selection method for the HP estimator. Then, a comparison between the performance of the HP and RP estimators is made using three different models.

3.4.1 Bandwidth selection method

Similar to the RP estimator, there are 2 bandwidths: $\bar{h}$ and h which correspond to the kernels $\bar{K}$ and K, respectively. We use $\bar{K}$ as the normal kernel and K as the infinite order sinc kernel. The bandwidth selection procedure is similar to that of the RP estimator. The first step is to estimate $\bar{h}$. Based on the data $(\bar{Z}_{(j)}, \bar{X}_{(j)})$, the bandwidth selection method of the nonparametric kernel type estimator of $f_{\bar{Z}|\bar{X}=x}$, denoted by $\widehat{f}_{\bar{Z}|\bar{X}=x}$, has been discussed in Bashtannyk and Hyndman (2001) and Hyndman and Yao (2002). We use the bandwidth of the $X_{(j)}$'s in $\widehat{f}_{\bar{Z}|\bar{X}=x}$ as our $\bar{h}$. Once $\bar{h}$ is determined, we apply the ERS method described in Section 2.5.2 on the characteristic function

$$\widehat{\phi}_{Y|X=x}(t) = \left\{ \frac{\widehat{\phi}_{\bar{Z}|\bar{X}=x}(ct)}{\phi_{\epsilon}(ct)} \right\}^{1/c}$$

in Equation (3.2).

3.4.2 Comparisons

We examine the performance of the RP and the HP estimators in Model 1 (see Section 2.5.1) and two additional models:

Model 2: $Y_{ij} \sim Gamma(9, 2X_{ij} + 3)$

Model 3: $Y_{ij} \sim Lognormal\{1 + 0.25X_{ij}, (0.5 + 0.25X_{ij})^2\}$,

where $X_{ij} \sim Unif(-0.5, 0.5)$, $i = 1, \ldots, c$ and $j = 1, \ldots, J$. Recall Y_{ij} can not be observed directly; $\bar{Z}_j = \bar{Y}_j + \epsilon_j$ is observed instead, where $\epsilon_j \sim Normal(0, 0.05^2)$. Model 1 specifies a Gaussian distribution for Y_{ij} and its mean is a linear function of x. Model 2 specifies a gamma distribution where its rate parameter is also a linear function of x. The decay rate of $\phi_{Y|X=x}(t)$ in both models does not depend on x. Different from the first two models, the decay rate of $\phi_{Y|X=x}(t)$ in Model 3 depends on x. We focus on the local linear estimator for RP and HP, denoted by $\widehat{f}_{RL}(y|x)$ and $\widehat{f}_{HL}(y|x)$, respectively. We simulate 500 data sets for each simulation setting. In

Table 3.1: Mean MISE of 500 simulations for $\widehat{f}_{RL}(y|x)$ and $\widehat{f}_{HL}(y|x)$ ($\times 10^{-2}$). For each method, we use group sizes $c = 2, 3, 4, 5$ and sample sizes $N = 1800, 3600$.

		Normal		Gamma		Lognormal	
c	Method	1800	3600	1800	3600	1800	3600
2	RL	2.067	1.635	1.032	0.810	1.020	0.723
	HL	1.749	1.322	0.992	0.754	1.002	0.774
3	RL	3.947	3.193	2.282	1.853	2.435	1.746
	HL	2.890	2.424	1.641	1.323	1.829	1.549
4	RL	5.749	4.814	3.612	2.949	3.988	2.983
	HL	3.777	3.106	2.105	1.802	2.451	2.149
5	RL	7.727	6.517	5.070	4.237	5.606	4.490
	HL	4.668	3.908	2.620	2.279	3.150	2.751

each simulation, we determine $\widehat{f}_{RL}(y|x)$ and $\widehat{f}_{HL}(y|x)$ using the bandwidth selection method described in Chapters 2 and 3 and calculate the MISE for group sizes $c = 2$ to 5.

Table 1 shows the mean MISE from $B = 500$ simulations. We can observe that as the group size c increases, the MISE increases for each model. This is within expectation because as c increases, the number of observations decreases and the convergence rate is reduced. It is also not surprising to see that $\widehat{f}_{HL}(y|x)$ generally performs better than $\widehat{f}_{RL}(y|x)$ and the gap between the MISE of $\widehat{f}_{HL}(y|x)$ and $\widehat{f}_{RL}(y|x)$ grows larger as c increases. This is in line with the theoretical result in Section 3.3.2 which shows the HP estimator's variance is c times smaller than that of the RP estimator. Lastly, consistent with the deconvolution theory, we can observe that Model 2 (which represents an ordinary smooth distribution) has smaller MISE than Model 1 (which represents a super smooth distribution).

Based on the 500 simulated data sets, Figures 3.1 and 3.2 display percentiles of $\widehat{f}_{RL}(y|x)$ and $\widehat{f}_{HL}(y|x)$ when $N = 3600$ and $c = 2$. Both RP and HP estimators can recover the shape of the distribution accurately whenever x is in the center (0) or near the boundary (± 0.3) in all three model settings. Furthermore, the performance indicated in the bottom two rows in each figure suggests our method can provide

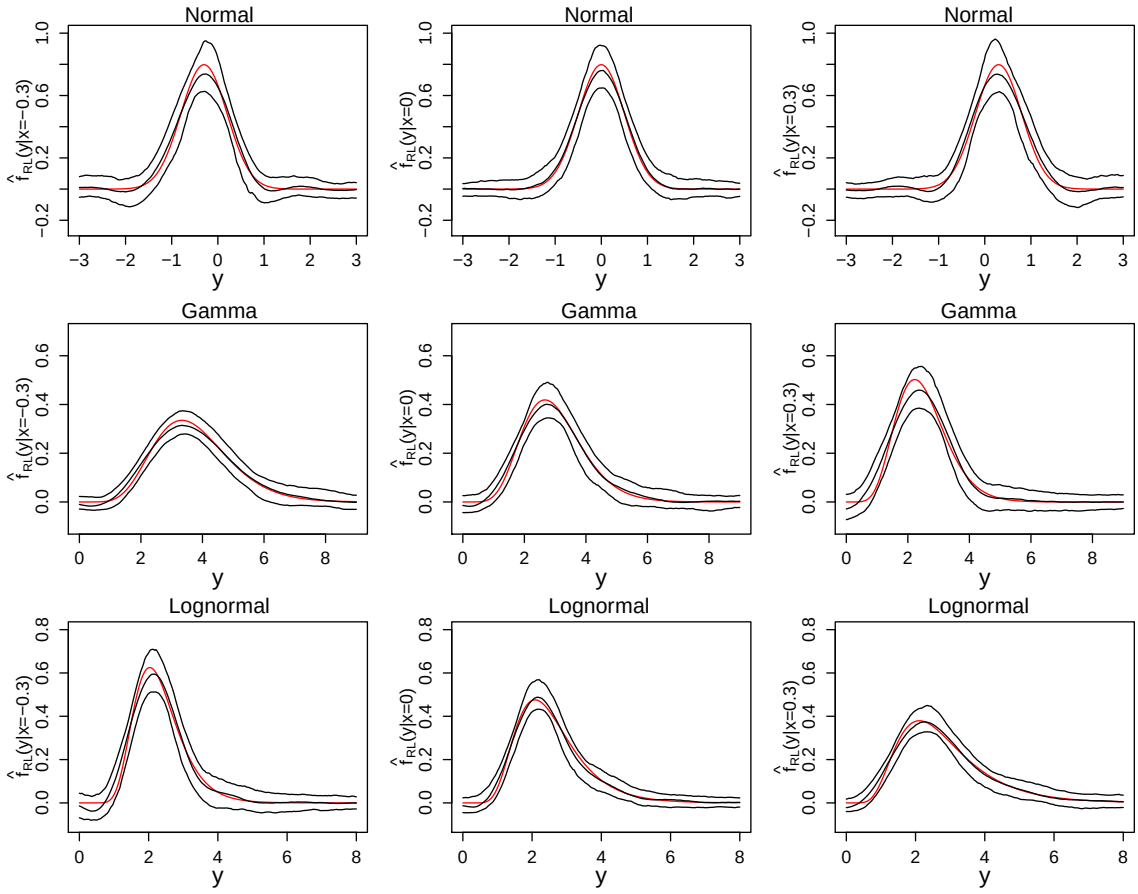

Figure 3.1: Random pooling. Plots of $\widehat{f}_{RL}(y|x)$ when $J = 3600$ and $c = 2$. Top: Model 1 (normal distribution); Middle: Model 2 (gamma distribution); Bottom: Model 3 (lognormal distribution). Left: $x = -0.3$; Middle: $x = 0$; Right: $x = 0.3$. For each y, we calculate the 2.5th, 50th, and 97.5th percentiles based on 500 data sets. The red curve is the true density.

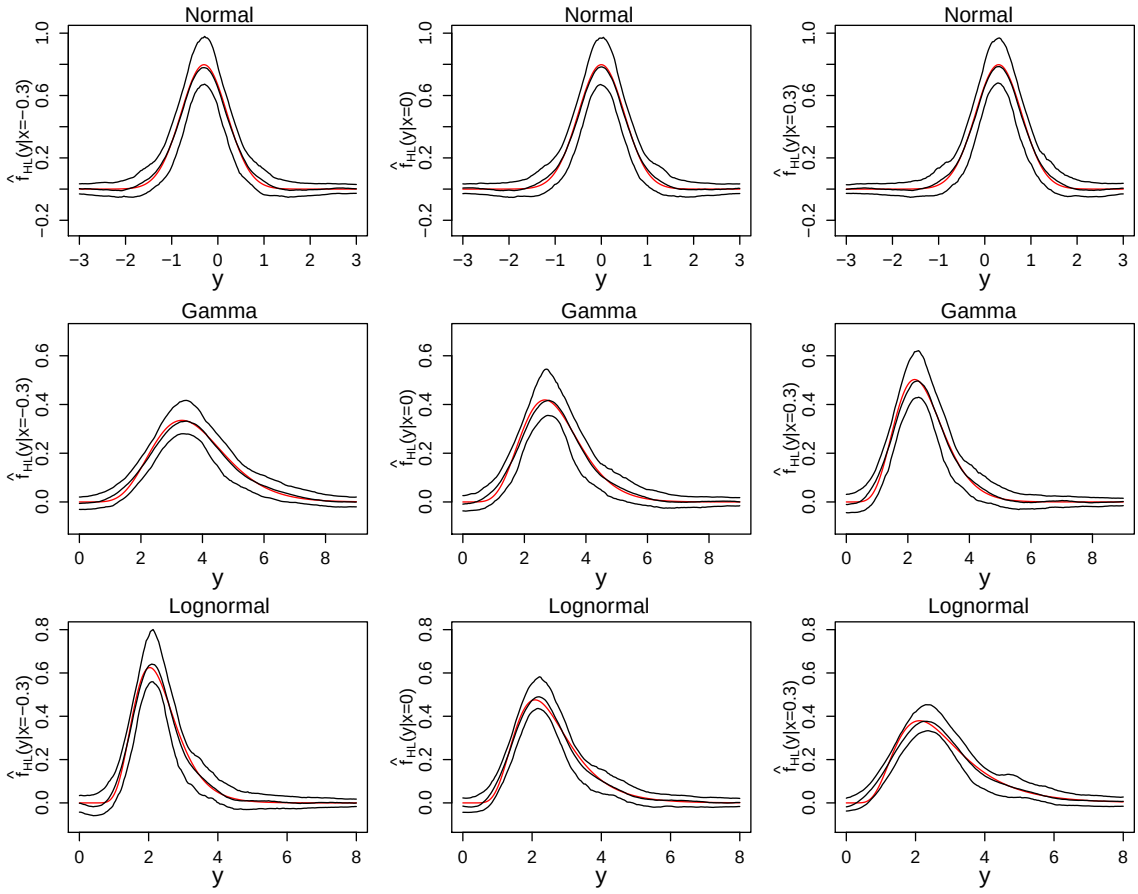

Figure 3.2: Homogeneous pooling. Plots of $\widehat{f}_{HL}(y|x)$ when $J = 3600$ and $c = 2$. Top: Model 1 (normal distribution); Middle: Model 2 (gamma distribution); Bottom: Model 3 (lognormal distribution). Left: $x = -0.3$; Middle: $x = 0$; Right: $x = 0.3$. For each y, we calculate the 2.5th, 50th, and 97.5th percentiles based on 500 data sets. The red curve is the true density.

accurate estimation when the underlying distribution is skewed.

3.5 Real Data Analysis

We illustrate our estimation methods using a data set collected between 2011 and 2012 which is available at the National Health and Nutrition Examination Survey (NHANES) website. Polyfluorochemicals (PFCs) are a family of artificial chemicals that are widely used in nonstick cookware and fabrics. Some PFCs, such as perfluorooctane sulfonate (PFOS) are difficult to break down naturally and have been found in humans. Due to the potential risk of bio-accumulation, it is meaningful to study the distribution of PFCs in the human population. Kärrman et al. (2006) studied the relationship between PFCs and age, gender, and geographic region using pooled serum samples from Australian residents.

We use the NHANES data set to explore the relationship between PFOS concentration level and age. This data set consists of individual measurements of PFOS from serum samples, and observations from $N = 1812$ individuals in the United States are used in the analysis. We utilize the age (year) as X and the concentration level of PFOS (ng/ml) as Y. To emulate pooling, we artifically create pools of size c to obtain $\bar{Y}_j$. Measurement error is further added to $\bar{Y}_j$, that is, $\bar{Z}_j = \bar{Y}_j + \epsilon_j$ where $\epsilon_j \sim Normal(0, 0.5^2)$. We implement this procedure for $B = 500$ times and average the 500 curves of $\widehat{f}_{RL}(y|x)$ and $\widehat{f}_{HL}(y|x)$ for each x. We do not know the true underlying distribution $f(y|x)$ for these data; therefore, we use the local linear estimator calculated from the original individual ($c = 1$) data, denoted by $f_{REF}(y|x)$, as a reference.

Figure 3.3 displays the estimates $\widehat{f}_{RL}(y|x)$ and $\widehat{f}_{HL}(y|x)$ for ages x between 16 and 75 years. The red curves serve as the reference distribution $f_{REF}(y|x)$, calculated from the individual data. One can observe the accumulation of PFOS is low at a young age; however, the PFOS level is higher on average and its distribution is

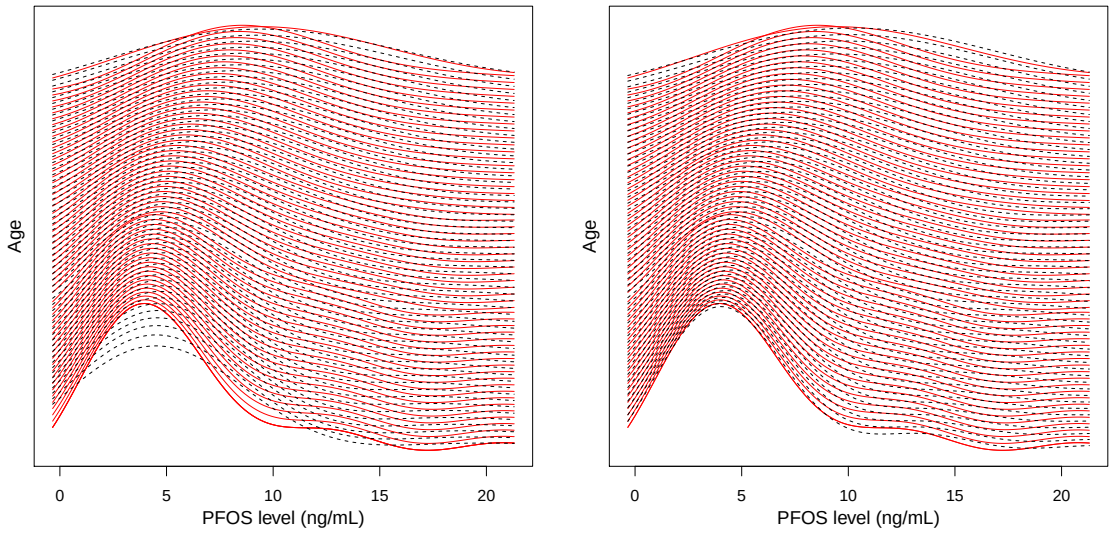

Figure 3.3: Data analysis. Left: $f_{REF}(y|x)$ (solid red) and $\widehat{f}_{RL}(y|x)$ (dotted black). Right: $f_{REF}(y|x)$ (solid red) and $\widehat{f}_{HL}(y|x)$ (dotted black). The y axis shows ages from 19 to 75 years. Each curve represents an estimated density function at a specific age x.

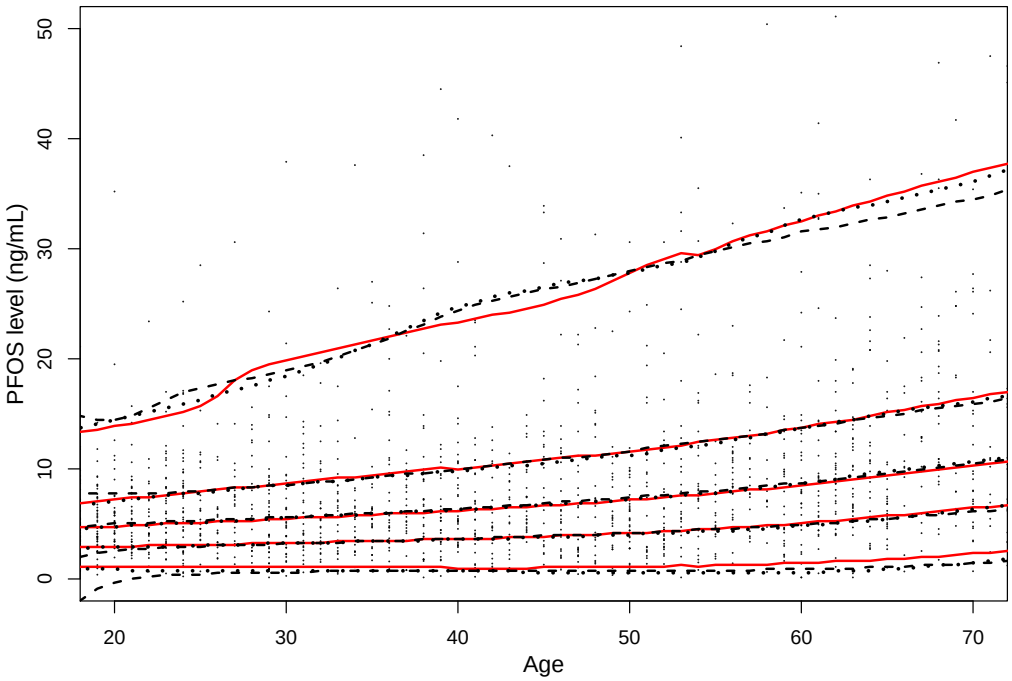

Figure 3.4: Quantile plot of X (age) and Y (PFOS level). The five lines shown are the 5th, 25th, 50th, 75th, and 95th quantiles, respectively. Dashed black: $f_{RL}(y|x)$; Dotted black: $f_{HL}(y|x)$; Solid red: $f_{REF}(y|x)$. Obervations larger than 50 ng/ml are omitted.

more spread out for larger ages. Comparing the estimators with $f_{REF}(y|x)$, we see $\widehat{f}_{RL}(y|x)$ fails to capture the shape of $f(y|x)$ at a young age, but approximates the reference distribution adequately. On the other hand, the HP estimate $\widehat{f}_{HL}(y|x)$ does an adequate job for nearly all ages. Treating $f_{REF}(y|x)$ as the true distribution, the MISE of $\widehat{f}_{RL}(y|x)$ and $\widehat{f}_{HL}(y|x)$ are 8.90×10^{-5} and 1.44×10^{-5}, respectively, which indicates the improved performance of the HP estimate.

Figure 3.4 shows the quantile plot of PFOS level against age. One notes the estimate $\widehat{f}_{RL}(y|x)$ is much more spread out at a young age (<25), which is also observed in Figure 3.3. Both $\widehat{f}_{RL}(y|x)$ and $\widehat{f}_{HL}(y|x)$ fit the 25th, 50th, and 75th quantiles well but underestimate the 5th quantile. This is expected because the data are sparse in the lower tail.

Chapter 4

Prevalence estimation for two-stage hierarchical group testing

4.1 Introduction

Group testing, or pooled testing, is a strategy that pools individual specimens (e.g., blood, urine, etc.) and tests the pools for the presence of a disease. The idea of group testing traces back to 1943 when Dorfman proposed pooling blood samples to detect syphilis among American soldiers in World War II. Since his seminal work, group testing has been used to screen for various diseases, including HIV (Pilcher et al., 2005), influenza virus (Van et al., 2011), and chlamydia and gonorrhea (Lewis, Lockary, and Kobic, 2012).

The group testing literature generally splits into two topics: case identification and estimation. The case identification literature describes algorithms to classify every individual as positive or negative; see Kim et al. (2007) for a review. One widely used protocol is two-stage hierarchical group testing, or Dorfman testing (DT). In the first stage, individual specimens are pooled together and tested. If the pool diagnosis is positive, a second stage of retesting is used to identify the status of each specimen inside the pool. When the disease prevalence is low, DT can save substantial costs when compared to testing specimens one by one. This strategy has been utilized by the State Hygienic Laboratory (SHL) at the University of Iowa and has saved millions of dollars since 1999 (Jirsa, 2008).

In the estimation problem, one aims to estimate the population-level prevalence

(Hepworth and Watson, 2009; Liu et al., 2012) or to build regression models which link the true individual disease status with covariates (Vansteelandt, Goetghebeur, and Verstraeten, 2000; McMahan et al., 2017). For the former goal, group testing can improve estimation precision in the presence of misclassification errors. The literature describing this phenomenon mainly focuses on the protocol that tests master pools only and does not involve further retests. We refer to this pooling protocol as master pool testing (MPT). Tu, Litvak, and Pagano (1995) and Liu et al. (2012) showed that MPT can be used to estimate the disease prevalence more precisely when compared to testing specimens individually. While their work assumes the assay sensitivity and specificity are known, Huang et al. (2017) proposed MPT designs that consist of three different pool sizes to estimate the disease prevalence (p), the sensitivity (S_e), and the specificity (S_p) simultaneously. Their designs are proved to be the best in terms of D-optimality and D_s-optimality criteria among all MPT configurations.

A natural question is whether DT, which adds a retesting stage after MPT, can be utilized in the estimation problem considered by Huang et al. (2017). In this chapter, we propose using maximum likelihood to estimate p, S_e, and S_p using DT. We also construct confidence sets for these parameters and four widely used operating characteristics in the classification problem. There are two advantages to the methods we propose. First, it is common to assume the true p, S_e, and S_p are known when evaluating the case identification operating characteristics of DT (e.g., expected number of tests, expected correct classification rate, etc.) However, these values are usually unknown in practice. Therefore, our work enables researchers to estimate p, S_e, S_p and the operating characteristics directly from the observed data. Second, for a large range of values of p, we show through theoretical calculation and simulation that the estimation precision using DT is higher than that using the optimal MPT design in Huang et al. (2017).

The rest of this chapter is organized as follows. In Section 4.2, we summarize

the group testing data collected from the DT procedure and describe our estimation methods for p, S_e, S_p and the case identification operating characteristics. In Section 4.3, we assess our estimation methods using simulation and then apply them to a chlamydia data set collected at the SHL. In Section 4.4, we provide the optimal DT design for a fixed budget and compare it with the optimal MPT design provided by Huang et al. (2017). In Section 4.5, we conclude with a summary and discuss future work. Derivations and additional results are in Appendix C.

4.2 Data and Methodology

4.2.1 Dorfman testing data

The two-stage hierarchical procedure proposed by Dorfman (1943) is described below.

Stage 1: Randomly assign individuals to non-overlapping master pools. Test the master pools.

Stage 2: Based on the diagnosis of the master pool,

- if the diagnosis is negative, then every individual in the master pool is diagnosed as negative;
- if the diagnosis is positive, then every individual within the pool is retested. Diagnoses are based on the results of the individual tests.

When p is low, most master pools in Stage 1 test negatively, which leads to a reduction in testing costs. When a master pool tests positive, the retests in Stage 2 further classify all the involving individuals. We refer to the collection of all the testing results from both stages as DT data. Our goal is to use DT data to estimate $\boldsymbol{\theta} = (p, S_e, S_p)^T$.

To present DT data more formally, we denote by N the total number of individuals. The N individuals are randomly assigned to n non-overlapping groups, each

of size $x = N/n$, where n divides N and $x > 1$. Let y denote the number of pools that have been diagnosed as positive in Stage 1. For $j = 1, \ldots, y$, we further denote by z_j the number of individuals in the jth positive pool that have been retested positive in Stage 2. The DT data can be formally presented by $(n, x, y, \boldsymbol{z})$, where $\boldsymbol{z} = \{z_1, \ldots, z_y\}$.

4.2.2 Estimation

We pursue maximum likelihood estimation of $\boldsymbol{\theta}$ and propose three ways to construct confidence regions for $\boldsymbol{\theta}$. We also write Wald confidence intervals for four widely used operating characteristics in DT.

Our estimation proceeds under two assumptions. First, the test results are independent conditional on the true statuses of the individuals. Second, there is no dilution effect, i.e., S_e and S_p are the same for different pool sizes. These assumptions are standard in literature; see, e.g., Tebbs, McMahan, and Bilder (2013). Under these two assumptions, the log-likelihood function of $\boldsymbol{\theta}$ can be derived as

$$\begin{aligned} l_{\mathcal{D}}(\boldsymbol{\theta}|n, x, y, \boldsymbol{z}) &= (n-y)\log\{1-\pi(x|\boldsymbol{\theta})\} \\ &+ \sum_{j=1}^{y} \log[\{\pi(x|\boldsymbol{\theta}) - S_e\}(1-S_p)^{z_j} S_p^{x-z_j} + \pi(1|\boldsymbol{\theta})^{z_j}\{1-\pi(1|\boldsymbol{\theta})\}^{x-z_j} S_e], \end{aligned} \tag{4.1}$$

where $\pi(x|\boldsymbol{\theta}) = (1-p)^x(1-S_p) + \{1-(1-p)^x\}S_e$ is the probability of a pool of size x being diagnosed as positive. See Appendix C for a detailed derivation.

Denote by $\hat{\boldsymbol{\theta}}_{\mathcal{D}}$ the maximum likelihood estimator (MLE) of $\boldsymbol{\theta}$; that is,

$$\hat{\boldsymbol{\theta}}_{\mathcal{D}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ l_{\mathcal{D}}(\boldsymbol{\theta}|n, x, y, \boldsymbol{z}).$$

Obtaining a closed-form expression for the MLE $\hat{\theta}_{\mathcal{D}}$ is not possible, but it is straightforward to maximize $\hat{\boldsymbol{\theta}}_{\mathcal{D}}$ numerically. We accomplish this in R by using the `maxLik` package.

We can construct confidence regions for $\boldsymbol{\theta}$ by inverting the Wald, score, and likelihood ratio tests. First, denote the score function by

$$U_{\mathcal{D}}(\boldsymbol{\theta}|n,x,y,\boldsymbol{z}) = \frac{\partial l_{\mathcal{D}}(\boldsymbol{\theta}|n,x,y,\boldsymbol{z})}{\partial \boldsymbol{\theta}}$$

and the Fisher information matrix

$$I_{\mathcal{D}}(\boldsymbol{\theta}|n,x) = E_{y,\boldsymbol{z}}\left\{-\frac{\partial^2 l_{\mathcal{D}}(\boldsymbol{\theta}|n,x,y,\boldsymbol{z})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\right\}. \tag{4.2}$$

Note that $U_{\mathcal{D}}(\boldsymbol{\theta}|n,x,y,\boldsymbol{z})$ and $I_{\mathcal{D}}(\boldsymbol{\theta}|n,x)$ involve the derivatives of the log-likelihood function in Equation (4.1). Because these derivatives are complex, we use R to derive $U_{\mathcal{D}}(\boldsymbol{\theta}|n,x,y,\boldsymbol{z})$ and $I_{\mathcal{D}}(\boldsymbol{\theta}|n,x)$ symbolically. Denote by $\chi^2_{1-\alpha,3}$ the upper α quantile of the chi-square distribution with three degrees of freedom. A large-sample $100(1-\alpha)\%$ Wald confidence region is given by

$$\{\boldsymbol{\theta} : (\hat{\boldsymbol{\theta}}_{\mathcal{D}} - \boldsymbol{\theta})^T I_{\mathcal{D}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}}|n,x)(\hat{\boldsymbol{\theta}}_{\mathcal{D}} - \boldsymbol{\theta}) < \chi^2_{1-\alpha,3}\}.$$

In addition, large-sample $100(1-\alpha)\%$ score and likelihood ratio confidence regions are given by

$$\{\boldsymbol{\theta} : U_{\mathcal{D}}(\boldsymbol{\theta}|n,x,y,\boldsymbol{z})^T I_{\mathcal{D}}(\boldsymbol{\theta}|n,x)\, U_{\mathcal{D}}(\boldsymbol{\theta}|n,x,y,\boldsymbol{z}) < \chi^2_{1-\alpha,3}\}$$

and

$$\{\boldsymbol{\theta} : 2l_{\mathcal{D}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}}|n,x,y,\boldsymbol{z}) - 2l_{\mathcal{D}}(\boldsymbol{\theta}|n,x,y,\boldsymbol{z}) < \chi^2_{1-\alpha,3}\},$$

respectively. These regions are also obtained numerically.

Along with estimating $\boldsymbol{\theta}$, we can estimate operating characteristics commonly seen in the group testing literature to measure the efficiency and accuracy of DT; see, e.g., Kim et al. (2007) and Malinovsky, Albert, and Roy (2016). We focus on four characteristics: the expected number of tests per individual $E(T)$, the expected correct classification rate $E(C)$, the pooling positive predictive value PPV and the

pooling negative predictive value NPV. These expressions are given by

$$
\begin{aligned}
E(T) &= S_e - (1-p)^x(S_e + S_p - 1) + \frac{1}{x} \\
E(C) &= (1-p)^x\{(1-S_p)(S_p + S_e - 1)\} + (1-p)(1 - S_e + S_pS_e - S_e^2) + S_e^2 \\
PPV &= \frac{pS_e(DT)}{(1-p)\{1 - S_p(DT)\} + pS_e(DT)} \\
NPV &= \frac{(1-p)S_p(DT)}{p\{1 - S_e(DT)\} + (1-p)S_p(DT)},
\end{aligned}
$$

where

$$
\begin{aligned}
S_e(DT) &= S_e^2 \\
S_p(DT) &= 1 - (1 - S_p)[(1 - S_p)(1-p)^{x-1} + S_e\{1 - (1-p)^{x-1}\}].
\end{aligned}
$$

The quantity $E(T)$ measures cost efficiency and $E(C)$, PPV, and NPV are measures of classification accuracy.

Note these four operating characteristics can be thought of as functions of $\boldsymbol{\theta}$, say $h(\boldsymbol{\theta})$, where $h : \mathbb{R}^3 \to \mathbb{R}^1$. Based on the large-sample properties of MLEs, when n is large, we have

$$\hat{\boldsymbol{\theta}}_{\mathcal{D}} \sim AN(\boldsymbol{\theta}, I_{\mathcal{D}}(\boldsymbol{\theta}|n,x)^{-1})$$

and

$$h(\hat{\boldsymbol{\theta}}_{\mathcal{D}}) \sim AN\left(h(\boldsymbol{\theta}), h_{\boldsymbol{\theta}}(\boldsymbol{\theta})^T I_{\mathcal{D}}(\boldsymbol{\theta}|n,x)^{-1} h_{\boldsymbol{\theta}}(\boldsymbol{\theta})\right),$$

where $h_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \partial h(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, by the Delta Method. For any continuously differentiable function h, a large-sample $100(1-\alpha)\%$ Wald confidence interval of $h(\boldsymbol{\theta})$ is given by

$$[h(\hat{\boldsymbol{\theta}}_{\mathcal{D}}) - z_{\alpha/2} sd\{h(\hat{\boldsymbol{\theta}}_{\mathcal{D}})\}, h(\hat{\boldsymbol{\theta}}_{\mathcal{D}}) + z_{\alpha/2} sd\{h(\hat{\boldsymbol{\theta}}_{\mathcal{D}})\}],$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution, and

$$sd\{h(\hat{\boldsymbol{\theta}}_{\mathcal{D}})\} = \sqrt{h_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}})^T I_{\mathcal{D}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}}|n,x)^{-1} h_{\boldsymbol{\theta}}(\hat{\boldsymbol{\theta}}_{\mathcal{D}})}.$$

Similar to $U_{\mathcal{D}}(\boldsymbol{\theta}|n,x,y,\boldsymbol{z})$ and $I_{\mathcal{D}}(\boldsymbol{\theta}|n,x)$, we derive $h_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ symbolically.

Table 4.1: Point estimation using DT with a fixed number of individuals $N = 5000$. BIAS denotes the average bias over $B = 1000$ Monte Carlo data sets. SD denotes the sample standard deviation of the 1000 estimates, and SE denotes the averaged standard error. The Mean-squared error (MSE) is also shown. All values are multiplied by 10^3.

	$x = 5$, $n = 1000$			$x = 10$, $n = 500$		
	p	S_e	S_p	p	S_e	S_p
BIAS	1.30	-6.57	1.01	0.71	-2.48	0.01
SD	6.90	43.36	10.74	5.66	39.47	8.24
SE	7.36	52.27	11.32	5.93	42.75	7.99
MSE	0.05	1.92	0.12	0.03	1.56	0.07

Table 4.2: Number of tests and coverage probabilities of $\boldsymbol{\theta}$ using DT with a fixed number of individuals $N = 5000$. MNT and SDNT denote the mean and standard deviation of the number of tests using DT over $B = 1000$ Monte Carlo data sets. W-CP, S-CP, and LR-CP denote the coverage probability of joint 95% Wald, score, and likelihood ratio confidence regions of $\boldsymbol{\theta}$.

	MNT	SDNT	W-CP	S-CP	LR-CP
$x = 5$, $n = 1000$	2272	71	0.948	0.937	0.949
$x = 10$, $n = 500$	2560	108	0.958	0.944	0.958

4.3 Simulation Evidence and Real Data Analysis

4.3.1 A simulation study

We use simulation to assess the finite-sample performance of our estimators. We set $N = 5000$, $p = 0.05$, and consider $S_e, S_p \in \{0.90, 0.95, 0.99\}$ and $x \in \{5, 10\}$. For each (p, S_e, S_p, N, x), we first generate the true disease statuses of the N individuals independently from a Bernoulli distribution with probability of success being p. Setting the pool size to be x, we then implement the two-stage hierarchical algorithm described in Section 4.2.1 to simulate a set of DT data. Finally, we apply our estimation methods to the DT data set. We repeat the process of data generation and estimation 1000 times. Results for $S_e = S_p = 0.95$ are summarized in Tables 4.1–4.3. Results for other considered S_e's and S_p's are included in Appendix C.

Tables 4.1 and 4.2 summarize the performance of the MLE of $\boldsymbol{\theta}$. We first see that, for both $x = 5$ and $x = 10$, the MLEs exhibit small bias and variance, indicating

Table 4.3: Operating characteristics for DT with a fixed number of individuals $N = 5000$. True value (TRUE), coverage probability (CP), and average length (LEN) of 95% Wald confidence intervals for $E(T)$, $E(C)$, PPV, NPV are shown.

	$x = 5$, $n = 1000$			$x = 10$, $n = 500$		
	TRUE	CP	LEN	TRUE	CP	LEN
$E(T)$	0.454	0.949	0.054	0.511	0.953	0.086
$E(C)$	0.985	0.979	0.019	0.977	0.959	0.020
PPV	0.814	0.960	0.159	0.713	0.946	0.150
NPV	0.995	0.970	0.024	0.995	0.997	0.019

that our method works well in estimating p, S_e, and S_p. It is interesting to see that the estimation of S_p outperforms the one of S_e. A possible reason is when p is low as 0.05, truly negative pools are much more than truly positive ones, which provides more information to the estimation of S_p. We also notice that the averaged standard errors are in agreement with the sample standard deviations of the estimates. This agreement is further reinforced by Table 4.2 where the coverage probability of joint 95% Wald, score, and likelihood ratio confidence regions of $\boldsymbol{\theta}$ are all at the nominal level.

When estimating the four operating characteristics for DT, Table 4.3 shows that the empirical coverage probabilities using our confidence interval estimates are all close to the nominal level except the ones for $E(C)$ and NPV. This is because that the true values of these two characteristics are very close to 1, the upper bound of the corresponding parameter space. In these cases, we might need a much larger N to reveal the asymptotic normality established by the Delta Method.

From the perspective of testing cost, Table 4.2 shows that using DT with $x = 5$ or $x = 10$ yields about 50% cost reduction when comparing to individual testing. It is important to note that using data collected from individual testing cannot identify p, S_e, and S_p simultaneously. When comparing $x = 5$ to $x = 10$, the average number of tests increased from 2272 to 2560. This is because that pools of size 10 have a larger chance to be truly positive than the ones of size 5. The larger number of tests

also provides more information to the estimation of $\boldsymbol{\theta}$. As a result, the MSEs of p, S_e, and S_p when $x = 10$ are all smaller than the ones when $x = 5$ (see Table 4.1). However, this pattern does not imply that large pool sizes are always preferable to small ones. In Section 4.4, we derive the optimal design of DT given a fixed number of tests, and then compare it with the optimal MPT design proposed by Huang et al. (2017).

4.3.2 Real data analysis

We assess our methods using a chlamydia data set collected at the SHL. Every year the SHL utilizes group testing to screen Iowa residents for chlamydia and gonorrhea, and have saved approximately $3.1 million dollars from 2009 to 2014. Collaborating with the SHL, McMahan et al. (2017) provided a group testing data set that was collected from screening female swab specimens using DT with pool size 4. There are 2273 pools in total.

We apply our estimation method to the DT data set and summarize the estimates in Table 4.4 and Figure 4.1, where values that exceed 1 are not truncated in order to show the whole range of the confidence intervals or region. We make several observations here. First, the estimates of p and S_p are consistent with the previous studies using the similar data sets (McMahan, Tebbs, and Bilder, 2012a, McMahan et al., 2017). Our estimator of S_e is exactly 1. McMahan et al. (2017) presents a smaller estimate of S_e but they utilize regression models with covariate information. Second, the joint confidence region of $\boldsymbol{\theta}$ is a skewed ellipsoid whose volume equals 1.91×10^{-5}. This is significantly smaller than the volume of the cuboid (3.94×10^{-5}) formed by the marginal confidence intervals of p, S_e, and S_p separately (shown in Table 4.4). This is not surprising since the joint confidence region accounts for the correlation between the parameter estimates. Finally, our estimates of the operating characteristics indicate that DT saves more than 40% of tests when compared to

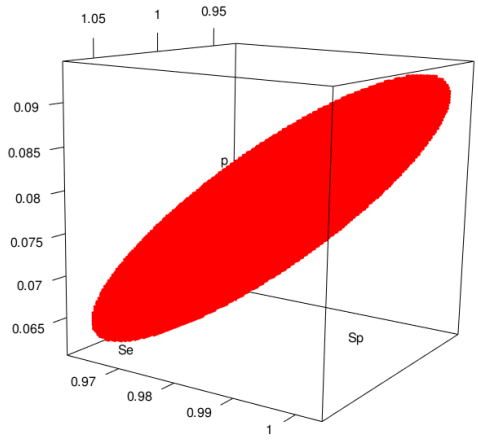

Figure 4.1: 95% Wald confidence region (red) of $\boldsymbol{\theta}$ in Iowa chlamydia data set.

Table 4.4: Parameter estimates in Iowa chlamydia data set. The results include parameter estimate, estimated standard error and 95% confidence interval of p, S_e, S_p, and operating characteristics $E(T)$, $E(C)$, PPV, NPV.

	Estimate	SE	95% CI
p	0.077	0.005	(0.068,0.087)
S_e	1.000	0.021	(0.958,1.042)
S_p	0.982	0.006	(0.970,0.994)
$E(T)$	0.539	0.009	(0.520,0.557)
$E(C)$	0.996	0.002	(0.992,1.001)
PPV	0.953	0.018	(0.918,0.989)
NPV	1.000	0.004	(0.992,1.008)

individual testing, and remains a high accuracy.

4.4 The Optimal DT Design

To find the optimal DT design for estimating $\boldsymbol{\theta}$, we focus on a widely used criterion, D-optimality, which seeks to minimize the determinant of the estimate's covariance matrix, or equivalently, maximizes the determinant of the Fisher information matrix for the estimate of $\boldsymbol{\theta}$. In our context, we seek for the (n_D, x_D) that maximizes $\log|I_{\mathcal{D}}(\boldsymbol{\theta}|n,x)|$ with respect to (n,x), where $I_{\mathcal{D}}(\boldsymbol{\theta}|n,x)$ is presented in (4.2). We assume $\boldsymbol{\theta}$ as known (or can be estimated in advance).

Recall from Section 4.2.2 that we have assumed S_e and S_p do not depend on the

pool size. This is often not true when the pool size becomes unrealistically large. The same concern has appeared in Huang et al. (2017), where, as a solution, the maximum allowable pool size has been set to be 61. However, a pool of size 61 is rarely used in practice; e.g., Van et al. (2011) used the real-time PCR methodology to detect Influenza virus in pools of size 10; Lewis, Lockary, and Kobic (2012) screened chlamydia and gonorrhea using pool size of 4 in the Infertility Prevention Project; American red cross applies a general group testing framework of pool size 16 to screen the blood donations for infectious diseases (Hepatitis B, Zika virus, etc.) Therefore, we choose the maximum allowable pool size as $x_U = 20$.

In addition, we set the maximum affordable number of tests to be a predetermined value m, because most screening practices only have a limited budget. Our search of (n_D, x_D) is conducted subject to this cost constraint as well. However, in DT it is impossible to determine the number of tests exactly before conducting the screening, because whether to test individuals or not is decided by the uncertain master pool diagnosis. Therefore, we consider the distribution of the number of tests instead. If n pools are tested using DT, when n is large, the number of tests approximately follows a normal distribution with mean $n\{1 + x\pi(x|\boldsymbol{\theta})\}$ and standard deviation $x\sqrt{n\pi(x|\boldsymbol{\theta})(1 - \pi(x|\boldsymbol{\theta}))}$. To make our search reasonable, for a given pool size x, we choose n to be the largest integer satisfying

$$n\{1 + x\pi(x|\boldsymbol{\theta})\} + 3x\sqrt{n\pi(x|\boldsymbol{\theta})\{1 - \pi(x|\boldsymbol{\theta})\}} \leq m, \tag{4.3}$$

i.e., the mean number of tests plus three standard deviations is still smaller than m. Even in the worst case, DT will almost certainly cost less tests than m.

Under the restrictions, $x \leq x_U$ and (4.3), we use the following steps to determine the optimal design for DT:

- Step 1: For each x from 2 to x_U, calculate the largest integer n that satisfies Equation (4.3).

- Step 2: Go through each pair of (n, x) identified by Step 1. Find the optimal setting

$$(n_D, x_D) = \underset{(n,x)}{\text{argmax}} |I_{\mathcal{D}}(\boldsymbol{\theta}|n, x)|.$$

The Fisher information under the optimal setting is $I_{\mathcal{D}}(\boldsymbol{\theta}|n_D, x_D)$.

4.4.1 Theoretical comparison with MPT

In MPT protocol, one must utilize data collected from pools of at least three different sizes to estimate $\boldsymbol{\theta}$ to avoid the identifiability issue. When the number of tests is fixed to be m, Huang et al. (2017) proved that the D-optimal MPT design is to test $m/3$ pools of size x_L, $m/3$ of size x_M, and $m/3$ of size x_U, where x_L is a predetermined minimum allowable pool size, x_M is a function of x_L and x_U; see Theorem 2 of Huang et al. (2017). Denote $\boldsymbol{n}_M = (n_L, n_M, n_U)^T = (m/3, m/3, m/3)^T$ and $\boldsymbol{x}_M = (x_L, x_M, x_U)^T$. The optimal MPT design is then $(\boldsymbol{n}_M, \boldsymbol{x}_M)$. It follows that the corresponding Fisher information is

$$I_{\mathcal{M}}(\boldsymbol{\theta}|\boldsymbol{n}_M, \boldsymbol{x}_M) = \sum_{i\in\{L,m,U\}} \frac{n_i}{\pi(x_i|\boldsymbol{\theta})\{1-\pi(x_i|\boldsymbol{\theta})\}} \frac{\partial \pi(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \frac{\partial \pi(x_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}^T,$$

where

$$\frac{\partial \pi(x|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left\{x(1-p)^{x-1}(S_e + S_p - 1), 1-(1-p)^x, -(1-p)^x\right\}^T.$$

Define

$$f(\boldsymbol{\theta}, m, x_L, x_U) = \log|I_{\mathcal{D}}(\boldsymbol{\theta}|n_D, x_D)| - \log|I_{\mathcal{M}}(\boldsymbol{\theta}|\boldsymbol{n}_M, \boldsymbol{x}_M)|.$$

We take logarithm to make the value of $f(\boldsymbol{\theta}, m, x_L, x_U)$ in an appropriate scale. If $f(\boldsymbol{\theta}, m, x_L, x_U) > 0$, DT provides a larger determinant of the Fisher information matrix (therefore, a more efficient estimator of $\boldsymbol{\theta}$) when compared to MPT and vice versa. Figure 4.2 shows the value $f(\boldsymbol{\theta}, m, x_L, x_U)$ under the setting: $0 < p < 0.2$, $S_e, S_p \in \{0.90, 0.95, 0.99\}$, $x_L = 1$, $x_U = 20$, and $m = 3000$. The selection of p, S_e,

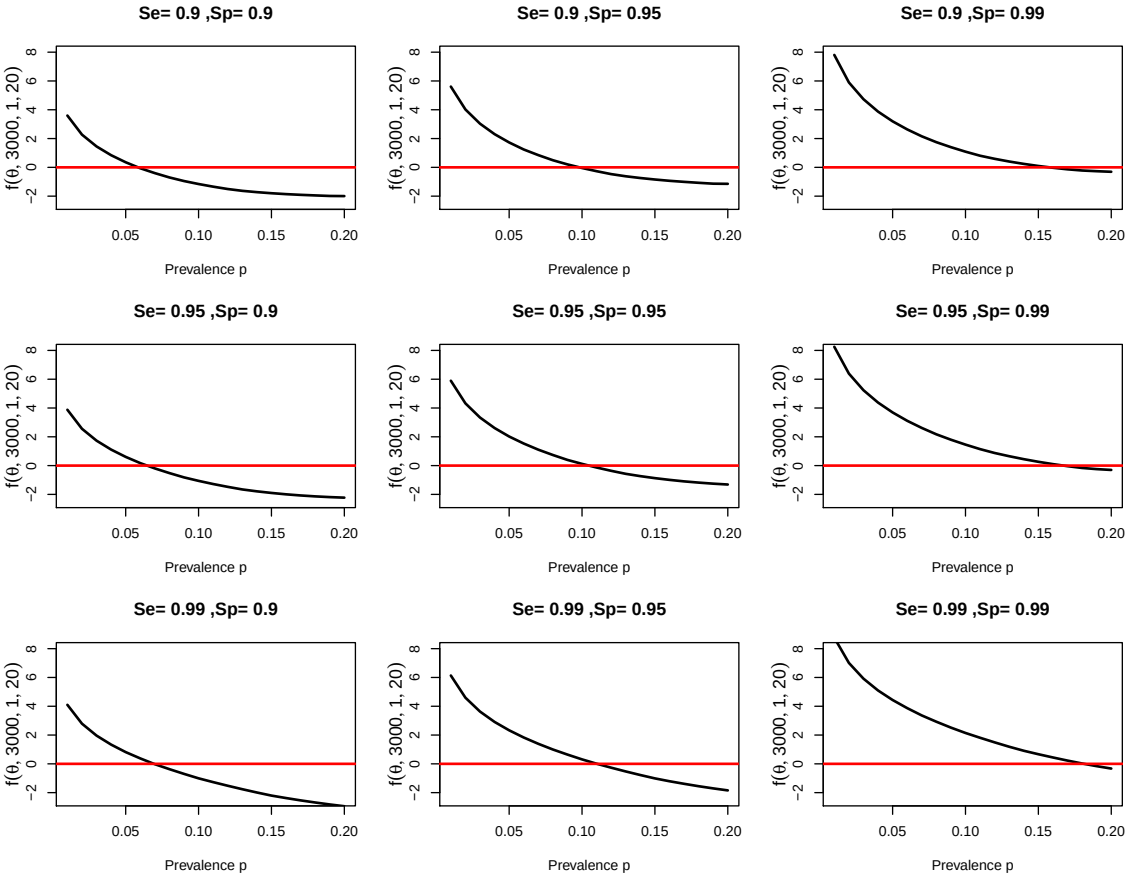

Figure 4.2: $f(\boldsymbol{\theta}, 3000, 1, 20)$ of different p, S_e, and S_p. The horizontal red line indicates the position of 0.

and S_p is comparable to the previous studies (e.g., McMahan, Tebbs, and Bilder, 2012a).

We make two observations on Figure 4.2. First, given S_e and S_p, f appears to be a decreasing, convex function of p. Second, the optimal DT design provides a better (theoretical) estimation efficiency than the one of MPT in a large range of values of the disease prevalence: from $p = 0$ to $p = 0.05$ in low sensitivity and specificity (e.g., $S_e = S_p = 0.9$), and up to $p = 0.2$ in high sensitivity and specificity (e.g., $S_e = S_p = 0.99$).

4.4.2 Numerical comparison with MPT

We also use simulation to verify the theoretical comparison. We focus on the setting of the center plot in Figure 4.2: $S_e = S_p = 0.95$. Other parameters are the same as in Figure 4.2: $x_L = 1$, $x_U = 20$, and $m = 3000$. Three representative values of the disease prevalence are selected: $p = 0.05$, 0.10, 0.15, corresponding to the scenarios that $f(\boldsymbol{\theta}, 3000, 1, 20)$ is greater than, approximately equal to, and less than 0, respectively. We generate group testing data using the optimal MPT and DT designs and implement the estimation procedure 1000 times.

The simulation results fit the center plot in Figure 4.2 very well. In Table 4.5, MSEs of estimates based on DT are significantly better, a bit better, generally worse than the ones based on MPT when $p = 0.05$, 0.10, 0.15, respectively. In Table 4.6, the number of tests is fixed to be 3000 for MPT, while mean number of tests for DT stays approximately three standard deviations below 3000, indicating the cost of DT almost never exceeds the budget. The performance of the joint 95% Wald confidence region is also within expectation; e.g., when $p = 0.05$, the DT coverage probability is close to 95% while the MPT coverage probability is significantly pinned down while the average volume of DT is only approximately 1/3 of the one of MPT. Lastly, Huang et al. (2017) also explored the D_s-optimal design, which only focuses on the

Table 4.5: Point estimation comparison between MPT and DT with a fixed number of tests $m = 3000$ and test accuracies $S_e = S_p = 0.95$. BIAS denotes the average bias over $B = 1000$ Monte Carlo data sets. SD denotes the sample standard deviation of the 1000 estimates, and SE denotes the averaged standard error. The Mean-squared error (MSE) is also shown. All values are multiplied by 10^3.

		MPT			DT		
		p	S_e	S_p	p	S_e	S_p
$p = 0.05$	BIAS	3.38	-20.45	2.05	0.73	-2.87	0.02
	SD	9.37	78.98	11.49	5.48	37.37	8.49
	SE	13.09	135.40	12.33	5.71	40.07	8.71
	MSE	0.10	6.66	0.14	0.03	1.40	0.07
$p = 0.10$	BIAS	0.39	2.36	0.60	0.44	0.43	-0.32
	SD	10.27	32.35	15.27	9.32	29.51	11.11
	SE	10.92	36.95	15.86	9.64	31.07	11.32
	MSE	0.11	1.05	0.23	0.09	0.87	0.12
$p = 0.15$	BIAS	-0.54	2.03	-0.37	1.17	-0.78	0.95
	SD	10.67	14.97	19.55	13.57	27.88	14.13
	SE	10.87	15.07	19.38	13.73	28.27	14.51
	MSE	0.11	0.23	0.38	0.19	0.78	0.20

Table 4.6: Test characteristics comparison between MPT and DT with a fixed number of tests $m = 3000$ and test accuracies $S_e = S_p = 0.95$. MNT and SDNT denote the mean and standard deviation of the number of tests over $B = 1000$ Monte Carlo data sets. CP and VOL denote the coverage probability and the average volume of joint 95% Wald confidence region of $\boldsymbol{\theta}$ over the 1000 simulations.

	MPT			DT		
	$p = 0.05$	$p = 0.10$	$p = 0.15$	$p = 0.05$	$p = 0.10$	$p = 0.15$
MNT	3000	3000	3000	2721	2779	2815
SDNT	0	0	0	94	74	64
CP	0.869	0.94	0.94	0.963	0.945	0.946
VOL($\times 10^5$)	25.591	14.537	11.939	9.431	19.132	15.983

estimation efficiency of p instead of $\boldsymbol{\theta}$. We follow a similar idea as in this section to compare D_s-optimal designs between DT and MPT. The result is in Section C.3 in Appendix C and the conclusion is similar.

4.5 Conclusion

We propose MLEs of p, S_e, and S_p in DT protocol. We also provide the confidence intervals for operating characteristics such as $E(T)$, $E(C)$, PPV, and NPV. These values can assist lab technicians to measure the efficiency of DT when lacking knowledge of the disease prevalence and the assay sensitivity and specificity. DT strongly competes with MPT for two reasons: First, DT can not only estimates p, S_e, and S_p but can identify each individual's disease status as well. Even when the prevalence estimation is the primary goal, it is usually beneficial to know the disease status of a patient. Second, from the perspective of estimation efficiency, our result shows DT outperforms MPT in a large range of values of p, e.g., for $S_e = S_p = 0.95$, the optimal design of DT is more efficient than the one of MPT when p is below 0.1.

The focus of this study is DT, which uses the two-stage hierarchical structure with one pool size. In case identification, McMahan, Tebbs, and Bilder (2012b) found that one can reduce the number of tests by using adaptive pool sizes based on different levels of risk for each individual. Black, Bilder, and Tebbs (2015) investigated the optimal configuration for multiple stages to minimize the cost. One future work is to explore whether these generalizations of DT can help to improve the estimation efficiency. Other extensions include relaxing the assumptions in Section 4.2.1 by taking dilution effects and conditional dependence into consideration.

Bibliography

Ai, Jing, Rong Zhang, Yue Li, Jielin Pu, Yanjie Lu, Jundong Jiao, Kang Li, Bo Yu, Zhuqin Li, Rongrong Wang, et al. (2010). "Circulating microRNA-1 as a potential novel biomarker for acute myocardial infarction". In: *Biochemical and Biophysical Research Communications* 391.1, pp. 73–77.

Bashtannyk, David and Rob Hyndman (2001). "Bandwidth selection for kernel conditional density estimation". In: *Computational Statistics and Data Analysis* 36.3, pp. 279–298.

Bates, Michael, Simon Buckland, Nick Garrett, Samuel Caudill, and Howard Ellis (2005). "Methodological aspects of a national population-based study of persistent organochlorine compounds in serum". In: *Chemosphere* 58.7, pp. 943–951.

Bilder, Christopher, Joshua Tebbs, and Peng Chen (2010). "Informative retesting". In: *Journal of the American Statistical Association* 105.491, pp. 942–955.

Black, Michael, Christopher Bilder, and Joshua Tebbs (2012). "Group testing in heterogeneous populations by using halving algorithms". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 61.2, pp. 277–290.

Black, Michael, Christopher Bilder, and Joshua Tebbs (2015). "Optimal retesting configurations for hierarchical group testing". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 64.4, pp. 693–710.

Carroll, Raymond and Peter Hall (1988). "Optimal rates of convergence for deconvolving a density". In: *Journal of the American Statistical Association* 83.404, pp. 1184–1186.

Caudill, Samuel (2012). "Use of pooled samples from the national health and nutrition examination survey". In: *Statistics in Medicine* 31.27, pp. 3269–3277.

Chen, Peng, Joshua Tebbs, and Christopher Bilder (2009). "Group testing regression models with fixed and random effects". In: *Biometrics* 65.4, pp. 1270–1278.

Chung, Kai-Lai (2001). *A Course in Probability Theory*. Academic press.

Delaigle, Aurore and Peter Hall (2012). "Nonparametric regression with homogeneous group testing data". In: *Annals of Statistics* 40.1, pp. 131–158.

Delaigle, Aurore and Peter Hall (2015). "Nonparametric methods for group testing data, taking dilution into account". In: *Biometrika* 102.4, pp. 871–887.

Delaigle, Aurore, Peter Hall, and Justin Wishart (2014). "New approaches to nonparametric and semiparametric regression for univariate and multivariate group testing data". In: *Biometrika* 101.3, pp. 567–585.

Delaigle, Aurore and Alexander Meister (2012). "Nonparametric regression analysis for group testing data". In: *Journal of the American Statistical Association* 106.494, pp. 640–650.

Delaigle, Aurore and Wen-Xin Zhou (2015). "Nonparametric and parametric estimators of prevalence from group testing data with aggregated covariates". In: *Journal of the American Statistical Association* 110.512, pp. 1785–1796.

Dhand, Navneet, Wesley Johnson, and Jenny-Ann Toribio (2010). "A Bayesian approach to estimate OJD prevalence from pooled fecal samples of variable pool size". In: *Journal of Agricultural, Biological, and Environmental Statistics* 15.4, pp. 452–473.

Diggle, Peter and Peter Hall (1993). "A Fourier approach to nonparametric deconvolution of a density estimate". In: *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 523–531.

Fahey, Jed, Philippe Ourisson, and Frederick Degnan (2006). "Pathogen detection, testing, and control in fresh broccoli sprouts". In: *Nutrition Journal* 5.1, p. 13.

Fan, Jianqing (1991a). "Asymptotic normality for deconvolution kernel density estimators". In: *Sankhyā: The Indian Journal of Statistics, Series A*, pp. 97–110.

Fan, Jianqing (1991b). "On the optimal rates of convergence for nonparametric deconvolution problems". In: *Annals of Statistics*, pp. 1257–1272.

Fan, Jianqing (2018). *Local Polynomial Modelling and Its Applications: Monographs on Statistics and Applied Probability 66*. Routledge.

Fan, Jianqing and Tsz Ho Yim (2004). "A crossvalidation method for estimating conditional densities". In: *Biometrika* 91.4, pp. 819–834.

Faraggi, David, Benjamin Reiser, and Enrique Schisterman (2003). "ROC curve analysis for biomarkers based on pooled assessments". In: *Statistics in Medicine* 22.15, pp. 2515–2527.

Farrington, Paddy (1992). "Estimating prevalence by group testing using generalized linear models". In: *Statistics in Medicine* 11.12, pp. 1591–1597.

Finkelstein, Mark, Howard Tucker, and Jerry Veeh (1999). "Extinguishing the distinguished logarithm problems". In: *Proceedings of the American Mathematical Society*, pp. 2773–2777.

Gastwirth, Joseph (2000). "The efficiency of pooling in the detection of rare mutations". In: *American Journal of Human Genetics* 67.4, p. 1036.

Hepworth, Graham and Ray Watson (2009). "Debiased estimation of proportions in group testing". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 58.1, pp. 105–121.

Hou, Peijie, Joshua Tebbs, Christopher Bilder, and Christopher McMahan (2017). "Hierarchical group testing for multiple infections". In: *Biometrics* 73.2, pp. 656–665.

Huang, Shih-Hao, Mong-Na Lo Huang, Kerby Shedden, and Weng Kee Wong (2017). "Optimal group testing designs for estimating prevalence with uncertain testing errors". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.5, pp. 1547–1563.

Hughes-Oliver, Jacqueline and William Swallow (1994). "A two-stage adaptive group-testing procedure for estimating small proportions". In: *Journal of the American Statistical Association* 89.427, pp. 982–993.

Hyndman, Rob and Qiwei Yao (2002). "Nonparametric estimation and symmetry tests for conditional density functions". In: *Journal of Nonparametric Statistics* 14.3, pp. 259–278.

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

Kärrman, Anna, Jochen Mueller, Bert Van Bavel, Fiona Harden, Leisa-Maree Toms, and Gunilla Lindström (2006). "Levels of 12 perfluorinated chemicals in pooled Australian serum, collected 2002- 2003, in relation to age, gender, and region". In: *Environmental Science and Technology* 40.12, pp. 3742–3748.

Kim, Hae-Young, Michael Hudgens, Jonathan Dreyfuss, Daniel Westreich, and Christopher Pilcher (2007). "Comparison of group testing algorithms for case identification in the presence of test error". In: *Biometrics* 63.4, pp. 1152–1163.

Lewis, Joanna, Vivian Lockary, and Sadika Kobic (2012). "Cost savings and increased efficiency using a stratified specimen pooling strategy for Chlamydia trachomatis and Neisseria gonorrhoeae". In: *Sexually Transmitted Diseases* 39.1, pp. 46–48.

Lin, Juexin and Dewei Wang (2018). "Single-index regression for pooled biomarker data". In: *Journal of Nonparametric Statistics*, pp. 1–21.

Linton, Oliver and Yoon-Jae Whang (2002). "Nonparametric estimation with aggregated data". In: *Econometric Theory* 18.2, pp. 420–468.

Liu, Aiyi and Enrique Schisterman (2003). "Comparison of diagnostic accuracy of biomarkers with pooled assessments". In: *Biometrical Journal: Journal of Mathematical Methods in Biosciences* 45.5, pp. 631–644.

Liu, Aiyi, Chunling Liu, Zhiwei Zhang, and Paul Albert (2012). "Optimality of group testing in the presence of misclassification". In: *Biometrika* 99.1, pp. 245–251.

Liu, Yan, Christopher McMahan, and Colin Gallagher (2017). "A general framework for the regression analysis of pooled biomarker assessments". In: *Statistics in Medicine* 36.15, pp. 2363–2377.

Ma, Chang-Xing, Albert Vexler, Enrique Schisterman, and Lili Tian (2011). "Cost-efficient designs based on linearly associated biomarkers". In: *Journal of Applied Statistics* 38.12, pp. 2739–2750.

Malinovsky, Yaakov, Paul Albert, and Anindya Roy (2016). "Reader reaction: A note on the evaluation of group testing algorithms in the presence of misclassification". In: *Biometrics* 72.1, pp. 299–302.

Malinovsky, Yaakov, Paul Albert, and Enrique Schisterman (2012). "Pooling designs for outcomes under a Gaussian random effects model". In: *Biometrics* 68.1, pp. 45–52.

McMahan, Christopher, Joshua Tebbs, and Christopher Bilder (2012a). "Informative Dorfman screening". In: *Biometrics* 68.1, pp. 287–296.

McMahan, Christopher, Joshua Tebbs, and Christopher Bilder (2012b). "Two-dimensional informative array testing". In: *Biometrics* 68.3, pp. 793–804.

McMahan, Christopher, Alexander McLain, Colin Gallagher, and Enrique Schisterman (2016). "Estimating covariate-adjusted measures of diagnostic accuracy based on pooled biomarker assessments". In: *Biometrical Journal* 58.4, pp. 944–961.

McMahan, Christopher, Joshua Tebbs, Timothy Hanson, and Christopher Bilder (2017). "Bayesian regression for group testing data". In: *Biometrics* 73.4, pp. 1443–1452.

Meister, Alexander (2007). "Optimal convergence rates for density estimation from grouped data". In: *Statistics and probability letters* 77.11, pp. 1091–1097.

Mitchell, Emily, Robert Lyles, Amita Manatunga, Michelle Danaher, Neil Perkins, and Enrique Schisterman (2014). "Regression for skewed biomarker outcomes subject to pooling". In: *Biometrics* 70.1, pp. 202–211.

Mitchell, Emily, Robert Lyles, Amita Manatunga, and Enrique Schisterman (2015). "Semiparametric regression models for a right-skewed outcome subject to pooling". In: *American Journal of Epidemiology* 181.7, pp. 541–548.

Mumford, Sunni, Enrique Schisterman, Albert Vexler, and Aiyi Liu (2006). "Pooling biospecimens and limits of detection: effects on ROC curve analysis". In: *Biostatistics* 7.4, pp. 585–598.

Parikh, Chirag, Jaya Mishra, Heather Thiessen-Philbrook, Belda Dursun, Qing Ma, Caitlin Kelly, Catherine Dent, Prasad Devarajan, and Charles Edelstein (2006). "Urinary IL-18 is an early predictive biomarker of acute kidney injury after cardiac surgery". In: *Kidney International* 70.1, pp. 199–203.

Pilcher, Christopher, Susan Fiscus, Trang Nguyen, Evelyn Foust, Leslie Wolf, Del Williams, Rhonda Ashby, Judy Owen O'dowd, J Todd McPherson, Brandt Stalzer, et al. (2005). "Detection of acute infections during HIV testing in North Carolina". In: *New England Journal of Medicine* 352.18, pp. 1873–1883.

Saha-Chaudhuri, Paramita and Clarice Weinberg (2013). "Specimen pooling for efficient use of biospecimens in studies of time to a common event". In: *American Journal of Epidemiology* 178.1, pp. 126–135.

Schisterman, Enrique and Albert Vexler (2008). "To pool or not to pool, from whether to when: applications of pooling to biospecimens subject to a limit of detection". In: *Paediatric and Perinatal Epidemiology* 22.5, pp. 486–496.

Schisterman, Enrique, Neil Perkins, Aiyi Liu, and Howard Bondell (2005). "Optimal cut-point and its corresponding Youden Index to discriminate individuals using pooled blood samples". In: *Epidemiology*, pp. 73–81.

Sobel, Milton and Robert Elashoff (1975). "Group testing with a new goal, estimation". In: *Biometrika* 62.1, pp. 181–193.

Tebbs, Joshua, Christopher McMahan, and Christopher Bilder (2013). "Two-stage hierarchical group testing for multiple infections with application to the infertility prevention project". In: *Biometrics* 69.4, pp. 1064–1073.

Thompson, Keith (1962). "Estimation of the proportion of vectors in a natural population of insects". In: *Biometrics* 18.4, pp. 568–578.

Tu, Xin-Ming, Eugene Litvak, and Marcello Pagano (1995). "On the informativeness and accuracy of pooled testing in estimating prevalence of a rare disease: application to HIV screening". In: *Biometrika* 82.2, pp. 287–297.

Van, Tam, Joseph Miller, David Warshauer, Erik Reisdorf, Daniel Jernigan, Rosemary Humes, and Peter Shult (2011). "Pooling nasopharyngeal/throat swab specimens to increase testing capacity for influenza viruses by PCR". In: *Journal of Clinical Microbiology*, JCM–05631.

Vansteelandt, Stijn, Els Goetghebeur, and Thomas Verstraeten (2000). "Regression models for disease prevalence with diagnostic tests on pools of serum samples". In: *Biometrics* 56.4, pp. 1126–1133.

Vexler, Albert, Aiyi Liu, and Enrique Schisterman (2010). "Nonparametric deconvolution of density estimation based on observed sums". In: *Journal of Nonparametric Statistics* 22.1, pp. 23–39.

Vexler, Albert, Enrique Schisterman, and Aiyi Liu (2008). "Estimation of ROC curves based on stably distributed biomarkers subject to measurement error and pooling mixtures". In: *Statistics in Medicine* 27.2, pp. 280–296.

Wang, Dewei, Christopher McMahan, Colin Gallagher, and Karunarathna Kulasekera (2013). "Semiparametric group testing regression models". In: *Biometrika* 101.3, pp. 587–598.

Warasi, Md, Joshua Tebbs, Christopher McMahan, and Christopher Bilder (2016). "Estimating the prevalence of multiple diseases from two-stage hierarchical pooling". In: *Statistics in Medicine* 35.21, pp. 3851–3864.

Warasi, Md, Christopher McMahan, Joshua Tebbs, and Christopher Bilder (2017). "Group testing regression models with dilution submodels". In: *Statistics in Medicine* 36.30, pp. 4860–4872.

Weinberg, Clarice and David Umbach (1999). "Using pooled exposure assessment to improve efficiency in case-control studies". In: *Biometrics* 55.3, pp. 718–726.

Appendix A

Proof for Chapter 2

A.1 Proof of Theorem 2.1

In this chapter, we present the proofs of Theorems 2.1 and 2.2. Lemmas are shown in the end of each proof. From Equations (2.7) and (2.10), our local polynomial estimator $\widehat{f}_{RP}(y|x)$ can be written as

$$\widehat{f}_{RP}(y|x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\frac{\widehat{\phi}_{\bar{Z}|X=x}(ct)}{\{\widehat{\phi}_{\bar{Z}}(ct)\}^{(c-1)/c}\{\phi_{\epsilon}(ct)\}^{1/c}}\phi_K(ht)dt, \tag{A.1}$$

where

$$\widehat{\phi}_{\bar{Z}|X=x}(t) = \frac{\sum_{j=1}^{J}\sum_{i=1}^{c}\exp(it\bar{Z}_j)w_{ij}(x)}{\sum_{j=1}^{J}\sum_{i=1}^{c}w_{ij}(x)},$$

in which the form of $w_{ij}(x)$ depends on the value of ℓ; i.e., the order of the (local) polynomial. When $\ell = 0$, $w_{ij}(x) = \bar{K}_{\bar{h}}(X_{ij} - x)$, and Equation (A.1) provides the local constant (Nadaraya-Watson) estimator of $f(y|x)$. When $\ell = 1$,

$$\begin{aligned} w_{ij}(x) = \bar{K}_{\bar{h}}(X_{ij} - x)\sum_j\sum_i \bar{K}_{\bar{h}}(X_{ij} - x)(X_{ij} - x)^2 \\ - \bar{K}_{\bar{h}}(X_{ij} - x)(X_{ij} - x)\sum_j\sum_i \bar{K}_{\bar{h}}(X_{ij} - x)(X_{ij} - x), \end{aligned}$$

and Equation (A.1) defines the local linear estimator. The proofs presented below focus on the local linear case. The Proofs for the local constant estimator follow a similar pattern and are hence omitted.

Before presenting the proofs, we need to introduce some notation. Recalling from Condition (C2), the support of ϕ_K is $[-1, 1]$. Denote by

$$\mathcal{B}_J = \{t \in \mathbb{R} : |t| \leq h^{-1}\}.$$

Then $\phi_K(ht)$ is not zero only when $t \in \mathcal{B}_J$. Let $\phi_{\bar{Y}}(t)$ be the CF of $\bar{Y}_j$. Noting that $\bar{Z}_j = \bar{Y}_j + \epsilon_j$, and ϵ_j and $\bar{Y}_j$ are independent, we define our empirical estimator of $\phi_{\bar{Y}}(t)$ to be $\widehat{\phi}_{\bar{Y}}(t) = \widehat{\phi}_{\bar{Z}}(t)/\phi_\epsilon(t)$. We write $\phi_{Y|X=x}(t)$ as $\phi_{Y|x}(t)$. In addition, letting $\phi_{\bar{Y}|x}(t) = E\{\exp(it\bar{Y}_j)|X_{ij} = x\} = \phi_{\bar{Z}|x}(t)/\phi_\epsilon(t)$, similarly, we have our estimator of $\phi_{\bar{Y}|x}(t)$ by $\widehat{\phi}_{\bar{Y}|x}(t) = \widehat{\phi}_{\bar{Z}|x}(t)/\phi_\epsilon(t)$, where we write $\phi_{\bar{Z}|X_{ij}=x}(t) = \phi_{\bar{Z}|x}(t)$ and $\widehat{\phi}_{\bar{Z}|X_{ij}=x}(t) = \widehat{\phi}_{\bar{Z}|x}(t)$ for the simplicity of notation. We also write $\phi_{\bar{Y}|X_{ij}}(t) = E\{\exp(it\bar{Y}_j)|X_{ij}\}$ and $\phi_{\bar{Z}|X_{ij}}(t) = E\{\exp(it\bar{Z}_j)|X_{ij}\}$. Lastly, the notation $f(n) \asymp g(n)$ means there exist $M_1, M_2 > 0$ such that $M_1 g(n) \le f(n) \le M_2 g(n)$.

Proof of Theorem 2.1. To obtain the asymptotic property of $\widehat{f}_{RP}(y|x) - f(y|x)$, we decompose it by

$$\widehat{f}_{RP}(y|x) - f(y|x) = \frac{1}{2\pi}\int e^{-ity}\left\{\widehat{\phi}_{Y|x}(t) - \phi_{Y|x}(t)\right\}\phi_K(ht)dt \tag{A.2}$$

$$+ \frac{1}{2\pi}\int e^{-ity}\phi_{Y|x}(t)\left\{\phi_K(ht) - 1\right\}dt, \tag{A.3}$$

where

$$\widehat{\phi}_{Y|x}(t) = \frac{\widehat{\phi}_{\bar{Z}|x}(ct)}{\widehat{\phi}_{\bar{Y}}(ct)^{(c-1)/c}\phi_\epsilon(ct)}. \tag{A.4}$$

The term (A.3) is exactly the last summand in $B_{\bar{h},h}(x,y)$ in Theorem 2.1. For term (A.2), because $\phi_K(ht)$ is not zero only when $t \in \mathcal{B}_J$, it suffices to look into the difference between $\widehat{\phi}_{Y|x}(t)$ and $\phi_{Y|x}(t)$ when $t \in \mathcal{B}_J$.

We start with the term $\widehat{\phi}_{\bar{Y}}(ct)^{-(c-1)/c}$ in $\widehat{\phi}_{Y|x}(t)$ for $t \in \mathcal{B}_J$. Noting that $\widehat{\phi}_{\bar{Y}}(ct)$ is complex, when $t \in \mathcal{B}_J$, we can expand $\widehat{\phi}_{\bar{Y}}(ct)^{-(c-1)/c}$ into power series as

$$\frac{1}{\widehat{\phi}_{\bar{Y}}(ct)^{(c-1)/c}} = \frac{1}{\phi_{\bar{Y}}(ct)^{(c-1)/c}}\left\{1 - \frac{c-1}{c}\frac{\Delta(ct)}{\phi_{\bar{Y}}(ct)} + \tilde{\lambda}_{2,r_0}(t)\frac{\Delta(ct)^2}{\phi_{\bar{Y}}(ct)^2}\right\}, \tag{A.5}$$

where $|\tilde{\lambda}_{2,r_0}(t)| \le 4$ and $\Delta(ct) = \widehat{\phi}_{\bar{Y}}(ct) - \phi_{\bar{Y}}(ct)$. To make this expansion valid, one needs to show that (i) $\widehat{\phi}_{\bar{Y}}(ct)/\phi_{\bar{Y}}(ct)$ is larger than a positive constant when $t \in \mathcal{B}_J$, and (ii) the cth root of $\widehat{\phi}_{\bar{Y}}(ct)$ exists. To prove (i), we show when $t \in \mathcal{B}_J$ and

$Jh^{2c\alpha_0+2\beta_0} \to \infty$, $|\Delta(ct)/\phi_{\bar{Y}}(ct)| < 1/2$ uniformly on t with probability 1. Note that

$$\begin{aligned} E|\Delta(ct)|^2 &= \frac{1}{J^2|\phi_\epsilon(ct)|^2} E\left[\sum_j \left\{e^{ict\bar{Z}_j} - \phi_{\bar{Z}}(ct)\right\} \sum_j \left\{e^{-ict\bar{Z}_j} - \phi_{\bar{Z}}(-ct)\right\}\right] \\ &= \frac{1-|\phi_{\bar{Z}}(ct)|^2}{J|\phi_\epsilon(ct)|^2}. \end{aligned}$$

When $t \in \mathcal{B}_J$,

$$E\left\{\frac{|\Delta(ct)|^2}{|\phi_{\bar{Y}}(ct)|^2}\right\} = \frac{1-|\phi_{\bar{Z}}(ct)|^2}{J|\phi_{\bar{Z}}(ct)|^2} = O\left\{\frac{1}{J|\phi_{\bar{Z}}(ct)|^2}\right\} = O\left(\frac{1}{Jh^{2c\alpha_0+2\beta_0}}\right).$$

The last equation is due to $\inf_{t\in\mathcal{B}_J} |\phi_{\bar{Z}}(ct)| = |\phi_Y(1/h)|^c |\phi_\epsilon(c/h)|$ for large J and ϕ_Y and ϕ_ϵ are ordinary smooth of orders α_0 and β_0, respectively. When $Jh^{2c\alpha_0+2\beta_0} \to \infty$, $|\Delta(t)/\phi_{\bar{Y}}(t)| = o_p(1)$ and thus is less than $1/2$ uniformly on $t \in \mathcal{B}_J$ with probability 1. For (ii), with the property $|\Delta(t)/\phi_{\bar{Y}}(t)| = o_p(1)$, $|\widehat{\phi}_{\bar{Y}}(ct)| \geq \inf_{t\in\mathcal{B}_J} |\phi_{\bar{Y}}(ct)|/2 > 0$, which means that $\widehat{\phi}_{\bar{Y}}(ct)$ will not vanish when $t \in \mathcal{B}_J$. According to Chung (2001), the cth root of $\widehat{\phi}_{\bar{Y}}(ct)$ exists. Therefore, we have established the expansion in Equation (A.5).

Let $\Delta_2(t) = \widehat{\phi}_{\bar{Z}|x}(t)/\phi_\epsilon(t) - \phi_{\bar{Y}|x}(t)$. Plugging Equation (A.5) into Equation (A.4) yields a decomposition of $\widehat{\phi}_{Y|x}(t) - \phi_{Y|x}(t)$; i.e.,

$$\begin{aligned} \widehat{\phi}_{Y|x}(t) - \phi_{Y|x}(t) = &\frac{\Delta_2(ct)}{\phi_{\bar{Y}}(ct)^{(c-1)/c}} + \frac{1-c}{c}\frac{\phi_{\bar{Y}|x}(ct)}{\phi_{\bar{Y}}(ct)^{(2c-1)/c}}\Delta(ct) + \frac{1-c}{c}\frac{\Delta(ct)\Delta_2(ct)}{\phi_{\bar{Y}}(ct)^{(2c-1)/c}} \\ &+ \tilde{\lambda}_{2,r_0}(t)\frac{\phi_{\bar{Y}|x}(ct)}{\phi_{\bar{Y}}(ct)^{(3c-1)/c}}\Delta(ct)^2 + \tilde{\lambda}_{2,r_0}(t)\frac{\Delta(ct)^2\Delta_2(ct)}{\phi_{\bar{Y}}(ct)^{(3c-1)/c}}. \end{aligned} \tag{A.6}$$

Let $\Delta_1(t) = \widehat{\phi}_{\bar{Z}}(t) - \phi_{\bar{Z}}(t)$. Noting that

$$\phi_{\bar{Y}}(ct)^{1/c} = \phi_Y(t) \text{ and } \phi_{\bar{Y}|x}(ct) = \phi_{Y|x}(t)\phi_Y(t)^{c-1},$$

we write Equation (A.6) as

$$\begin{aligned} \widehat{\phi}_{Y|x}(t) - \phi_{Y|x}(t) = &\frac{\Delta_2(ct)}{\phi_Y(t)^{c-1}} + \frac{1-c}{c}\frac{\phi_{Y|x}(t)\Delta_1(ct)}{\phi_Y(t)^c\phi_\epsilon(ct)} + \frac{1-c}{c}\frac{\Delta_1(ct)\Delta_2(ct)}{\phi_Y(t)^{2c-1}\phi_\epsilon(ct)} \\ &+ \tilde{\lambda}_{2,r_0}(t)\frac{\phi_{Y|x}(t)\Delta_1(ct)^2}{\phi_Y(t)^{2c}\phi_\epsilon(ct)^2} + \tilde{\lambda}_{2,r_0}(t)\frac{\Delta_1(ct)^2\Delta_2(ct)}{\phi_Y(t)^{3c-1}\phi_\epsilon(ct)^2}. \end{aligned} \tag{A.7}$$

Plugging Equation (A.7) into term (A.2) yields 5 terms, which are denoted by T_1 to T_5. We list them below and evaluate them one by one. The most important term is

$$\begin{aligned}
T_1 &= \frac{1}{2\pi}\int \exp(-ity)\frac{\Delta_2(ct)}{\phi_Y(t)^{c-1}}\phi_K(ht)dt \\
&= \frac{1}{2\pi}\int \exp(-ity)\left[\frac{\sum_{j=1}^J\sum_{i=1}^c w_{ij}(x)\{\phi_{\bar{Z}|X_{ij}}(ct)-\phi_{\bar{Z}|x}(ct)\}}{\sum_{j=1}^J\sum_{i=1}^c w_{ij}(x)\phi_\epsilon(ct)}\right]\frac{\phi_K(ht)}{\phi_Y(t)^{c-1}}dt \\
&\quad + \frac{1}{2\pi}\int \exp(-ity)\left[\frac{\sum_{j=1}^J\sum_{i=1}^c w_{ij}(x)\{e^{ic\bar{Z}_j t}-\phi_{\bar{Z}|X_{ij}}(ct)\}}{\sum_{j=1}^J\sum_{i=1}^c w_{ij}(x)\phi_\epsilon(ct)}\right]\frac{\phi_K(ht)}{\phi_Y(t)^{c-1}}dt \\
&= T_1^* + T_1^{**}. \qquad \text{(A.8)}
\end{aligned}$$

In the following, we show that term T_1^*, together with (A.3), provides $B_{\bar{h},h}(x,y)$; term T_1^{**} provides $V_{\bar{h},h}^{1/2}(x,y)$; the rest terms $T_2,\ldots,T_5$ are all $o_p(T_1^{**})$.

For T_1^*, we have

$$\begin{aligned}
T_1^* &= \frac{1}{2\pi}\int \exp(-ity)\left[\frac{\sum_{j=1}^J\sum_{i=1}^c w_{ij}(x)\{\phi_{Y|X_{ij}}(t)-\phi_{Y|x}(t)\}}{\sum_{j=1}^J\sum_{i=1}^c w_{ij}(x)}\right]\phi_K(ht)dt \\
&= \frac{\sum_{j=1}^J\sum_{i=1}^c w_{ij}(x)\{f(y|X_{ij})-f(y|x)\}}{\sum_{j=1}^J\sum_{i=1}^c w_{ij}(x)}\{1+o_p(1)\} \\
&= \frac{\sum_{j=1}^J\sum_{i=1}^c w_{ij}(x)}{\{\sum_{j=1}^J\sum_{i=1}^c w_{ij}(x)\}}\left\{\frac{\partial f(y|x)}{\partial x}(X_{ij}-x)+\frac{\partial^2 f(y|x)}{2\partial x^2}(X_{ij}-x)^2+O_p(X_{ij}-x)^3\right\} \\
&\quad \times\{1+o_p(1)\} \\
&= \bar{h}^2 b_1(x)\frac{\partial^2 f(y|x)}{2\partial x^2}+o_p(\bar{h}^2), \qquad \text{(A.9)}
\end{aligned}$$

where $b_1(x)=\mu_2$ and $\mu_2=\int u^2K(u)du$. The last equation follows the typical argument of local polynomial estimator; see, e.g., Fan (2018). Combining term (A.9) with term (A.3) yields $B_{\bar{h},h}(x,y)$.

For T_1^{**}, letting

$$W_i(x)=\sum_{j=1}^J w_{ij}(x)/\sum_{i=1}^c\sum_{j=1}^J w_{ij}(x),$$

we have $T_1^{**}=\sum_{i=1}^c W_i(x)T_{1i}^{**}$, where

$$T_{1i}^{**}=\frac{1}{2\pi}\int \exp(-ity)\left[\frac{\sum_{j=1}^J w_{ij}(x)\{e^{ic\bar{Z}_j t}-\phi_{\bar{Z}|X_{ij}}(ct)\}}{\sum_{j=1}^J w_{ij}(x)\phi_\epsilon(ct)}\right]\frac{\phi_K(ht)}{\phi_Y(t)^{c-1}}dt.$$

Denote by $\mathcal{X}_i = \{X_{i1}, \ldots, X_{iJ}\}$ for $i = 1, \ldots, c$. Using Lemmas A.1 and A.2 and Parseval's identity, the order of $E(T_{1i}^{**2}|\mathcal{X}_i)$ is the same as the order of

$$\frac{\sum_{j=1}^J w_{ij}(x)^2}{\{\sum_{j=1}^J w_{ij}(x)\}^2} \int \left\{ \frac{1}{2\pi h} \int \exp\left(-it\frac{y - c\bar{Z}_j}{h}\right) \frac{\phi_K(t)}{\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)} dt \right\}^2 f_{\bar{Z}|X_{ij}}(\bar{Z}_j) d\bar{Z}_j,$$

which is

$$\frac{c^{2\beta_0} v_1(x) f_{c\bar{Z}|x}(y)}{2\pi J \bar{h} h^{2(c-1)\alpha_0 + 2\beta_0 + 1}} \int \frac{|\phi_K(t)|^2 t^{2(c-1)\alpha_0 + 2\beta_0}}{A_1^{2(c-1)} A_2^2} dt\{1 + o_p(1)\},$$

where $v_1(x) = \nu_0 / f_X(x)$, $\nu_0 = \int K^2(u)du$, and $f_{\bar{Z}|X_{ij}}$ and $f_{\bar{Z}|x}$ are the density functions of $\bar{Z}_j$ given X_{ij} and $X_{ij} = x$, respectively. Combining the above result with the definition of $W_i(x)$, the order of T_1^{**} equals

$$O_p\{c^{-1/2+\beta_0} J^{-1/2} \bar{h}^{-1/2} h^{-(c-1)\alpha_0 - \beta_0 - 1/2}\},$$

which is $V_{\bar{h},h}^{1/2}(x, y)$.

Now we show that $T_2, \ldots, T_5$ are all $o_p(T_1^{**})$. Starting from T_2, we rewrite it as

$$\begin{aligned} T_2 &= \frac{1-c}{c} \frac{1}{2\pi} \int \exp(-ity) \frac{\phi_{Y|x}(t)}{\phi_Y(t)^c} \Delta(ct) \phi_K(ht) dt \\ &= \frac{1-c}{c} \frac{1}{2\pi J} \sum_{j=1}^J \int \exp(-ity) \frac{\phi_{Y|x}(t)\phi_K(ht)}{\phi_Y(t)^c \phi_\epsilon(ct)} \{e^{ic\bar{Z}_j t} - \phi_{\bar{Z}}(ct)\} dt. \end{aligned} \tag{A.10}$$

Then following the similar argument in deriving T_1^{**}, we have

$$T_2 = O_p\{J^{-1/2} h^{-c\alpha_0 + \rho_0(x) - \beta_0 - 1/2}\}.$$

As $\bar{h}^{-1} h^{2\{\alpha_0 - \rho_0(x)\}} \to \infty$, we have $T_2 = o_p(T_1^{**})$. The T_3, T_4, and T_5 are written by

$$T_3 = \frac{1}{2\pi} \int \exp(-ity) \frac{1-c}{c} \frac{\Delta_1(ct)\Delta_2(ct)}{\phi_Y(t)^{2c-1}\phi_\epsilon(ct)} \phi_K(ht) dt \tag{A.11}$$

$$T_4 = \frac{1}{2\pi} \int \exp(-ity) \tilde{\lambda}_{2,r_0}(t) \frac{\phi_{Y|x}(t)\Delta_1(ct)^2}{\phi_Y(t)^{2c}\phi_\epsilon(ct)^2} \phi_K(ht) dt \tag{A.12}$$

$$T_5 = \frac{1}{2\pi} \int \exp(-ity) \tilde{\lambda}_{2,r_0}(t) \frac{\Delta_1(ct)^2\Delta_2(ct)}{\phi_Y(t)^{3c-1}\phi_\epsilon(ct)^2} \phi_K(ht) dt. \tag{A.13}$$

We conclude

$$T_3 = O_p\{J^{-1} \bar{h}^{-1/2} h^{-(2c-1)\alpha_0 - 2\beta_0 - 1/2}\}$$

from Lemmas A.3. Comparing the order of T_1^{**}, T_2, and T_3, we have as

$$\bar{h}^{-1}h^{2\{\alpha_0-\rho_0(x)\}} \to \infty \text{ and } Jh^{2c\alpha_0+2\beta_0} \to \infty,$$

$T_2 = o_p(T_1^{**})$ and $T_3 = o_p(T_1^{**})$, respectively. For T_4 and T_5, as $|\Delta(ct)|/|\phi_{\bar{Y}}(ct)| < 1/2$ when $t \in \mathcal{B}_J$, we have $T_4 = O_p(T_2)$ and $T_5 = O_p(T_3)$. Consequently, $T_4 = o_p(T_1^{**})$ and $T_5 = o_p(T_1^{**})$, respectively.

In summary, we have completed the proof of Theorem 2.1 by showing that

$$\begin{aligned}\widehat{f}_{RP}(y|x) - f(y|x) &= \frac{1}{2\pi}\int e^{-ity}\phi_{Y|x}(t)\left\{\phi_K(ht)-1\right\}dt + T_1^* + T_1^{**}\{1+o_p(1)\} \\ &= B_{\bar{h},h}(x,y) + V_{\bar{h},h}^{1/2}(x,y),\end{aligned}$$

where

$$\begin{aligned}B_{\bar{h},h}(x,y) &= (2\pi)^{-1}\int e^{-ity}\phi_{Y|x}(t)\left\{\phi_K(ht)-1\right\}dt + T_1^* \\ V_{\bar{h},h}(x,y) &= T_1^{**2}\{1+o_p(1)\}.\end{aligned}$$

□

A.1.1 Lemmas for Theorem 2.1

Lemma A.1 is from Lemma 2.1 in Fan (1991a). Lemma A.2 verifies the condition of Lemma A.1. Lemmas A.3 uses techniques similar to Delaigle and Zhou (2015) to show the order of T_3.

Lemma A.1. *Suppose that $F_n(\cdot)$ is a sequence of Borel functions satisfying $F_n(y) \to F(y)$ and $sup_n|F_n(y)| \le F^*(y)$, where $F^*(y)$ satisfies*

$$\int F^*(y)dy \le \infty \text{ and } \lim_{y\to\infty}|yF^*(y)| = 0.$$

If x is a continuous point of a density $f(\cdot)$, then for any sequence $h_n \to 0$, we have

$$\lim_{n\to\infty}\frac{1}{h_n}\int F_n\Big(\frac{x-y}{h_n}\Big)f(y)dy = f(x)\int F(y)dy.$$

Lemma A.2. *Under Conditions OO and (C2),*

$$|h^{(c-1)\alpha_0+\beta_0}K^*(x)| \le \min(C_1, C_2/x),$$

where C_1 and C_2 are constants.

Proof. Note under Condition OO, there exists a large M such that

$$|t^{\alpha_0}\phi_Y(t)| > A_1/2 \text{ and } |t^{\beta_0}\phi_\epsilon(t)| > A_2/2, \text{ when } |t| > M.$$

Then for $c > 1$,

$$\begin{aligned}
|h^{(c-1)\alpha_0+\beta_0}K^*(x)| &\le \int_{-1}^{1}\left|\frac{h^{(c-1)\alpha_0+\beta_0}\phi_K(t)}{\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)}\right|dt \\
&\le \int_{Mh\le|t|\le 1}\left|\frac{t^{(c-1)\alpha_0+\beta_0}\phi_K(t)}{(A_1/2)^{c-1}A_2/2}\right|dt \\
&\quad + \int_{0\le|t|\le Mh}\frac{\sup_t|\phi_K(t)|}{\min_{|t|\le M}|\phi_Y(t)^{c-1}\phi_\epsilon(ct)|}dt = O(1). \qquad \text{(A.14)}
\end{aligned}$$

On the other hand, by integration by parts, we have, as $J \to \infty$ and $x \neq 0$,

$$|K^*(x)| = \left|\frac{1}{2\pi}\int_{-1}^{1}\exp(-itx)\frac{\phi_K(t)}{\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)}dt\right| \le F_1 + F_2, \qquad \text{(A.15)}$$

where

$$\begin{aligned}
F_1 &= \left|\frac{1}{x\pi}\frac{\phi_K(1)}{\phi_Y(1/h)^{c-1}\phi_\epsilon(c/h)}\right| \\
F_2 &= \left|\frac{1}{2x\pi}\int_{-1}^{1}\exp(-itx)\left\{\frac{\phi_K(t)}{\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)}\right\}'dt\right|.
\end{aligned}$$

For F_1, when J is large, $F_1 = O\{h^{-(c-1)\alpha_0-\beta_0}/x\}$. For F_2,

$$\begin{aligned}
F_2 &= \left|\frac{1}{2x\pi}\int_{-1}^{1}\exp(-itx)\frac{\phi_K'(t)}{\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)}dt\right| \\
&\quad + \left|\frac{1}{2x\pi}\int_{-1}^{1}\exp(-itx)\frac{\phi_K(t)\{\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)\}'}{\{\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)\}^2}dt\right| \\
&\le \frac{1}{2x\pi}\int_{-1}^{1}\left|\frac{\phi_K'(t)}{\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)}\right|dt + \frac{(c-1)h^{-1}}{2x\pi}\int_{-1}^{1}\left|\frac{\phi_K(t)\phi_Y'(t/h)}{\phi_Y(t/h)^{c}\phi_\epsilon(ct/h)}\right|dt \\
&\quad + \frac{ch^{-1}}{2x\pi}\int_{-1}^{1}\left|\frac{\phi_K(t)\phi_\epsilon'(ct/h)}{\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)^2}\right|dt \\
&= F_{21} + F_{22} + F_{23}.
\end{aligned}$$

For F_{21},

$$F_{21} \leq \frac{1}{2x\pi}\int_{0<|t|<Mh}\frac{|\sup_t \phi_K'(t)|}{\min_{|t|\leq M}|\phi_Y(t)^{c-1}\phi_\epsilon(ct)|}dt$$
$$+\frac{h^{-(c-1)\alpha_0-\beta_0}}{2x\pi}\int_{Mh<|t|<1}\frac{t^{(c-1)\alpha_0+\beta_0}\sup_t|\phi_K'(t)|}{(A_1/2)^{c-1}A_2/2}dt = O\{h^{-(c-1)\alpha_0-\beta_0}/x\}.$$

For F_{22}, provided that $\phi_Y'(t)t^{\alpha_0+1} = O(1)$ in Condition OO,

$$F_{22} \leq \frac{(c-1)h^{-1}}{2x\pi}\int_{0<|t|<Mh}\sup_t|\phi_K'(t)|\min_{|t|\leq M}\frac{|\phi_Y'(t)|}{|\phi_Y(t)^c\phi_\epsilon(ct)|}dt+$$
$$const.\frac{(c-1)h^{-(c-1)\alpha_0-\beta_0}}{2x\pi}\int_{|t|>Mh}\frac{t^{(c-1)\alpha_0+\beta_0-1}|\phi_K(t)|}{(A_1/2)^cA_2/2}dt = O\{h^{-(c-1)\alpha_0-\beta_0}/x\}.$$

For F_{23}, provided that $\phi_\epsilon'(t)t^{\beta_0+1} = O(1)$ in Condition OO,

$$F_{23} \leq \frac{ch^{-1}}{2x\pi}\int_{0<|t|<Mh}\sup_t|\phi_K(t)|\min_{|t|\leq M}\frac{|\phi_\epsilon'(ct)|}{|\phi_Y(t)^{c-1}\phi_\epsilon(ct)^2|}dt$$
$$+const.\frac{ch^{-(c-1)\alpha_0-\beta_0}}{2x\pi}\int_{|t|>Mh}\frac{t^{(c-1)\alpha_0+\beta_0-1}\phi_K(t)}{(A_1/2)^{c-1}(A_2/2)^2}dt = O\{h^{-(c-1)\alpha_0-\beta_0}/x\}.$$

Combining F_1, F_{21}, F_{22}, and F_{23}, we have $|h^{(c-1)\alpha_0+\beta_0}K^*(x)| = O(1/x)$. Further combining this with Equation (A.14), we finish our proof. □

Lemma A.3. *Under Conditions OO and (C1)-(C3), when* $(2c-1)\alpha_0+2\beta_0 > 1$,

$$T_3 = O_p\{J^{-1}\bar{h}^{-1/2}h^{-(2c-1)\alpha_0-2\beta_0-1/2}\}.$$

Proof. Recall that $|\tilde{\lambda}_{2,r_0}(t)| \leq 4$ when $t \in \mathcal{B}_J$,

$$E(T_3^2) \leq 16E\left\{\left|\int \exp(-ity)\frac{\phi_K(ht)\Delta_2(ct)\Delta_1(ct)}{\phi_Y(t)^{2c-1}\phi_\epsilon(ct)}dt\right|^2\right\}$$
$$= 16\int\int \exp\{-i(u-v)y\}\frac{\phi_K(hu)\phi_K(-hv)}{\phi_Y(u)^{2c-1}\phi_Y(-v)^{2c-1}\phi_\epsilon(cu)^2\phi_\epsilon(-cv)^2}$$
$$\times E\{\Delta_2(cu)\phi_\epsilon(cu)\Delta_2(-cv)\phi_\epsilon(-cv)\Delta_1(cu)\Delta_1(-cv)\}dudv. \qquad (A.16)$$

To keep notation concise, further define

$$\omega_j(t) = e^{it\bar{Z}_j} - \phi_{\bar{Z}}(t)$$
$$\tau_j(t) = \sum_{i=1}^c w_{ij}(x)\{e^{it\bar{Z}_j} - \phi_{\bar{Z}|x}(t)\}/\sum_{j=1}^J\sum_{i=1}^c w_{ij}(x).$$

Then we have,

$$
\begin{aligned}
& E\left\{\Delta_2(cu)\phi_\epsilon(cu)\Delta_2(-cv)\phi_\epsilon(-cv)\Delta_1(cu)\Delta_1(-cv)\right\} \\
&= \frac{1}{J^2}E\left\{\sum_{j=1}^{J}\tau_j(cu)\sum_{j=1}^{J}\tau_j(-cv)\sum_{j=1}^{J}\omega_j(cu)\sum_{j=1}^{J}\omega_j(-cv)\right\} \\
&= \frac{1}{J^2}E\left\{\sum_{j_1\neq j_2}\tau_{j_1}(cu)\tau_{j_1}(-cv)\omega_{j_2}(cu)\omega_{j_2}(-cv)\right\} \\
&\quad + \frac{1}{J^2}E\left\{\sum_{j_1\neq j_2}\tau_{j_1}(cu)\tau_{j_2}(-cv)\omega_{j_1}(cu)\omega_{j_2}(-cv)\right\} \\
&\quad + \frac{1}{J^2}E\left\{\sum_{j_1\neq j_2}\tau_{j_1}(cu)\tau_{j_2}(-cv)\omega_{j_2}(cu)\omega_{j_1}(-cv)\right\} \\
&\quad + \frac{1}{J^2}E\left\{\sum_{j=1}^{J}\tau_j(cu)\tau_j(-cv)\omega_j(cu)\omega_j(-cv)\right\} + R_J(u,v), \qquad \text{(A.17)}
\end{aligned}
$$

where $R_J(u,v)$ is negligible to the other four terms and thus is omitted in the following calculation. Plugging Equation (A.17) into Equation (A.16) yields 4 terms, which are denoted by F_1, F_2, F_3, and F_4. We focus on

$$
\begin{aligned}
F_1 = \frac{16}{J^2}\int\int \exp\left\{-i(u-v)y\right\}\frac{\phi_K(hu)\phi_K(-hv)}{\phi_Y(u)^{2c-1}\phi_Y(-v)^{2c-1}\phi_\epsilon(cu)^2\phi_\epsilon(-cv)^2} \\
\times\sum_{j_1\neq j_2}E\left\{\tau_{j_1}(cu)\tau_{j_1}(-cv)\omega_{j_2}(cu)\omega_{j_2}(-cv)\right\}dudv
\end{aligned}
$$

to illustrate this procedure. Let $\tilde{\mathcal{X}} = \{X_1,\ldots,X_N\}$. To calculate F_1, first we have

$$
\begin{aligned}
& E\left\{\tau_{j_1}(cu)\tau_{j_1}(-cv)\omega_{j_2}(cu)\omega_{j_2}(-cv)\right\} \qquad \text{(A.18)} \\
&= E\left[E\left\{\frac{\sum_{i=1}^{c}w_{ij_1}(x)\{e^{iu\bar{Z}_{j_1}}-\phi_{\bar{Z}|x}(u)\}\sum_{i=1}^{c}w_{ij_1}(x)\{e^{-iv\bar{Z}_{j_1}}-\phi_{\bar{Z}|x}(-v)\}}{\{\sum_{j=1}^{J}\sum_{i=1}^{c}w_{ij}(x)\}^2}\middle|\tilde{\mathcal{X}}\right\}\right] \\
&\quad\times\left[\phi_{\bar{Z}}\{c(u-v)\}-\phi_{\bar{Z}}(cu)\phi_{\bar{Z}}(-cv)\right],
\end{aligned}
$$

which is of the same order as

$$
\begin{aligned}
& E\left\{\frac{\sum_{i=1}^{c}w_{ij_1}^2(x)\left[\phi_{\bar{Z}|X_{ij_1}}\{c(u-v)\}-\phi_{\bar{Z}|x}(cu)\phi_{\bar{Z}|X_{1j_1}}(-cv)\right]}{\{\sum_{j=1}^{J}\sum_{i=1}^{c}w_{ij}(x)\}^2}\right. \\
&\quad\left.+\frac{\sum_{i=1}^{c}w_{ij_1}^2(x)\left[-\phi_{\bar{Z}|X_{ij_1}}(cu)\phi_{\bar{Z}|x}(-cv)+\phi_{\bar{Z}|x}(cu)\phi_{\bar{Z}|x}(-cv)\right]}{\{\sum_{j=1}^{J}\sum_{i=1}^{c}w_{ij}(x)\}^2}\right\} \\
&\quad\times\left[\phi_{\bar{Z}}\{c(u-v)\}-\phi_{\bar{Z}}(cu)\phi_{\bar{Z}}(-cv)\right]. \qquad \text{(A.19)}
\end{aligned}
$$

The dominant term inside term (A.19) is

$$E\left[\frac{\sum_{i=1}^{c} w_{ij_1}^2(x)\phi_{\bar{Z}|X_{ij_1}}\{c(u-v)\}}{\{\sum_{j=1}^{J}\sum_{i=1}^{c} w_{ij}(x)\}^2}\right]\phi_{\bar{Z}}\{c(u-v)\}. \tag{A.20}$$

Thus, replacing (A.18) with its dominant term (A.20), we have

$$\begin{aligned}
|F_1| &= \frac{16(J-1)}{J^2}\left|\int\int \exp\{-i(u-v)y\}\,\frac{\phi_{\bar{Z}}\{c(u-v)\}\phi_K(hu)\phi_K(-hv)}{\phi_Y(u)^{2c-1}\phi_Y(-v)^{2c-1}\phi_\epsilon(cu)^2\phi_\epsilon(-cv)^2}\right.\\
&\quad\left.\times E\left[\frac{\sum_{j=1}^{J}\sum_{i=1}^{c} w_{ij}^2(x)\phi_{\bar{Z}|X_{ij}}\{c(u-v)\}}{\{\sum_{j=1}^{J}\sum_{i=1}^{c} w_{ij}(x)\}^2}\right]dudv\right|\\
&= \frac{16(J-1)}{J^2}\left|\int\int \exp\{-i(u-v)y\}\,\frac{\phi_Y(u-v)^{2c-1}\phi_\epsilon(cu-cv)^2\phi_K(hu)\phi_K(-hv)}{\phi_Y(u)^{2c-1}\phi_Y(-v)^{2c-1}\phi_\epsilon(cu)^2\phi_\epsilon(-cv)^2}\right.\\
&\quad\left.\times E\left[\frac{\sum_{j=1}^{J}\sum_{i=1}^{c} w_{ij}^2(x)\phi_{Y|X_{ij}}(u-v)}{\{\sum_{j=1}^{J}\sum_{i=1}^{c} w_{ij}(x)\}^2}\right]dudv\right|\\
&\le \frac{16}{J}\int\int \frac{|\phi_Y(u-v)^{2c-1}||\phi_\epsilon(cu-cv)^2|}{|\phi_Y(u)^{2c-1}||\phi_Y(-v)^{2c-1}||\phi_\epsilon(cu)^2||\phi_\epsilon(-cv)^2|}dudv\\
&\quad\times E\left[\frac{\sum_{j=1}^{J}\sum_{i=1}^{c} w_{ij}^2(x)}{\{\sum_{j=1}^{J}\sum_{i=1}^{c} w_{ij}(x)\}^2}\right],
\end{aligned}$$

which is of the same order as

$$\frac{v_1(x)}{J^2c\bar{h}}\int\int \frac{|\phi_Y(u-v)^{2c-1}||\phi_\epsilon(cu-cv)^2|}{|\phi_Y(u)^{2c-1}||\phi_Y(-v)^{2c-1}||\phi_\epsilon(cu)^2||\phi_\epsilon(-cv)^2|}dudv. \tag{A.21}$$

To evaluate the integral, we partition $\mathbb{R}^2$ into eight regions: $\mathbb{R}^2 = \cup_{r=1}^{8}\mathcal{A}_r$, and write

$$\int\int_{\mathbb{R}} = \sum_{r=1}^{8}\int\int_{\mathcal{A}_r}$$

where

$$\mathcal{A}_1 = \{(u,v)\in\mathbb{R}^2 : 2|u-v| > M, |u| > M, |v| > m\}$$

$$\mathcal{A}_2 = \{(u,v)\in\mathbb{R}^2 : 2|u-v| > M, |u| > M, |v| \le m\}$$

$$\mathcal{A}_3 = \{(u,v)\in\mathbb{R}^2 : 2|u-v| > M, |u| \le M, |v| > m\}$$

$$\mathcal{A}_4 = \{(u,v)\in\mathbb{R}^2 : 2|u-v| > M, |u| \le M, |v| \le m\}$$

$$\mathcal{A}_5 = \{(u,v)\in\mathbb{R}^2 : 2|u-v| \le M, |u| > M, |v| > m\}$$

$$\mathcal{A}_6 = \{(u,v)\in\mathbb{R}^2 : 2|u-v| \le M, |u| > M, |v| \le m\}$$

$$\mathcal{A}_7 = \{(u,v)\in\mathbb{R}^2 : 2|u-v| \le M, |u| \le M, |v| > m\}$$

$$\mathcal{A}_8 = \{(u,v)\in\mathbb{R}^2 : 2|u-v| \le M, |u| \le M, |v| \le m\}.$$

The leading term among these 8 integrals is

$$\Big(\int\int_{\mathcal{A}_1}+\int\int_{\mathcal{A}_5}\Big)\frac{|\phi_K(hu)||\phi_K(hv)||\phi_Y(u-v)^{2c-1}||\phi_\epsilon(cu-cv)^2|}{|\phi_Y(u)^{2c-1}||\phi_Y(-v)^{2c-1}||\phi_\epsilon(cu)^2||\phi_\epsilon(-cv)^2|}dudv. \tag{A.22}$$

The integral with respect to $\mathcal{A}_1$ is of the same order as

$$\int\int_{\mathcal{A}_1}|\phi_K(hu)||\phi_K(hv)|\frac{|u|^{(2c-1)\alpha_0+2\beta_0}|v|^{(2c-1)\alpha_0+2\beta_0}}{|u-v|^{(2c-1)\alpha_0+2\beta_0}}dudv,$$

and the one with respect to $\mathcal{A}_5$ is of the same order as

$$\int\int_{\mathcal{A}_5}|\phi_K(hu)||\phi_K(hv)||\phi_Y(u-v)^{2c-1}||\phi_\epsilon(cu-cv)^2||u|^{(2c-1)\alpha_0+2\beta_0}|v|^{(2c-1)\alpha_0+2\beta_0}dudv.$$

When $|u-v|\geq M$, we have

$$\frac{|u-v|^{(2c-1)\alpha_0+2\beta_0}}{\{1+|u-v|\}^{(2c-1)\alpha_0+2\beta_0}}\geq\frac{M^{(2c-1)\alpha_0+2\beta_0}}{\{1+M\}^{(2c-1)\alpha_0+2\beta_0}},$$

and thus

$$\{|u-v|^{(2c-1)\alpha_0+2\beta_0}\}^{-1}\leq const.\{1+|u-v|\}^{-(2c-1)\alpha_0-2\beta_0}.$$

When $|u-v|\leq M$, we can find a large L such that

$$|\phi_Y(u-v)^{2c-1}||\phi_\epsilon(cu-cv)^2|\leq L(1+|u-v|)^{-(2c-1)\alpha_0-2\beta_0};$$

e.g., let $L=\sup_{|t|\leq M}\{|\phi_Y(t)^{2c-1}||\phi_\epsilon(ct)^2|\}(1+M)^{(2c-1)\alpha_0+2\beta_0}$. Thus, (A.22) is bounded by a constant times

$$\begin{aligned}
&\int\int_{\mathcal{A}_5\cup\mathcal{A}_1}|\phi_K(hu)||\phi_K(hv)|\frac{|u|^{(2c-1)\alpha_0+2\beta_0}|v|^{(2c-1)\alpha_0+2\beta_0}}{(1+|u-v|)^{(2c-1)\alpha_0+2\beta_0}}dudv\\
\leq\ & const.\int\int|\phi_K(hu)||\phi_K(hv)|\frac{|u|^{(2c-1)\alpha_0+2\beta_0}|v|^{(2c-1)\alpha_0+2\beta_0}}{(1+|u-v|)^{(2c-1)\alpha_0+2\beta_0}}dudv\\
\leq\ & const.\int\int|\phi_K(hu)||\phi_K(hv)|\frac{|u|^{(2c-1)\alpha_0+2\beta_0}(1+|v|)^{(2c-1)\alpha_0+2\beta_0}}{(1+|u-v|)^{(2c-1)\alpha_0+2\beta_0}}dudv\\
\leq\ & const.\int\int|\phi_K(hu)||\phi_K(hv)|\frac{|u|^{(2c-1)\alpha_0+2\beta_0}(|u|+1+|u-v|)^{(2c-1)\alpha_0+2\beta_0}}{(1+|u-v|)^{(2c-1)\alpha_0+2\beta_0}}dudv\\
\leq\ & const.\int\int|\phi_K(hu)||\phi_K(hv)|\frac{|u|^{(2c-1)\alpha_0+2\beta_0}(|u|+1+|u-v|)^{(2c-1)\alpha_0+2\beta_0}}{(1+|u-v|)^{(2c-1)\alpha_0+2\beta_0}}dudv\\
\leq\ & const.\int\int|\phi_K(hu)||\phi_K(hv)|\frac{|u|^{2(2c-1)\alpha_0+4\beta_0}}{(1+|u-v|)^{(2c-1)\alpha_0+2\beta_0}}dudv\\
&+const.\int\int|\phi_K(hu)||\phi_K(hv)||u|^{(2c-1)\alpha_0+2\beta_0}dudv=G_1+G_2.
\end{aligned}$$

The last inequality is due to the fact that when $a > 1$,

$$(x_1 + x_2)^a \leq 2^{a-1}x_1^a + 2^{a-1}x_2^a.$$

For G_1, noting that $\sup_v |\phi_K(hv)| \leq 1$ and substituting $u - v$ with s, we have

$$G_1 \leq \int |\phi_K(hu)||u|^{2(2c-1)\alpha_0+4\beta_0}du \int \frac{1}{(1+|s|)^{(2c-1)\alpha_0+2\beta_0}}ds,$$

When $(2c-1)\alpha_0+2\beta_0 > 1$, the order of G_1 is $h^{-2(2c-1)\alpha_0-4\beta_0-1}$. The term G_2 is of order $h^{-(2c-1)\alpha_0-2\beta_0-2}$, which is smaller than $h^{-2(2c-1)\alpha_0-4\beta_0-1}$. Therefore, the order of the integral inside term (A.21) equals $h^{-2(2c-1)\alpha_0-4\beta_0-1}$. Then the order of term (A.21), which is the leading term of F_1, equals $O_p\{J^{-2}\bar{h}^{-1}h^{-2(2c-1)\alpha_0-4\beta_0-1}\}$. Following the similar arguments, we can show the terms F_2, F_3, and F_4 are negligible to term F_1. Then we can see T_3 is of order $O_p\{J^{-1}\bar{h}^{-1/2}h^{-(2c-1)\alpha_0-2\beta_0-1/2}\}$. □

A.2 Proof of Theorem 2.2

Proof of Theorem 2.2. The proof is similar to Theorem 2.1. The main step is to plug Equation (A.7) into term (A.2) and to analyze the 5 terms yielded. Before this step, we need to verify that it is valid to expand $1/\{\widehat{\phi}_{\bar{Y}}(ct)\}^{(c-1)/c}$ into power series at $\phi_{\bar{Y}}(ct)$. This can be done by showing: As $t \in \mathcal{B}_J$ and

$$Jh^{c_2}\exp\{-2c\gamma^{-1}h^{-\alpha_2} - 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\} \to \infty,$$

for any c_2, we have $|\Delta(ct)/\phi_{\bar{Y}}(ct)| < 1/2$ uniformly on t with probability 1 and $\widehat{\phi}_{\bar{Y}}(t)$ does not vanish.

Once we establish Equation (A.5), by plugging it into Equation (A.4), we obtain a decomposition of $\widehat{\phi}_{Y|x}(t) - \phi_{Y|x}(t)$, i.e., Equation (A.7). Plugging Equation (A.7) into term (A.2) yields 5 terms T_1 to T_5 defined in (A.8), (A.10)–(A.13). We will evaluate them one by one.

The calculation of T_1^* is the same as Equation (A.9) in Theorem 2.1. We have

$$T_1^* = \bar{h}^2 b_1(x)\frac{\partial^2 f(y|x)}{2\partial x^2} + o_p(\bar{h}^2).$$

Term T_1^*, together with term (A.3), makes up the bias $B_{\bar{h},h}(x,y)$. Now we focus on the evaluation of variance $V_{\bar{h},h}(x,y)$, which is related to T_1^{**}, T_2, T_3, T_4, and T_5. We will show that T_1^{**} provides $V_{\bar{h},h}^{1/2}(x,y)$, and the rest terms T_2, T_3, T_4, and T_5 are all $o_p(T_1^{**})$. For the proof below, we assume $\alpha_2 \geq \beta_2$, i.e., the CF of Y decays faster than the CF of ϵ. For $\alpha_2 < \beta_2$, the proof is in the same pattern. Now we start our proof, which will check the order of T_1^{**}, T_2, T_3, T_4, and T_5.

The most challenging part is T_1^{**}, where we further split into two cases: $\alpha_2 < 1$ and $\alpha_2 \geq 1$. In Case 1, where $\alpha_2 < 1$, with Condition (C4), we can calculate the accurate order of T_1^{**}. In Case 2, where $\alpha_2 \geq 1$, it's hard to calculate the exact order of T_1^{**} and we estimate the upper bound instead. We include the conclusion of the two cases in Lemma A.4. Note that from (A.42) in the proof of Lemma A.4, we show whenever in Case 1 or Case 2,

$$T_1^{**} = O_p[J^{-1/2}\bar{h}^{-1/2}h^{c_4}\exp\{(c-1)\gamma^{-1}h^{-\alpha_2} + c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}], \tag{A.23}$$

for some constant c_4. Equation (A.23) will be used to compare the order of T_1^{**} with the rest terms T_2–T_5 below.

For T_2,

$$\begin{aligned} E(T_2^2) &\asymp \frac{(1-c)^2}{c^2}\frac{1}{J}E\left(\left[\frac{1}{2\pi}\int \exp\left\{-it\left(y - c\bar{Z}_j\right)\right\}\frac{\phi_{Y|x}(t)\phi_K(ht)}{\phi_Y(t)^c\phi_\epsilon(ct)}dt\right]^2\right) \\ &= \frac{(1-c)^2}{c^2}\frac{1}{Jh^2}E\left(\left[\frac{1}{2\pi}\int_{-1}^{1}\exp\left\{-it\left(\frac{y - c\bar{Z}_j}{h}\right)\right\}\frac{\phi_{Y|x}(t/h)\phi_K(t)}{\phi_Y(t/h)^c\phi_\epsilon(ct/h)}dt\right]^2\right), \end{aligned}$$

which is less than a constant times

$$\frac{1}{Jh^2}\int_0^1 \frac{|\phi_{Y|x}(t/h)\phi_K(t)|}{|\phi_Y(t/h)^c\phi_\epsilon(ct/h)|}dt.$$

We partition $\int_0^1$ into $\int_0^{Mh}$ and $\int_{Mh}^1$, where M is a large number such that when $t > M$, for some positive constant C_{Y1}, C_{Y2}, $C_{\epsilon 1}$ and $C_{\epsilon 2}$,

$$C_{Y1}|t|^{\alpha_1}\exp(-|t|^{\alpha_2}/\gamma) \leq \phi_Y(t) \leq C_{Y2}|t|^{\alpha_1}\exp(-|t|^{\alpha_2}/\gamma)$$

$$C_{\epsilon 1}|t|^{\beta_1}\exp(-|t|^{\beta_2}/\zeta) \le \phi_\epsilon(t) \le C_{\epsilon 2}|t|^{\beta_1}\exp(-|t|^{\beta_2}/\zeta),$$

according to Condition SS. Then we have

$$\int_0^{Mh} \frac{|\phi_{Y|x}(t/h)\phi_K(t)|}{|\phi_Y(t/h)^c\phi_\epsilon(ct/h)|}dt \le \frac{Mh}{\min_{t\in(0,M)}|\phi_Y(t)^c\phi_\epsilon(ct)|} = O(h), \tag{A.24}$$

and

$$\begin{aligned}
&\int_{Mh}^1 \frac{|\phi_{Y|x}(t/h)\phi_K(t)|}{|\phi_Y(t/h)^c\phi_\epsilon(ct/h)|}dt \\
&\le const.\Big[h^{c\alpha_1-\rho_1(x)+\beta_1}\int_{Mh}^1 t^{-c\alpha_1+\rho_1(x)-\beta_1}\times \\
&\quad \exp\{c\gamma^{-1}h^{-\alpha_2}t^{\alpha_2} - \varrho(x)^{-1}h^{-\rho_2(x)}t^{\rho_2(x)} + c^{\beta_2}\zeta^{-1}h^{-\beta_2}t^{\beta_2}\}dt\Big] \\
&= O[h^{const.}\exp\{c\gamma^{-1}h^{-\alpha_2} + c^{\beta_2}\zeta^{-1}h^{-\beta_2} - \varrho(x)^{-1}h^{-\rho_2(x)}\}].
\end{aligned} \tag{A.25}$$

Combining (A.24) and (A.25), we get

$$T_2 = O_p[J^{-1/2}h^{const.}\exp\{c\gamma^{-1}h^{-\alpha_2} + c^{\beta_2}\zeta^{-1}h^{-\beta_2} - \varrho(x)^{-1}h^{-\rho_2(x)}\}], \tag{A.26}$$

For T_3, by hölder inequality,

$$E(|\Delta_1(ct)\Delta_2(ct)|) \le \sqrt{E|\Delta_1(ct)|^2}\sqrt{E|\Delta_2(ct)|^2} \asymp \frac{1}{J\bar{h}^{1/2}\phi_\epsilon(ct)}.$$

Thus,

$$\begin{aligned}
E|T_3| &\le const.\int \frac{E|\Delta_1(ct)\Delta_2(ct)|}{|\phi_Y(t)|^{2c-1}|\phi_\epsilon(ct)|}|\phi_K(ht)|dt \\
&\asymp \frac{1}{J\bar{h}^{1/2}}\int \frac{|\phi_K(ht)|}{|\phi_Y(t)|^{2c-1}|\phi_\epsilon(ct)|^2}dt = \frac{1}{J\bar{h}^{1/2}h}\int \frac{|\phi_K(t)|}{|\phi_Y(t/h)|^{2c-1}|\phi_\epsilon(ct/h)|^2}dt.
\end{aligned}$$

Following the argument similar to Equation (A.24) and (A.25), by partitioning the integral into $\int_0^{Mh}$ and $\int_{Mh}^1$, we have

$$T_3 = O_p[J^{-1}\bar{h}^{-1/2}h^{const.}\exp\{(2c-1)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}]. \tag{A.27}$$

Comparing the order of T_2 in Equation (A.26) and T_3 in Equation (A.27) with the order of T_1^{**} in Equation (A.23), we have when

$$\bar{h}^{-1}h^{c_1}\exp\{2\varrho(x)^{-1}h^{-\rho_2} - 2\gamma^{-1}h^{-\alpha_2}\} \to \infty$$

$$Jh^{c_2}\exp\{-2c\gamma^{-1}h^{-\alpha_2} - 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\} \to \infty,$$

for any c_1, c_2, $T_2 = o_p(T_1^{**})$ and $T_3 = o_p(T_1^{**})$ hold, respectively. For T_4 and T_5, noting that $|\Delta(ct)/\phi_Y(t)^c| < 1/2$ when $t \in \mathcal{B}_J$, $T_4 = O_p(T_2) = o_p(T_1^{**})$ and $T_5 = O_p(T_3) = o_p(T_1^{**})$, respectively.

In summary, we have completed the proof of Theorem 2.2 by showing that

$$\begin{aligned}\widehat{f}_{RP}(y|x) - f(y|x) &= (2\pi)^{-1} \int e^{-ity} \phi_{Y|x}(t) \{\phi_K(ht) - 1\} \, dt + T_1^* + T_1^{**}\{1 + o_p(1)\} \\ &= B_{\bar{h},h}(x,y) + V_{\bar{h},h}^{1/2}(x,y),\end{aligned}$$

where

$$\begin{aligned}B_{\bar{h},h}(x,y) = &\ (2\pi)^{-1} \int e^{-ity} \phi_{Y|x}(t) \{\phi_K(ht) - 1\} \, dt + T_1^* \\ V_{\bar{h},h}(x,y) = &\ T_1^{**2}\{1 + o_p(1)\}.\end{aligned}$$

When $\alpha_2 < \beta_2$, the pattern is the same but the position of α_2 and β_2 is switched in the proof. □

A.2.1 Lemmas for Theorem 2.2

Lemma A.4 presents the rate of term T_1^{**2} in the proof of Theorem 2.2.

Lemma A.4. *Under Conditions SS, (C1)–(C4), and the assumption that $\alpha_2 > \beta_2$, if $\alpha_2 < 1$, the term T_1^{**2} in the proof of Theorem 2.2 equals*

$$\frac{v_1(x) f_{c\bar{Z}|x}(y)}{cJ\bar{h}h} \int K^{*2}(u) du \{1 + o(1)\}.$$

If $\alpha_2 \geq 1$, then

$$T_1^{**2} = O_p[c^{-1-2\beta_1} J^{-1} \bar{h}^{-1} h^{c_3} \exp\{2(c-1)\gamma^{-1} h^{-\alpha_2} + 2c^{\beta_2} \zeta^{-1} h^{-\beta_2}\}],$$

for some constant c_3.

Proof. To calculate the order of T_1^{**}, recalling in the proof of Theorem 2.1, we have shown $T_1^{**} = \sum_{i=1}^{c} W_i(x) T_{1i}^{**}$, where

$$W_i(x) = \sum_{j=1}^{J} w_{ij}(x) / \sum_{i=1}^{c} \sum_{j=1}^{J} w_{ij}(x)$$

$$T_{1i}^{**} = \frac{1}{2\pi}\int \exp(-ity)\left[\frac{\sum_{j=1}^{J} w_{ij}(x)\{e^{ic\bar{Z}_j t} - \phi_{\bar{Z}|X_{ij}}(ct)\}}{\sum_{j=1}^{J} w_{ij}(x)\phi_\epsilon(ct)}\right]\frac{\phi_K(ht)}{\phi_Y(t)^{c-1}}dt.$$

Denote by $\mathcal{X}_i = \{X_{i1}, \ldots, X_{iJ}\}$, for $i = 1, \ldots, c$. We have also shown that, using Lemmas A.1 and A.2 and Parseval's identity, the order of $E(T_{1i}^{**2}|\mathcal{X}_i)$ is the same as the order of

$$\frac{\sum_{j=1}^{J} w_{ij}(x)^2}{\{\sum_{j=1}^{J} w_{ij}(x)\}^2}\int\left\{\frac{1}{2\pi h}\int \exp\left(-it\frac{y-c\bar{Z}_j}{h}\right)\frac{\phi_K(t)}{\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)}dt\right\}^2 f_{\bar{Z}|X_{ij}}(\bar{Z}_j)d\bar{Z}_j,$$

which can be written as

$$\frac{\sum_{j=1}^{J} w_{ij}(x)^2}{\{\sum_{j=1}^{J} w_{ij}(x)\}^2}\frac{1}{h^2}\int K^{*2}(\frac{y-c\bar{Z}_j}{h})f_{\bar{Z}|X_{ij}}(\bar{Z}_j)d\bar{Z}_j, \tag{A.28}$$

where

$$K^*(u) = \frac{1}{2\pi}\int_{-1}^{1} e^{-itu}\frac{\phi_K(t)}{\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)}dt$$

is defined in Equation (2.11).

We discuss the order of (A.28) by considering two cases: Case 1, where $\alpha_2 < 1$, and Case 2, where $\alpha_2 \geq 1$. In Case 1, we derive the order of $\int |K^*(x)|^2 dx$ and $\int |xK^*(x)|^2 dx$ and use them to calculate the accurate order of (A.28). According to Condition (C4), $\phi_K(t)$ is $I_{[-1,1]}(t)$ or a more complex piecewise function. When $\phi_K(t)$ is the piecewise function, by Parseval's identity, for large J,

$$\begin{aligned}\int |K^*(x)|^2 dx &= \frac{1}{\pi}\int_0^1 \frac{|\phi_K(t)|^2}{|\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)|^2}dt \\ &= \frac{1}{\pi}\left\{\int_0^{Mh} + \int_{Mh}^{\tau} + \int_\tau^1\right\}\frac{|\phi_K(t)|^2}{|\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)|^2}dt. \end{aligned}\tag{A.29}$$

The first two terms are less than a constant times

$$h + h^{(2c-2)\alpha_1+2\beta_1}\exp\{(2c-2)\gamma^{-1}(\tau/h)^{\alpha_2} + 2\zeta^{-1}(c\tau/h)^{\beta_2}\}\int_{Mh}^{\tau} t^{-(2c-2)\alpha_1-2\beta_1}dt.$$

For the last term, we can first write it as

$$\frac{1}{\pi c^{2\beta_1}}h^{(2c-2)\alpha_1+2\beta_1}\int_\tau^1 \phi_K(t)^2\exp\{(2c-2)\gamma^{-1}(t/h)^{\alpha_2} + 2\zeta^{-1}(ct/h)^{\beta_2}\}t^{-(2c-2)\alpha_1-2\beta_1}dt. \tag{A.30}$$

Substituting t with $1-h^{\alpha_2}s$, we have

$$\begin{aligned}
&\int_\tau^1 \phi_K(t)^2 \exp\{(2c-2)\gamma^{-1}(t/h)^{\alpha_2} + 2\zeta^{-1}(ct/h)^{\beta_2}\} t^{-(2c-2)\alpha_1-2\beta_1} dt \\
&= h^{\alpha_2} \int_0^{(1-\tau)/h^{\alpha_2}} \Bigg[\phi_K(1-h^{\alpha_2}s)^2 (1-h^{\alpha_2}s)^{-(2c-2)\alpha_1-2\beta_1} \\
&\quad \times \exp\{(2c-2)\gamma^{-1}(1-h^{\alpha_2}s)^{\alpha_2}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}(1-h^{\alpha_2}s)^{\beta_2}h^{-\beta_2}\} \Bigg] ds \\
&= h^{(2b+1)\alpha_2} \exp\{(2c-2)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\} \\
&\quad \times \int_0^{(1-\tau)/h^{\alpha_2}} \Bigg[\frac{\phi_K(1-h^{\alpha_2}s)^2}{(h^{\alpha_2}s)^{2b}} s^{2b} (1-h^{\alpha_2}s)^{-(2c-2)\alpha_1-2\beta_1} \\
&\quad \times \exp\left\{ (2c-2)\gamma^{-1}\frac{(1-h^{\alpha_2}s)^{\alpha_2}-1}{h^{\alpha_2}s} s + 2c^{\beta_2}\zeta^{-1}\frac{(1-h^{\alpha_2}s)^{\beta_2}-1}{h^{\alpha_2}s} h^{\alpha_2-\beta_2} s \right\} \Bigg] ds \\
&= a_0^2 h^{(2b+1)\alpha_2} \exp\{(2c-2)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\} \int_0^\infty s^{2b} \exp\{-(2c-2)\gamma^{-1}\alpha_2 s\} ds.
\end{aligned} \tag{A.31}$$

For the last equation, noting that $dx^{\alpha_2}/dx = \alpha_2 x^{\alpha_2-1}$ and $\phi_K(1-t)/t^b \to a_0$ as $t \to 0$, for large enough J, we have the function inside the integral is bounded by

$$const.|s|^{2b} \exp\{-(c-1)\gamma^{-1}\alpha_2 s\},$$

which is integrable. Then we can use the dominated convergence theorem to guarantee the last equation hold. Plugging Equation (A.31) into (A.30), and noting term (A.30) is the dominant term of (A.29), we have

$$\int |K^*(x)|^2 dx \asymp c^{-2\beta_1} h^{(2b+1)\alpha_2+(2c-2)\alpha_1+2\beta_1} \exp\{(2c-2)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}. \tag{A.32}$$

In addition, similar to (A.15), using Parseval's identity and integration by parts, we have

$$\begin{aligned}
\int |xK^*(x)|^2 dx \leq const. \Bigg\{ &\int_0^1 \left| \frac{\phi_K'(t)}{\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)} \right|^2 dt \\
&+ \frac{1}{h^2}\int_0^1 \left| \frac{\phi_K(t)\phi_Y'(t/h)}{\phi_Y(t/h)^{c}\phi_\epsilon(ct/h)} \right|^2 dt \\
&+ \frac{1}{h^2}\int_0^1 \left| \frac{\phi_K(t)\phi_\epsilon'(ct/h)}{\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)^2} \right|^2 dt \Bigg\}.
\end{aligned} \tag{A.33}$$

Partitioning the integral to $\int_0^{Mh} + \int_{Mh}^{\tau} + \int_{\tau}^{1}$, we can see all three terms in (A.33) are bounded by

$$h^{(2b-1)\alpha_2+(2c-2)\alpha_1+2\beta_1} \exp\{(2c-2)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}.$$

That is,

$$\int |xK^*(x)|^2 dx = O[h^{(2b-1)\alpha_2+(2c-2)\alpha_1+2\beta_1} \exp\{(2c-2)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}]. \tag{A.34}$$

When $\phi_K(t) = I_{[-1,1]}(t)$, following the similar argument as above,

$$\int |K^*(x)|^2 dx = O[c^{-2\beta_1} h^{\alpha_2+(2c-2)\alpha_1+2\beta_1} \exp\{(2c-2)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}], \tag{A.35}$$

and

$$\int |xK^*(x)|^2 dx = O\left[h^{-\alpha_2+(2c-2)\alpha_1+2\beta_1} \exp\{(2c-2)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}\right]. \tag{A.36}$$

Note that Equations (A.35) and (A.36) are compatible with Equations (A.32) and (A.34) by setting $b^* = b$ when $\phi_K(t)$ is the piecewise function and $b^* = 0$ when $\phi_K(t) = I_{[-1,1]}(t)$. Combining Equations (A.32) and (A.34), and Equations (A.35) and (A.36), we can see that no matter $\phi_K(t)$ is the piecewise function or $I_{[-1,1]}(t)$, (A.28) can be written as

$$\begin{aligned}
&\frac{\sum_{j=1}^J w_{ij}(x)^2}{\{\sum_{j=1}^J w_{ij}(x)\}^2} \frac{1}{h^2} \int K^{*2}\left(\frac{y - c\bar{Z}_j}{h}\right) f_{\bar{Z}|X_{ij}}(\bar{Z}) d\bar{Z} \\
&= \frac{\sum_{j=1}^J w_{ij}(x)^2}{\{\sum_{j=1}^J w_{ij}(x)\}^2} \frac{1}{h} \int K^{*2}(u) f_{c\bar{Z}|X_{ij}}(y - hu) du \\
&= \frac{\sum_{j=1}^J w_{ij}(x)^2}{\{\sum_{j=1}^J w_{ij}(x)\}^2} \frac{1}{h} \left\{ f_{c\bar{Z}|X_{ij}}(y) \int K^{*2}(u) du + h^2 \int u^2 K^{*2}(u) f''_{c\bar{Z}|X_{ij}}(y - \theta_y hu) du \right\} \\
&= \frac{v_1(x)}{J\bar{h}h} f_{c\bar{Z}|x}(y) \int K^{*2}(u) du \{1 + o(1)\} \qquad \text{(A.37)} \\
&= \frac{v_1(x)}{\pi c^{2\beta_1} J\bar{h}} a_0^2 f_{c\bar{Z}|x}(y)\ h^{(2b^*+1)\alpha_2+(2c-2)\alpha_1+2\beta_1-1} \exp\{(2c-2)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\} \\
&\quad \times \int_0^1 s^{2b} \exp\{-(2c-2)\gamma^{-1}\alpha_2 s\} ds \{1 + o(1)\}, \qquad \text{(A.38)}
\end{aligned}$$

which is of the same order as

$$c^{-2\beta_1} J^{-1} \bar{h}^{-1} h^{(2b^*+1)\alpha_2+(2c-2)\alpha_1+2\beta_1-1} \exp\{(2c-2)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}.$$

Combining the definition of $W_i(x)$, the order of T_1^{**2} equals

$$O_p[c^{-1-2\beta_1} J^{-1} \bar{h}^{-1} h^{(2b^*+1)\alpha_2+(2c-2)\alpha_1+2\beta_1-1} \exp\{(2c-2)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}]. \tag{A.39}$$

Note that T_{1i}^{**2} can be written as (A.37). Thus, the T_1^{**2} can be further written as

$$\frac{v_1(x) f_{c\bar{Z}|x}(y)}{cJ\bar{h}h} \int K^{*2}(u)du\{1+o(1)\}.$$

In Case 2, where $\alpha_2 \geq 1$, it's hard to calculate the exact rate of T_1^{**}. Instead we calculate the upper bound. For large enough J, $\sup_y |K^*(y)|$ is less than

$$\begin{aligned}
&const. \int_0^1 \frac{|\phi_K(t)|}{|\phi_Y(t/h)^{c-1}\phi_\epsilon(ct/h)|}dt \\
&\leq const. \left[h + h^{(c-1)\alpha_1+\beta_1} \int_{Mh}^1 t^{-(c-1)\alpha_1-\beta_1} \exp\{(c-1)\gamma^{-1}h^{-\alpha_2}t^{\alpha_2} + c^{\beta_2}\zeta^{-1}h^{-\beta_2}t^{\beta_2}\}dt\right] \\
&= O[h^{const.} c^{-\beta_1} \exp\{(c-1)\gamma^{-1}h^{-\alpha_2} + c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}].
\end{aligned} \tag{A.40}$$

Thus, the order of T_1^{**2} equals

$$O_p[J^{-1}c^{-1-2\beta_1}\bar{h}^{-1}h^{c_3} \exp\{2(c-1)\gamma^{-1}h^{-\alpha_2} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}], \tag{A.41}$$

for some constant c_3. Note that no matter in Case 1 or Case 2, by combining (A.39) and (A.41), T_1^{**} can be written as

$$T_1^{**} = O_p[J^{-1/2}\bar{h}^{-1/2}h^{c_4} \exp\{(c-1)\gamma^{-1}h^{-\alpha_2} + c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}], \tag{A.42}$$

for some constant c_4. □

Appendix B

Proof for Chapter 3

In this chapter, we present the conditions and proofs for Theorems 3.1 and 3.2. To keep notation concise, we use X_{ij}, $\bar{X}_j$, and $\bar{Y}_j$ to denote $X_{(ij)}$, $\bar{X}_{(j)}$, and $\bar{Y}_{(j)}$ throughout this chapter. Letting $q(x,t) = \phi_{Y|X=x}(t)$, we also denote $\partial q(x,t)/\partial x$ as $q'(x,t)$, $\partial^2 q(x,t)/\partial x^2$ as $q''(x,t)$, and $\partial^3 q(x,t)/\partial x^3$ as $q'''(x,t)$. Lastly, the notation $f(n) \asymp g(n)$ means there exist $M_1, M_2 > 0$ such that $M_1 g(n) \leq f(n) \leq M_2 g(n)$.

B.1 Conditions

In this section, we list the conditions that are used in the proof of Chapter 3. Condition H is similar to the conditions for local polynomial estimators in Delaigle and Meister (2012): For a given x and fixed c, when $J \to \infty$,

1. $\sum_{j=1}^{J} w_j(x)(\bar{X}_j - x)/\sum_{j=1}^{J} w_j(x) = 0$
2. $\sum_{j=1}^{J} w_j(x)(\bar{X}_j - x)^2/\{\sum_{j=1}^{J} w_j(x)\}^2 = b_2(x)\bar{h}^2 + o_p(\bar{h}^2)$
3. $\sum_{j=1}^{J} w_j(x)^2/\{\sum_{j=1}^{J} w_j(x)\}^2 = v_2(x)/(J\bar{h}) + O_p\{1/(J\bar{h})\}$
4. $\sum_{j_1 \neq j_2} w_{j_1}(x) w_{j_2}(x)/\{\sum_{j=1}^{J} w_j(x)\}^2 = o_p(1)$
5. $\sum_{j=1}^{J} w_j(x)^k/\{\sum_{j=1}^{J} w_j(x)\}^k = O_p\{1/(J\bar{h})^{k-1}\}$
6. $w_j(x) = 0$ when $|\bar{X}_j - x| > C\bar{h}$, where C is a constant.

The quantities $b_2(x)$, $v_2(x)$ are continuous functions and are related to $w_j(x)$. The weight function $w_j(x)$ is the generalized weight and is related to $\bar{K}_{\bar{h}}(\cdot)$, $\bar{X}_j$, and

the type of estimator. Note that condition (i) does not hold for the local constant estimator. Like in the Theorems 2.1 and 2.2, we only focus on the proof of the local linear estimator. The proof for local constant estimator is omitted.

Condition TO describes the tail behavior of $q(x,t)$ and its derivatives in the ordinary smooth case: When $q(x,t)$ is ordinary smooth, as $t \to \infty$,

1. $|q'(x,t)| \asymp t^{-\rho_0(x)} \log(t)$
2. $|q''(x,t)| \asymp t^{-\rho_0(x)} \log(t)^2$
3. $|q'''(x,t)| \asymp t^{-\rho_0(x)} \log(t)^3$.

Condition TS describes the tail behavior of $q(x,t)$ and its derivatives in the super smooth case: When $q(x,t)$ is super smooth, as $t \to \infty$,

1. $|q'(x,t)| \asymp \exp\{-|t|^{\rho_2(x)}/\varrho(x)\}|t|^{\rho_1(x)+\rho_2(x)} \log(t)$
2. $|q''(x,t)| \asymp \exp\{-|t|^{\rho_2(x)}/\varrho(x)\}|t|^{\rho_1(x)+2\rho_2(x)} \log(t)^2$
3. $|q'''(x,t)| \asymp \exp\{-|t|^{\rho_2(x)}/\varrho(x)\}|t|^{\rho_1(x)+3\rho_2(x)} \log(t)^3$.

B.2 Proof of Theorem 3.1

Proof of Theorem 3.1. The homogeneous estimator $\widehat{f}_{HP}(y)$ can be written as

$$\widehat{f}_{HP}(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ity} \widehat{U}(ct)^{1/c} \phi_K(ht) dt,$$

where

$$\widehat{U}(ct) = \frac{\sum_{j=1}^{J} w_j(x) \exp(ict\bar{Z}_j)}{\sum_{j=1}^{J} w_j(x) \phi_\epsilon(ct)}.$$

To obtain the asymptotic property of $\widehat{f}_{HP}(y|x) - f(y|x)$ under Condition OO, we start with the following decomposition,

$$\widehat{f}_{HP}(y|x) - f(y|x) = \frac{1}{2\pi} \int e^{-ity} \left\{ \widehat{U}(ct)^{1/c} - q(x,t) \right\} \phi_K(ht) dt \tag{B.1}$$

$$+ \frac{1}{2\pi} \int e^{-ity} q(x,t) \left\{ \phi_K(ht) - 1 \right\} dt. \tag{B.2}$$

The term (B.2) is exactly the last summand in $B_{\bar{h},h}(x,y)$ in Theorem 3.1. It suffices to check (B.1), which can be written as

$$\frac{1}{2\pi}\int e^{-ity}\left[\{\tilde{U}(ct)+\Delta_3(ct)\}^{1/c}-q(x,t)\right]\phi_K(ht)dt,$$

where

$$\tilde{U}(ct)=E[\widehat{U}(ct)|\mathcal{X}]=\frac{\sum_{j=1}^J w_j(x)E[\exp(ict\bar{Y}_j)|\mathcal{X}]}{\sum_{j=1}^J w_j(x)}, \tag{B.3}$$

and $\Delta_3(ct)=\widehat{U}(ct)-\tilde{U}(ct)$. We want to expand $\widehat{U}(ct)^{1/c}=\{\tilde{U}(ct)+\Delta_3(ct)\}^{1/c}$ to power series at $\tilde{U}(ct)$. To achieve this goal, recalling that $\phi_K(ht)\neq 0$ only on $\mathcal{B}_J=(-1/h,1/h)$, we need (i) $|\Delta_3(ct)/\tilde{U}(ct)|<1/2$ on $\mathcal{B}_J$, and (ii) $\widehat{U}(ct)$ dost not vanish on $\mathcal{B}_J$. To prove (i), first we have

$$E\{|\Delta_3(ct)|^2|\mathcal{X}\}\leq\frac{\sum_{j=1}^J w_j(x)^2}{|\phi_\epsilon(ct)|^2\{\sum_{j=1}^J w_j(x)\}^2}=O_p\left\{\frac{1}{J\bar{h}|\phi_\epsilon(ct)|^2}\right\}.$$

We show in Lemma B.1 that $\tilde{U}(ct)=q(x,t)^c\{1+o_p(1)\}$ for a fixed t. Noting that the order of $q(x,t)$ is $t^{-\rho_0(x)}$ under Condition OO, we have

$$E\left\{\frac{|\Delta_3(ct)|^2}{|\tilde{U}(ct)|^2}\middle|\mathcal{X}\right\}=O_p\left\{\frac{1}{J\bar{h}h^{2c\rho_0(x)+2\beta_0}}\right\}.$$

Thus, $|\Delta_3(ct)|^2/|\tilde{U}(ct)|=o_p(1)$ as $J\bar{h}h^{2c\rho_0(x)+2\beta_0}\to\infty$. That is, $|\Delta_3(ct)/\tilde{U}(ct)|<1/2$ for all $t\in\mathcal{B}_J$ uniformly with probability 1. To prove (ii), we can see $|\widehat{U}(ct)|>\inf_{t\in\mathcal{B}_J}|\tilde{U}(ct)|/2>0$ as $t\in\mathcal{B}_J$ and thus $\widehat{U}(ct)$ does not vanish on $\mathcal{B}_J$. Therefore, we have when $t\in\mathcal{B}_J$,

$$\begin{aligned}\widehat{U}(ct)^{1/c}&=\{\tilde{U}(ct)+\Delta_3(ct)\}^{1/c}\\&=\tilde{U}(ct)^{1/c}+\frac{1}{c}\tilde{U}(ct)^{1/c-1}\Delta_3(ct)+\tilde{\lambda}_2^{**}(t)\tilde{U}(ct)^{1/c-2}\Delta_3(ct)^2,\end{aligned} \tag{B.4}$$

where $|\tilde{\lambda}_2^{**}(t)|\leq 4$. Plugging Equation (B.14) in Lemma B.1 into (B.4), we have

$$\begin{aligned}\widehat{U}(ct)^{1/c}-q(x,t)=&\frac{1}{c}q(x,t)^{1-c}R_{\tilde{U}}(t)+\frac{1}{c}\tilde{U}(ct)^{1/c-1}\Delta_3(ct)+\tilde{\lambda}_2^{*}(t)q(x,t)^{1-2c}R_{\tilde{U}}(t)^2\\&+\tilde{\lambda}_2^{**}(t)\tilde{U}(ct)^{1/c-2}\Delta_3(ct)^2,\end{aligned} \tag{B.5}$$

where $R_{\tilde{U}}(t)$ and $\tilde{\lambda}_2^*(t)$ are defined in Equations (B.20) and (B.14) in Lemma B.1. Plugging Equation (B.5) into (B.1), we get 4 terms denoted by $T_1, \ldots, T_4$. We will show that term T_1, together with (B.2), makes up the bias for $\widehat{f}_{HP}(y|x)$; the square of term T_2 makes up the variance; terms T_3 and T_4 are negligible to T_1 and T_2, respectively. We list them below and evaluate them one by one. Term T_1 is

$$T_1 = \frac{1}{2\pi c}\int_{-\infty}^{\infty} e^{-ity} q(x,t)^{1-c} R_{\tilde{U}}(t)\phi_K(ht)dt.$$

Plugging $R_{\tilde{U}}(t)$ in Equation (B.20), we have

$$T_1 = \bar{h}^2 \frac{b_2(x)}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\left\{\frac{1}{2}q''(x,t) + \frac{c-1}{2}\frac{q'(x,t)^2}{q(x,t)}\right\}\phi_K(ht)dt \tag{B.6}$$

$$+ O_p(J^{2\epsilon-2})\frac{1}{2\pi c}\int_{-\infty}^{\infty} e^{-ity} q(x,t)^{1-c}\mathcal{M}^*(t)\log(t)^2\phi_K(ht)dt \tag{B.7}$$

$$+ O_p(\bar{h}^3)\frac{1}{2\pi c}\int_{-\infty}^{\infty} e^{-ity} q(x,t)^{1-c}\mathcal{M}^{**}(t)\log(t)^3\phi_K(ht)dt. \tag{B.8}$$

Note that to obtain the accurate order of T_1, we need to know the order of the integral in term (B.6). Under Condition TO, there exists a large $M > 0$ such that when $t > M$,

$$\left|\frac{1}{2}q''(x,t) + \frac{c-1}{2}\frac{q'(x,t)^2}{q(x,t)}\right| \leq const.t^{-\rho_0(x)}\log(t)^2.$$

Thus,

$$\begin{aligned}
&\left|\int_{-\infty}^{\infty}\exp\left(-ity\right)\left\{\frac{1}{2}q''(x,t) + \frac{c-1}{2}\frac{q'(x,t)^2}{q(x,t)}\right\}\phi_K(ht)dt\right| \\
&= \left|\frac{1}{h}\int_{-1}^{1}\exp(-\frac{ity}{h})\left\{\frac{1}{2}q''(x,t/h) + \frac{c-1}{2}\frac{q'(x,t/h)^2}{q(x,t/h)}\right\}\phi_K(t)dt\right| \\
&\leq const.\left(\int_0^{Mh} + \int_{Mh}^{1}\right)\left|\frac{1}{h}\left\{\frac{1}{2}q''(x,t/h) + \frac{c-1}{2}\frac{q'(x,t/h)^2}{q(x,t/h)}\right\}\right|dt \\
&= O(1) + O\{h^{\rho_0(x)-1}\log(h)^2\}.
\end{aligned}$$

When $\rho_0(x) > 1$, we have the integral in term (B.6) is bounded. Thus, the order of term (B.6) is $\bar{h}^2$. Following the similar argument, we can see terms (B.7) and (B.8)

are $O(J^{2\epsilon-2})$ and $O(\bar{h}^3)$, respectively. Note that when $J^{1-\epsilon}\bar{h} \to \infty$, $J^{2-2\epsilon} = o_p(\bar{h}^2)$. Thus,

$$T_1 = \bar{h}^2 \frac{b_2(x)}{2\pi} \int_{-\infty}^{\infty} e^{-ity} \left\{ \frac{1}{2} q''(x,t) + \frac{c-1}{2} \frac{q'(x,t)^2}{q(x,t)} \right\} \phi_K(ht) dt + o_p(\bar{h}^2). \tag{B.9}$$

Term T_2 is

$$T_2 = \frac{1}{2\pi c} \int_{-\infty}^{\infty} e^{-ity} \tilde{U}(ct)^{1/c-1} \Delta_3(ct) \phi_K(ht) dt.$$

Similar to Lemma B.1, when $t \in \mathcal{B}_J$, we can expand $\tilde{U}(ct)^{1/c-1}$ into power series

$$\tilde{U}(ct)^{1/c-1} = q(x,t)^{1-c} + \tilde{\lambda}_1^*(t) q(x,t)^{1-2c} R_{\tilde{U}}(t),$$

where $|\tilde{\lambda}_1^*(t)| \leq 4$. Plugging this into T_2, we have

$$T_2 = \frac{1}{2\pi c} \int_{-\infty}^{\infty} e^{-ity} q(x,t)^{1-c} \Delta_3(ct) \phi_K(ht) dt \tag{B.10}$$

$$+ \frac{1}{2\pi c} \int_{-\infty}^{\infty} e^{-ity} \tilde{\lambda}_1^*(t) q(x,t)^{1-2c} R_{\tilde{U}}(t) \Delta_3(ct) \phi_K(ht) dt = T_2^* + T_2^{**}. \tag{B.11}$$

For T_2^*, letting $\bar{\mathcal{X}} = \{\bar{X}_1, \bar{X}_2, \ldots, \bar{X}_J\}$, we have $E[T_2^{*2} | \bar{\mathcal{X}}]$ is of the same order as

$$\frac{\sum_{j=1}^J w_j(x)^2}{\{\sum_{j=1}^J w_j(x)\}^2} \frac{1}{c^2 h^2} E\left\{ \left[\frac{1}{2\pi} \int \exp\left(-it \frac{y - c\bar{Z}_j}{h} \right) \frac{\phi_K(t)}{q(x, t/h)^{c-1} \phi_\epsilon(ct/h)} dt \right]^2 \middle| \bar{\mathcal{X}} \right\},$$

which equals

$$\frac{c^{2\beta_0 - 2} v_2(x) \bar{f}_{c\bar{Z}|\bar{X}=x}(y)}{2\pi J \bar{h} h^{2(c-1)\rho_0(x) + 2\beta_0 + 1}} \int \frac{|\phi_K(t)|^2 t^{2(c-1)\rho_0(x) + 2\beta_0}}{A_3(x)^{2(c-1)} A_2^2} dt \{1 + o_p(1)\},$$

where $\bar{f}_{c\bar{Z}|\bar{X}=x}(y) = \lim_{J\to\infty} \sum_1^J f_{c\bar{Z}_j|\bar{X}_j=x}(y)/J$. Then we have

$$T_2^* = O_p\{c^{-1+\beta_0} J^{-1/2} \bar{h}^{-1/2} h^{-(c-1)\rho_0(x) - \beta_0 - 1/2}\}.$$

Following the similar argument as in T_2^*, it can be shown that

$$E[T_2^{**2} | \bar{\mathcal{X}}] = O_p\{(\bar{h}^4 + 1/J^{4-4\epsilon}) \log(h)^4 J^{-1} \bar{h}^{-1} h^{-2(c-1)\rho_0(x) - 2\beta_0 - 1}\}.$$

Thus, when $J^{1-\epsilon}\bar{h} \to \infty$ and $\bar{h}\log(h) \to 0$, $T_2^{**} = o_p(T_2^*)$. Combing the order of T_2^* and T_2^{**}, we have

$$T_2 = O_p\{c^{-1+\beta_0} J^{-1/2} \bar{h}^{-1/2} h^{-(c-1)\rho_0(x) - \beta_0 - 1/2}\}.$$

Term T_3 is

$$T_3 = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\tilde{\lambda}_2^*(t)\frac{R_{\tilde{U}}(t)^2}{q(x,t)^{2c-1}}\phi_K(ht)dt \tag{B.12}$$
$$= \int_{-\infty}^{\infty}\Bigg[\left\{\frac{c}{2}q''(x,t)q(x,t)^{c-1} + \frac{c(c-1)}{2}q'(x,t)^2q(x,t)^{c-2}\right\}b_2(x)\bar{h}^2$$
$$+ \mathcal{M}^{**}(t)\log(t)^3O_p(\bar{h}^3) + \mathcal{M}^*(t)\log(t)^2O_p(J^{2\epsilon-2})\Bigg]^2\frac{\phi_K(ht)}{q(x,t)^{2c-1}}dt,$$

where $\mathcal{M}^*$ and $\mathcal{M}^{**}$ are defined in Equations (B.16) and (B.17) in Lemma B.1. Note that $|\tilde{\lambda}_2^*(t)| \leq 4$ and order of the term inside integral is $t^{-\rho_0(x)}\log(t)^4$ with respect to t. Following the similar argument as in T_1, we can show that when $\rho_0(x) > 1$, if $J^{1-\epsilon}\bar{h} \to \infty$ and $\bar{h}log(h) \to 0$, $T_3 = o(T_1)$.

Term T_4 is

$$T_4 = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\tilde{\lambda}_2^{**}(t)\tilde{U}(ct)^{1/c-2}\Delta_3(ct)^2\phi_K(ht)dt.$$

Similar to Lemma B.1, when $t \in \mathcal{B}_J$, we can expand $\tilde{U}(ct)^{1/c-2}$ to power series

$$\tilde{U}(ct)^{1/c-2} = q(x,t)^{1-2c} + \tilde{\lambda}_1^{**}(t)q(x,t)^{1-3c}R_{\tilde{U}}(t), \tag{B.13}$$

where $|\tilde{\lambda}_1^{**}(t)| \leq 4$. Then

$$T_4 = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\tilde{\lambda}_2^{**}(t)q(x,t)^{1-2c}\Delta_3^2(ct)\phi_K(ht)dt$$
$$+ \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-ity}\tilde{\lambda}_2^{**}(t)\tilde{\lambda}_1^{**}(t)q(x,t)^{1-3c}R_{\tilde{U}}(t)\Delta_3^2(ct)\phi_K(ht)dt = T_4^* + T_4^{**}.$$

To evaluate T_4^*, noting that $|\tilde{\lambda}_2^{**}(t)| < 4$, we have $E(|T_4^*|^2|\mathcal{X})$ is less than or equal to

$$16E\left\{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-iy(u-v)}\frac{\Delta_3(cu)^2}{q(x,u)^{2c-1}}\frac{\Delta_3(-cv)^2}{q(x,-v)^{2c-1}}\phi_K(hu)\phi_K(-hv)dudv\middle|\mathcal{X}\right\},$$

which is of the same order as

$$\frac{\sum_{j_1\neq j_2} w_{j_1}(x)^2w_{j_2}(x)^2}{\{\sum w_j(x)\}^4}\int\int\frac{|\phi_K(hu)||\phi_K(hv)||\phi_{Y|x}^{2c}(u-v)||\phi_\epsilon^2(cu-cv)|}{|\phi_{Y|x}(u)^{2c-1}||\phi_{Y|x}(-v)^{2c-1}||\phi_\epsilon(cu)^2||\phi_\epsilon(-cv)^2|}dudv.$$

Following the similar argument as in Lemma A.3, as $(2c-1)\rho_0(x) + 2\beta_0 > 1$,

$$T_4^* = O_p\{J^{-1}\bar{h}^{-1}h^{-(2c-1)\rho_0(x)-2\beta_0-1/2}\}.$$

For T_4^{**}, note that $|R_{\tilde{U}}(t)/q(x,t)^c| < 1/2$ in probability when $t \in \mathcal{B}_J$. Thus, $T_4^{**} = O_p(T_4^*)$. Combining the order of T_4^* and T_4^{**}, we have

$$T_4 = O_p\{J^{-1}\bar{h}^{-1}h^{-(2c-1)\rho_0(x)-2\beta_0-1/2}\}.$$

When $J\bar{h}h^{2c\rho_0(x)+2\beta_0} \to \infty$, $T_4 = o_p(T_2)$.

Combining the order of T_1, T_2^*, T_2^{**}, T_3, and T_4, together with term (B.2), we finish to show that

$$\begin{aligned}\widehat{f}_{HP}(y|x) - f(y|x) &= \frac{1}{2\pi}\int e^{-ity}q(x,t)\left\{\phi_K(ht)-1\right\}dt + T_1 + T_2^*\{1+o_p(1)\} \\ &= B_{\bar{h},h}(x,y) + V_{\bar{h},h}^{1/2}(x,y),\end{aligned}$$

where

$$\begin{aligned}B_{\bar{h},h}(x,y) &= \frac{1}{2\pi}\int \exp(-ity)q(x,t)\{\phi_K(ht)-1\}dt + T_1 \\ V_{\bar{h},h}(x,y) &= T_2^{*2}\{1+o_p(1)\}.\end{aligned}$$

□

B.2.1 Lemmas for Theorem 3.1

Lemma B.1 shows the expression of $\tilde{U}(ct)$ and $\tilde{U}(ct)^{1/c}$ which are used in the proof of Theorem 3.1.

Lemma B.1. *Under Conditions OO, H, TO, and (C1)–(C3), for a fixed t, $\tilde{U}(ct) = q(x,t)^c + R_{\tilde{U}}(ct)$ where $R_{\tilde{U}}(ct) = o_p(1)$ with respect to J. Furthermore, when $t \in \mathcal{B}_J$, if $J^{1-\epsilon}\bar{h} \to \infty$ and $\bar{h}log(h) \to 0$,*

$$\tilde{U}(ct)^{1/c} = q(x,t) + \frac{1}{c}q(x,t)^{1-c}R_{\tilde{U}}(t) + \tilde{\lambda}_2^*(t)q(x,t)^{1-2c}R_{\tilde{U}}(t)^2, \tag{B.14}$$

where $|\tilde{\lambda}_2^*(t)| \le 4$.

Proof. For $\tilde{U}(ct)$, first we look into the conditional expectation $E[\exp(ict\bar{Y}_j)|\mathcal{X}]$ inside $\tilde{U}(ct)$. Let D_j denote the maximum of $|X_{ij} - \bar{X}_j|$ over $i = 1, 2, \ldots, c$. According to

Delaigle and Meister (2012), for each $\epsilon > 0$, $D_j = O_p(c/N^{1-\epsilon})$ uniformly in j such that $|\bar{X}_j - x| \leq \bar{C}\bar{h}$ and $x \in I$, where $\bar{C} > 0$ and I is a given compact, non-degenerate interval. Note that

$$E\{\exp(ict\bar{Y}_j)|\mathcal{X}\} = \prod_{i=1}^{c} q(X_{ij}, t).$$

For any given t, Taylor expand $q(X_{ij}, t)$ on $\bar{X}_j$, we have

$$\begin{aligned} q(X_{ij}, t) = q(\bar{X}_j, t) + q'(\bar{X}_j, t)(X_{ij} - \bar{X}_j) \\ + [Re\{q''(\varepsilon^*_{ij}, t)\} + iIm\{q''(\eta^*_{ij}, t)\}](X_{ij} - \bar{X}_j)^2, \end{aligned} \tag{B.15}$$

where ε^*_{ij} and η^*_{ij} are between X_{ij} and $\bar{X}_j$. Then

$$\begin{aligned} \prod_{i=1}^{c} q(X_{ij}, t) = q(\bar{X}_j, t)^c + c(c-1)q(\bar{X}_j, t)^{c-2}q'(\bar{X}_j, t)^2(X_{ij} - \bar{X}_j)^2/2 + \\ cq(\bar{X}_j, t)^{c-1}\sum_{i=1}^{c}[Re\{q''(\varepsilon^*_{ij}, t)\} + iIm\{q''(\eta^*_{ij}, t)\}](X_{ij} - \bar{X}_j)^2\{1 + o(1)\}. \end{aligned}$$

Noting that $|X_{ij} - \bar{X}_j| \leq D_j$, under Condition TO, we write the equation above as

$$\prod_{i=1}^{c} q(X_{ij}, t) = q(\bar{X}_j, t)^c + \mathcal{M}^*(t)\log(t)^2 O_p(D_j^2), \tag{B.16}$$

where $\mathcal{M}^*(t)$ is of order $t^{-c\rho_0(x)}$ with respect to t. Further Taylor expanding $q(\bar{X}_j, t)$ on x, for any given t, we have

$$\begin{aligned} q(\bar{X}_j, t) = q(x, t) + q'(x, t)(\bar{X}_j - x) + q''(x, t)(\bar{X}_j - x)^2/2 \\ + [Re\{q'''(\varepsilon^{**}_{ij}, t)\} + iIm\{q'''(\eta^{**}_{ij}, t)\}](X_{ij} - \bar{X}_j)^3/6, \end{aligned}$$

where ε^{**}_{ij} and η^{**}_{ij} are between $\bar{X}_j$ and x. Then

$$\begin{aligned} q(\bar{X}_j, t)^c = q(x, t)^c + cq(x, t)^{c-1}q'(x, t)(\bar{X}_j - x) + \frac{c}{2}q''(x, t)q(x, t)^{c-1}(\bar{X}_j - x)^2 \\ + \frac{c(c-1)}{2}q'(x, t)^2 q(x, t)^{c-2}(\bar{X}_j - x)^2 + \mathcal{M}^{**}(t)\log(t)^3 O_p(|\bar{X}_j - x|^3), \end{aligned} \tag{B.17}$$

where $\mathcal{M}^{**}(t)$ is of order $t^{-c\rho_0(x)}$ with respect to t, according to Condition TO. Plugging (B.17) into (B.16), we have

$$\begin{aligned}\prod_{j=1}^{c} q(X_{ij},t) = q(x,t)^c &+ cq(x,t)^{c-1}q'(x,t)(\bar{X}_j - x) \\ &+ \frac{c}{2}q''(x,t)q(x,t)^{c-1}(\bar{X}_j - x)^2 \\ &+ \frac{c(c-1)}{2}q'(x,t)^2 q(x,t)^{c-2}(\bar{X}_j - x)^2 \\ &+ \mathcal{M}^{**}(t)\log(t)^3 O_p(|\bar{X}_j - x|^3) + \mathcal{M}^{*}(t)log(t)^2 O_p(D_j^2). \end{aligned} \tag{B.18}$$

Noting that $E[\exp(ict\bar{Y}_j)|\mathcal{X}] = \prod_{j=1}^{c} q(X_{ij},t)$, we plug Equation (B.18) into Equation (B.3) and have

$$\begin{aligned}\tilde{U}(ct) = q(x,t)^c &+ cq(x,t)^{c-1}q'(x,t)\frac{\sum_{j=1}^{J} w_j(x)(\bar{X}_j - x)}{\sum_{j=1}^{J} w_j(x)} \\ &+ \left\{\frac{c}{2}q''(x,t)q(x,t)^{c-1} + \frac{c(c-1)}{2}q'(x,t)^2 q(x,t)^{c-2}\right\}\frac{\sum_{j=1}^{J} w_j(x)(\bar{X}_j - x)^2}{\sum_{j=1}^{J} w_j(x)} \\ &+ \frac{\sum_{j=1}^{J} w_j(x)[\mathcal{M}^{**}(t)\log(t)^3 O_p(|\bar{X}_j - x|^3) + \mathcal{M}^{*}(t)\log(t)^2 O_p(D_j^2)]}{\sum_{j=1}^{J} w_j(x)}. \end{aligned} \tag{B.19}$$

According to Condition H, Equation (B.19) can be written as

$$\begin{aligned}\tilde{U}(ct) &= q(x,t)^c + \left\{\frac{c}{2}q''(x,t)q(x,t)^{c-1} + \frac{c(c-1)}{2}q'(x,t)^2 q(x,t)^{c-2}\right\} b_2(x)\bar{h}^2 \\ &\quad + \mathcal{M}^{**}(t)\log(t)^3 O_p(\bar{h}^3) + \mathcal{M}^{*}(t)\log(t)^2 O_p(J^{2\epsilon-2}) \\ &= q(x,t)^c + R_{\tilde{U}}(t), \end{aligned} \tag{B.20}$$

where $R_{\tilde{U}}(t)$ is the rest part in the right-hand side of the first equation except $q(x,t)^c$. We can see for any fixed t, $R_{\tilde{U}}(t) = O_p(\bar{h}^2 + J^{2\epsilon-2})$. Further, under Condition TO, letting $t = 1/h$, we can see $|R_{\tilde{U}}(1/h)/q(x,1/h)^c|$ is of the same order as $(\bar{h}^2 + 1/J^{2-2\epsilon})\log(1/h)^2$ with respect to J. Thus, when $t \in \mathcal{B}_J$, if $J^{1-\epsilon}\bar{h} \to \infty$ and $\bar{h}\log(h) \to 0$, we have $|R_{\tilde{U}}(t)/q(x,t)^c| < 1/2$ uniformly with probability 1. Therefore, when $t \in \mathcal{B}_J$, we can further expand $\tilde{U}(ct)^{1/c}$ to be

$$\tilde{U}(ct)^{1/c} = q(x,t) + \frac{1}{c}q(x,t)^{1-c}R_{\tilde{U}}(t) + \tilde{\lambda}_2^*(t)q(x,t)^{1-2c}R_{\tilde{U}}(t)^2, \tag{B.21}$$

where $|\tilde{\lambda}_2^*(t)| \leq 4$. □

B.3 Proof of Theorem 3.2

Proof of Theorem 3.2. The proof of Theorem 3.2 is similar to that of Theorem 3.1, where we rely on Equation (B.5) to decompose $\widehat{f}_{HP}(y|x) - f(y|x)$ into 4 terms. The technical argument to deal with CFs of super smooth distributions is similar to that of Theorem 2.2. We only show the key steps. There are two assumptions that are frequently used throughout the proof. Assumption 1 is

$$J\bar{h}h^{d_1} \exp\{-2c\varrho(x)^{-1}h^{-\rho_2(x)} - 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\} \to \infty,$$

for any d_1. Assumption 2 is $J^{1-\epsilon}\bar{h} \to \infty$ and $\bar{h}h^{d_2} \to 0$, for any d_2.

First, it can be shown that when $t \in \mathcal{B}_J$ and Assumption 1 holds, we can establish the expansion of power series of $\widehat{U}(ct)^{1/c}$ to obtain Equation (B.5). Then we plug Equation (B.5) into (B.1) and obtain 4 terms, $T_1, \ldots, T_4$. Same as in Theorem 3.1, term T_1, together with (B.2), makes up the bias for $\widehat{f}_{HP}(y|x)$; the square of term T_2 makes up the variance; terms T_3 and T_4 are negligible to T_1 and T_2, respectively.

For term T_1, under Condition TS, we can show the integrals in the terms (B.6), (B.7), and (B.8) are bounded. Thus, when $J^{1-\epsilon}h \to \infty$, we have T_1 the same as Equation (B.9). For term T_2, similar to Equation (B.10), we decompose $T_2 = T_2^* + T_2^{**}$. Following the similar argument as in Lemma A.4, we have

$$T_2^{*2} = O_p[c^{-2-2\beta_1}J^{-1}\bar{h}^{-1}h^{d_3} \exp\{2(c-1)\varrho(x)^{-1}h^{-\rho_2(x)} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}],$$

for some constant d_3. We also have when Assumption 2 holds, $T_2^{**} = o_p(T_2^*)$, and term $T_3 = o_p(T_1)$. For term T_4, it can be shown that under Assumption 2,

$$T_4 = O_p[J^{-1}\bar{h}^{-1}h^{d_4} \exp\{(2c-1)\varrho(x)^{-1}h^{-\rho_2(x)} + 2c^{\beta_2}\zeta^{-1}h^{-\beta_2}\}],$$

for some constant d_4. In addition, when Assumption 1 holds, $T_4 = o_p(T_2^*)$.

Combining the order of T_1, T_2^*, T_2^{**}, T_3, and T_4, together with term (B.2), we finish to show that

$$\begin{aligned}\widehat{f}_{HP}(y|x) - f(y|x) &= \frac{1}{2\pi}\int e^{-ity} q(x,t)\left\{\phi_K(ht) - 1\right\} dt + T_1 + T_2^*\{1 + o_p(1)\} \\ &= B_{\bar{h},h}(x,y) + V_{\bar{h},h}^{1/2}(x,y),\end{aligned}$$

where

$$\begin{aligned}B_{\bar{h},h}(x,y) &= \frac{1}{2\pi}\int \exp(-ity) q(x,t)\{\phi_K(ht) - 1\} dt + T_1 \\ V_{\bar{h},h}(x,y) &= T_2^{*2}\{1 + o_p(1)\}.\end{aligned}$$

□

Appendix C

Appendix for Chapter 4

C.1 Derivation of the Log-likelihood Function

In this section, we derive the log-likelihood function under the Dorfman testing protocol. Let x denote the pool size and G denote the pool diagnosis. The true status of the ith individual in the pool is $\tilde{A}_i$, and the corresponding diagnosis is A_i, where $i = 1, \ldots, x$. Let g, a_i, and $\tilde{a}_i$ denote the outcome of G, A_i, and $\tilde{A}_i$, respectively, where $g, a_i, \tilde{a}_i \in \{0, 1\}$. Lastly, denote by $z = \sum_{i=1}^{x} a_i$. We first introduce Lemma C.1, and the log-likelihood function follows by direct calculation.

Lemma C.1. *Assume that the test results are independent conditional on the true statuses of the individuals and there is no dilution effect. Then*

$$\begin{aligned} &P(G = 1, A_1 = a_1, \ldots, A_x = a_x) \\ =&[\pi(x|\boldsymbol{\theta}) - S_e](1 - S_p)^z S_p^{x-z} + \pi(1|\boldsymbol{\theta})^z [1 - \pi(1|\boldsymbol{\theta})]^{x-z} S_e, \end{aligned}$$

where $\pi(x|\boldsymbol{\theta})$ is defined in Section 4.2.1.

Proof. Under the two assumptions stated in the lemma,

$$
\begin{aligned}
& P(G=1, A_1=a_1, \ldots, A_x=a_x) \\
= & \sum_{\tilde{a}_i \in \{0,1\}} P(G=1, A_1=a_1, \ldots, A_x=a_x | \tilde{A}_1=\tilde{a}_1, \ldots, \tilde{A}_x=\tilde{a}_x) \\
& \qquad \times P(\tilde{A}_1=\tilde{a}_1, \ldots, \tilde{A}_x=\tilde{a}_x) \\
= & \sum_{\tilde{a}_i \in \{0,1\}} P(G=1 | \tilde{A}_1=\tilde{a}_1, \ldots, \tilde{A}_x=\tilde{a}_x) \\
& \qquad \times P(A_1=a_1, \tilde{A}_1=\tilde{a}_1) \ldots P(A_x=a_x, \tilde{A}_x=\tilde{a}_x) \\
= & \, P(G=1 | \tilde{A}_1=0, \ldots, \tilde{A}_x=0) P(A_1=a_1, \tilde{A}_1=0) \ldots P(A_x=a_x, \tilde{A}_x=0) \\
& + \sum_{\{\tilde{a}_1, \ldots, \tilde{a}_x\} \neq \{0, \ldots, 0\}} P(G=1 | \tilde{A}_1=\tilde{a}_1, \ldots, \tilde{A}_x=\tilde{a}_x) \\
& \qquad \times P(A_1=a_1, \tilde{A}_1=\tilde{a}_1) \ldots P(A_x=a_x, \tilde{A}_x=\tilde{a}_x).
\end{aligned}
$$

The first term equals

$$(1-S_p)^{z+1} S_p^{x-z} (1-p)^x.$$

The second term equals

$$S_e[pS_e + (1-p)(1-S_p)]^z [(1-p)S_p + p(1-S_e)]^{x-z} - S_e(1-S_p)^z S_p^{x-z}(1-p)^x.$$

We finish the proof by adding up the two terms. □

The log-likelihood function of DT directly follows Lemma C.1.

C.2 Additional Simulation Evidence

Table C.1: Point estimation for DT with a fixed number of individuals $N = 5000$. Two settings are considered: pool size $x = 5$ with $n = 1000$ pools and pool size $x = 10$ with $n = 500$ pools. In each setting, $B = 1000$ simulations are implemented. BIAS denotes the average bias over the 1000 Monte Carlo data sets. SD denotes the sample standard deviation of the 1000 estimates, and SE denotes the averaged standard error. The Mean-squared error (MSE) is also shown. All values are multiplied by 10^3.

			$x = 5$			$x = 10$		
			p	S_e	S_p	p	S_e	S_p
$S_p = 0.90$	$S_e = 0.90$	BIAS	0.83	-0.15	-1.09	2.68	-10.38	-0.67
		SD	9.74	67.87	16.59	12.44	86.94	11.51
		SE	9.96	75.85	16.78	16.76	96.88	12.79
		MSE	0.10	4.61	0.28	0.16	7.67	0.13
	$S_e = 0.95$	BIAS	1.55	-6.47	0.94	2.27	-14.59	-0.58
		SD	8.01	52.57	15.18	8.77	67.22	10.86
		SE	9.14	69.20	15.36	9.12	80.52	10.80
		MSE	0.07	2.81	0.23	0.08	4.73	0.12
	$S_e = 0.99$	BIAS	2.66	-21.28	1.37	2.65	-22.67	-0.44
		SD	6.69	38.67	14.69	6.67	46.24	10.48
		SE	8.73	65.18	14.60	8.11	74.56	10.45
		MSE	0.05	1.95	0.22	0.05	2.65	0.11
$S_p = 0.95$	$S_e = 0.90$	BIAS	0.88	-2.54	0.37	0.82	-2.00	0.02
		SD	7.55	53.27	11.57	6.63	51.39	8.15
		SE	7.92	56.07	12.05	6.62	50.27	8.41
		MSE	0.06	2.84	0.13	0.05	2.65	0.07
	$S_e = 0.95$	BIAS	1.30	-6.57	1.01	0.71	-2.48	0.01
		SD	6.90	43.36	10.74	5.66	39.47	8.24
		SE	7.36	52.27	11.32	5.93	42.75	7.99
		MSE	0.05	1.92	0.12	0.03	1.56	0.07
	$S_e = 0.99$	BIAS	2.27	-15.20	2.54	1.34	-9.31	0.51
		SD	5.40	29.81	9.37	4.74	24.50	7.32
		SE	6.92	48.18	10.67	5.56	38.29	7.69
		MSE	0.03	1.12	0.09	0.02	0.69	0.05
$S_p = 0.99$	$S_e = 0.90$	BIAS	-0.20	4.89	-0.74	0.20	0.59	-0.18
		SD	5.92	39.26	7.58	5.07	30.59	5.86
		SE	5.96	39.81	8.23	5.08	30.54	5.88
		MSE	0.04	1.57	0.06	0.03	0.94	0.03
	$S_e = 0.95$	BIAS	0.18	0.85	-0.11	0.19	0.29	0.05
		SD	5.16	31.34	6.93	4.66	24.07	5.67
		SE	5.57	35.80	7.83	4.67	24.04	5.69
		MSE	0.03	0.98	0.05	0.02	0.58	0.03
	$S_e = 0.99$	BIAS	1.06	-6.47	1.46	0.83	-3.02	0.89
		SD	4.00	18.94	5.20	4.05	13.57	5.29
		SE	5.25	31.36	7.38	4.48	20.11	5.50
		MSE	0.02	0.40	0.03	0.02	0.19	0.03

Table C.2: Operating characteristics estimation for DT with a fixed number of individuals $N = 5000$. Two settings are considered: pool size $x = 5$ with $n = 1000$ pools and pool size $x = 10$ with $n = 500$ pools. In each setting, $B = 1000$ simulations are implemented. Summarized results include: true value (TRUE), coverage probability (CP95), and average length (LEN) of 95% Wald confidence intervals for $E(T)$, $E(C)$, PPV, NPV.

			$x = 5$			$x = 10$		
			TRUE	CP95	LEN	TRUE	CP95	LEN
$S_p = 0.90$	$S_e = 0.90$	$E(T)$	0.481	0.944	0.056	0.520	0.950	0.086
		$E(C)$	0.967	0.969	0.034	0.952	0.962	0.056
		PPV	0.632	0.964	0.239	0.518	0.958	0.176
		NPV	0.990	0.995	0.035	0.989	0.993	0.052
	$S_e = 0.95$	$E(T)$	0.492	0.961	0.056	0.541	0.938	0.087
		$E(C)$	0.971	0.988	0.029	0.956	0.991	0.043
		PPV	0.648	0.948	0.206	0.534	0.964	0.155
		NPV	0.995	0.994	0.032	0.995	1.000	0.039
	$S_e = 0.99$	$E(T)$	0.501	0.966	0.057	0.557	0.932	0.087
		$E(C)$	0.974	0.988	0.027	0.958	0.999	0.037
		PPV	0.661	0.941	0.189	0.546	0.960	0.143
		NPV	0.999	0.991	0.030	0.999	1.000	0.032
$S_p = 0.95$	$S_e = 0.90$	$E(T)$	0.442	0.944	0.053	0.491	0.948	0.085
		$E(C)$	0.981	0.966	0.022	0.973	0.939	0.025
		PPV	0.804	0.959	0.179	0.701	0.958	0.166
		NPV	0.990	0.984	0.026	0.990	0.939	0.024
	$S_e = 0.95$	$E(T)$	0.454	0.949	0.054	0.511	0.953	0.086
		$E(C)$	0.985	0.979	0.019	0.977	0.959	0.020
		PPV	0.814	0.960	0.159	0.713	0.946	0.150
		NPV	0.995	0.970	0.024	0.995	0.997	0.019
	$S_e = 0.99$	$E(T)$	0.463	0.953	0.055	0.527	0.954	0.087
		$E(C)$	0.988	0.988	0.017	0.980	0.982	0.017
		PPV	0.821	0.954	0.146	0.722	0.956	0.140
		NPV	0.999	0.989	0.022	0.999	0.998	0.016
$S_p = 0.99$	$S_e = 0.90$	$E(T)$	0.411	0.941	0.051	0.467	0.954	0.084
		$E(C)$	0.989	0.932	0.014	0.987	0.935	0.013
		PPV	0.961	0.984	0.125	0.926	0.973	0.162
		NPV	0.990	0.940	0.018	0.990	0.943	0.014
	$S_e = 0.95$	$E(T)$	0.423	0.946	0.052	0.487	0.949	0.085
		$E(C)$	0.993	0.985	0.011	0.992	0.934	0.010
		PPV	0.963	0.997	0.113	0.930	0.971	0.149
		NPV	0.995	0.993	0.016	0.995	0.937	0.011
	$S_e = 0.99$	$E(T)$	0.432	0.963	0.052	0.503	0.958	0.086
		$E(C)$	0.997	0.993	0.009	0.995	0.921	0.007
		PPV	0.964	1.000	0.105	0.933	0.983	0.140
		NPV	0.999	0.993	0.014	0.999	0.991	0.008

Table C.3: Point estimation for DT with a fixed number of tests $m = 3000$. BIAS denotes the average bias over $B = 1000$ Monte Carlo data sets. SD denotes the sample standard deviation of the 1000 estimates, and SE denotes the averaged standard error. The Mean-squared error (MSE) is also shown. All values are multiplied by 10^3.

			$S_p = 0.90$			$S_p = 0.95$			$S_p = 0.99$		
			p	S_e	S_p	p	S_e	S_p	p	S_e	S_p
$S_e = 0.90$	$p = 0.05$	BIAS	1.10	-4.21	-0.51	0.32	1.91	-0.27	-0.06	1.88	-0.47
		SD	8.05	67.79	12.84	6.13	44.27	8.82	4.47	27.73	5.87
		SE	8.38	72.28	12.98	6.01	43.76	9.04	4.57	27.97	6.01
		MSE	0.07	4.61	0.17	0.04	1.96	0.08	0.02	0.77	0.04
	$p = 0.10$	BIAS	1.63	-2.67	0.14	0.67	-0.34	0.13	0.13	0.68	-0.18
		SD	12.64	46.80	14.86	10.10	33.30	11.59	7.84	23.37	7.66
		SE	13.01	47.58	14.73	10.27	33.96	11.64	8.35	24.73	9.08
		MSE	0.16	2.20	0.22	0.10	1.11	0.13	0.06	0.55	0.06
	$p = 0.15$	BIAS	0.44	1.78	-0.72	1.07	-0.62	0.22	-0.92	2.43	-1.46
		SD	17.73	40.71	18.13	14.76	30.56	14.69	11.36	23.71	10.50
		SE	17.77	40.35	17.74	14.70	30.96	14.90	12.29	24.41	12.52
		MSE	0.32	1.66	0.33	0.22	0.93	0.22	0.13	0.57	0.11
$S_e = 0.95$	$p = 0.05$	BIAS	1.37	-6.38	0.67	0.73	-2.87	0.02	0.23	-1.38	-0.03
		SD	6.99	53.19	11.95	5.48	37.37	8.49	4.25	23.92	5.85
		SE	7.59	64.35	12.14	5.71	40.07	8.71	4.37	24.29	5.89
		MSE	0.05	2.87	0.14	0.03	1.40	0.07	0.02	0.57	0.03
	$p = 0.10$	BIAS	0.85	-0.81	-0.90	0.44	0.43	-0.32	0.05	0.40	-0.58
		SD	10.70	36.52	12.94	9.32	29.51	11.11	7.42	20.72	7.81
		SE	11.95	43.60	14.11	9.64	31.07	11.32	7.92	21.85	8.99
		MSE	0.12	1.34	0.17	0.09	0.87	0.12	0.06	0.43	0.06
	$p = 0.15$	BIAS	1.66	-1.62	0.21	1.17	-0.78	0.95	-0.38	1.45	-0.64
		SD	15.79	34.35	16.67	13.57	27.88	14.13	10.13	18.77	9.69
		SE	16.54	36.92	17.10	13.73	28.27	14.51	11.59	21.73	12.34
		MSE	0.25	1.18	0.28	0.19	0.78	0.20	0.10	0.35	0.09
$S_e = 0.99$	$p = 0.05$	BIAS	2.18	-19.29	-0.24	1.16	-8.97	0.41	0.66	-3.35	0.74
		SD	6.18	39.85	12.26	4.56	23.48	7.88	3.68	13.49	4.91
		SE	7.31	61.90	11.90	5.39	36.70	8.07	4.18	19.30	5.53
		MSE	0.04	1.96	0.15	0.02	0.63	0.06	0.01	0.19	0.03
	$p = 0.10$	BIAS	3.03	-10.64	0.67	2.10	-5.98	1.39	1.13	-2.45	0.96
		SD	8.94	26.62	12.24	8.12	19.08	9.48	6.27	12.16	6.45
		SE	11.41	41.39	12.75	9.05	26.83	10.42	7.56	18.54	8.80
		MSE	0.09	0.82	0.15	0.07	0.40	0.09	0.04	0.15	0.04
	$p = 0.15$	BIAS	3.85	-8.87	1.75	2.55	-4.74	2.29	1.17	-1.63	0.62
		SD	12.63	23.08	15.57	10.37	16.04	12.10	7.84	11.17	7.50
		SE	15.69	34.26	16.63	12.69	23.51	13.33	11.04	19.03	12.13
		MSE	0.17	0.61	0.25	0.11	0.28	0.15	0.06	0.13	0.06

Table C.4: Point estimation for MPT with a fixed number of tests $m = 3000$. BIAS denotes the average bias over $B = 1000$ Monte Carlo data sets. SD denotes the sample standard deviation of the 1000 estimates, and SE denotes the averaged standard error. The Mean-squared error (MSE) is also shown. All values are multiplied by 10^3.

			$S_p = 0.90$			$S_p = 0.95$			$S_p = 0.99$		
			p	S_e	S_p	p	S_e	S_p	p	S_e	S_p
$S_e = 0.90$	$p = 0.05$	BIAS	2.49	-1.43	1.58	2.22	-1.92	1.36	1.41	0.77	0.44
		SD	11.58	94.77	13.54	10.77	94.96	11.47	9.02	89.81	7.73
		SE	15.09	148.91	14.30	13.92	146.97	12.23	12.93	145.07	10.04
		MSE	0.14	8.98	0.19	0.12	9.02	0.13	0.08	8.07	0.06
	$p = 0.10$	BIAS	-0.08	5.24	0.20	0.03	4.57	0.29	-0.98	6.48	-1.63
		SD	12.35	37.77	17.12	11.62	37.63	15.33	9.86	36.10	10.94
		SE	12.62	39.78	17.30	11.84	39.63	15.74	11.22	39.97	14.23
		MSE	0.15	1.45	0.29	0.14	1.44	0.24	0.10	1.35	0.12
	$p = 0.15$	BIAS	-0.78	2.71	-0.43	-0.68	2.61	-0.21	-2.04	3.62	-3.40
		SD	12.59	17.22	20.68	11.82	17.36	19.44	9.68	16.68	13.79
		SE	12.74	17.47	20.49	12.02	17.52	19.27	11.42	17.67	18.13
		MSE	0.16	0.30	0.43	0.14	0.31	0.38	0.10	0.29	0.20
$S_e = 0.95$	$p = 0.05$	BIAS	4.43	-25.33	2.33	3.38	-20.45	2.05	2.40	-15.81	0.90
		SD	10.72	81.67	12.71	9.37	78.98	11.49	7.71	73.98	7.60
		SE	14.33	135.95	14.26	13.09	135.40	12.33	12.23	135.75	10.20
		MSE	0.13	7.31	0.17	0.10	6.66	0.14	0.07	5.72	0.06
	$p = 0.10$	BIAS	0.34	2.76	0.57	0.39	2.36	0.60	-0.60	4.35	-1.45
		SD	10.80	31.74	17.05	10.27	32.35	15.27	8.62	30.90	10.73
		SE	11.55	36.78	17.38	10.92	36.95	15.86	10.39	37.44	14.42
		MSE	0.12	1.02	0.29	0.11	1.05	0.23	0.08	0.97	0.12
	$p = 0.15$	BIAS	-0.63	2.12	-0.60	-0.54	2.03	-0.37	-1.70	2.90	-3.46
		SD	11.27	14.77	20.67	10.67	14.97	19.55	8.98	14.54	14.12
		SE	11.45	14.97	20.55	10.87	15.07	19.38	10.40	15.25	18.31
		MSE	0.13	0.22	0.43	0.11	0.23	0.38	0.08	0.22	0.21
$S_e = 0.99$	$p = 0.05$	BIAS	5.66	-43.78	3.05	4.51	-37.55	2.75	3.55	-33.22	1.45
		SD	9.12	67.46	12.58	8.10	65.23	11.41	6.51	59.96	7.42
		SE	13.54	124.85	14.29	12.38	125.30	12.36	11.64	126.44	10.31
		MSE	0.12	6.47	0.17	0.09	5.66	0.14	0.06	4.70	0.06
	$p = 0.10$	BIAS	2.46	-6.12	2.38	2.42	-6.57	2.32	1.46	-4.98	-0.08
		SD	7.94	20.17	16.20	7.66	20.83	14.67	6.24	19.32	10.16
		SE	10.60	32.40	17.35	10.08	32.82	15.88	9.63	33.41	14.49
		MSE	0.07	0.44	0.27	0.06	0.48	0.22	0.04	0.40	0.10
	$p = 0.15$	BIAS	1.00	-0.66	1.54	0.93	-0.71	1.38	-0.25	0.14	-1.94
		SD	9.01	9.53	19.18	8.58	9.70	17.95	7.26	9.23	12.64
		SE	10.32	12.24	20.20	9.89	12.46	19.11	9.52	12.72	18.11
		MSE	0.08	0.09	0.37	0.07	0.10	0.32	0.05	0.09	0.16

Table C.5: Test characteristics using DT with a fixed number of tests $m = 3000$. MNT and SDNT denote the mean and standard deviation of the number of tests using DT over $B = 1000$ Monte Carlo data sets. W-CP, S-CP, and LR-CP denote the coverage probability of joint 95% Wald, score, and likelihood ratio confidence regions of $\boldsymbol{\theta}$ over the 1000 estimates. W-VOL denotes the average volume of joint 95% Wald confidence region. The value of W-VOL is multiplied by 10^5.

		$S_p = 0.90$			$S_p = 0.95$			$S_p = 0.99$		
		$p = 0.01$	$p = 0.05$	$p = 0.10$	$p = 0.01$	$p = 0.05$	$p = 0.10$	$p = 0.01$	$p = 0.05$	$p = 0.10$
$S_e = 0.90$	MNT	2717	2781	2813	2718	2777	2813	2714	2775	2807
	SDNT	94	72	60	97	71	64	94	77	63
	W-CP	0.948	0.945	0.937	0.932	0.951	0.948	0.959	0.957	0.96
	S-CP	0.954	0.938	0.948	0.944	0.953	0.947	0.952	0.956	0.945
	LR-CP	0.958	0.94	0.949	0.942	0.958	0.949	0.963	0.961	0.954
	W-VOL	42.37	30.10.	75.03	13.60	21.47	28.60	5.19	8.45	11.32
$S_e = 0.95$	MNT	2723	2783	2819	2721	2779	2815	2714	2778	2813
	SDNT	94	74	60	94	74	64	93	75	63
	W-CP	0.959	0.965	0.958	0.963	0.945	0.946	0.956	0.962	0.96
	S-CP	0.949	0.949	0.951	0.96	0.941	0.942	0.948	0.952	0.955
	LR-CP	0.961	0.955	0.959	0.962	0.945	0.943	0.954	0.957	0.961
	W-VOL	33.55	31.05	42.94	9.43	19.13	15.98	3.06	6.40	10.20
$S_e = 0.99$	MNT	2724	2769	2817	2698	2763	2798	2695	2780	2816
	SDNT	96	77	61	102	83	65	101	76	61
	W-CP	0.963	0.967	0.959	0.965	0.958	0.972	0.946	0.951	0.952
	S-CP	0.931	0.951	0.934	0.944	0.947	0.961	0.935	0.957	0.95
	LR-CP	0.948	0.958	0.947	0.962	0.955	0.973	0.954	0.968	0.967
	W-VOL	21.86	24.50	32.51	8.15	10.95	13.55	2.08	4.05	6.19

Table C.6: Test characteristics using MPT with a fixed number of tests $m = 3000$. MNT and SDNT denote the mean and standard deviation of the number of tests using MPT over $B = 1000$ Monte Carlo data sets. W-CP, S-CP, and LR-CP denote the coverage probability of joint 95% Wald, score, and likelihood ratio confidence regions of $\boldsymbol{\theta}$ over the 1000 estimates. W-VOL denotes the average volume of joint 95% Wald confidence region. The value of W-VOL is multiplied by 10^5.

		$S_p = 0.90$			$S_p = 0.95$			$S_p = 0.99$		
		$p = 0.01$	$p = 0.05$	$p = 0.10$	$p = 0.01$	$p = 0.05$	$p = 0.10$	$p = 0.01$	$p = 0.05$	$p = 0.10$
$S_e = 0.90$	MNT	3000	3000	3000	3000	3000	3000	3000	3000	3000
	SDNT	0	0	0	0	0	0	0	0	0
	W-CP	0.844	0.916	0.942	0.851	0.924	0.934	0.894	0.946	0.955
	S-CP	0.962	0.95	0.939	0.958	0.953	0.938	0.954	0.952	0.942
	LR-CP	0.971	0.952	0.937	0.968	0.952	0.94	0.976	0.963	0.953
	W-VOL	35.38	19.78	17.58	27.38	16.79	15.49	20.23	14.22	13.75
$S_e = 0.95$	MNT	3000	3000	3000	3000	3000	3000	3000	3000	3000
	SDNT	0	0	0	0	0	0	0	0	0
	W-CP	0.846	0.937	0.935	0.869	0.94	0.94	0.913	0.968	0.959
	S-CP	0.953	0.946	0.934	0.951	0.956	0.944	0.962	0.95	0.949
	LR-CP	0.964	0.955	0.936	0.967	0.956	0.945	0.98	0.968	0.958
	W-VOL	32.51	16.86	13.35	25.59	14.54	11.94	19.30	12.50	10.75
$S_e = 0.99$	MNT	3000	3000	3000	3000	3000	3000	3000	3000	3000
	SDNT	0	0	0	0	0	0	0	0	0
	W-CP	0.861	0.961	0.96	0.884	0.961	0.969	0.919	0.98	0.984
	S-CP	0.952	0.957	0.957	0.947	0.953	0.958	0.953	0.958	0.96
	LR-CP	0.968	0.972	0.962	0.964	0.97	0.963	0.977	0.979	0.976
	W-VOL	30.32	14.42	9.42	24.17	12.62	8.60	18.49	11.01	7.87

C.3 Comparison of D_s-Optimal Designs

We investigate the comparison between MPT and DT under D-optimality in Section 4.4.1. This section presents the comparison under D_s-optimality. Assume the budget is m assays. We denote $\boldsymbol{x}^*_{\mathcal{M}} = (1, x^*_M, 20)$ the corresponding D_s-optimal design, where x^*_M is the best middle pool size calculated based on x_L, x_U, m, p, S_e, and S_p according to Huang et al. (2017). In the D_s-optimal design, the number of pools is not evenly divided. Denote by $\boldsymbol{n}^*_{\mathcal{M}}$ the D_s-optimal number of pools corresponding to $\boldsymbol{x}^*_{\mathcal{M}}$ following Huang et al. (2017). In DT, similar to Section 4.4.1, we use the following steps to determine the optimal design:

- Step 1: For each x from 2 to x_U, calculate the largest integer n that satisfies Equation (4.3).
- Step 2: Go through each pair of (n, x) identified by Step 1. Find $(n^*_D, x^*_D) = \underset{(n,x)}{\operatorname{argmin}} I^-_{\mathcal{D}}(\boldsymbol{\theta}|n, x)_{[1,1]}$, where $I^-_{\mathcal{D}}(\boldsymbol{\theta}|n, x)$ is the generalized inverse of $I_{\mathcal{D}}(\boldsymbol{\theta}|n, x)$ and $[1, 1]$ means the first element in $I^-_{\mathcal{D}}(\boldsymbol{\theta}|n, x)$.

Define

$$g(\boldsymbol{\theta}, m, x_L, x_U) = \log I^-_{\mathcal{M}}(\boldsymbol{\theta}|\boldsymbol{n}^*_M, \boldsymbol{x}^*_M)_{[1,1]} - \log I^-_{\mathcal{D}}(\boldsymbol{\theta}|n^*_D, x^*_D)_{[1,1]}.$$

We take logarithm to make the value of $g(\boldsymbol{\theta}, m, x_L, x_U)$ in an appropriate scale. If $g(\boldsymbol{\theta}, m, x_L, x_U) > 0$, DT achieves higher estimation efficiency in terms of D_s-optimality than MPT and vice versa. We can draw the similar conclusion as in Section 4.4.1 from figure C.1.

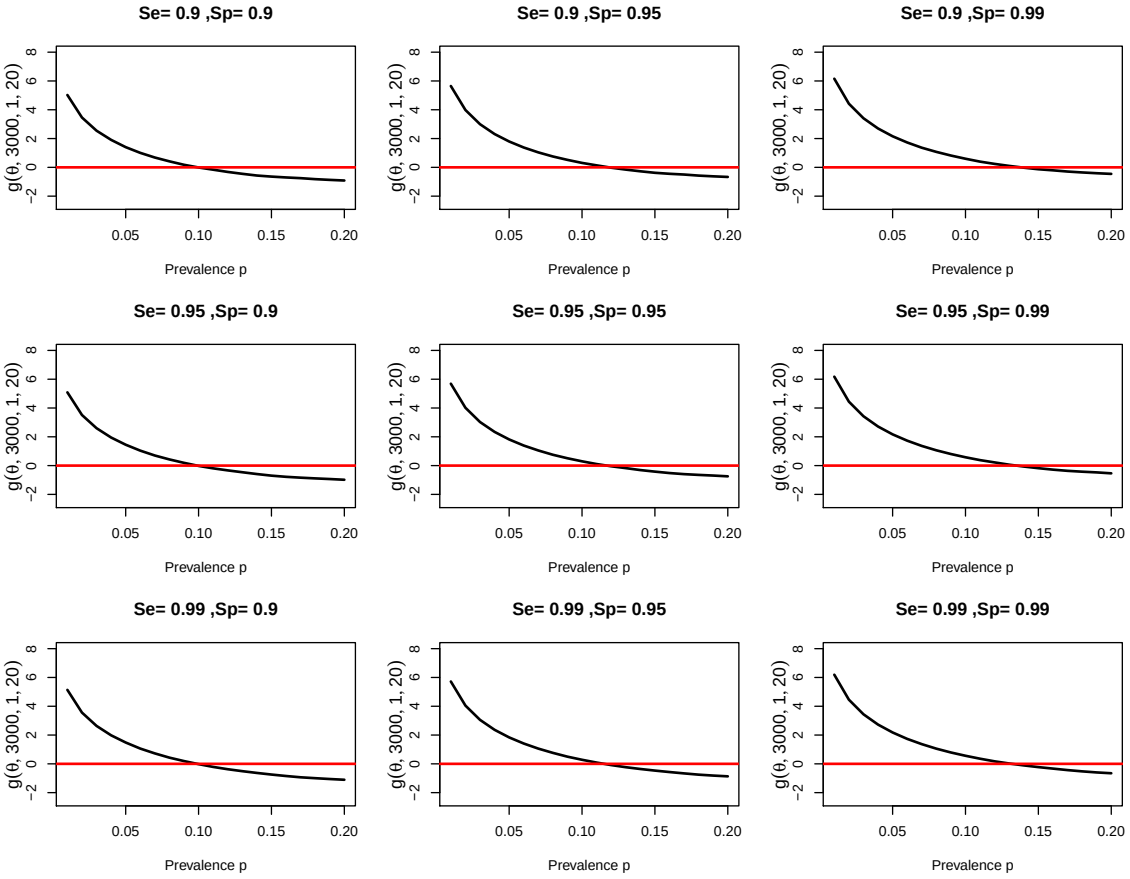

Figure C.1: $g(\boldsymbol{\theta}, 3000, 1, 20)$ of different p, S_e, and S_p. The horizontal red line indicates the position of 0.