Topics on nonparametric calibration, kernel ridge regression imputation and nonparametric propensity score estimation

by

Hengfang Wang

A dissertation submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Jae Kwang Kim, Co-major Professor
Zhengyuan Zhu, Co-major Professor
Emily Berg
Lynna Chu
Cindy L. Yu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation. The Graduate College will ensure this dissertation is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2021

DEDICATION

To my loving parents Yan Zhang and Weihu Wang,

and my girl friend Chuying Huang. Thank you for all of your love.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

ACKNOWLEDGMENTS

This dissertation cannot be complete without the support of our Department of Statistics, my friends and family.

I would like to take this opportunity to express my thanks to people who helped me a lot in many aspects of life. First and foremost, my co-major professor Dr. Jae Kwang Kim for his guidance, patience and support throughout this research and the writing of this dissertation. His insights and words of encouragement have often inspired me and renewed my hopes for research. I would also want to thank my co-major professor Dr, Zhengyuan Zhu, who provide me valuable advices for career and academia. My previous committee member, Dr. Song Xi Chen, I want to thank you for your generous mentorship for my initial graduate career. Further, I also want to thank other committee members, Dr. Emily Berg, Dr. Cindu L. Yu and Dr. Lynna Chu, for your efforts and contribution for my research.

ABSTRACT

This dissertation focuses on statistical issues arising in survey data and item nonresponse. In particular, it covers topics on nonparametric calibration in survey data, kernel ridge regression imputation and density ratio estimation in propensity score approach.

The first project is about nonparametric calibration in survey sampling. Estimation of a finite population mean or total is important in survey sampling. Calibration estimation is a popular method to address this issue by adjusting the sampling weights to match the unknown population totals of auxiliary variables. When the auxiliary vairbales are observed for all units in the finite population, one can apply the model calibration using the working outcome model. Traditional parametric calibration approach might not be robust in practice. We develope a nonparametric calibration method employing infinite-dimensional reproducing kernel Hilbert space (RKHS) that does not require an explicit outcome model. Under mild assumptions, the proposed calibration estimator attains the Godambe-Joshi lower bound asymptotically.

The second project is about handling missing data using kernel ridge regression method. Missing data is frequently encountered in practice. In some cases, missingness is planned to reduce the cost or the response burden. Ignoring the cases with missing values can lead to misleading results. To avoid the potential problem with missing data, imputation is commonly used. Kernel Ridge Regression (KRR) is a modern nonparametric regression technique based on the theory of Reproducing Kernel Hilbert Space, which enjoys the model robustness. We consider such method to imputation. Specifically, we establish the root-n consistency of the KRR imputation estimators and show that it is optimal in the sense that it achieves the lower bound of the semiparametric asymptotic variance. We further consider propensity score weighting method using kernel ridge regression and discuss its asymptotic properties.

The third project is about propensity score estimation using density ration function approach. The propensity score approach is also a popular tool for handling item nonresponse. The propensity score is often developed using the model for the response probability. In practice, regression models for binary response, e.g., logistic regression, can be utilized to model the response probability given the observed auxiliary information. An inverse probability weighting estimator can then be constructed to get an unbiased estimation of the target parameter. We consider an alternative approach of estimating the inverse of the propensity scores using density ratio function. Density ratio estimation can be obtained by applying the maximum entropy method which uses the Kullback-Leibler distance measure. By including the covariates for the outcome regression models only into the density ratio model, we can achieve efficient propensity score estimation. We further extend the proposed approach to handling the multivariate missing case.

CHAPTER 1. GENERAL INTRODUCTION

In this dissertation, we develope nonparametric approaches to address some issues in survey data and item nonresponse. Specifically, the whole dissertation includes three papers, the first one covers the topic on nonparametric calibration, the second one is about nonparametric imputation and the third one is related to density ratio estimation in propensity score approach.

An important topic on survey sampling is estimation of a finite population mean or total. Horvitz-Thompson (HT) estimator, in addition with a probability sample, can be used to achieve the design-based unbiased estimation and does not require any model assumptions although it is not efficient. We can improve the efficiency of the HT estimator in the estimation stage by using auxiliary information available in the population level. One approach is parametric calibration and it has been well studied in the literature. Due to the strong assumption of such parametric form, it might not be robust in practice. Chapter 2 presents a functional calibration employing infinite-dimensional reproducing kernel Hilbert space (RKHS). Due to the infinite-dimensional space, the traditional calibration equations can no longer work in this scenario. As a twist, we utilize a validity measure to quantify the distance between a nonlinear transformation of auxiliary information within sample and that in the finite population in a RKHS and construct a finite-dimensional objective function to solve the optimization problem. Numerical algorithms are developed and implemented to solve the optimization problem in the functional calibration. Furthermore, under the nonparametric working model, the proposed calibration estimator attains the Godambe-Joshi lower bound asymptotically.

Chapter 3 consider an issue in item nonresponse. Missing data is very common in practice. Imputation is a popular technique to avoid the potential problem with missing data. After imputation, the imputed dataset can serve as a complete dataset that has no missing values, which in turn makes results from different analysis methods consistent. How to make statistical

inferences with imputed point estimators is an important statistical problem. On the other hand, Kernel Ridge Regression (KRR) is a modern regression technique based on the theory of reproducing kernel Hilbert space. In this chapter, we consider KRR as a nonparametric imputation method. Under regularity conditions, we establish the root-n consistency of the KRR imputation estimators and show that it is optimal in the sense that it achieves the lower bound of the semiparametric asymptotic variance. A nonparametric propensity score estimator using the KRR method is also developed by the maximum entropy method of the density ratio function. Variance estimation for KRR imputation estimator is then developed using the nonparametric propensity score weights.

Propensity score (PS) is a popular approach to handling the missing data problem using inverse weighting. However, correct specification of the propensity score model can be challenging and we often do not have a good understanding of the response mechanism to specify the propensity model correctly. The existing methods for propensity score estimation are either based on maximum likelihood method or calibration method with some penalization in the calibration equation. The calibration method gives a doubly robust flavour, but the choice of the objective function for calibration estimation is not fully agreed. In Chapter 4, we consider an alternative approach of estimating the inverse of the propensity scores using density ratio function. By partitioning the sample into two groups based on the response status of the elements, we can apply the density ratio function estimation method and obtain the inverse propensity scores. Density ratio estimation can be obtained by applying the maximum entropy method essentially do the maximization of the lower bound of the Kullback-Leibler distance measure. We can achieve efficient propensity score estimation by including the covariates for the outcome regression models only into the density ratio model. We also extend this framework to the high dimensional scenario where redundant covariates present. We further extend the proposed approach to the multivariate missing case.

CHAPTER 2. NONPARAMETRIC FUNCTIONAL CALIBRATION ESTIMATION IN SURVEY SAMPLING

Hengfang Wang, Jae Kwang Kim and Zhengyuan Zhu

Iowa State University

Modified from a manuscript to be submitted to *Scandinavian Journal of Statistics*

2.1 Abstract

Calibration estimation, a technique of adjusting the sampling weights to match the unknown population totals of auxiliary variables, is a popular method of estimation in survey sampling. When the auxiliary vairbales are observed for all units in the finite population, one can apply the model calibration of Wu and Sitter (2001) using the working outcome model. In this paper, we develope a kernel-based nonparametric calibration method that does not require an explicit outcome model. The proposed method achieves the approximate calibration for all functions in the infinite-dimensional reproducing kernel Hilbert space (RKHS). Numerical algorithms are developed and implemented to solve the optimization problem in the function calibration, and some asymptotic results are presented as well. Furthermore, under the nonparametric working model, the proposed calibration estimator attains the Godambe-Joshi lower bound asymptotically. Simulation results are presented to compare the proposed method with other calibration methods. Empirical study illustrates the performance of our proposed estimator.

2.2 Introduction

Estimation of finite population means or totals is an important problem in survey sampling. Horvitz-Thompson (HT) estimator combined with a probability sample is used to achieve the design-based unbiased estimation which does not require any model assumptions. However, HT

estimator is not necessarily efficient. Using auxiliary information available in the population level, we can improve the efficiency of the HT estimator in the estimation stage. To incorporate the auxiliary information, one idea is to use a relationship between the study variable y and the auxiliary variable x to construct a prediction-type estimator. Generalized regression (GREG) estimator, discussed by Cassel et al. (1976) and Huang (1978), is a classical example of using the regression model to improve the efficiency of the HT estimator. Deville and Särndal (1992) viewed the GREG estimator as a special case of the calibration estimator whose weights are obtained by minimizing a distance measure between the design weights and the final weights subject to calibration equations. See Fuller (2009)(Chapter 2) for a more rigorous treatment of the GREG and calibration estimation.

Calibration estimation involves a working model either implicitly or explicitly. The GREG estimator implicitly uses a linear regression model for calibration. Isaki and Fuller (1982) developed a unified theory of regression estimator under a linear regression model. Firth and Bennett (1998) used non-linear regression models in the calibration estimation. Wu and Sitter (2001) formalized the idea and developed the so-called model calibration estimation and established its design consistency. Kim and Park (2010) generalized the idea further to discuss the functional-form calibration.

For nonparametric case, Breidt and Opsomer (2000) introduced the local polynomial regression estimator as a nonparametric calibration estimator, which uses nonparametric basis functions in the calibration estimation. The idea is further extended by Breidt et al. (2005) using penalised spline method. Also, Goga (2005) proposed B-spline approach as a nonparametric calibration estimation. Meanwhile, Montanari and Ranalli (2005) imposed neural network method and developed a second stage calibration procedure to get an efficient estimator. Most of the aforementioned nonparametric calibration methods involve the choice of bandwidth selection (or model complexity parameter) to obtain the best prediction (or best selection for the calibration functions), which implicitly assume that the study variable is univariate. For the multivariate case, each study variable will have different models and a single choice of the

bandwidth parameter cannot achieve the optimality uniformly. Breidt and Opsomer (2017) provided a comprehensive overview of he modern prediction methods for calibration.

In this paper, rather than using a prespecified calibration equation, we employ an infinite-dimensional RKHS in the optimization problem for calibration. Due to the infinite-dimensional space, the traditional calibration equations can no longer work in this scenario. As a twist, borrowing the idea of Wong and Chan (2018), we utilize a validity measure to quantify the distance between a nonlinear transformation of auxiliary information within sample and that in the finite population in a RKHS and construct a finite-dimensional objective function to solve the optimization problem. Also, if we use a kernel ridge regression as a working model, under mild conditions, the proposed method can achieve the Godambe-Joshi lower bound asymptotically.

Compared with other nonparametric approaches, our proposed estimator has the following advantages: (i) there is no need to pre-specify the form of calibration equation; (ii) multivariate case can be easily handled; (iii) our proposed estimator usually performs better than other nonparametric calibration estimators under complex settings and are comparable with other methods under simple settings.

2.3 Basic Setup

Suppose that we have a finite population of size N and we denote the population as $\mathcal{F}_N = \{(\mathbf{x}_i, y_i) : i \in \mathcal{U}_N\}$, where $\mathbf{x}_i \in \mathbb{R}^d$ and $\mathcal{U}_N = \{1, \ldots, N\}$ is the index set of the finite population. Here, y_i is the study variable and $\mathbf{x}_i$ is the corresponding covariate for unit i. Further, we assume that $\mathbf{x}_i$'s are available throughout the finite population. We are interested in estimating the finite population mean $\mathbb{E}(Y) = \theta$.

From the finite population, suppose we select a probability sample with index set $\mathcal{S}_N \subset \mathcal{U}_N$ of size n_N. Let $\pi_{iN} = \mathbb{P}(i \in \mathcal{S}_N)$ be the first order inclusion probability of unit i. For simplicity, we will suppress subscript N for n_N, $\mathcal{S}_N$, $\mathcal{U}_N$ and π_{iN} from now on. Moreover, we can define $\delta_i = \mathbb{I}\{i \in \mathcal{S}\}$ as the sample membership indicator for unit i. To estimate $\mathbb{E}(Y)$, Horvitz-Tompson

estimator can be used as

$$\widehat{\mathbb{E}(Y)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\pi_i}\delta_i y_i.$$

The Horvitz-Tompson estimator is design unbiased but is not necessarily efficient. To improve efficiency, Deville and Särndal (1992) proposed using

$$\widehat{\mathbb{E}(Y)} = \frac{1}{N}\sum_{i=1}^{N}\delta_i w_i y_i,$$

where w_i's are determined to minimize $Q_N(\mathbf{w}, \mathbf{d})$, a distance measurement between $\mathbf{w}$ and $\mathbf{d}$, subject to

$$\frac{1}{N}\sum_{i=1}^{N} w_i\delta_i\mathbf{U}(\mathbf{X}_i) = \frac{1}{N}\sum_{i=1}^{N}\mathbf{U}(\mathbf{X}_i), \tag{2.1}$$

where $\mathbf{U}(\cdot) = (u_1(\cdot), \ldots, u_L(\cdot))^{\mathrm{T}}$ and $u_j(\cdot) : \mathcal{X} \to \mathbb{R}$ is measurable for $j = 1, \ldots, L$ such that $\mathbb{E}[u(\boldsymbol{X})]$ is finite. If we use

$$Q_N(\mathbf{w}, \mathbf{d}) = \frac{1}{N}\sum_{i=1}^{N}\delta_i\frac{(w_i - d_i)^2}{d_i q_i}, \tag{2.2}$$

the resulting calibration estimator is algebraically equivalent to the following GREG estimator.

$$\hat{\theta}_{greg} = t_{y\pi} + (\mathbf{t}_{\mathbf{x}} - \hat{\mathbf{t}}_{\mathbf{x}\pi})^{\mathrm{T}}\hat{\mathbf{B}}_s$$

where $t_{y\pi} = N^{-1}\sum_{i=1}^{N}\delta_i d_i y_i$, $\hat{\mathbf{t}}_{\mathbf{x}\pi}N^{-1} = \sum_{i=1}^{N}\delta_i d_i\mathbf{x}_i$, $\mathbf{t}_{\mathbf{x}} = N^{-1}\sum_{i=1}^{N}\mathbf{x}_i$ and $\hat{\mathbf{B}}_s = (\sum_{i=1}^{N}\delta_i d_i q_i\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}})^{-1}\sum_{i=1}^{N}\delta_i d_i q_i\mathbf{x}_i y_i$. Equation (2.1) forms the constraints in the optimization problem for calibration. A good choice of the calibration function $\mathbf{U}(\cdot)$ can improve the efficiency of the resulting calibration estimator. For example, Isaki and Fuller (1982) used a model $y_i = \mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \epsilon_i$ and showed that the generalized regression estimator using $u(\mathbf{x}_i) = \mathbf{x}_i$ achieves the Godambe-Joshi lower bound, which is the lower bound of the anticipated variance under the superpopulation model. The nonparametric calibration estimator, considered in Breidt et al. (2005) and Montanari and Ranalli (2005), implicitly assume that $m(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X})$ is well approximated by a linear combination of the L calibration functions.

Suppose we have the additional superpopulation model ξ assumption: $m(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X})$, $y_i = m(\mathbf{x}_i) + \epsilon_i$ for $i = 1, \ldots, N$. Then we have the following decomposition:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^{N}\delta_i w_i y_i - \frac{1}{N}\sum_{i=1}^{N} y_i &= \frac{1}{N}\sum_{i=1}^{N}\delta_i w_i \{m(\mathbf{x}_i) + \epsilon_i\} - \frac{1}{N}\sum_{i=1}^{N}\{m(\mathbf{x}_i) + \epsilon_i\} \\ &= \underbrace{\frac{1}{N}\sum_{i=1}^{N}\{(\delta_i w_i - 1)m(\mathbf{x}_i)\}}_{:=T_1} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\delta_i(w_i - d_i)\epsilon_i}_{:=T_2} \\ &\quad + \underbrace{\left\{\frac{1}{N}\sum_{i=1}^{N}(\delta_i d_i - 1)\epsilon_i\right\}}_{:=T_3}. \end{aligned} \tag{2.3}$$

If we can control T_1 and T_2, then consistency of the calibration estimator can be obtained by the consistency of term T_3. As we can see, the term T_1 is negligible if the function $m(\cdot)$ lies in $\mathcal{H}_L$ where $\mathcal{H}_L =: \text{span}\{u_1, \ldots, u_L\}$. In addition, for term T_2, we can control it by minimizing $Q_N(\mathbf{w}, \mathbf{d})$. Once T_1 and T_2 are controlled, T_3 is the leading term in (2.3).

To facilitate our description, we first consider the simple setup of $m \in \mathcal{H}_L$. Define the empirical validity measure for calibration estimator which satisfies (2.1) as the following,

$$S_N(\mathbf{w}, u) = \left\{\frac{1}{N}\sum_{i=1}^{N}(\delta_i w_i - 1)u(\boldsymbol{X}_i)\right\}^2, \tag{2.4}$$

where $\mathbf{w} = (w_1, \ldots, w_N)^{\mathrm{T}}$ and $u(\cdot)$ is known. If $\mathcal{H}_L$ has finite basis with $L \ll n$, there always exists $\mathbf{w}_u$ such that $S_N(\mathbf{w}_u, u) = 0$, i.e., $\min_{\mathbf{w}} \sup_{u \in \mathcal{H}_L} S_N(\mathbf{w}, u) = 0$.

Note that the calibration estimation of Deville and Särndal (1992) can be written in the following optimization problem version

$$(\hat{\mathbf{w}}, \hat{\boldsymbol{\lambda}}) = \underset{(\mathbf{w}^{\mathrm{T}}, \boldsymbol{\lambda}^{\mathrm{T}})^{\mathrm{T}}}{\arg\min}\left\{Q_N(\mathbf{w}, \mathbf{d}) + \sum_{j=1}^{L}\lambda_j S_N(\mathbf{w}, u_j)\right\}, \tag{2.5}$$

where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_L)^{\mathrm{T}}$ is a vector of Lagrange multipliers. The consistency of calibration estimator obtained from (2.5) can be found in Deville and Särndal (1992) and Fuller (2009). If $m(\cdot)$ lies in $\mathcal{H}_L$, T_1 in this scenario equals to 0, while $T_2 = o_p\left(n^{-1/2}\right)$ and T_3 is the main term. Note the T_3 is design unbiased to zero and the variance of T_3 is equal to the Godambe-Joshi lower bound (Isaki and Fuller, 1982).

Literally, if $m(\cdot)$ is unknown, both number of L and form of $\{u_j : j = 1, \ldots, L\}$ remain high flexibility and are hard to be predetermined. Also, as pointed out by Hellerstein and Imbens (1999), such finite approximation by $\mathcal{H}_L$ might not be consistent without additional assumptions on the superpopulation model. Here we adopt the covariate functional balancing idea from Wong and Chan (2018) to do such calibration via an approximation by RKHS.

2.4 Proposed Method

Rather than employing a finite-basis functional space $\mathcal{H}_L$, we relax the functional space to be an infinite-dimension functional space $\mathcal{H}$. In this case, the quantity $\min_{\mathbf{w}} \sup_{u \in \mathcal{H}} S_N(\mathbf{w}, u)$ can be larger than 0. To discuss the Hilbert space property, we denote its inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and the induced norm $\|\cdot\|_{\mathcal{H}}$. One canonical example of such a space is the Sobolev space. Specifically, assuming that the domain of such functional space is $[0, 1]$, the Sobolev space of order l can be denoted as

$$\mathcal{W}_2^l = \left\{ f : [0,1] \to \mathbb{R} \mid f, f^{(1)}, \ldots, f^{(l-1)} \text{ are absolute continuous and } f^{(l)} \in L^2[0,1] \right\}.$$

One possible norm for this space can be

$$\|f\|_{\mathcal{W}_2^l}^2 = \sum_{q=0}^{l-1} \left\{ \int_0^1 f^{(q)}(t) dt \right\}^2 + \int_0^1 \left\{ f^{(l)}(t) \right\}^2 dt$$

Readers can refer to Wahba (1990) for a thorough treatment of the RKHS technique. In this section, we employ the Sobolev space of second order as the approximation space. Given the Hilbert space $\mathcal{H}$, the objective function for calibration estimation in the spirit of (2.5) is

$$\min_{\mathbf{w}} \sup_{u \in \mathcal{H}} \left\{ Q_N(\mathbf{w}; \mathbf{d}) + \lambda S_N(\mathbf{w}, u) \right\}, \tag{2.6}$$

where $S_N(\mathbf{w}, u)$ is defined in (2.4).

However, two issues would appear if we wish to optimize (2.6) directly. The first one is the scaling issue, i.e., for a constant c, we would have: $S_N(\mathbf{w}, cu) = c^2 S_N(\mathbf{w}, u)$. To get out of this dilemma, we can do the standardization procedure by scaling $S_N(\mathbf{w}, u)$ with the corresponding

empirical norm. By the following Cauchy-Schwarz inequality, we have

$$S_N(\mathbf{w}, u) = \left\{ \frac{1}{N} \sum_{i=1}^{N} (\delta_i w_i - 1) u(\mathbf{X}_i) \right\}^2 \leq \|u\|_N^2 \left\{ \frac{1}{N} \sum_{i=1}^{N} (\delta_i w_i - 1)^2 \right\},$$

where $\|u\|_N^2 = \frac{1}{N} \sum_{i=1}^{N} u(\mathbf{X}_i)^2$. Therefore, we can restrict our interest of $u(\cdot)$ within a normed sphere: $\tilde{\mathcal{H}}_N = \{u \in \mathcal{H} : \|u\|_N = 1\}$. The other issue is the overfitting problem. The infinite-dimensional reproducing kernel Hillbert space is too broad to measure such a distance and is very sensitive to the observations we have in the sample. Such issue can be handled by a penalized method. Specifically, we can penalize $\|\cdot\|_{\mathcal{H}}$ to control the possible overfitting of u. As long as the original survey weights do not wiggle too much, such penalization would work well. All discussions above lead to the following mini-max type optimization:

$$\min_{\mathbf{w} \geq \nu} \left[Q_N(\mathbf{w}; \mathbf{d}) + \lambda \sup_{u \in \tilde{\mathcal{H}}_N} \left\{ S_N(\mathbf{w}, u) - \tau \|u\|_{\mathcal{H}}^2 \right\} \right], \tag{2.7}$$

where $\lambda, \tau > 0$ are two tuning parameters. Here, $\mathbf{w} \geq \nu$ indicates $w_i \geq \nu$, for $i \in \mathcal{I}$. ν is a small positive value to ensure the positiveness of $\{w_i : i \in \mathcal{I}\}$. The equation (2.7) shows the similarities and differences between our proposed method and the original calibration method in (2.5). Roughly speaking, the traditional calibration approach can be understood as 'hard calibration', and our proposed method is similar to 'soft calibration', as mentioned in Davies (2018). If we divide (2.7) by λ, which would not affect the opimized result, and reparameterize the tuning paramters, we would have:

$$\min_{\mathbf{w} \geq \nu} \left[\sup_{u \in \tilde{\mathcal{H}}_N} \left\{ S_N(\mathbf{w}, u) - \lambda_1 \|u\|_{\mathcal{H}}^2 \right\} + \lambda_2 Q_N(\mathbf{w}; \mathbf{d}) \right], \tag{2.8}$$

where $\lambda_1, \lambda_2 > 0$ are 2 tuning parameters. We'll focus on (2.8) to formulate our estimator.

2.4.1 Theoretical Results

First of all, we'll list the technical assumptions for our theoretical results.The first four assumptions are about design part and the last three assumptions are relavant to superpopulation model.

(A1) Under Poisson sampling, for the survey weight, we assume that for $i \in \mathcal{S}$,

$$d_i = \mathcal{O}\left(\frac{N}{n}\right).$$

(A2) We assume that for any $u \in \mathcal{H}$,

$$\frac{\frac{1}{N}\sum_{i\in\mathcal{S}} d_i u(\mathbf{X}_i) - \frac{1}{N}\sum_{i\in\mathcal{U}} u(\mathbf{X}_i)}{\|u\|_N} = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right).$$

(A3) The quantity $\frac{d}{l} < 2$, where d is the dimension of $\mathbf{x}$ and l is the order of $\mathcal{H} = \mathcal{W}_2^l$.

(A4) The regression function $m(\cdot) \in \mathcal{H}$.

(A5) The error terms $\{\epsilon_i\}_{i=1}^N$ are uncorrelated, $\mathbb{E}(\epsilon_i) = 0$ and $\text{var}(\epsilon_i) = \sigma_i^2 \leq \sigma^2$. In addition, $\{\epsilon_i\}_{i=1}^N$ are independent with $\{\mathbf{x}_i\}_{i=1}^N$ and $\{\delta_i\}_{i=1}^N$.

Remark 2.1. *The first condition is a regular condition for survey weights order. The second condition is general for Horvitz-Tompson estimator. In our scenario, it's reasonable to assume the normalized difference between HT estimator the population mean is of* $\mathcal{O}_p(n^{-1/2})$

Remark 2.2. *The third condition is a regularity condition which facilitates our technical proof with entropy theory. The forth condition is for the superpopulation model, which lies in a reproducing kernel Hilbert space. The last assumption is a quite mild assumption for residuals.*

Theorem 2.1. *Suppose* $(A1) \sim (A5)$ *hold, if* $\lambda_1 \asymp n^{-1}$ *and* $\lambda_2 \asymp n^{\epsilon-1}$*, then*

$$\frac{1}{N}\sum_{i=1}^N \delta_i \hat{w}_i Y_i - \mathbb{E}(Y) = \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right). \tag{2.9}$$

where ϵ *is a constant larger than* 0.

During the derivation of Theorem 4.1, we get the following two facts:

$$T_1 = \frac{1}{N}\sum_{i=1}^N \{(\delta_i \hat{w}_i - 1)m(\mathbf{X}_i)\} \quad = \quad O_p\left(\frac{1}{\sqrt{n}}\right), \tag{2.10}$$

$$T_2 = \frac{1}{N}\sum_{i=1}^N \delta_i(\hat{w}_i - d_i)\epsilon_i \quad = \quad o_p\left(\frac{1}{\sqrt{n}}\right) \tag{2.11}$$

where (2.10) indicates the nonparametric approximation order and (2.11) shows the strengh of our proposed method. The resulting $\mathcal{O}_p(n^{-1/2})$ is due to the consistent survey design.

2.4.2 Nonparametric Regression Estimator

We can utilize the superpopulation model to extend the original estimator with more efficiency. In order to ensure the superpopulation model $m(\cdot)$ is in a reproducing-kernel Hilbert space, kernel ridge regression or smoothing spline mentioned in Gu (2013) might be a good candidate. Follow the difference estimator idea of Särndal et al. (2003), a modified estimator can be written in the form:

$$\hat{\theta}_{Nreg} := \frac{1}{N}\left[\sum_{i\in\mathcal{S}} \hat{w}_i \left\{Y_i - \hat{m}(\mathbf{X}_i)\right\} + \sum_{i\in\mathcal{U}} \hat{m}(\mathbf{X}_i)\right]. \tag{2.12}$$

Note that

$$\begin{aligned}\hat{\theta}_{Nreg} - \frac{1}{N}\sum_{i=1}^{N} y_i = \underbrace{\frac{1}{N}\sum_{i=1}^{N}(\delta_i\hat{w}_i - 1)\{m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i)\}}_{:=T_1^*} + \underbrace{\frac{1}{N}\sum_{i=1}^{N}\delta_i(\hat{w}_i - d_i)\epsilon_i}_{:=T_2} \\ + \underbrace{\left\{\frac{1}{N}\sum_{i=1}^{N}(\delta_i d_i - 1)\epsilon_i\right\}}_{:=T_3}.\end{aligned} \tag{2.13}$$

Thus, it remains to show (or investigate) the order of

$$T_1^* = \frac{1}{N}\sum_{i=1}^{N}(\delta_i\hat{w}_i - 1)h(\mathbf{x}_i) = o_p(n^{-1/2}) \tag{2.14}$$

where $h(\mathbf{x}_i) = m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i)$.

Theorem 2.2. *Suppose assumptions (A1) $\sim$ (A5), (B1) and the tuning parameter assumption in Lemma 2.6 hold, further assume $\lambda_1 \asymp n^k$, $\lambda_2 \asymp n^{\epsilon-1}$, where $-\frac{2l^2+ld+d}{(2l+1)d} < k < -1$, and $\epsilon < \frac{2l}{2l+1}$. Then $\tilde{t}_y$ asymptotically attains Godambe-Joshi lower bound in the sense that*

$$n\mathbb{E}\left(\hat{\theta}_{Nreg} - \frac{1}{N}\sum_{i=1}^{N} Y_i\right)^2 = \frac{n}{N^2}\sum_{i\in\mathcal{U}}\sigma_i^2\frac{1-\pi_i}{\pi_i} + o(1). \tag{2.15}$$

By (2.15), the asymptotic variance of $\hat{\theta}_{Nreg}$ is approximated by

$$V(\hat{e}_{HT}) = \frac{1}{N^2}\sum_{i\in U}\sum_{j\in U}(\pi_{ij} - \pi_i\pi_j)\frac{\epsilon_i}{\pi_i}\frac{\epsilon_j}{\pi_j},$$

where $\epsilon_i = y_i - m(\mathbf{x}_i)$. Thus, we can obtain the following linearization variance estimator

$$\hat{V}(\hat{\theta}_{Nreg}) = \frac{1}{N^2}\sum_{i\in A}\sum_{j\in A}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\frac{\hat{\epsilon}_i}{\pi_i}\frac{\hat{\epsilon}_j}{\pi_j},$$

where $\hat{e}_i = y_i - \hat{m}(\mathbf{x}_i)$, as a consistent variance estimator of $\hat{\theta}_{Nreg}$.

2.5 Computational Details

2.5.1 Optimization

Specifically, the inner part of (2.8) can be rewritten as:

$$\sup_{u\in\mathcal{H}}\left\{\frac{S_N(\mathbf{w},u)}{\|u\|_N^2} - \lambda_1\frac{\|u\|_{\mathcal{H}}^2}{\|u\|_N^2}\right\}. \tag{2.16}$$

By representer theorem in Wahba (1990), we can easily arrive the conclusion that the optimization of the above objective function have finite-dimensional representation by $\text{span}\{K(\mathbf{X}_j,\cdot) : j \in \mathcal{U}\}$. In addition, we can define the Gram matrix: $M = (K(\mathbf{X}_i,\mathbf{X}_j))_{N\times N} \in \mathbb{R}^{N\times N}$. For notational convenience, we further have the eigenvalue decompostion for M:

$$M = \begin{pmatrix} P_1 & P_2 \end{pmatrix}\begin{pmatrix} Q_1 & 0 \\ 0 & Q_2 \end{pmatrix}\begin{pmatrix} P_1^{\mathrm{T}} \\ P_2^{\mathrm{T}} \end{pmatrix},$$

where Q_1 is the diagonal matrix with all r non-zero diagonal elements. In addition, $Q_2 = \mathbf{0}$. Also notice that if $r = N$, such Q_2 would disappear.

By representer theorem, we can express (2.16) as:

$$\sup_{\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_N)^{\mathrm{T}}\in\mathbb{R}^N}\left[\frac{S_N\left\{\mathbf{w},\sum_{j\in\mathcal{U}}\alpha_j K(\mathbf{X}_j,\cdot)\right\}}{\alpha^{\mathrm{T}}M^2\boldsymbol{\alpha}/N} - \lambda_1\frac{\boldsymbol{\alpha}^{\mathrm{T}}M\boldsymbol{\alpha}}{\boldsymbol{\alpha}^{\mathrm{T}}M^2\boldsymbol{\alpha}/N}\right]. \tag{2.17}$$

In addition, the empirical validity measure in this case can be represented as:

$$S_N\left\{\mathbf{w},\sum_{j\in\mathcal{U}}\alpha_j K(\mathbf{X}_j,\cdot)\right\} = \frac{1}{N^2}\boldsymbol{\alpha}^{\mathrm{T}}MA(\mathbf{w})M\boldsymbol{\alpha},$$

where $A(\mathbf{w}) = a(\mathbf{w})a(\mathbf{w})^{\mathrm{T}}$ with $a(\mathbf{w}) = (\delta_1 w_1 - 1,\ldots,\delta_N w_N - 1)^{\mathrm{T}}$. Additionally, define $\boldsymbol{\beta} = Q_1 P_1^{\mathrm{T}}\boldsymbol{\alpha}$, the optimization problem in (2.17) can be expressed as:

$$\sup_{\boldsymbol{\beta}\in\mathbb{R}^r:\|\boldsymbol{\beta}\|\le 1}\boldsymbol{\beta}^{\mathrm{T}}\left\{\frac{1}{N}P_1^{\mathrm{T}}A(\mathbf{w})P_1 - N\lambda_1 Q_1^{-1}\right\}\boldsymbol{\beta}.$$

Therefore, the final optimization problem can be rewritten as:

$$\min_{\mathbf{w}\geq\nu}\left[\sigma_{\max}\left\{\frac{1}{N}P_1^{\mathrm{T}}A(\mathbf{w})P_1 - N\lambda_1 Q_1^{-1}\right\} + \lambda_2 Q_N(\mathbf{w};\mathbf{d})\right]. \tag{2.18}$$

It should be noted that we only do the optimization for $\{w_i : i \in \mathcal{I}\}$. In addition, as $P_1^T a(\mathbf{w})$ is an affine transformation of $\mathbf{w}$, Slater's condition and convexity analysis can confirm the global minimum of the above function. If we only have largest eigenvalue with multiplicity 1, the inner part of (2.18) is differentiable and the corresponding gradient has a closed form. As a reult, the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm with bound constraints (L-BFGS-B) can be directly applied. Further, mathematically there might be two largest eigenvalues. A two-part computational strategy can be employed. Further computational details can be found in Overton (1992) and Wong and Chan (2018).

2.5.2 Tuning Parameter Selection

We modify the tuning parameter selection idea from Wong and Chan (2018). First of all, to lower the computational burden, we won't tune a 2D grid for $\lambda_1 \times \lambda_2$. Comparing with the traditional calibration Lagrangian multiplier, we would set $\lambda_2 = n^{\epsilon-1}$, where $\epsilon = 0.1$. After the above argument, what we need to tune is just the parameter λ_1. As a result, we might use the functional approximation measure to find a reasonable solution.

Still, consider the inner optimization part in (2.8): $\sup_{u\in\tilde{\mathcal{H}}_N}\left\{S_N(\mathbf{w},u) - \lambda_1\|u\|_{\mathcal{H}}^2\right\}$. The Largrangian multiplier implies that the above optimization is equvalent to $\sup_{\{u\in\tilde{\mathcal{H}}_N:\|u\|_{\mathcal{H}}\leq\gamma\}} S_N(\mathbf{w},u)$ for some γ and there's relationship between λ_1 and γ. The following functional approximation measure:

$$B_N(\mathbf{w}) = \sup_{\{u\in\tilde{\mathcal{H}}_N:\|u\|_{\mathcal{H}}\leq\gamma\}} S_N(\mathbf{w},u) \tag{2.19}$$

can be viewed as the measure of errors on the set $\mathcal{A}_\gamma := \{u \in \tilde{\mathcal{H}}_N : \|u\|_{\mathcal{H}} \leq \gamma\}$ given $\mathbf{w}$. When γ is large, i.e., λ_1 is small, $\mathcal{A}_\gamma$ is fairly large. Therefore, in this scenario, doing the maximization over large set implies that $B_N(\mathbf{w})$ is also fairly large. On the other hand, as γ goes down towards 0, $B_N(\mathbf{w})$ also goes down to 0, where λ_1 goes to ∞. Therefore, we can select a λ_1 such that

$B_N(\mathbf{w})$ is closed to 0 enough. Additionally, suppose the above situation happens, further decrease of γ or increase of λ won't bring us significant balancing results. Therefore, practically, we can select a sequence of $\lambda_1 : 0 < \lambda_1^{(1)} < \lambda_1^{(2)} < \cdots < \lambda_1^{(J)}$ and the optimal index j^* can be select as:

$$j^* = \inf_j \left\{ j \in \{1, \ldots, J-1\} : \frac{B_N(\hat{\mathbf{w}}^{(j+1)}) - B_N(\hat{\mathbf{w}}^{(j)})}{\lambda_{j+1} - \lambda_j} \geq \alpha \right\}, \tag{2.20}$$

where $\hat{\mathbf{w}}^{(j+1)}$ is the optimization results from (2.18) whose $\lambda_1 = \lambda_1^{(j)}$, for $j = 1, \ldots, J$. In our numerical experiment, we set $\alpha = -10^{-6}$.

2.6 Simulation Study

Simulation studies have been done to measure the finite sample performance of the proposed estimator. In this section we basically have three simulation setups. The first part is in favor of traditional calibration approach. In the second simulation study, we compare the results between the two aforementioned estimators when the traditional calibration estimators are jeopardized to show the robustness of our kernel-based method.

2.6.1 Simulation Setting I

Specifically, we assume the finite poplation size $N = 2000$, the expectation of probability sample size is $n = 100$ and $n = 200$. In addition, assume our superpopulation model:

$$y_i = 270 + 27.4x_{i1} + 13.7x_{i2} + 13.7x_{i3} + 13.7x_{i4} + \sigma\epsilon_i, i = 1, \ldots, N, \tag{2.21}$$

where $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})^{\mathrm{T}} \overset{i.i.d}{\sim} \mathcal{N}(\mathbf{0}, I_{4\times4})$. Also, such $\epsilon \overset{i.i.d}{\sim} \mathcal{N}(0, 1)$ and $\sigma = 5$. The parameter we're interested in is $\mu_y = N^{-1}\sum_{i=1}^{N} y_i$. For the corresponding probability sample $\mathcal{S}$, we generate them by Poisson sampling with inclusion probability $\pi_i = \frac{n}{N}$.

2.6.2 Simulation Setting II

In this simulation setup, we keep the same population size and sample size as above. Here we assume our superpopulation model as:

$$y_i = 210 + 27.4m_1(\mathbf{x}_i) + 13.7m_2(\mathbf{x}_i) + 13.7m_3(\mathbf{x}_i) + 13.7m_4(\mathbf{x}_i) + \sigma\epsilon_i, i = 1, \ldots, N, \tag{2.22}$$

where $\mathbf{X}_i = (X_{i1}, X_{i2}, X_{i3}, X_{i4})^{\mathrm{T}} \overset{i.i.d}{\sim} \mathcal{N}(\mathbf{0}, I_{4\times4})$. Additionally, $\{\epsilon_i\}_{i=1}^N$ and σ is choosen the same as before. Additionally, we have

$m_1(\mathbf{x}_i) = \exp(x_{i1}/2), m_2(\mathbf{x}_i) = \frac{x_{i2}}{1+\exp(x_{1i})}, m_3(\mathbf{x}_i) = \left(\frac{x_{i1}x_{i3}}{25} + 0.6\right)^3$, and $m_4(\mathbf{x}_i) = (x_{i2} + x_{i4} + 5)^2$.

For the corresponding probability sample $\mathcal{S}$, we generate them by Poisson sampling method with inclusion probability same as the previous simulation settings.

2.6.3 Simulation Results

The simulation results are presented at the following table 2.1:

Table 2.1: Results for Simulation I and II: bias, standard error, and root mean square error for nonlinear case based on sample of size $n = 100$ and $n = 200$ of a fixed population of size $N = 2000$

Simulation	Estimator	n					
		100			200		
		Bias	SE	RMSE	Bias	SE	RMSE
I	HT	0.1649	3.4999	3.5038	0.1115	2.3206	2.3233
	Calibration	0.0304	0.5254	0.5263	0.0123	0.3414	0.3417
	$\hat{\mu}_1$	0.0299	0.6516	0.6523	0.0121	0.4686	0.4687
	$\hat{\mu}_2$	0.0394	0.6453	0.6465	0.0160	0.4676	0.4679
	Neural_Net	0.0171	0.7821	0.7823	0.0058	0.4532	0.4532
	Smoothing	0.0335	0.5262	0.5272	0.0142	0.3435	0.3438
II	HT	0.3527	20.4784	20.4814	0.4447	13.7036	13.7108
	Calibration	-0.5080	3.7909	3.8248	-0.1923	2.6303	2.6373
	$\hat{\mu}_1$	-0.6216	1.0380	1.2099	-0.2199	0.5463	0.5889
	$\hat{\mu}_2$	-0.4460	0.7370	0.8614	-0.1725	0.4924	0.5218
	Neural_Net	-0.7990	3.1193	3.2200	-0.3498	1.6897	1.7255
	Smoothing	-0.1520	3.2435	3.2471	-0.0175	2.0224	2.0225

where *HT* denotes the Horvitz-Tompson estimator, *Calibration* denotes the traditional calibration estimator. $\hat{\mu}_1$ denotes the estimator generated by (2.18) and $\hat{\mu}_2$ denotes the estimator generated by (2.12). *Neural_Net* denotes the neural network estimator mentioned in Montanari and Ranalli (2005) tuned with five-fold cross validation. *Smoothing* denote the penalized splines estimators in Breidt et al. (2005), where we applied generalized additive model with cubic spline with knots $\min\{n/(4d), 35\}$ for each covariate.

As we can see, in first simulation setting, as it's in favor of traditional calibration method, also, the true model is a sub-model of penalized spline method, that's why *Calibration* and *Smoothing* behave best. Our proposed estimators are fairly comparable with those best two in the sense of RMSE. Also, for HT estimator, as there are no auxiliary information, it behaves worst among them. Also, the neural network method behaves worst except HT estimator. In second simulation setting, which is a more general case, the true underlying model is additive with respect to unknown nonlinear transformation of each dimension of the auxiliary information we have. In the sense of RMSE, the smoothing spline method behaves best, while our method is comparable. In third simulation setting, which is a more complex case, the true underlying model is unknown nonlinear transformation of the auxiliary information we have. In the sense of RMSE, our proposed estimators behave best.

2.7 Application

We compare our proposed method with other methods in Swiss municipalities population dataset, which was collected in 2003. We use Poisson sampling method with equal probability with expectation sample size 500 to sample from 2896 municipalities. We are interested in estimating the total population of the whole nation with other covariates. In particular, for each municipality, we have their municipality area, wood area, area under cultivation, mountain pasture area, area with buildings and industrial area. Table 2.2 shows the estimated national population and the corresponding 95% confidence interval via each method. The true population is 7288010. Correspondingly, Figure 2.1 shows the results graphically. Nearly all methods covers this number except smoothing spline method. Additionally, our proposed method show less bias and narrower confidence interval, which indicates the possibly nonlinear relationship among population and other covariates.

Table 2.2: Esimated national population

Estimator	Estimate	95%Confidence Interval
Hortvitz-Tompson	8369370	(4299597, 12439144)
Calibration	7466609	(6809103, 8124114)
Smoothing	6624187	(6348489, 6899886)
Neural_Net	7196561	(6612246, 7780877)
$\hat{\mu}_2$	7255057	(7063286, 7446827)

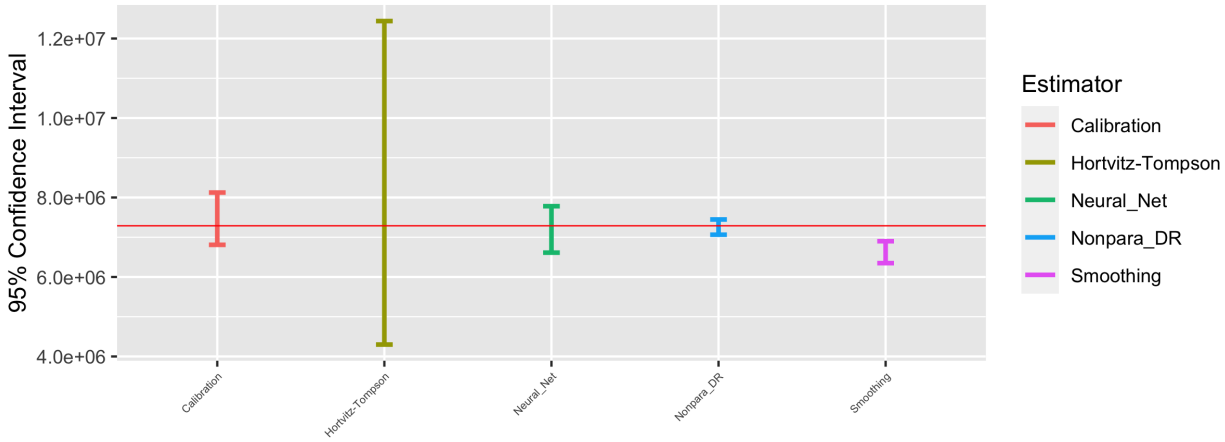


Figure 2.1: Esimated national population with 95% confidence interval, where horizontal red line is the true population.

2.8 References

Birman, M. S. and Solomyak, M. Z. (1967). Piecewise-polynomial approximations of functions of the classes w_pˆα. *Matematicheskii Sbornik*, 115(3):331–355.

Breidt, F., Claeskens, G., and Opsomer, J. (2005). Model-assisted estimation for complex surveys using penalised splines. *Biometrika*, 92(4):831–846.

Breidt, F. J. and Opsomer, J. D. (2000). Local polynomial regresssion estimators in survey sampling. *The Annals of Statistics*, 28(4):1026–1053.

Breidt, F. J. and Opsomer, J. D. (2017). Model-assisted survey estimation with modern prediction techniques. *Statistical Science*, 32(2):190–205.

Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1976). Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63(3):615–620.

Davies, G. (2018). *Examination of approaches to calibration in survey sampling*. PhD thesis, Cardiff University.

Deville, J.-C. and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87(418):376–382.

Firth, D. and Bennett, K. (1998). Robust models in probability sampling. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(1):3–21.

Fuller, W. A. (2009). *Introduction to statistical time series*, volume 428. John Wiley & Sons.

Goga, C. (2005). Réduction de la variance dans les sondages en présence d'information auxiliarie: Une approache non paramétrique par splines de régression. *Canadian Journal of Statistics*, 33(2):163–180.

Gu, C. (2013). *Smoothing spline ANOVA models*, volume 297. Springer Science & Business Media.

Hellerstein, J. K. and Imbens, G. W. (1999). Imposing moment restrictions from auxiliary data by weighting. *Review of Economics and Statistics*, 81(1):1–14.

Huang, E. T.-J. H. (1978). Nonnegative regression estimation for sample survey data.

Isaki, C. T. and Fuller, W. A. (1982). Survey design under the regression superpopulation model. *Journal of the American Statistical Association*, 77(377):89–96.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, 78(1):21–39.

Lin, Y. (2000). Tensor product space anova models. *The Annals of Statistics*, 28(3):734–755.

Montanari, G. E. and Ranalli, M. G. (2005). Nonparametric model calibration estimation in survey sampling. *Journal of the American Statistical Association*, 100(472):1429–1442.

Overton, M. L. (1992). Large-scale optimization of eigenvalues. *SIAM Journal on Optimization*, 2(1):88–120.

Särndal, C.-E., Swensson, B., and Wretman, J. (2003). *Model assisted survey sampling*. Springer Science & Business Media.

van de Geer, S. A. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge university press.

Wahba, G. (1990). *Spline models for observational data*, volume 59. SIAM.

Wong, R. K. and Chan, K. C. G. (2018). Kernel-based covariate functional balancing for observational studies. *Biometrika*, 105(1):199–213.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96(453):185–193.

Zhang, Y., Duchi, J., and Wainwright, M. (2013). Divide and conquer kernel ridge regression. In *Conference on learning theory*, pages 592–617.

2.9 Appendix: Technical Details

For convenience, we'll use c with subscript as a constant that is larger than 0. Lemma 2.1 would automatically hold after we get Lemma 2.5. Then Theorem 2.1 can be obtained naturally. The proof of Theorem 2.2 is at the end of this document.

2.9.1 Proof of Theorem 2.1

Lemma 2.1. *Let* $\hat{\mathbf{w}}$ *be the solution to* (2.8)*. Assume* $\lambda_1 \asymp n^{-1}$ *and* $\lambda_2 \asymp n^{\epsilon-1}$ *for some* $\epsilon > 0$*. Then, under assumptions* $[A1] \sim [A4]$*, we have*

1. $S_N(\hat{\mathbf{w}}, m) = \mathcal{O}_p(\frac{1}{n}) \|m\|_N$.
2. *There exist constants* $W > 0$ *and* $S^2 > 0$ *such that* $\mathbb{E}\{n^{\epsilon} Q_N(\hat{\mathbf{w}}; \mathbf{d})\} \leq W$ *and* $\mathbb{E}\{n S_N(\hat{\mathbf{w}}, m)\} \leq S^2$.

Lemma 2.1 states the convergece rate for $(T1)$ in (2.3), and the boundedness of expectation for $(T2)$ therein. Literally, Lemma 2.1 just makes all ingredients ready for derivation of Theorem 4.1. To prove Lemma 2.1, we use the following Lemmas.

Lemma 2.2. *Suppose* $\lambda_1 = \mathcal{O}\left(n^{-1}\right)$ *and* $\lambda_2 \asymp n^{\epsilon-1}$ *for* $\epsilon > 0$*, a legitimate solution* $\{\hat{w}_i : i \in \mathcal{S}\}$ *should satisfy*

$$\hat{w}_i = d_i + o_p\left(\frac{N}{n}\right). \tag{2.23}$$

Moreover, we have $\hat{w}_i = d_i + \mathcal{O}_p(n^{-\epsilon/2-1}N)$

Proof of Lemma 2.2. First of all, we claim that $\{d_i : i \in \mathcal{S}\}$ is a good candidate of our solution. Define the function $F_{N,\lambda_1,\lambda_2}(\cdot)$ as the following form:

$$F_{N,\lambda_1,\lambda_2}(\mathbf{w}) = \sup_{u\in\tilde{\mathcal{H}}_N} \left\{ S_N(\mathbf{w},u) - \lambda_1 \|u\|_{\mathcal{H}}^2 \right\} + \lambda_2 Q_N(\mathbf{w};\mathbf{d}) \tag{2.24}$$

Then, we have

$$\begin{aligned} F_{N,\lambda_1,\lambda_2}(\mathbf{d}) &= \sup_{u\in\tilde{\mathcal{H}}_N} \left\{ S_N(\mathbf{d},u) - \lambda_1 \|u\|_{\mathcal{H}}^2 \right\} \\ &\lesssim \sup_{u\in\mathcal{H}} \frac{S_N(\mathbf{d},u)}{\|u\|_N^2} + \mathcal{O}\left(\frac{1}{n}\right) \\ &= \sup_{u\in\mathcal{H}} \left[\frac{\frac{1}{N}\sum_{i\in\mathcal{S}} d_i u(\mathbf{X}_i) - \frac{1}{N}\sum_{i\in\mathcal{U}} u(\mathbf{X}_i)}{\|u\|_N} \right]^2 + \mathcal{O}\left(\frac{1}{n}\right) \\ &= \mathcal{O}\left(\frac{1}{n}\right), \end{aligned} \tag{2.25}$$

where the last inequality holds by Assumption 3. However, suppose for any $\mathbf{w}$ satisfying have $w_i = d_i + \mathcal{O}_p\left(n^{-1}N\right)$, we would have

$$\lambda_2 Q_N(\mathbf{w};\mathbf{d}) \;=\; \mathcal{O}\left(n^{\epsilon-1}\right) \times \frac{1}{N} \times \mathcal{O}\left(n\right) \times \mathcal{O}_p\left(\frac{N}{n}\right) = \mathcal{O}_p\left(n^{\epsilon-1}\right),$$

which implies that such $\mathbf{w}$ is not a good solution compared with (2.25) , whose effect cannot match with $\mathbf{d}$. Therefore, we arrive the conclusion that $\hat{w}_i = d_i + o_p\left(n^{-1}N\right)$, for $i \in \mathcal{S}$. Specifically, one can observe that a good candidate $\mathbf{w}$ has to satisfy $\lambda_2 Q_N(\mathbf{w};\mathbf{d}) \lesssim \mathcal{O}_p\left(n^{-1}\right)$, which implies $\hat{w}_i = d_i + \mathcal{O}_p\left(n^{-\epsilon/2-1}N\right)$. □

Lemma 2.3. *Suppose assumptions [A1] ∼ [A3] hold. Let* $\mathbf{w}^\star = (w_1^\star, \ldots, w_N^\star)^{\mathrm{T}} = (1/\pi(\mathbf{x}_1), \ldots, 1/\pi(\mathbf{x}_N))^{\mathrm{T}}$. *Then there exists a constant* $c > 0$ *such that* $\forall T \geq c$,

$$\mathbb{P}\left\{ \sup_{u\in\tilde{\mathcal{H}}_N} \frac{N S_N(\mathbf{w}^*,u)}{\|u\|_{\mathcal{H}}^{\frac{d}{l}}} \geq T^2 \right\} \leq c\exp\left(-\frac{T^2}{c}\right). \tag{2.26}$$

Proof of Lemma 2.3 . For given $\mathbf{w}^*$, let $\gamma_i = \delta_i w_i^* - 1$, for $i = 1, \ldots, N$. Then, we have $\mathbb{E}(\gamma_i|\mathbf{X}_i) = \mathbb{E}(\delta_i w_i^* - 1|\mathbf{X}_i) = 0$. Further, by Poisson sampling scheme, we have the conditional

independence between In addition, by the order of d_i and w_i^*, we may assume that there exists constant $\tilde{c}_1$ and $\tilde{c}_2$ such that $\mathbb{P}(\delta_i = 1) \leq \tilde{c}_1 n/N$ and $\sup_{i=1,\ldots,N} w_i^* \leq c_2 N/n$. Then we have

$$\begin{aligned}
\mathbb{E}(\gamma_i^{2k}) &= \mathbb{E}_{\mathbf{X}_i}\left[\mathbb{E}\left\{(\delta_i w_i^* - 1)^{2k} | \mathbf{X}_i\right\}\right] \\
&= \mathbb{E}_{\mathbf{X}_i}\left[\mathbb{E}\left\{(\delta_i w_i^*)^{2k} | \mathbf{X}_i\right\}\right] \\
&= \mathbb{E}_{\mathbf{X}_i}\left[\mathbb{P}(\delta_i = 1)\mathbb{E}\left\{(w_i^*)^{2k}\right\}\right] \\
&\leq \tilde{c}_1 \tilde{c}_2^{2k} \left(\frac{N}{n}\right)^{2k-1} \\
&\leq \frac{(2k)!}{2^k k!}\left\{\frac{N}{n}(\tilde{c}_1 \vee 1)(\tilde{c}_2 \vee 1)\right\}^{2k}
\end{aligned} \tag{2.27}$$

Therefore, $\{\gamma_i\}_{i=1}^N$ are uniformly subgaussian. Thus, there exists constants K and σ_0^2, such that:

$$\max_{i \in U} K^2 \left[\mathbb{E}\left(e^{\frac{\gamma_i^2}{K^2}} | \{\mathbf{X}_i\}_{i=1}^N\right) - 1\right] \leq \sigma_0^2 \tag{2.28}$$

By (A4) and Birman and Solomyak (1967), there exists a constant A, for any $\xi > 0$, s.t:

$$H_\infty(\xi, \{u \in \mathcal{H} : \|u\|_{\mathcal{H}} \leq 1\}) \leq A\xi^{-\frac{d}{l}}, \tag{2.29}$$

where $H_\infty(\delta, \mathcal{G})$ is the $\delta - uniform\ entropy$ of $\mathcal{G}$, a set of functions. Specifically, let $\delta > 0$ and $N_\infty(\delta, \mathcal{G})$ be the smallest values of N such that there exists $\{g_j\}_{j=1}^\infty \subset \mathcal{G}$ such that $\sup_{g \in \mathcal{G}} \min_{j=1,\ldots,N} |g_i - g_j|_\infty \leq \delta$, then $H_\infty(\delta, \mathcal{G}) = \log N_\infty(\delta, \mathcal{G})$. Thus, by Lemma 2.1 of (Lin, 2000) and lemma 8.4 in (van de Geer, 2000), for some constants, w.l.o.g., we can obtain that there exists a positive constant c, depending on $A, d, l, R, K, \sigma_0^2$, for all $T > 0$, satisfy

$$\mathbb{P}\left(\sup_{g \in \mathcal{H}} \frac{|\frac{1}{\sqrt{N}} \sum_{i=1}^N \gamma_i g(\mathbf{X}_i)|}{\|g\|_N^{1-\frac{d}{2l}}} \geq T \,\middle|\, \{\mathbf{X}_i\}_{i=1}^N\right) \leq c \exp\left(-\frac{T^2}{c^2}\right). \tag{2.30}$$

Further, as $\{\mathbf{X}_i\}_{i=1}^N$ is independent of c, we have

$$\mathbb{P}\left(\sup_{g \in \mathcal{H}} \frac{|\frac{1}{\sqrt{N}} \sum_{i=1}^N \gamma_i g(\mathbf{X}_i)|}{\|g\|_N^{1-\frac{d}{2l}}} \geq T\right) = \mathbb{P}\left(\sup_{g \in \mathcal{H}} \frac{|\frac{1}{\sqrt{N}} \sum_{i=1}^N \gamma_i g(\mathbf{X}_i)|}{\|g\|_N^{1-\frac{d}{2l}}} \geq T \,\middle|\, \{\mathbf{X}_i\}_{i=1}^N\right). \tag{2.31}$$

Additionally, by definition of $\{\gamma_i, i \in \mathcal{S}\}$, $\tilde{\mathcal{H}}_N$ and $S_N(\cdot,\cdot)$, we have,

$$\begin{aligned}\left\{\sup_{g\in\mathcal{H}} \frac{|\frac{1}{\sqrt{N}}\sum_{i=1}^N \gamma_i g(\mathbf{X}_i)|}{\|g\|_N^{1-\frac{d}{2l}}} \geq T\right\} &= \left\{\sup_{u\in\tilde{\mathcal{H}}_N} \frac{|\frac{1}{\sqrt{N}}\sum_{i=1}^N \gamma_i u(\mathbf{X}_i)|}{\|u\|_{\mathcal{H}}^{\frac{d}{2l}}} \geq T\right\} \\ &= \left\{\sup_{u\in\tilde{\mathcal{H}}_N} \frac{\frac{1}{N}\left\{\sum_{i=1}^N(\delta_i w_i^* - 1)u(\mathbf{X}_i)\right\}^2}{\|u\|_{\mathcal{H}}^{\frac{d}{l}}} \geq T^2\right\} \\ &= \left\{\sup_{u\in\tilde{\mathcal{H}}_N} \frac{NS_N(\mathbf{w}^*,u)}{\|u\|_{\mathcal{H}}^{\frac{d}{l}}} \geq T^2\right\}. \end{aligned} \tag{2.32}$$

□

Lemma 2.4. *Suppose assumptions [A1]∼[A4] hold, and* $\lambda_1 \asymp n^{-1}$ *and* $\lambda_2 \asymp n^{\epsilon-1}$*, for* $\epsilon > 0$*, we have* $S_N(\hat{\mathbf{w}}, u) = \mathcal{O}_p(n^{-1})$ *and* $Q_N(\hat{\mathbf{w}}; \mathbf{d}) = o_p(1)$*. Then, there exists a constant* $W > 0$*, s.t.* $\mathbb{E}\{n^\epsilon Q_N(\hat{\mathbf{w}}, \mathbf{d})\} \leq W$.

Proof of Lemma 2.4 . First of all, we introduce the basic inequality for our penalized method. Let $u^* = \arg\max_{u\in\tilde{\mathcal{H}}_N}\left\{S_N(\mathbf{w}^*, u) - \lambda_1\|u\|_{\mathcal{H}}^2\right\}$. For $f \in \tilde{\mathcal{H}}_N$, we have

$$\begin{aligned} S_N(\hat{\mathbf{w}}, f) - \lambda_1\|f\|_{\mathcal{H}}^2 + \lambda_2 Q_N(\hat{\mathbf{w}}; \mathbf{d}) &\leq F_{N,\lambda_1,\lambda_2}(\hat{\mathbf{w}}) \\ &\leq F_{N,\lambda_1,\lambda_2}(\mathbf{w}^*) \\ &\leq S_N(\mathbf{w}^*, u^*) - \lambda_1\|u^*\|_{\mathcal{H}}^2 + \lambda_2 Q_N(\mathbf{w}^*; \mathbf{d}) \end{aligned} \tag{2.33}$$

Additionally, if $\|u\|_N = 0$, we would have: $S_N(\hat{\mathbf{w}}, u) = 0$ for any $u \in \mathcal{H}$. Therefore, $\forall u \in \mathcal{H}$, we would have

$$S_N(\hat{\mathbf{w}}, u) - \lambda_1\|u\|_{\mathcal{H}}^2 + \lambda_2 Q_N(\hat{\mathbf{w}}; \mathbf{d})\|u\|_N^2 \leq \left\{S_N(\mathbf{w}^*, u^*) - \lambda_1\|u^*\|_{\mathcal{H}}^2 + \lambda_2 Q_N(\mathbf{w}^*; \mathbf{d})\right\}\|u\|_N^2 \tag{2.34}$$

The basic inequality (2.33) can be rearranged as:

$$S_N(\hat{\mathbf{w}}, f) + \lambda_1\|u^*\|_{\mathcal{H}}^2 + \lambda_2 Q_N(\hat{\mathbf{w}}; \mathbf{d}) \leq S_N(\mathbf{w}^*, u^*) + \lambda_1\|f\|_{\mathcal{H}}^2 + \lambda_2 Q_N(\mathbf{w}^*; \mathbf{d}). \tag{2.35}$$

We'll define a few events partition for the whole event space. First of all, let

$$\begin{aligned}
\tilde{G}_{N,1} &= \{S_N(\mathbf{w}^*,u^*) \text{ is the largest among } S_N(\mathbf{w}^*,u^*), \lambda_1\|f\|_{\mathcal{H}}^2 \text{ and } \lambda_2 Q_N(\mathbf{w}^*;\mathbf{d})\} \\
\tilde{G}_{N,2} &= \{\lambda_1\|f\|_{\mathcal{H}}^2 \text{ is the largest among } S_N(\mathbf{w}^*,u^*), \lambda_1\|f\|_{\mathcal{H}}^2 \text{ and } \lambda_2 Q_N(\mathbf{w}^*;\mathbf{d})\} \\
\tilde{G}_{N,3} &= \{\lambda_2 Q_N(\mathbf{w}^*;\mathbf{d}) \text{ is the largest among } S_N(\mathbf{w}^*,u^*), \lambda_1\|f\|_{\mathcal{H}}^2 \text{ and } \lambda_2 Q_N(\mathbf{w}^*;\mathbf{d})\}
\end{aligned} \tag{2.36}$$

Additionally, we define $G_{N,1} = \tilde{G}_{N,1}$, $G_{N,2} = \tilde{G}_{N,1} \setminus \tilde{G}_{N,2}$, $G_{N,3} = \tilde{G}_{N,3} \setminus \left(\tilde{G}_{N,1} \bigcup \tilde{G}_{N,2}\right)$ We'll further divide $G_{N,1}$ into two parts:

$$G_{N,1,T} = G_{N,1} \bigcap \left\{ S_N(\mathbf{w}^*,u^*) \le \frac{1}{N} T^2 \|u^*\|_{\mathcal{H}}^{\frac{d}{l}} \right\}$$
$$\mathring{G}_{N,1,T} = G_{N,1} \bigcap \left\{ S_N(\mathbf{w}^*,u^*) > \frac{1}{N} T^2 \|u^*\|_{\mathcal{H}}^{\frac{d}{l}} \right\}$$

Finally, we arrive a partition of the whole space $\Omega = G_{N,1,T} \dot{\bigcup} \mathring{G}_{N,1,T} \dot{\bigcup} G_{N,2} \dot{\bigcup} G_{N,3}$

Then, we'll discuss 3 scenarios on $G_{N,1,T}$, $G_{N,2}$ and $G_{N,3}$.

Case 1. On $G_{N,1,T}$, by (2.35), we have

$$\begin{aligned}
& \lambda_1\|u^*\|_{\mathcal{H}}^2 \le 3S_N(\mathbf{w}^*,u^*) \le 3\frac{1}{N}T^2\|u^*\|_{\mathcal{H}}^{\frac{d}{l}} \\
\Leftrightarrow \quad & \|u^*\|_{\mathcal{H}} \le 3^{\frac{l}{2l-d}} (\lambda_1 N)^{\frac{l}{d-2l}} T^{\frac{2l}{2l-d}} \\
\Leftrightarrow \quad & \|u^*\|_{\mathcal{H}}^{\frac{d}{l}} \le 3^{\frac{d}{2l-d}} (\lambda_1 N)^{\frac{d}{d-2l}} T^{\frac{2d}{2l-d}} \\
& := c_1 (\lambda_1 N)^{\frac{d}{d-2l}} T^{\frac{2d}{2l-d}}
\end{aligned} \tag{2.37}$$

On the other hand, we have

$$\begin{aligned}
S_N(\hat{\mathbf{w}},f) &\le 3S_N(\mathbf{w}^*,u^*) \\
&\le 3N^{-1}T^2\|u^*\|_{\mathcal{H}}^{\frac{d}{l}} \\
&\le 3c_1\lambda_1^{-\frac{d}{2l-d}} N^{\frac{2l}{d-2l}} T^{\frac{4l}{2l-d}} \\
&:= c_2\lambda_1^{-\frac{d}{2l-d}} N^{\frac{2l}{d-2l}} T^{\frac{4l}{2l-d}}
\end{aligned} \tag{2.38}$$

Similarly, we have:

$$\lambda_2 Q_N(\hat{\mathbf{w}};\mathbf{d}) \le c_2\lambda_1^{-\frac{d}{2l-d}} N^{\frac{2l}{d-2l}} T^{\frac{4l}{2l-d}} \tag{2.39}$$

Case 2. In a similar fashion, on $G_{N,2}$, we would have:

$$\begin{aligned} S_N(\hat{\mathbf{w}}, f) &\leq 3\lambda_1 \|f\|_{\mathcal{H}}^2 \\ \|u^*\|_{\mathcal{H}} &\leq \sqrt{3}\|f\|_{\mathcal{H}} \\ \lambda_2 Q_N(\hat{\mathbf{w}}; \mathbf{d}) &\leq 3\lambda_1 \|f\|_{\mathcal{H}}^2 \end{aligned} \tag{2.40}$$

Case 3. On $G_{N,3}$, we have:

$$\begin{aligned} S_N(\hat{\mathbf{w}}, f) &\leq 3\lambda_2 Q_N(\mathbf{w}^*; \mathbf{d}) \\ \lambda_1 \|u^*\|_{\mathcal{H}}^2 &\leq 3\lambda_2 Q_N(\mathbf{w}^*; \mathbf{d}) \\ \lambda_2 Q_N(\hat{\mathbf{w}}; \mathbf{d}) &\leq 3\lambda_2 Q_N(\mathbf{w}^*; \mathbf{d}) \end{aligned} \tag{2.41}$$

By Lemma 2.3, for some constant $c > 0$, $\forall T \geq c$, we have:

$$\mathbb{P}\left\{ S_N(\mathbf{w}^*, u^*) \leq \frac{1}{N} T^2 \|u^*\|_{\mathcal{H}}^{\frac{d}{l}} \right\} \geq 1 - c\exp\left(-\frac{T^2}{c^2}\right). \tag{2.42}$$

Therefore, we would have:

$$\begin{aligned} & \mathbb{P}\left[S_N(\hat{\mathbf{w}}, f) \leq \max\left\{ c_2 \lambda_1^{-\frac{d}{2l-d}} N^{\frac{2l}{d-2l}} T^{\frac{4l}{2l-d}}, 3\lambda_1 \|f\|_{\mathcal{H}}^2, 3\lambda_2 Q_N(\mathbf{w}^*; \mathbf{d}) \right\} \right] \\ = & \sum_{i=1}^{3} \mathbb{P}\left(\left[S_N(\hat{\mathbf{w}}, f) \leq \max\left\{ c_2 \lambda_1^{-\frac{d}{2l-d}} N^{\frac{2l}{d-2l}} T^{\frac{4l}{2l-d}}, 3\lambda_1 \|f\|_{\mathcal{H}}^2, 3\lambda_2 Q_N(\mathbf{w}^*; \mathbf{d}) \right\} \right] \bigcap G_{N,i} \right) \\ \geq & \mathbb{P}(G_{N,1,T}) + \mathbb{P}\left[\left\{ S_N(\hat{\mathbf{w}}, f) \leq c_2 \lambda_1^{-\frac{d}{2l-d}} N^{\frac{2l}{d-2l}} T^{\frac{4l}{2l-d}} \right\} \bigcap \mathring{G}_{N,1,T} \right] + \mathbb{P}(G_{N,2}) + \mathbb{P}(G_{N,3}) \\ = & 1 - \mathbb{P}\left[\left\{ S_N(\hat{\mathbf{w}}, f) > c_2 \lambda_1^{-\frac{d}{2l-d}} N^{\frac{2l}{d-2l}} T^{\frac{4l}{2l-d}} \right\} \cap \mathring{G}_{N,1,T} \right] \\ = & 1 - \mathbb{P}(\mathring{G}_{N,1,T}) \\ \geq & 1 - \mathbb{P}\left\{ S_N(\mathbf{w}^*, u^*) > \frac{1}{N} T^2 \|u\|_{\mathcal{H}}^{\frac{d}{l}} \right\} \\ \geq & 1 - c\exp\left(-\frac{T^2}{c^2}\right). \end{aligned} \tag{2.43}$$

Further, by Lemma 1, it is reasonable to assume that $n^{\epsilon} Q_N(\mathbf{w}^*; \mathbf{d}) \leq M$, as $n \to \infty$ and $N \to \infty$ for a constant M. Therefore, under the condition that $\lambda_1 \asymp \mathcal{O}\left(n^{-1}\right)$ and $\lambda_2 = \mathcal{O}\left(n^{\epsilon-1}\right)$, by (2.43), we would arrive the conclusion that:

$$S_N(\hat{\mathbf{w}}, f) = \mathcal{O}_p\left(\frac{1}{n}\right). \tag{2.44}$$

Also, with similar arguments to $\lambda_2 Q_N(\hat{\mathbf{w}};\mathbf{d})$ as above and if $\lambda_1 \asymp n^{-1}$ and $\lambda_2 \asymp n^{\epsilon-1}$ we would arive the conclusion that

$$\lambda_2 Q_N(\hat{\mathbf{w}};\mathbf{d}) = \mathcal{O}_p\left(\frac{1}{n}\right) \Leftrightarrow Q_N(\hat{\mathbf{w}};\mathbf{d}) = \mathcal{O}_p\left(n^{-\epsilon}\right) \tag{2.45}$$

Now, we consider $\mathbb{E}\left[n^{\epsilon} Q_N(\hat{\mathbf{w}};\mathbf{d})\right]$. As $\lambda_1 \asymp n^{-1}$ and $\lambda_2 \asymp n^{\epsilon-1}$, we would have:

$$\begin{aligned} \underline{B}_1 n^{-1} \leq \lambda_1 \leq \bar{B}_1 n^{-1} \\ \underline{B}_2 n^{\epsilon-1} \leq \lambda_2 \leq \bar{B}_2 n^{\epsilon-1} \end{aligned} \tag{2.46}$$

for some positive constants $\bar{B}_1 > \underline{B}_1 > 0$, $\bar{B}_2 > \underline{B}_2 > 0$. Then we might have the following decomposition:

$$\begin{aligned} \mathbb{E}\left\{n^{\epsilon} Q_N(\hat{\mathbf{w}};\mathbf{d})\right\} &= \underbrace{\mathbb{E}\left\{n^{\epsilon} Q_N(\hat{\mathbf{w}};\mathbf{d})|G_{N,1}\right\}\mathbb{P}(G_{N,1})}_{(S1)} + \underbrace{\mathbb{E}\left\{n^{\epsilon} Q_N(\hat{\mathbf{w}};\mathbf{d})|G_{N,2}\right\}\mathbb{P}(G_{N,2})}_{(S2)} \\ &\quad + \underbrace{\mathbb{E}\left\{n^{\epsilon} Q_N(\hat{\mathbf{w}};\mathbf{d})|G_{N,3}\right\}\mathbb{P}(G_{N,3})}_{(S3)} \end{aligned}$$

For term $S1$, we might choose: $c_3 = \max\{c, 2\underline{B}_1^{-\frac{d}{2l-d}}\underline{B}_2^{-1}, 2\}$ and sufficiently small $a > 0$ such that $c_3 > \underline{B}_1^{-\frac{d}{2l-d}}\underline{B}_2^{-1} c_3^{\frac{4la}{2l-d}} c_2$ and $\frac{4la}{2l-d} < 1$. The above also implies that a fulfills $\min\left[\frac{(2l-d)\log(\frac{c_3}{c_2}\underline{B}_1^{\frac{d}{2l-d}}\bar{B}_2)}{4d\log c_3}, \frac{2l-d}{4l}\right] > a > 0$. Then, we'll have

$$\begin{aligned} & \mathbb{E}\left\{n^{\epsilon} Q_N(\hat{\mathbf{w}};\mathbf{d})|G_{N,1}\right\}\mathbb{P}(G_{N,1}) \\ = & \int_0^{\infty} \mathbb{P}\left\{n^{\epsilon} Q_N(\hat{\mathbf{w}};\mathbf{d}) > t|G_{N,1}\right\}\mathbb{P}(G_{N,1})dt \\ = & \int_0^{\infty} \mathbb{P}\left\{n^{\epsilon} Q_N(\hat{\mathbf{w}};\mathbf{d}) > t \cap G_{N,1}\right\} dt \\ \leq & c_3 + \int_{c_3}^{\infty} \mathbb{P}\left\{n^{\epsilon} Q_N(\hat{\mathbf{w}};\mathbf{d}) > t \cap G_{N,1,t^a}\right\} dt + \int_{c_3}^{\infty} \mathbb{P}\left\{n^{\epsilon} Q_N(\hat{\mathbf{w}};\mathbf{d}) > t \cap \mathring{G}_{N,1,t^a}\right\} dt \end{aligned} \tag{2.47}$$

Specifically, we have

$$\begin{aligned} & \int_{c_3}^{\infty} \mathbb{P}\left\{n^{\epsilon} Q_N(\hat{\mathbf{w}};\mathbf{d}) > t \cap G_{N,1,t^a}\right\} dt \\ \leq & \int_{c_3}^{\infty} \mathbb{P}\left\{t^{\frac{4la}{2l-d}}\underline{B}_1^{\frac{-d}{2l-d}}\underline{B}_2^{-1} c_2 \geq n^{\epsilon} Q_N(\hat{\mathbf{w}};\mathbf{d}) > t\right\} = 0 \end{aligned} \tag{2.48}$$

In addition, we have

$$\begin{aligned}\int_{c_3}^{\infty}\mathbb{P}\left\{n^{\epsilon}Q_N(\hat{\mathbf{w}};\mathbf{d})>t\cap\mathring{G}_{N,1,t^a}\right\}dt &\leq \int_{c_3}^{\infty}\mathbb{P}(\mathring{G}_{N,1,t^a})dt\\ &\leq \int_{c_3}^{\infty}c\exp\left(-\frac{t^{2a}}{c^2}\right)\\ &= -c\frac{c^{\frac{1}{a}}\Gamma\left(\frac{1}{2^a},\frac{t^{2a}}{c^2}\right)}{2a}|_{t=c_3}^{\infty}\leq c_4 \end{aligned} \tag{2.49}$$

The last inequality is implied by the fact: $\frac{\Gamma(s,x)}{x^{s-1}e^{-x}}\to 1$ as $x\to 1$. As a result, we have:

$$\mathbb{E}\left\{n^{\epsilon}Q_N(\hat{\mathbf{w}};\mathbf{d})|G_{N,1}\right\}\mathbb{P}(G_{N,1})<\infty \tag{2.50}$$

As for term $S2$, we have:

$$\mathbb{E}\left\{n^{\epsilon}Q_N(\hat{\mathbf{w}};\mathbf{d})|G_{N,2}\right\}\mathbb{P}(G_{N,2})\leq 3\bar{B}_1\underline{B}_2^{-1}\left\|f\right\|_{\mathcal{H}}^2<\infty \tag{2.51}$$

Also, for term $S3$, we have:

$$\begin{aligned}\mathbb{E}\left\{n^{\epsilon}Q_N(\hat{\mathbf{w}};\mathbf{d})|G_{N,3}\right\}\mathbb{P}(G_{N,3}) &\leq 3\bar{B}_2\mathbb{E}\left\{n^{\epsilon}Q_N(\mathbf{w}^*;\mathbf{d})|G_{N,3}\right\}\mathbb{P}(G_{N,3})\\ &\leq 3\bar{B}_2\mathbb{E}\left\{n^{\epsilon}Q_N(\mathbf{w}^*;\mathbf{d})\right\}\\ &\leq 3\bar{B}_2M<\infty \end{aligned} \tag{2.52}$$

Combining the above results, we finally arrive the conclusion that there exists a constant $W>0$ such that $\mathbb{E}\left\{n^{\epsilon}Q_N(\hat{\mathbf{w}};\mathbf{d})\right\}\leq W$. □

Lemma 2.5. *Suppose assumptions [A1] ∼ [A5] hold, if* $\lambda_1\asymp n^{-1}$ *and* $\lambda_2=\mathcal{O}\left(n^{\epsilon-1}\right)$*, then* $S_N(\hat{\mathbf{w}},m)=\mathcal{O}_p\left(n^{-1}\right)\left\|m\right\|_N$*. Further, if* $\lambda_2\asymp n^{\epsilon-1}$*, there exists constant* $S>0$*, such that* $\mathbb{E}\left\{nS_N(\hat{\mathbf{w}},m)\right\}\leq S^2$*.*

Proof of Lemma 2.5 . By (2.34), we have

$$\begin{aligned}&S_N(\hat{\mathbf{w}},m)+\lambda_1\left\|u^*\right\|_{\mathcal{H}}^2\left\|m\right\|_N^2+\lambda_2Q_N(\hat{\mathbf{w}};\mathbf{d})\left\|m\right\|_N^2\\ \leq&S_N(\mathbf{w}^*,u^*)\left\|m\right\|_N^2+\lambda_1\left\|m\right\|_{\mathcal{H}}^2+\lambda_2Q_N(\mathbf{w}^*;\mathbf{d})\left\|m\right\|_N^2\end{aligned}$$

By Lemma 2.1 in Lin (2000), we have: $S_N(\mathbf{w}^*,u^*)=\mathcal{O}_p\left(n^{-1}\right)\left\|u^*\right\|_{\mathcal{H}}^{\frac{d}{l}}$. Similarly, we'll discuss each scenarios in (2.53) as in Lemma 2.5.

Case 1. $S_N(\mathbf{w}^*, u^*) \|m\|_N^2$ is the largest of the right hand side of (2.53). If $\|m\|_N \neq 0$, we would have:

$$\begin{aligned} & \lambda_1 \|u^*\|_{\mathcal{H}}^2 \leq \mathcal{O}_p\left(n^{-1}\right) \|u^*\|_{\mathcal{H}}^{\frac{d}{l}} \\ \Leftrightarrow \quad & \|u^*\|_{\mathcal{H}} \leq \lambda_1^{-\frac{l}{2l-d}} \mathcal{O}_p\left(n^{\frac{l}{d-2l}}\right) \\ \Leftrightarrow \quad & \|u^*\|_{\mathcal{H}}^{\frac{d}{l}} \leq \lambda_1^{-\frac{d}{2l-d}} \mathcal{O}_p\left(n^{\frac{d}{d-2l}}\right). \end{aligned} \tag{2.53}$$

Thus, we have

$$S_N(\hat{\mathbf{w}}, m) \leq \lambda_1^{-\frac{d}{2l-d}} \mathcal{O}_p\left(n^{\frac{2l}{d-2l}}\right) \|m\|_N^2. \tag{2.54}$$

Also, if $\|m\|_N^2 = 0$, we natually have the above inequality.

Case 2. Suppose $\lambda_1 \|m\|_{\mathcal{H}}^2$ is the largest in the right hand side, we have:

$$S_N(\hat{\mathbf{w}}, m) \leq 3\lambda_1 \|m\|_{\mathcal{H}}^2. \tag{2.55}$$

Case 3. Suppose that $\lambda_2 Q_N(\mathbf{w}^*; \mathbf{d}) \|m\|_N^2$ is the largest in the right hand side of (2.53), we can obtain that:

$$S_N(\hat{\mathbf{w}}, m) \leq 3\lambda_2 n^{-\epsilon} M \|m\|_N^2 \tag{2.56}$$

Due to Lemma 2.1 in Lin (2000), we have: $\|m\|_N \leq R \|m\|_H < \infty$.

Therefore, we have

$$S_N(\hat{\mathbf{w}}, m) = \mathcal{O}_p\left[\max\left\{\lambda_1^{-\frac{d}{2l-d}} n^{\frac{2l}{d-2l}} \|m\|_N^2, \lambda_1 \|m\|_{\mathcal{H}}, \lambda_2 n^{-\epsilon} M \|m\|_N^2\right\}\right] \tag{2.57}$$

Since $S_N(\hat{\mathbf{w}}, m) = 0$ if $\|m\|_N = 0$, we have $S_N(\hat{\mathbf{w}}, m) = \mathcal{O}_p\left(\frac{1}{n}\right) \|m\|_N^2$.

Based on the Lemma 2.3, with similar arguments in Lemma 2.4, we might have that there exists a constant $\tilde{S}^2 > 0$ such that $\mathbb{E}\left\{n^2 \tilde{S}_N^2(\hat{\mathbf{w}}, m)\right\} \leq \tilde{S}^2$, where

$$\tilde{S}_N(\hat{\mathbf{w}}, m) = \begin{cases} S_N\left(\hat{\mathbf{w}}, \frac{m}{\|m\|_N}\right), & \text{if } \|m\|_{AB} \neq 0 \\ 0, & \text{Otherwise} \end{cases} \tag{2.58}$$

Moreover,

$$\begin{aligned}\mathbb{E}\left\{nS_N(\hat{m},m)\right\} &= \mathbb{E}\left\{n\tilde{S}_N(\hat{m},\frac{m}{\|m\|_N})\,\|m\|_N^2\right\}\\ &\leq \frac{1}{2}\left[\mathbb{E}\left\{n^2\tilde{S}_N^2(\hat{\mathbf{w}},m)\right\}+\frac{n\int m^4d\mathbb{P}}{N^2}+\frac{n(N-1)}{N^2}\left(\int m^2d\mathbb{P}\right)^2\right]\\ &\leq \frac{1}{2}\left\{\tilde{S}^2+\frac{n}{N}\int m^4d\mathbb{P}+\frac{n}{N}\left(\int m^2d\mathbb{P}\right)^2\right\} \qquad (2.59)\end{aligned}$$

By Lemma 2.1 of Lin (2000), we have $\int m^4d\mathbb{P}<\infty$ and $\left(\int m^2d\mathbb{P}\right)^2<\infty$. That is, $\mathbb{E}\left\{nS_N(\hat{m},m)\right\}$ is bounded. □

Proof of Theorem 1. Recall the decomposition:

$$\begin{aligned}\frac{1}{N}\sum_{i=1}^N\delta_iw_iY_i-\frac{1}{N}\sum_{i=1}^NY_i &= \underbrace{\frac{1}{N}\sum_{i=1}^N\left\{(\delta_iw_i-1)m(\mathbf{X}_i)\right\}}_{(T1)}+\underbrace{\frac{1}{N}\sum_{i=1}^N\delta_i(w_i-d_i)\epsilon_i}_{(T2)}\\ &\quad+\underbrace{\left\{\frac{1}{N}\sum_{i=1}^N(\delta_id_i-1)\epsilon_i\right\}}_{(T3)} \qquad (2.60)\end{aligned}$$

By Lemma 2.1 of Lin (2000), we have $\|m\|_2^2=\left(\int m^2d\mathbb{P}\right)^2<\infty$. Since $\boldsymbol{X}_1,\ldots,\boldsymbol{X}_N$ are $i.i.d$, we can show that

$$\|m\|_N=\int m^2d\mathbb{P}+o_p(1) \qquad (2.61)$$

Therefore, the term $T1$ can be bounded by

$$\left|\frac{1}{N}\sum_{i=1}^N\left\{(\delta_iw_i-1)m(\mathbf{X}_i)\right\}\right|=\sqrt{S_N(\hat{\mathbf{w}},m)}=\mathcal{O}_p\left\{\left(\frac{1}{n}\right)^{\frac{1}{2}}\right\}\|m\|_2+o_p\left\{\left(\frac{1}{n}\right)^{\frac{1}{2}}\right\} \qquad (2.62)$$

via Lemma (2.4). In addition, $\mathbb{E}\left\{nS_N(\hat{\mathbf{w}},m)\right\}<\infty$. As for term $T2$, we first define $\hat{\psi}_i=\delta_i(\hat{w}_i-d_i)$. By assumption, we have $\mathbb{E}[\epsilon_i|\hat{\psi}_1,\ldots,\hat{\psi}_N]=0$. By EVVE formula, we would have:

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^N\hat{\psi}_i\epsilon_i\right) &= \mathbb{E}\left[\mathrm{Var}\left\{\frac{1}{N}\sum_{i=1}^N\hat{\psi}_i\epsilon_i\middle|\hat{\psi}_1,\ldots,\hat{\psi}_N\right\}\right]\\ &= \mathbb{E}\left[\mathbb{E}\left\{\left(\frac{1}{N}\sum_{i=1}^N\hat{\psi}_i\epsilon_i\right)^2\middle|\hat{\psi}_1,\ldots,\hat{\psi}_N\right\}\right]\\ &\lesssim \frac{\sigma^2}{N}\mathbb{E}\left\{Q_N(\hat{\mathbf{w}};\mathbf{d})\right\}\mathcal{O}(d_i)\lesssim\frac{\sigma^2Wn^{-\epsilon}}{n}=o\left(\frac{1}{n}\right). \qquad (2.63)\end{aligned}$$

Therefore, we have $N^{-1}\sum_{i=1}^{N}\delta_i(\hat{w}_i - d_i)\epsilon_i = o_p\left(n^{-1/2}\right)$, which implies $\mathbb{E}\left\{n^{1/2}N^{-1}\sum_{i=1}^{N}\delta_i\left(\hat{w}_i - d_i\right)^2\epsilon_i\right\}^2 < \infty$. Finally, term $T3$ can be directly handled by sample design assumption (A1), i.e., $N^{-1}\sum_{i\in\mathcal{S}} d_i\epsilon_i - N^{-1}\sum_{i\in\mathcal{U}}\epsilon_i = \mathcal{O}_p\left(n^{-1/2}\right)$. In a nutshell, we have $N^{-1}\sum_{i=1}^{N}\delta_i\hat{w}_i Y_i - \mathbb{E}(Y) = \mathcal{O}_p\left(n^{-1/2}\right)$. □

2.9.2 Proof of Theorem 2.2

The additional assumption for Theorem 2 is

[B1 For some $k \geq 2$, there is a constant $\rho < \infty$ such that $\mathbb{E}[\phi_j(\mathbf{X})^{2k}] \leq \rho^{2k}$ for all $j \in \mathbb{N}$, where $\{\phi_j\}_{j=1}^{\infty}$ are orthonormal basis by expansion from Mercer's theorem.

Lemma 2.6. *Suppose assumptions [A4], [A5] and [B1] hold, and the kernel ridge regression for a ℓ-th order Sobolev space $\mathcal{H}$ with tuning parameter of order $n^{-2\ell/(2\ell+1)}$, then*

$$\mathbb{E}(\|m - \hat{m}\|_N^2) = \mathcal{O}\left\{\left(\frac{\sigma^2}{n}\right)^{\frac{2\ell}{2\ell+1}}\right\}. \tag{2.64}$$

Proof of Lemma 2.6. The proof can be found in Corollary 4 in Zhang et al. (2013). □

Lemma 2.7. *Suppose assumptions (A1) $\sim$ (A5), (B1) and the tuning parameter assumption in Lemma 2.6 hold, further assume $\lambda_1 \asymp n^k$, $\lambda_2 \asymp n^{\epsilon-1}$, where $-\frac{2\ell^2+ld+d}{(2l+1)d} < k < -1$, and $\epsilon < \frac{2l}{2l+1}$, then $S_N(\hat{\mathbf{w}}, h) = o_p\left(n^{-1}\right)$.*

Proof of Lemma 2.7. Let $h = m - \hat{m}$. Obviously, as $m, \hat{m} \in \mathcal{H}$. we have $h \in \mathcal{H}$. By Lemma 2.6, we have $\|h\|_N = \mathcal{O}_p(n^{-\ell/(2\ell+1)}) = o_p(1)$. It is also easy to verify that $\lambda_1^{-1}\|h\|_N^{\frac{2(2l-d)}{d}} = o_p(n)$ and $\lambda_2\|h\|_N^2 = o_p(n^{-1})$.

Rearranging the terms in (2.34), we would immediately get:

$$\begin{aligned} &S_N(\hat{\mathbf{w}}, h) + \lambda_1\|u^*\|_{\mathcal{H}}^2\|h\|_N^2 + \lambda_2 Q_N(\hat{\mathbf{w}}; \mathbf{d})\|h\|_N^2 \\ \leq &S_N(\mathbf{w}^*, u^*)\|h\|_N^2 + \lambda_1\|h\|_{\mathcal{H}}^2 + \lambda_2 Q_N(\mathbf{w}^*; \mathbf{d})\|h\|_N^2. \end{aligned} \tag{2.65}$$

The proof is similar to Lemma 2.4 and Lemma 2.5. □

Proof of Theorem 2.2. We have the following decomposition

$$\begin{aligned}
&\frac{1}{N}\left[\sum_{i\in\mathcal{S}}\hat{w}_i\left\{Y_i-\hat{m}(\mathbf{X}_i)\right\}+\sum_{i\in\mathcal{U}}\hat{m}(\mathbf{X}_i)\right]-\frac{1}{N}\sum_{i\in\mathcal{U}}Y_i \\
&=\underbrace{\frac{1}{N}\sum_{i=1}^{N}(\delta_i\hat{w}_i-1)h(\mathbf{X}_i)}_{:=a_N}+\underbrace{\frac{1}{N}\sum_{i=1}^{N}\delta_i(\hat{w}_i-d_i)\epsilon_i}_{:=b_N}+\underbrace{\frac{1}{N}\sum_{i=1}^{N}\delta_i d_i\epsilon_i}_{:=c_N}
\end{aligned} \tag{2.66}$$

Apparently, we have

$$\begin{aligned}
n\mathbb{E}\left(\frac{\tilde{t}_y}{N}-\frac{1}{N}\sum_{i=1}^{N}Y_i\right)^2 &= n\left\{\mathbb{E}(a_N^2)+\mathbb{E}(b_N^2)+\mathbb{E}(c_N^2)+2\mathbb{E}(a_Nb_N)\right. \\
&\quad \left.+2\mathbb{E}(a_Nc_N)+2\mathbb{E}(b_Nc_N)\right\}
\end{aligned} \tag{2.67}$$

By Lemma 2.7, the first term is $o_p(n^{-1/2})$. By dominated convergence theorem and Skorohod representation theorem, we have

$$\mathbb{E}(a_N^2)=n^{-1}\mathbb{E}(nS_N(\hat{\mathbf{w}},h))=o(n^{-1}) \tag{2.68}$$

Therefore, $n\mathbb{E}(a_N^2)=o(1)$. Moreover, with similar argument in *Theorem 2*, we have $b_N=o_p(n^{-1/2})$, which implies $n\mathbb{E}(b_N^2)=o(1)$. Next,

$$n\mathbb{E}(c_N^2)=\frac{n}{N^2}\sum_{i\in\mathcal{U}}\frac{1-\pi_i}{\pi_i}\sigma_i^2 \tag{2.69}$$

Additionally, the cross terms can be handled by Cauchy-Schwarz inequality. Therefore, we arrive the conclusion that our estimator attains Godambe-Joshi lower bound.

□

CHAPTER 3. STATISTICAL INFERENCE AFTER KERNEL RIDGE REGRESSION IMPUTATION UNDER ITEM NONRESPONSE

Hengfang Wang and Jae Kwang Kim

Iowa State University

Modified from a manuscript to be submitted to in *Biometrika*

3.1 Abstract

Imputation and propensity score weighting are two popular techniques for handling missing data. We consider a fully nonparametric approach to these methods using kernel ridge regression. Kernel ridge regression is a modern regression technique based on the theory of reproducing kernel Hilbert space. We first use the kernel ridge regression to develop imputation for handling item nonresponse. While this nonparametric approach is potentially promising for imputation, its statistical properties are not fully investigated in the literature. Under some conditions on the order of the tuning parameter, we first establish the root-n consistency of the kernel ridge regression imputation estimators and show that it achieves the lower bound of the semiparametric asymptotic variance. A nonparametric propensity score estimator using the kernel ridge regression is also developed by a novel application of the maximum entropy method for the density ratio function estimation. The resulting propensity score estimator is shown to achieve the same asymptotic variance as the kernel ridge regression imputation estimator. Results from a limited simulation study are also presented to confirm our theory.

3.2 Introduction

Missing data is a universal problem in statistics. Ignoring the cases with missing values can lead to misleading results (Kim and Shao, 2013; Little and Rubin, 2019). Two popular approaches

for handling missing data are imputation and propensity score weighting. Both approaches are based on some assumptions about the data structure and the response mechanism. In the statistical point of view, instead of using strong parametric model assumptions, nonparametric approaches are more attractive as they does not depend on explicit model assumptions.

In principle, any prediction techniques can be used to predict missing values using the responding units as a training sample. However, statistical inference with imputed estimator is not straightforward. Treating imputed data as if observed and applying the standard estimation procedure may result in misleading inference, leading to underestimation of the variance of imputed point estimators. How to incorporate the uncertainty of the estimated parameters in the final inference is challenging especially for nonparametric imputation because the model parameter is implicitly defined.

For nonparametric imputation, Cheng (1994) used the kernel-based nonparametric regression for imputation and established the root n-consistency of the imputed estimator. Wang and Chen (2009) employed the kernel smoothing approach to do empirical likelihood inference with missing values. Kim et al. (2014) proposed Bayesian multiple imputation using the Dirichlet process mixture. Sang et al. (2020) proposed semiparametric fractional imputation using Gaussian mixtures.

For nonparametric propensity score estimation, Hainmueller (2012) proposed the so-called entropy balancing to find the propensity score weights using the Kullback-Leibler information criterion. Chan et al. (2016) generalize this idea further to develop a general calibration weighting method that satisfies the covariance balancing property with increasing dimensions of the control variables. They further showed the global efficiency of the proposed calibration weighting estimator. Zhao (2019) generalized the idea further and developed a unified approach of covariate balancing propensity score method using tailored loss functions. Tan (2020) developed regularized calibrated estimation of propensity scores with high dimensional covariates.

In this paper, we consider kernel ridge regression as a tool for nonparametric function estimation for imputation and propensity score estimation. Kernel ridge regression (Friedman

et al., 2001; Shawe-Taylor and Cristianini, 2004) is a modern regression technique which can alleviate the effect of model assumption. By using a regularized M-estimator in reproducing kernel Hilbert space, kernel ridge regression can estimate the regression mean function with complex reproducing kernel Hilbert space while a regularized term makes the original infinite dimensional estimation problem viable (Wahba, 1990). van de Geer (2000); Mendelson (2002); Zhang (2005); Koltchinskii (2006); Steinwart et al. (2009) studied the error bounds for the estimates of kernel ridge regression method.

While the kernel ridge regression is a promising tool for handling missing data, its statistical inference is not fully investigated in the literature. Specifically, we obtain root-n consistency of the kernel ridge regression imputation estimator under some popular functional Hilbert spaces. Because the kernel ridge regression is a general tool for nonparametric regression with flexible assumptions, the proposed imputation method can be used widely to handle missing data without employing parameteric model assumptions. Variance estimation after the kernel ridge regression imputation is a challenging but important problem. To the best of our knowledge, this is the first paper which considers kernel ridge regression technique for imputation and discusses its variance estimation rigorously.

The kernel ridge regression is also used to obtain nonparametric propensity score weights for handling missing data. To do this, we use a novel application of density ratio function estimation in the same reproducing kernel Hilbert space. Maximum entropy method of Nguyen et al. (2010) is used for density ratio estimation, which is further applied to get the kernel ridge regression-based propensity score estimators. We further show the asymptotic equivalence of the resulting propensity score estimator with the kernel ridge regression-based imputation estimator. These theoretical findings can be used to make valid statistical inferences with the propensity score estimator.

3.3 Kernel Ridge Regression Imputation

Consider the problem of estimating $\theta = E(Y)$ from an independent and identically distributed sample $\{(\mathbf{x}_i, y_i), i = 1, \cdots, n\}$ of random vector $(\mathbf{X}, Y)$. Instead of always observing y_i, suppose that we observe y_i only if $\delta_i = 1$, where δ_i is the response indicator function of unit i taking values on $\{0, 1\}$. The auxiliary variable $\mathbf{x}_i$ are always observed. We assume that the response mechanism is missing at random in the sense of Rubin (1976).

Under missing-at-random, we can develop a nonparametric estimator $\hat{m}(\mathbf{x})$ of $m(\mathbf{x}) = E(Y \mid \mathbf{x})$ and construct the following imputation estimator:

$$\hat{\theta}_I = \frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_i y_i + (1 - \delta_i) \hat{m}(\mathbf{x}_i) \right\}. \tag{3.1}$$

If $\hat{m}(\mathbf{x})$ is constructed by the kernel-based nonparametric regression method, we can express

$$\hat{m}(\mathbf{x}) = \frac{\sum_{i=1}^{n} \delta_i K_h(\mathbf{x}_i, \mathbf{x}) y_i}{\sum_{i=1}^{n} \delta_i K_h(\mathbf{x}_i, \mathbf{x})} \tag{3.2}$$

where $K_h(\cdot)$ is the kernel function with bandwidth h. Under some suitable choice of the bandwidth h, Cheng (1994) first established the root-n consistency of the imputation estimator (3.1) with nonparametric function in (3.2). However, the kernel-based regression imputation in (3.2) is applicable only when the dimension of x is small.

In this paper, we extend the work of Cheng (1994) by considering a more general type of the nonparametric imputation, called kernel ridge regression imputation. The kernel ridge regression can be understood using the reproducing kernel Hilbert space theory (Aronszajn, 1950) and can be described as

$$\hat{m} = \arg\min_{m \in \mathcal{H}} \left[\sum_{i=1}^{n} \delta_i \left\{ y_i - m(\mathbf{x}_i) \right\}^2 + \lambda \left\| m \right\|_{\mathcal{H}}^2 \right], \tag{3.3}$$

where $\|m\|_{\mathcal{H}}^2$ is the norm of m in the Hilbert space $\mathcal{H}$ and $\lambda (> 0)$ is a tuning parameter for regularization. Here, the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is induced by such a kernel function, i.e.,

$$\langle f, K(\cdot, \mathbf{x}) \rangle_{\mathcal{H}} = f(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X}, f \in \mathcal{H},$$

namely, the reproducing property of $\mathcal{H}$. Naturally, this reproducing property implies the $\mathcal{H}$ norm of f: $\|f\|_{\mathcal{H}} = \langle f, f\rangle_{\mathcal{H}}^{1/2}$. Scholkopf and Smola (2002) provides a comprehensive overview of the machine learning techniques using the reproducing kernel functions.

One canonical example of such a functional Hilbert space is the Sobolev space. Specifically, assuming that the domain of such functional space is $[0,1]$, the Sobolev space of order ℓ can be denoted as

$$\mathcal{W}_2^{\ell} = \left\{ f : [0,1] \to \mathbb{R} | f, f^{(1)}, \ldots, f^{(\ell-1)} \subset \mathbb{C}[0,1], \quad f^{(\ell)} \in L^2[0,1] \right\},$$

where $\mathbb{C}[0,1]$ denotes the absolutely continuous function on $[0,1]$. One possible norm for this space can be

$$\|f\|_{\mathcal{W}_2^{\ell}}^2 = \sum_{q=0}^{\ell-1} \left\{ \int_0^1 f^{(q)}(t)dt \right\}^2 + \int_0^1 \left\{ f^{(\ell)}(t) \right\}^2 dt.$$

In this section, we employ the Sobolev space of second order as the approximation function space. For Sobolev space of order ℓ, we have the kernel function

$$K(x,y) = \sum_{q=0}^{\ell-1} k_q(x)k_q(y) + k_\ell(x)k_\ell(y) + (-1)^{\ell} k_{2\ell}(|x-y|),$$

where $k_q(x) = (q!)^{-1}B_q(x)$ and $B_q(\cdot)$ is the Bernoulli polynomial of order q. Smoothing spline method is a special case of the kernel ridge regression method.

By the representer theorem for reproducing kernel Hilbert space (Wahba, 1990), the estimate in (3.3) lies in the linear span of $\{K(\cdot, \mathbf{x}_i), i = 1, \ldots, n\}$. Specifically, we have

$$\hat{m}(\cdot) = \sum_{i=1}^{n} \hat{\alpha}_{i,\lambda} K(\cdot, \mathbf{x}_i), \tag{3.4}$$

where

$$\hat{\boldsymbol{\alpha}}_\lambda = \left(\boldsymbol{\Delta}_n \mathbf{K} + \lambda \mathbf{I}_n\right)^{-1} \boldsymbol{\Delta}_n \mathbf{y},$$

$\boldsymbol{\Delta}_n = \mathrm{diag}(\delta_1, \ldots, \delta_n)$, $\mathbf{K} = (K(\mathbf{x}_i, \mathbf{x}_j))_{ij}$, $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}}$ and $\mathbf{I}_n$ is the $n \times n$ identity matrix.

The tuning parameter λ is selected via generalized cross-validation in kernel ridge regeression, where the criterion for λ is

$$\mathrm{GCV}(\lambda) = \frac{n^{-1} \left\| \{\boldsymbol{\Delta}_n - \mathbf{A}(\lambda)\} \mathbf{y} \right\|_2^2}{n^{-1} \mathrm{tr}(\boldsymbol{\Delta}_n - \mathbf{A}(\lambda))}, \tag{3.5}$$

and $\mathbf{A}(\lambda) = \boldsymbol{\Delta}_n \mathbf{K}(\boldsymbol{\Delta}_n \mathbf{K} + \lambda \mathbf{I}_n)^{-1} \boldsymbol{\Delta}_n$. The value of λ minimizing the criterion (3.5) is used for the selected tuning parameter.

Using the kernel ridge regression imputation in (3.3), we can obtain the imputed estimator in(3.1). Because $\hat{m}(\mathbf{x})$ in (3.4) is a nonparametric regression estimator of $m(\mathbf{x}) = E(Y \mid \mathbf{x})$, we can expect that this imputation estimator in (3.1) is consistent for $\theta = E(Y)$ under missing at random, as long as $\hat{m}(\mathbf{x})$ is a consistent estimator of $m(\mathbf{x})$. Surprisingly, it turns out that the consistency of $\hat{\theta}_I$ to θ is of order $O_p(n^{-1/2})$, while the point-wise convergence rate for $\hat{m}(\mathbf{x})$ to $m(\mathbf{x})$ is slower.

We aim to establish two goals: (i) find the sufficient conditions for the root-n consistency of the imputation estimator $\hat{\theta}_I$ in (3.1) and give a formal proof; (ii) find a linearization variance formula for the imputation estimator $\hat{\theta}_I$ using the kernel ridge regression imputation. The first part is formally presented in Theorem 3.1 in Section 3.4. For the second part, we employ the density ratio estimation method of Nguyen et al. (2010) to get a consistent estimator of $\omega(\mathbf{x}) = \{\pi(\mathbf{x})\}^{-1}$ in the linearized version of $\hat{\theta}_I$. Estimation of $\omega(\mathbf{x})$ will be presented in Section 3.5.

3.4 Main Theory

Before we develop our main theory, we first introduce Mercer's theorem.

Lemma 3.1 (Mercer's theorem)**.** *Given a continuous, symmetric, positive definite kernel function $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$. For $\mathbf{x}, \mathbf{z} \in \mathcal{X}$, under some regularity conditions, Mercer's theorem characterizes K by the following expansion*

$$K(\mathbf{x}, \mathbf{z}) = \sum_{j=1}^{\infty} \lambda_j \phi_j(\mathbf{x}) \phi_j(\mathbf{z}),$$

where $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ are a non-negative sequence of eigenvalues and $\{\phi_j\}_{j=1}^{\infty}$ is an orthonormal basis for $L^2(\mathbb{P})$.

Furthermore, we make the following assumptions.

Assumption 3.1. *For some $k \geq 2$, there is a constant $\rho < \infty$ such that $E[\phi_j(X)^{2k}] \leq \rho^{2k}$ for all $j \in \mathbb{N}$, where $\{\phi_j\}_{j=1}^{\infty}$ are orthonormal basis by expansion from Mercer's theorem.*

Assumption 3.2. *The function $m \in \mathcal{H}$, and for $\mathbf{x} \in \mathcal{X}$, we have $E[\{Y - m(\mathbf{x})\}^2] \leq \sigma^2$, for some $\sigma^2 < \infty$.*

Assumption 3.3. *The response mechanism is missing at random. Furthermore, the propensity score $\pi(\mathbf{x}) = \mathrm{pr}(\delta = 1 \mid \mathbf{x})$ is uniformly bounded away from zero. In particular, there exists a positive constant $c > 0$ such that $\pi(\mathbf{x}_i) \geq c$, for $i = 1, \ldots, n$.*

The first assumption is a technical assumption which controls the tail behavior of $\{\phi_j\}_{j=1}^{\infty}$. Assumption 3.2 indicates that the noises have bounded variance. Assumption 3.1 and Assumption 3.2 together aim to control the error bound of the kernel ridge regression estimate $\hat{m}$. Furthermore, Assumption 3.3 means that the support for the respondents should be the same as the original sample support. Assumption 2.4.1 is a standard assumption for missing data analysis. We furhter introduce the following lemma. Let $\mathbf{S}_\lambda = (\mathbf{I}_n + \lambda \mathbf{K}^{-1})^{-1}$ be the linear smoother for the kernel ridge regression method. That is, $\hat{m} = \mathbf{S}_\lambda \mathbf{y}$ be the vector of regression predictor of $\mathbf{y}$ using the kernel ridge regression method.

Lemma 3.2 (modified Lemma 7 in Zhang et al. (2013))**.** *Suppose Assumption 3.1 and 3.2 hold, for a random vector $\mathbf{z} = E(\mathbf{z}) + \sigma\boldsymbol{\varepsilon}$, we have*

$$\mathbf{S}_\lambda \mathbf{z} = E(\mathbf{z} \mid \mathbf{x}) + \mathbf{a}_n,$$

where $\mathbf{a}_n = (a_1, \ldots, a_n)^{\mathrm{T}}$ and

$$a_i = \mathcal{O}_p\left(\lambda^{1/2} + \{\gamma(\lambda)\}^{1/2} n^{-1/2}\right), \tag{3.6}$$

for $i = 1, \ldots, n$, as long as $E(\|z_i\|_{\mathcal{H}})$ and σ^2 is bounded from above, for $i = 1, \ldots, n$, where $\boldsymbol{\varepsilon}$ are noise vector with mean zero and bounded variance and

$$\gamma(\lambda) = \sum_{j=1}^{\infty} (1 + \lambda/\mu_j)^{-1},$$

is the effective dimension and $\{\mu_j\}_{j=1}^{\infty}$ are the eigenvalues of kernel K used in $\hat{m}(\mathbf{x})$.

The first term in (3.6) denotes the order of bias term and the second term denotes the square root of the variance term. Specifically, we have the asymptotic mean square error for $\hat{m}$,

$$\text{AMSE}(\hat{m}) = O(1) \times \left\{ \lambda \|m\|_{\mathcal{H}}^2 + n^{-1}\gamma(\lambda) \right\}. \tag{3.7}$$

For the ℓ-th order of Sobolev space, we have $\mu_j \leq Cj^{-2\ell}$ and

$$\gamma(\lambda) = \sum_{j=1}^{\infty} (1 + j^{2\ell}\lambda)^{-1} \leq O\left(\lambda^{-1/(2\ell)}\right). \tag{3.8}$$

Note that (3.7) is minimized when $\lambda \asymp \gamma(\lambda)/n$, which is equivalent to $\lambda \asymp n^{-2\ell/(2\ell+1)}$ under (3.8). The optimal rate $\lambda \asymp n^{-2\ell/(2\ell+1)}$ leads to

$$\text{AMSE}(\hat{m}) = O(n^{-2\ell/(2\ell+1)}) \tag{3.9}$$

which is the optimal rate in Sobolev space, as discussed by Stone (1982).

To investigate the asymptotic properties of the kernel ridge regression imputation estimator, we express

$$\begin{aligned}
\hat{\theta}_I &= \frac{1}{n}\sum_{i=1}^{n} \{\delta_i y_i + (1-\delta_i)\hat{m}(\mathbf{x}_i)\} \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n} m(\mathbf{x}_i)}_{R_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \delta_i \{y_i - m(\mathbf{x}_i)\}}_{S_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n} (1-\delta_i) \{\hat{m}(\mathbf{x}_i) - m(\mathbf{x}_i)\}}_{T_n}.
\end{aligned}$$

Therefore, as long as we show

$$T_n = \frac{1}{n}\sum_{i=1}^{n} \delta_i \left\{ \frac{1}{\pi(\mathbf{x}_i)} - 1 \right\} \{y_i - m(\mathbf{x}_i)\} + o_p(n^{-1/2}), \tag{3.10}$$

then we can establish the root-n consistency. The following theorem formally states the theoretical result.

Theorem 3.1. *Suppose Assumption 3.1-3.3 hold for a Sobolev kernel of order ℓ, as long as*

$$n\lambda \to 0, \quad n\lambda^{1/2\ell} \to \infty, \tag{3.11}$$

we have

$$n^{1/2}\left(\hat{\theta}_I - \theta\right) \xrightarrow{\mathcal{L}} N(0, \sigma^2),$$

where

$$\sigma^2 = \mathrm{var}\{E(Y \mid \mathbf{x})\} + E\{\mathrm{var}(Y \mid \mathbf{x})/\pi(\mathbf{x})\} = \mathrm{var}(\eta)$$

with

$$\eta = m(\mathbf{x}) + \delta \frac{1}{\pi(\mathbf{x})} \{y - m(\mathbf{x})\}. \tag{3.12}$$

Remark 3.1. *Note that the optimal rate $\lambda \asymp n^{-2\ell/(2\ell+1)}$ does not satisfy the first part of (3.11). To control the bias part, we need a smaller λ such as $\lambda = n^{-\kappa}$ with $\kappa > 1$. Similar conditions are used for bandwidth selection for nonparametric kernel regression*

$$nh \to \infty \text{ and } n^{1/2}h^2 \to 0.$$

for dim$(\mathbf{x}) = 1$. See Wang and Chen (2009) for details.

Remark 3.2. *Theorem 1 is presented for a Sololev kernel. For sub-Gaussian kernel whose eigenvalues satisfy that*

$$\mu_j \le c_1 \exp(-c_2 j^2),$$

where c_1, c_2 are positive constants, we can establish similar results. To see this, note that

$$\begin{aligned}
\gamma(\lambda) &= \sum_{j=1}^{\infty} \frac{\mu_j}{\mu_j + \lambda} \\
&\le c_2^{-1/2}\{-\log(\lambda)\}^{1/2} + \frac{1}{\lambda}\int_{c_2^{-1/2}\{-\log(\lambda)\}^{1/2}} \exp(-c_2 z^2) dz \\
&\le c_2^{-1/2}\{-\log(\lambda)\}^{1/2} + O(1),
\end{aligned}$$

where the second term in the last equation can be obtained by the Gaussian tail bound inequality. Therefore, as long as $n\lambda \to 0$ and $n\{-\log(\lambda)\}^{-1/2} \to \infty$, we have $n^{-1}\mathbf{1}_n^{\mathrm{T}}\mathbf{a}_n = o_p(n^{-1/2})$ and the root-n consistency can be established.

Note that the asymptotic variance of the imputation estimator is equal to $n^{-1}\sigma^2$, which is the lower bound of the semiparametric asymptotic variance discussed in Robins et al. (1994). Thus, the kernel ridge regression imputation is asymptotically optimal. The influence function in (3.12)

can be used for variance estimation of $\hat{\theta}_I$. The idea is to estimate the influence function $\eta_i = m(\mathbf{x}_i) + \delta_i\{\pi(\mathbf{x}_i)\}^{-1}\{y_i - m(\mathbf{x}_i)\}$ and apply the standard variance estimator using $\hat{\eta}_i$. To estimate η_i, we need an estimator of $\pi(\mathbf{x})$. In the next section, we will consider a version of kernel ridge regression to estimate $\omega(x) = \{\pi(\mathbf{x})\}^{-1}$ directly. Once $\hat{\omega}_i(\mathbf{x})$ is obtained, we can use

$$\hat{\mathrm{V}} = \frac{1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(\hat{\eta}_i - \bar{\eta}_n)^2$$

as a variance estimator of $\hat{\theta}_I$ in (3.1), where

$$\hat{\eta}_i = \hat{m}(\mathbf{x}_i) + \delta_i\hat{\omega}_i(\mathbf{x}_i)\{y_i - \hat{m}(\mathbf{x}_i)\}$$

and $\bar{\eta}_n = n^{-1}\sum_{i=1}^{n}\hat{\eta}_i$.

3.5 Propensity Score Estimation

We now consider estimation of $\omega(x) = \{\pi(\mathbf{x})\}^{-1}$ using kernel ridge regression . In order to estimate $\omega(\mathbf{x}) = \{\pi(\mathbf{x})\}^{-1}$, we wish to develop a nonparametric method of estimating $\omega(\mathbf{x})$ using the same reproducing kernel Hilbert space theory. To do this, first define the following density ratio function

$$g(\mathbf{x}) = \frac{f(\mathbf{x} \mid \delta = 0)}{f(\mathbf{x} \mid \delta = 1)}, \tag{3.13}$$

and, by Bayes theorem, we have

$$\omega(\mathbf{x}) = \frac{1}{\pi(\mathbf{x})} = 1 + \frac{n_0}{n_1}g(\mathbf{x})$$

where $n_0 = n - n_1$. Thus, to estimate $\omega(\mathbf{x})$, we have only to estimate the density ration function $g(\mathbf{x})$ in (3.13). Now, to estimate $g(\mathbf{x})$ nonparametrically, we use the maximum entropy method (Nguyen et al., 2010) for density ratio function estimation.

For convenience, let $f_k(\mathbf{x}) = f(\mathbf{x} \mid \delta = k)$, for $k = 0, 1$. To explain the kernel ridge estimation estimation of $g(\mathbf{x})$, note that $g(\mathbf{x})$ can be understood as the maximizer of the Kullback-Leibler

divergence between f_0 and f_1, i.e.,

$$\begin{aligned} D_{KL}(f_0, f_1) &\geq \arg\max_{g>0} Q(g) \\ &= \arg\max_{g>0} \int \log\{g(\mathbf{x})\} f_0(\mathbf{x}) d\mu(\mathbf{x}) - \int g(x) f_1(\mathbf{x}) d\mu(x) \\ &= \arg\max_{g>0} \int g(\mathbf{x})[\log\{g(\mathbf{x})\} - 1] f_1(\mathbf{x}) d\mu(\mathbf{x}). \end{aligned} \tag{3.14}$$

That is, by (3.14), a sample version of $Q(g)$ can be written as

$$\hat{Q}(g) = \frac{1}{n_1} \sum_{i=1}^{n} \delta_i g(\mathbf{x}_i)[\log\{g(\mathbf{x}_i)\} - 1],$$

where $n_1 = \sum_{i=1}^{n} \delta_i$.

Since $g(\mathbf{x})$ is unknown, we want to impose constraints to formulate an M-estimation problem for $g(\mathbf{x})$. Given $\hat{m}(\cdot)$, using the idea of model calibration (Wu and Sitter, 2001), we would like to use

$$\frac{1}{n_1} \sum_{i=1}^{n} \delta_i g(\mathbf{x}_i) \hat{m}(\mathbf{x}_i) = \frac{1}{n_0} \sum_{i=1}^{n} (1 - \delta_i) \hat{m}(\mathbf{x}_i)$$

as a constraint for density ratio estimation. Note that it is algebraically equivalent to

$$\frac{1}{n} \sum_{i=1}^{n} \delta_i \left\{ 1 + \frac{n_0}{n_1} \cdot g(\mathbf{x}_i) \right\} \hat{m}(\mathbf{x}_i) = \frac{1}{n} \sum_{i=1}^{n} \hat{m}(\mathbf{x}_i).$$

Now, as we have $m \in \mathcal{H}$, and by the representer theorem in kernel ridge regression, we know that $\hat{m} \in \text{span}\{K(\cdot, \mathbf{x}_1), \ldots, K(\cdot, \mathbf{x}_n)\}$. Thus, the calibration constraint is

$$\frac{1}{n_1} \sum_{i=1}^{n} \delta_i g(\mathbf{x}_i)(K(\cdot, \mathbf{x}_1), \ldots, K(\cdot, \mathbf{x}_n))^{\mathrm{T}} = \frac{1}{n_0} \sum_{i=1}^{n} (1 - \delta_i)(K(\cdot, \mathbf{x}_1), \ldots, K(\cdot, \mathbf{x}_n))^{\mathrm{T}}. \tag{3.15}$$

This calibration property is also called covariate-balancing property (Imai and Ratkovic, 2014). Further, we want to incorporate with the normalization constraint $\sum_{i=1}^{n} \delta_i \omega(\mathbf{x}_i) = n$, i.e.,

$$\frac{1}{n_1} \sum_{i=1}^{n} \delta_i g(\mathbf{x}_i) = \frac{1}{n_0} \sum_{i=1}^{n} (1 - \delta_i). \tag{3.16}$$

Minimizing $\hat{Q}(g)$ subject to (3.15) and (3.16) is called the maximum entropy method. Using Lagrangian multiplier method, the solution to this optimization problem can be written as

$$\log\{g(\mathbf{x})\} \equiv \log\{g(\mathbf{x}; \boldsymbol{\phi})\} = \phi_0 + \sum_{i=1}^{n} \phi_i K(\mathbf{x}, \mathbf{x}_i) \tag{3.17}$$

for some $\boldsymbol{\phi} = (\phi_0, \ldots, \phi_n)^{\mathrm{T}} \in \mathbb{R}^{n+1}$. Thus, using the parametric form in (3.17), the optimization problem can be expressed as a dual form

$$\hat{Q}_0(\boldsymbol{\phi}) = \frac{1}{n_0}\sum_{i=1}^{n}(1-\delta_i)\log\{g(\mathbf{x}_i;\boldsymbol{\phi})\} - \frac{1}{n_1}\sum_{i=1}^{n}\delta_i g(\mathbf{x}_i;\boldsymbol{\phi}),$$

to formulate a legitimate estimation of $g(\cdot)$. Further, define $h(\mathbf{x};\boldsymbol{\phi}_1) = \log\{g(\mathbf{x};\boldsymbol{\phi}) - \phi_0\}$, where $\boldsymbol{\phi}_1 = (\phi_1, \ldots, \phi_n)^{\mathrm{T}}$. In our problem, to ensure the Representer theorem, we wish to find h that minimizes

$$-\hat{Q}_0(g;\boldsymbol{\phi}) + \tau \left\|h\right\|_{\mathcal{H}}^2 \tag{3.18}$$

over $\boldsymbol{\phi}$.

Hence, using the representer theorem again, the solution to (3.18) can be obtained as

$$\min_{\boldsymbol{\phi}_1 \in \mathbb{R}^n}\left\{\frac{1}{n_1}\sum_{i=1}^{n}\delta_i g(\mathbf{x}_i;\boldsymbol{\phi}) - \frac{1}{n_0}\sum_{i=1}^{n}(1-\delta_i)\log\{g(\mathbf{x}_i;\boldsymbol{\phi})\} + \tau\boldsymbol{\phi}^{\mathrm{T}}\mathbf{K}\boldsymbol{\phi}\right\} \tag{3.19}$$

and ϕ_0 is a normalizing constant satisfying

$$n_1 = \sum_{i=1}^{n}\delta_i \exp\{\phi_0 + \sum_{j=1}^{n}\hat{\phi}_j K(\mathbf{x}_i, \mathbf{x}_j)\}. \tag{3.20}$$

Thus, we use

$$\hat{g}(x) = \exp\{\hat{\phi}_0 + \sum_{j=1}^{n}\hat{\phi}_j K(\mathbf{x}, \mathbf{x}_j)\}$$

as the maximum entropy estimator of the density ratio function $g(\mathbf{x})$ using kernel method. Also,

$$\hat{\omega}(\mathbf{x}) = 1 + \frac{n_0}{n_1}\hat{g}(\mathbf{x})$$

is the maximum entropy estimator of $\omega(x) = \{\pi(\mathbf{x})\}^{-1}$. The estimator of $\omega(x)$ satisfies the calibration property by construction. That is, for any function $f(\mathbf{x}) \in \mathcal{H}$, we have

$$n^{-1}\sum_{i=1}^{n}\delta_i\hat{\omega}(\mathbf{x}_i)f(\mathbf{x}_i) = n^{-1}\sum_{i=1}^{n}f(\mathbf{x}_i).$$

The tuning parameter τ is chosen to minimize

$$D(\tau) = \left\|\frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{1 + \frac{n_0}{n_1}\cdot\hat{g}_\tau(x_i)\right\}\hat{m}(x_i) - \frac{1}{n}\sum_{i=1}^{n}\hat{m}(x_i)\right\|,$$

where $\hat{m}(x)$ is determined by kernel ridge regression estimation. Thus, we can use the following two-step procedure to determine the tuning parameter τ.

1. Use the kernel ridge regression to obtain $\hat{m}(\mathbf{x})$.

2. Given $\hat{m}(\mathbf{x})$, find $\hat{\tau}$ that minimizes $D(\tau)$.

As the objective function in (3.19) is convex, we apply the limited-memory Broyden-Fletcher-Goldfarb-Shanno algorithm to solve the optimization problem with the following first order partial derivatives:

$$\begin{aligned}
\frac{\partial U}{\partial \phi_0} =& \frac{1}{n_1}\sum_{i=1}^{n}\delta_i \exp\left(\phi_0 + \sum_{j=1}^{n}\phi_j K(\mathbf{x}_i,\mathbf{x}_j)\right) - 1, \\
\frac{\partial U}{\partial \phi_k} =& \frac{1}{n_1}\sum_{i=1}^{n}\delta_i K(\mathbf{x}_i,\mathbf{x}_k)\exp\left(\phi_0 + \sum_{j=1}^{n}\phi_j K(\mathbf{x}_i,\mathbf{x}_j)\right) - \frac{1}{n_0}\sum_{i=1}^{n}(1-\delta_i)K(\mathbf{x}_i,\mathbf{x}_k) \\
& + 2\tau\sum_{i=1}^{n}K(\mathbf{x}_i,\mathbf{x}_k)\phi_i, \quad k = 1,\ldots,n,
\end{aligned}$$

where U is the objective function in (3.19).

Further, we can also obtain the propensity score estimator based the above procedure, i.e.,

$$\hat{\theta}_{PS} = \frac{1}{n}\sum_{i=1}^{n}\delta_i\hat{\omega}(\mathbf{x}_i)y_i. \tag{3.21}$$

We have

Theorem 3.2. *Under regularity conditions stated in the supplementary material, we have*

$$n^{1/2}\left(\hat{\theta}_{PS} - \theta\right) \xrightarrow{\mathcal{L}} N(0,\sigma^2), \tag{3.22}$$

where $\sigma^2 = \mathrm{var}(\eta)$ *and*

$$\eta = m(\mathbf{x}) + \delta\left\{1 + \frac{n_0}{n_1}g(\mathbf{x})\right\}\{y - m(\mathbf{x})\}.$$

Theorem 3.2 implies that shown the propensity score estimator in (3.21) using the above procedure achieves the same asymptotic variance as the kernel ridge regression imputation estimator. The regularity conditions and the sketch of proof of Theorem 3.2 are presented in the Appendix. We can use a linearized variance estimator to get a valid variance estimate based on Theorem 3.2, similar to Theorem 3.1.

3.6 Simulation Study

To evaluate the performance of the proposed imputation method and its variance estimator, we conduct a limited simulation study. We consider the continuous study variable with three different data generating models. In the three models, we keep the response rate around 60% and $\text{var}(Y) \approx 10$. Also, $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}, x_{i4})^{\mathrm{T}}$ are generated independently element-wise from the uniform distribution on the support $(1, 3)$. In the first model A, we use a linear regression model $y_i = 3 + 2.5x_{i1} + 2.75x_{i2} + 2.5x_{i3} + 2.25x_{i4} + \sigma\epsilon_i$ to obtain y_i, where $\{\epsilon_i\}_{i=1}^n$ are generated from standard normal distribution and $\sigma = 3^{1/2}$. In the model B, we use $y_i = 3 + (1/35)x_{i1}^2 x_{i2}^3 x_{i3} + 0.1x_{i4} + \sigma\epsilon_i$ to generate data with a nonlinear structure. The model C for generating the study variable is $y_i = 3 + (1/180)x_{i1}^2 x_{i2}^3 x_{i3} x_{i4}^2 + \sigma\epsilon_i$.

In addition to $\{(\mathbf{x}_i, y_i), i = 1, \ldots, n\}$, we consider two response mechanisms. The response indicator variable δ's for both mechanisms are independently generated from the Bernoulli distribution. In the first missing mechanism, the probability for the Bernoulli distribution is $\text{logit}\,(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\beta} + 2.5)$, where $\boldsymbol{\beta} = (-1.1, 0.5, -0.25, -0.1)^{\mathrm{T}}$ and $\text{logit}(p) = \log\{p/(1-p)\}$. In the secondthe probability for the Bernoulli distribution is $\text{logit}\,(-0.3 + 0.7x_1^2 - 0.5x_2 - 0.25x_3 - 0.25x_4)$. We considered two sample sizes $n = 500$ and $n = 1,000$ with 1,000 Monte Carlo replications. The reproducing kernel Hilbert space we employed in the simulation study is the second-order Sobolev space. From each sample, we consider four imputation methods: imputation and propensity score methods related to kernel ridge regression and the others are B-spline and linear regression. For B-spline method, we employ the generalized additive model by R package 'mgcv' (Wood, 2012) . Specifically, we used cubic spine with 15 knots for each coordinate with restricted maximum likelihood estimation method.

The simulation results in Figure 3.1 and Figure 3.2 show that four methods show similar results under the linear model (model A), but both kernel ridge regression imputation estimators and propensity score estimators show robust performance under the nonlinear models (models B and C). All kernel ridge regression related methods provide neglgible biases in all scenarios.

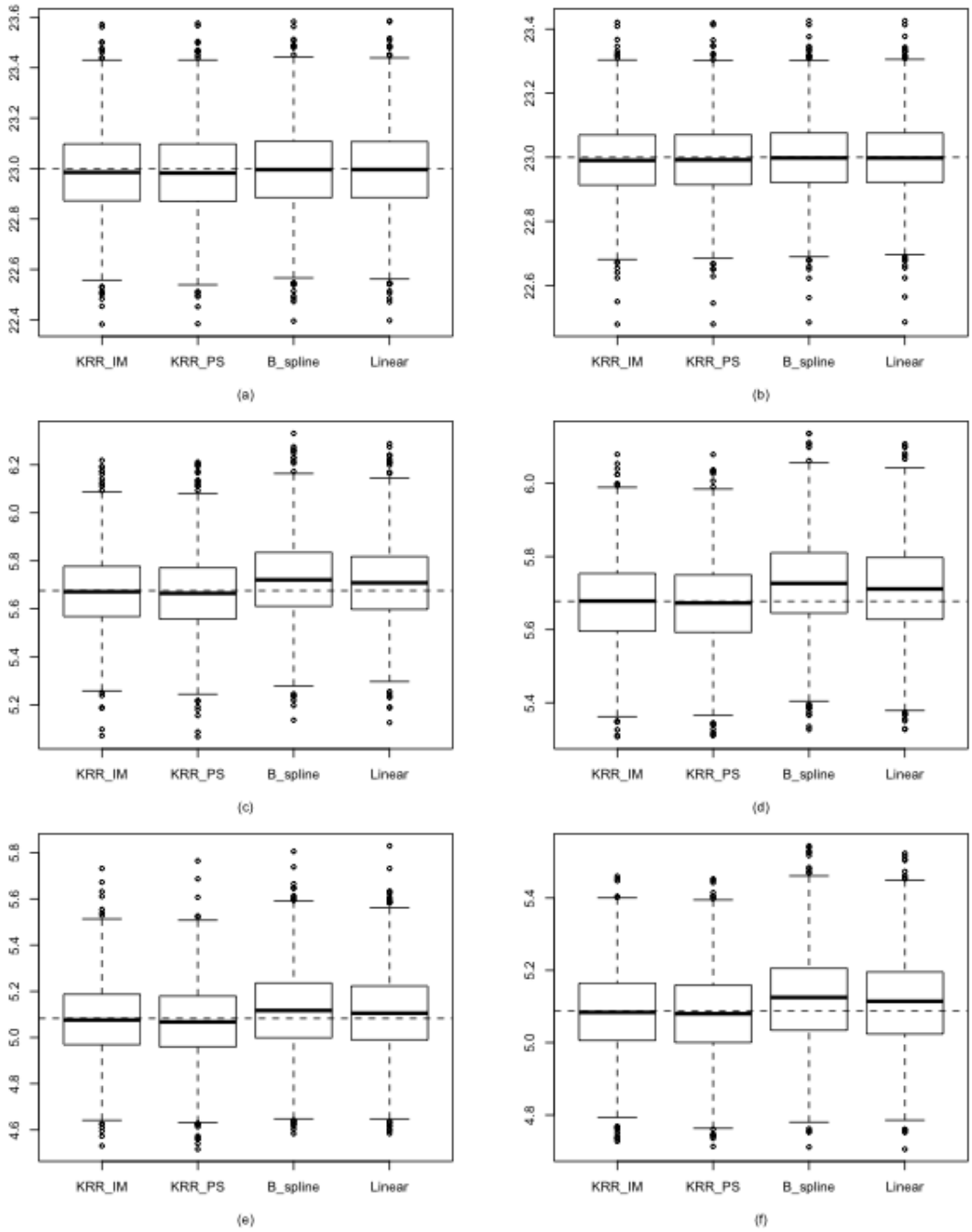


Figure 3.1: Boxplots with four estimators for model A ((a) for $n = 500$ and (b) for $n = 1000$), model B ((c) for $n = 500$ and (d) for $n = 1000$) and model C ((e) for $n = 500$ and (f) for $n = 1000$) under first response mechanism with true values (dashes). KRR_IM, kernel ridge regression imputation estimator; KRR_PS, kernel ridge regression propensity score estimator.

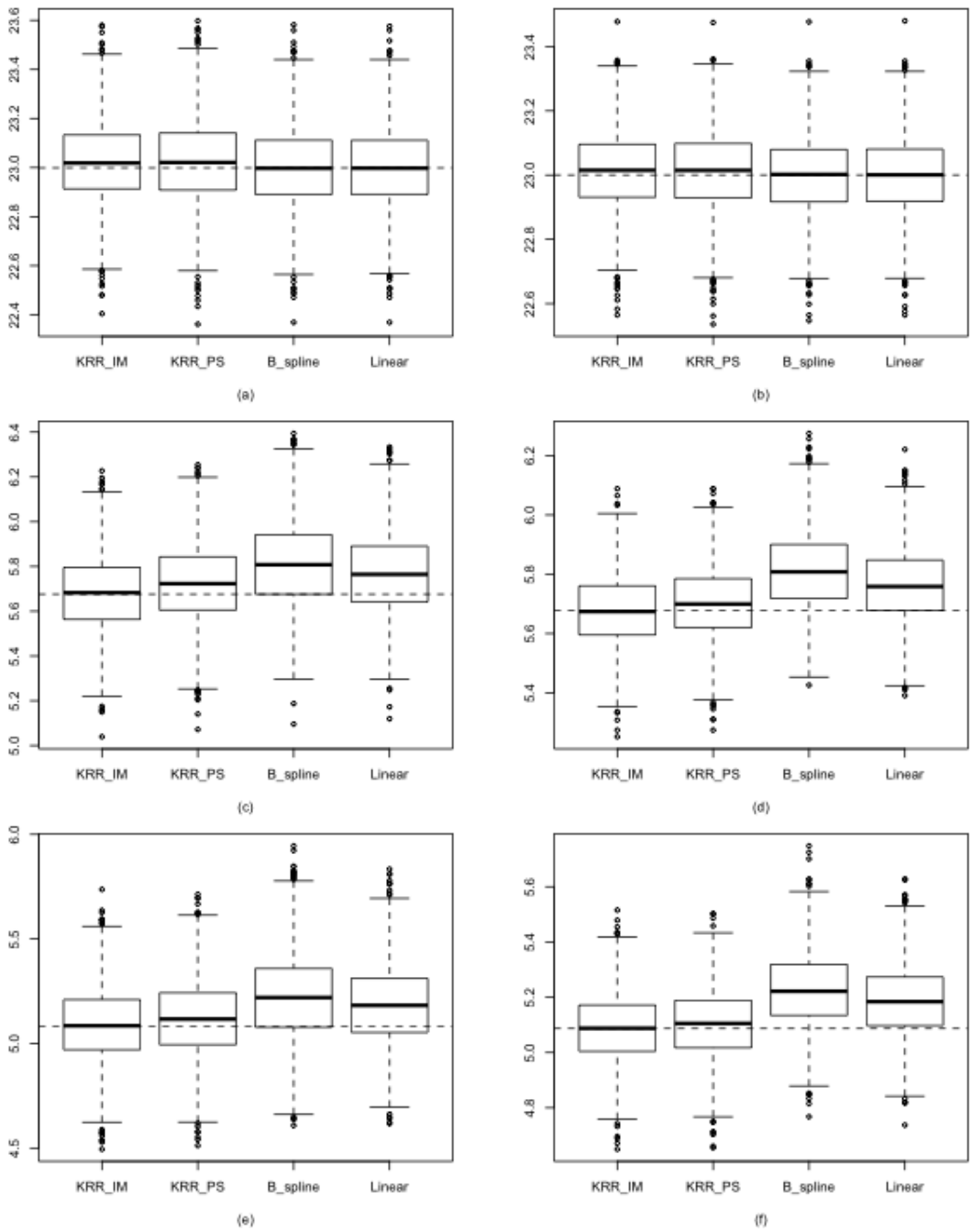


Figure 3.2: Boxplots with four estimators for model A ((a) for $n = 500$ and (b) for $n = 1000$), model B ((c) for $n = 500$ and (d) for $n = 1000$) and model C ((e) for $n = 500$ and (f) for $n = 1000$) under second response mechanism with true values (dashes). KRR_IM, kernel ridge regression imputation estimator; KRR_PS, kernel ridge regression propensity score estimator.

Table 3.1: Relative biases (R.B.) of the proposed variance estimator, coverage rates (C.R.) of the 90% and 95% confidence intervals for imputed estimators and propensity score estimators under kernel ridge regression with second-order Sobolev kernel and Gaussian kernel for continuous responses

Model	Criteria	First Missing Mechanism				Second Missing Mechanism			
		KRR_IM		KRR_PS		KRR_IM		KRR_PS	
		n=500	n=1000	n=500	n=1000	n=500	n=1000	n=500	n=1000
A	R.B(%)	0.09%	-2.8%	0.15%	-3.14%	3.4%	2.74%	-1.68%	-1.9%
	C.R.(90%)	90.3%	89.95%	90.3%	89.75%	90.25%	90.6%	89.15%	89.85%
	C.R.(95%)	95.5%	94.95%	95.7%	95%	95.2%	95.45%	94.65%	94.8%
B	R.B(%)	-2.77%	-5.42%	-5.77%	-6.6%	-6.07%	-3.42%	-11.25%	-6.23%
	C.R.(90%)	89.55%	89.7%	89.2%	89.2%	88.05%	90.05%	87.75%	89.3%
	C.R.(95%)	94.25%	94.55%	93.85%	94.1%	94.15%	94.7%	93.35%	94.1%
C	R.B(%)	-7.43%	-3.97%	-12.24%	-6.22%	-9.38%	-2.29%	-13.62%	-4.34%
	C.R.(90%)	87.95%	88.7%	86.7%	88.75%	88.8%	89.5%	87.5%	89.75%
	C.R.(95%)	93.35%	94.2%	92.35%	93.7%	93.95%	95.15%	93.25%	94.7%

In addition, we have computed the proposed variance estimators under kernel ridge regression imputation with the corresponding kernel. In Table 3.1, the relative biases (in percentage) of the proposed variance estimator and the coverage rates of two interval estimators under 90% and 95% nominal coverage rates are presented. The relative bias of the variance estimator are relatively low, which confirms the validity of the proposed variance estimator. Furthermore, the interval estimators show good performances in terms of the coverage rates.

3.7 Application

We applied the kernel ridge regression with the kernel of second-order Sobolev space to study the $PM_{2.5}(\mu g/m^3)$ concentration measured in Beijing, China (Liang et al., 2015). Hourly weather conditions: temperature, air pressure, cumulative wind speed, cumulative hours of snow and cumulative hours of rain are available from 2011 to 2015. Meanwhile, the averaged sensor response is subject to missingness. In December 2012, the missing rate of $PM_{2.5}$ is relatively high with missing rate 17.47%. We are interested in estimating the mean $PM_{2.5}$ in December with both imputed and propensity score kernel ridge regression estimates. The point estimates and

their 95% confidence intervals are presented in the Table 3.2. As a benchmark, the confidence interval computed from complete cases and confidence intervals for the imputed estimator under linear model (Kim and Rao, 2009) are also presented there. In addition, KRR_IM denotes the kernel ridge regression imputation estimator and KRR_PS denotes thes kernel ridge regression propensity score estimator.

Table 3.2: Point estimates (P.E.), standard error (S.E.) and 95% confidence intervals (C.I.) for imputed mean $PM_{2.5}$ in December, 2012 under kernel ridge regression

Estimator	P.E.	S.E.	95% C.I.
Complete	109.20	3.91	(101.53, 116.87)
Linear	99.61	3.68	(92.39, 106.83)
KRR_IM	101.92	3.50	(95.06, 108.79)
KRR_PS	102.25	3.50	(95.39, 109.12)

As we can see, the performances of kernel ridge regression imputation estimators are similar and created narrower 95% confidence intervals. Furthermore, the imputed $PM_{2.5}$ concentration during the missing period is relatively lower than the fully observed weather conditions on average. Therefore, if we only utilize the complete cases to estimate the mean of $PM_{2.5}$, the severeness of air pollution would be over-estimated.

3.8 Discussion

We consider kernel ridge regression as a tool for nonparametric imputation and establish its asymptotic properties. The proposed kernel ridge regression imputation can be used as a general tool for nonparametric imputation. By choosing different kernel functions, different nonparametric imputation methods can be developed. Asymptotic properties of the propensity score estimator are also established. The unified theory developed in this paper can cover various type of the kernel ridge regression imputation and enables us to make valid statistical inferences about the population means.

There are several possible extensions of the research. First, the theory can be directly applicable to other nonparametric imputation methods, such as smoothing splines (Claeskens

et al., 2009) or deep kernel learning (Bohn et al., 2019). Second, instead of using ridge-type penalty term, one can also consider other penalty functions such as the smoothly clipped absolute deviation penalty (Fan and Li, 2001) or adaptive lasso (Zou, 2006). Also, the proposed method can be used for causal inference, including estimation of average treatment effect from observational studies (Morgan and Winship, 2014; Yang and Ding, 2020). Developing tools for causal inference using the kernel ridge regression-based propensity score method will be an important extension of this research.

3.9 References

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404.

Bohn, B., Rieger, C., and Griebel, M. (2019). A represented theorem for deep kernel learning. *Journal of Machine Learning Research*, 20:1–32.

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78:673–700.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, 89(425):81–87.

Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96:529–544.

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

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20:25–46.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 243–263.

Kim, H. J., Reiter, J. P., Wang, Q., Cox, L. H., and Karr, A. F. (2014). Multiple imputation of missing or faulty values under linear constraints. *Journal of Business & Economic Statistics*, 32(3):375–386.

Kim, J. K. and Rao, J. (2009). A unified approach to linearization variance estimation from survey data after imputation for item nonresponse. *Biometrika*, 96(4):917–932.

Kim, J. K. and Shao, J. (2013). *Statistical methods for handling incomplete data*. CRC press.

Koltchinskii, V. (2006). Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656.

Liang, X., Zou, T., Guo, B., Li, S., Zhang, H., Zhang, S., Huang, H., and Chen, S. X. (2015). Assessing beijing's pm2. 5 pollution: severity, weather impact, apec and winter heating. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 471(2182):20150257.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Mendelson, S. (2002). Geometric parameters of kernel machines. In *International Conference on Computational Learning Theory*, pages 29–43. Springer.

Morgan, S. L. and Winship, C. (2014). *Counterfactuals and Causal Inference*. Cambridge University Press.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861.

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

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Sang, H., Kim, J. K., and Lee, D. (2020). Semiparametric fractional imputation using gaussian mixture models for handling multivariate missing data. *Journal of the American Statistical Association*, pages 1–10.

Scholkopf, B. and Smola, A. J. (2002). *Learning with Kernels*. The MIT Press.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.

Steinwart, I., Hush, D. R., and Scovel, C. (2009). Optimal rates for regularized least squares regression. In *COLT*, pages 79–93.

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

Tan, Z. (2020). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107(1):137–158.

van de Geer, S. A. (2000). *Empirical Processes in M-estimation*, volume 6. Cambridge university press.

Wahba, G. (1990). *Spline models for observational data*, volume 59. SIAM.

Wang, D. and Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, 37(1):490–517.

Wood, S. (2012). mgcv: Mixed gam computation vehicle with gcv/aic/reml smoothness estimation.

Wu, C. and Sitter, R. R. (2001). A model-calibration approach to using complete auxiliary information from survey data. *Journal of the American Statistical Association*, 96:185–193.

Yang, P. and Ding, P. (2020). Combining multiple observational data sources to estimate causal effects. *Journal of the American Statistical Association*, 115:1540–1554.

Zhang, T. (2005). Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098.

Zhang, Y., Duchi, J., and Wainwright, M. (2013). Divide and conquer kernel ridge regression. In *Conference on learning theory*, pages 592–617.

Zhao, Q. (2019). Covariate balancing propensity score by tailored loss functions. *The Annals of Statistics*, 47:965–993.

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

3.10 Appendix: Technical Details

3.10.1 Proof for Theorem 3.1

To prove our main theorem, we write

$$\begin{aligned}\hat{\theta}_I &= \frac{1}{n}\sum_{i=1}^{n}\{\delta_i y_i + (1-\delta_i)\hat{m}(\mathbf{x}_i)\} \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^{n} m(\mathbf{x}_i)}_{R_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}\delta_i\{y_i - m(\mathbf{x}_i)\}}_{S_n} + \underbrace{\frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\{\hat{m}(\mathbf{x}_i) - m(\mathbf{x}_i)\}}_{T_n}.\end{aligned}$$

Therefore, as long as we show

$$T_n = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{\frac{1}{\pi(\mathbf{x}_i)} - 1\right\}\{y_i - m(\mathbf{x}_i)\} + o_p(n^{-1/2}), \tag{3.23}$$

then the main theorem automatically holds.

To show (3.10), note that

$$\begin{aligned}\hat{\mathbf{m}} &= \mathbf{K}\left(\boldsymbol{\Delta}_n\mathbf{K} + \lambda\mathbf{I}_n\right)^{-1}\boldsymbol{\Delta}_n\mathbf{y} \\ &= \mathbf{K}\left\{\left(\boldsymbol{\Delta}_n + \lambda\mathbf{K}^{-1}\right)\mathbf{K}\right\}^{-1}\boldsymbol{\Delta}_n\mathbf{y} \\ &= \left(\boldsymbol{\Delta}_n + \lambda\mathbf{K}^{-1}\right)^{-1}\boldsymbol{\Delta}_n\mathbf{y},\end{aligned}$$

where $\hat{\mathbf{m}} = (\hat{m}(\mathbf{x}_1), \ldots, \hat{m}(\mathbf{x}_n))^{\mathrm{T}}$. Let $\mathbf{S}_\lambda = (\mathbf{I}_n + \lambda\mathbf{K}^{-1})^{-1}$, we have

$$\hat{\mathbf{m}} = \left(\boldsymbol{\Delta}_n + \lambda\mathbf{K}^{-1}\right)^{-1}\boldsymbol{\Delta}_n\mathbf{y} = \mathbf{C}_n^{-1}\mathbf{d}_n,$$

where

$$\begin{aligned}\mathbf{C}_n &= \mathbf{S}_\lambda\left(\boldsymbol{\Delta}_n + \lambda\mathbf{K}^{-1}\right), \\ \mathbf{d}_n &= \mathbf{S}_\lambda\boldsymbol{\Delta}_n\mathbf{y}.\end{aligned}$$

By Lemma 2, we obtain

$$\begin{aligned}\mathbf{C}_n &= E(\boldsymbol{\Delta}_n \mid \mathbf{x}) + \mathbf{a}_n \\ &= \boldsymbol{\Pi} + \mathbf{a}_n,\end{aligned}$$

where $\mathbf{\Pi} = \mathrm{diag}(\pi(\mathbf{x}_1), \ldots, \pi(\mathbf{x}_n))$ and $\gamma(\lambda)$ is the effective dimension of kernel K. Similarly, we have

$$\begin{aligned}\mathbf{d}_n &= E(\mathbf{\Delta}_n \mathbf{y} \mid \mathbf{x}) + \mathbf{a}_n \\ &= \mathbf{\Pi}\mathbf{m} + \mathbf{a}_n.\end{aligned}$$

Consequently, by Taylor expansion, we have

$$\begin{aligned}\hat{\mathbf{m}} &= \mathbf{m} + \mathbf{\Pi}^{-1}\left(\mathbf{d}_n - \mathbf{C}_n\mathbf{m}\right) + o_p\left(\mathbf{a}_n\right) \\ &= \mathbf{m} + \mathbf{\Pi}^{-1}\left\{\mathbf{S}_\lambda \mathbf{\Delta}_n \mathbf{y} - \mathbf{S}_\lambda\left(\mathbf{\Delta}_n + \lambda\mathbf{K}^{-1}\right)\mathbf{m}\right\} + o_p\left(\mathbf{a}_n\right) \\ &= \mathbf{m} + \mathbf{\Pi}^{-1}\mathbf{S}_\lambda \mathbf{\Delta}_n\left(\mathbf{y} - \mathbf{m}\right) + O_p(\mathbf{a}_n),\end{aligned}$$

where the last equality holds because

$$\begin{aligned}\mathbf{S}_\lambda \lambda \mathbf{K}^{-1}\mathbf{m} &= \mathbf{S}_\lambda\left\{\left(\mathbf{I}_n + \lambda\mathbf{K}^{-1}\right) - \mathbf{I}_n\right\}\mathbf{m} \\ &= \mathbf{m} - \mathbf{S}_\lambda\mathbf{m} = O_p(\mathbf{a}_n).\end{aligned}$$

Therefore, we have

$$\begin{aligned}T_n &= n^{-1}\mathbf{1}_n^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{\Delta}_n\right)\left(\hat{\mathbf{m}} - \mathbf{m}\right) \\ &= n^{-1}\mathbf{1}_n^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{\Delta}_n\right)\mathbf{\Pi}^{-1}\mathbf{S}_\lambda\mathbf{\Delta}_n\left(\mathbf{y} - \mathbf{m}\right) + O_p(n^{-1}\mathbf{1}_n^{\mathrm{T}}\mathbf{a}_n) \\ &= n^{-1}\mathbf{1}_n^{\mathrm{T}}\left(\mathbf{I}_n - \mathbf{\Pi}\right)\mathbf{\Pi}^{-1}\mathbf{\Delta}_n\left(\mathbf{y} - \mathbf{m}\right) + O_p(n^{-1}\mathbf{1}_n^{\mathrm{T}}\mathbf{a}_n) \\ &= n^{-1}\mathbf{1}_n^{\mathrm{T}}\left(\mathbf{\Pi}^{-1} - \mathbf{I}_n\right)\mathbf{\Delta}_n\left(\mathbf{y} - \mathbf{m}\right) + O_p(n^{-1}\mathbf{1}_n^{\mathrm{T}}\mathbf{a}_n).\end{aligned}$$

For ℓ-th order of Sobolev space, we have

$$\begin{aligned}\gamma(\lambda) &= \sum_{j=1}^{\infty}\frac{1}{1 + j^{2\ell}\lambda} \\ &\leq \lambda^{-\frac{1}{2\ell}} + \sum_{\{j:j>\lambda^{-\frac{1}{2\ell}}\}}\frac{1}{1 + j^{2\ell}\lambda} \\ &\leq \lambda^{-\frac{1}{2\ell}} + \lambda^{-1}\int_{\lambda^{-\frac{1}{2\ell}}}^{\infty} z^{-2\ell}dz \\ &= \lambda^{-\frac{1}{2\ell}} + \frac{1}{2\ell - 1}\lambda^{-\frac{1}{2\ell}} \\ &= O\left(\lambda^{-\frac{1}{2\ell}}\right).\end{aligned}$$

Additionally,

$$n^{-1}\mathbf{1}_n^{\mathrm{T}}\mathbf{a}_n = O_p\left(\lambda^{1/2} + n^{-1}\{\gamma(\lambda)\}^{1/2}\right),$$

which implies that, as long as $n\lambda \to 0$ and $n\lambda^{1/2\ell} \to \infty$, holds, we have $n^{-1}\mathbf{1}_n^{\mathrm{T}}\mathbf{a}_n = o_p(n^{-1/2})$ and (3.10) is established.

3.10.2 Regularity Conditions and Proof for Theorem 3.2

To prove Theorem 3.2, we need the following lemma for consistency of our propensity score estimation.

Let $\mathbb{P}_0$ be the distribution of whose pdf is $f(\mathbf{x} \mid \delta = 0)$, $\mathbb{P}_1$ be the distribution of whose pdf is $f(\mathbf{x} \mid \delta = 1)$ and

$$g^\star(\mathbf{x}) = \frac{f(\mathbf{x} \mid \delta = 0)}{f(\mathbf{x} \mid \delta = 1)}.$$

Let $\hat{g}$ be the estimated density ratio function generated by (3.18) in the main article. Further, let $\mathcal{G} = \{\exp(h) : h \in \mathcal{H}\}$ be the function space generated by $\mathcal{H}$. Let $G = \sup_{g\in\mathcal{G}} |g(\mathbf{x})|$.

Lemma 3.3 (modified Theorem 1 in Nguyen et al. (2010))**.** *Suppose the Kullback-Leibler divergence $D_{KL}(\mathbb{P}_0||\mathbb{P}_1)$ is bounded. Further, suppose that there exists a $g \in \mathcal{G}$ such that $g = g^\star$ almost surely. Additionally, assume the envelope condition*

$$\int G d\mathbb{P}_1 < \infty,$$

and further for all $\delta > 0$,

$$\frac{1}{n}\mathcal{H}_\delta(\mathcal{G} - g^\star, L_1(\mathbb{P}_{1,n})) \xrightarrow{\mathbb{P}_1} 0,$$
$$\frac{1}{n}\mathcal{H}_\delta\left(\log\frac{\mathcal{G} + g^\star}{2g^\star}, L_1(\mathbb{P}_{0,n})\right) \xrightarrow{\mathbb{P}_0} 0,$$

where $\mathcal{H}_\delta(\mathcal{M}, L_1(\mathbb{Q}))$ denotes the δ-entropy of $\mathcal{M}$ for the $L_1(\mathbb{Q})$-metric and $\mathbb{P}_{0,n}$ is the empirical distribution for $\mathbb{P}_0$. Then $h_{\mathbb{P}_1}(g^\star, \hat{g}) \to 0$ almost surely, where $h_{\mathbb{P}_1}(g^\star, \hat{g})$ is the Hellinger distance between $g^\star$.

The proof of Lemma 3.3 is presented in Nguyen et al. (2010). Thus, the convergence of propensity score is obtained.

The regularity conditions for Theorem 3.2 are essentially the conditions for Theorem 3.1 and Lemma 3.3 in addition with $n^{1/2}\tau \to 0$, $\|m\|_{\mathcal{H}} = O_p(1)$ and $\|h\|_{\mathcal{H}} = O_p(1)$. Still, we take Sobolev space of order ℓ as an example, the results for other kernels can be modified accordingly. Now we introduce the proof of Theorem 2.

Proof. We have

$$\begin{aligned}\hat{\theta}_{PS} &= \frac{1}{n}\sum_{i=1}^{n} m(\mathbf{x}_i) + \frac{1}{n}\sum_{i=1}^{n} \delta_i \hat{w}(\mathbf{x}_i) e_i + \frac{1}{n}\sum_{i=1}^{n} \{\delta_i \hat{w}(\mathbf{x}_i) - 1\}\hat{m}(\mathbf{x}_i) \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} \{\delta_i \hat{w}(\mathbf{x}_i) - 1\}\{m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i)\} \\ &= P_n + Q_n + R_n + S_n.\end{aligned}$$

Due to monotone transformation among g, h and w, the convergence of $\hat{g}$ in Lemma 3.3 implies the convergence of $\hat{h}$ and $\hat{w}$. As $\mathbb{E}(e_i|\mathbf{x}_i) = 0$, by convergence of $\hat{w}$, we have

$$Q_n = \frac{1}{n}\sum_{i=1}^{n} \delta_i w(\mathbf{x}_i) e_i + o_p(n^{-1/2}).$$

Note that

$$\delta_i \hat{w}(\mathbf{x}_i) - 1 = \frac{n_0}{n_1}\hat{r}(\mathbf{x}_i) - (1 - \delta_i),$$

by construction of propensity score estimator, the Fréchet derivative of (3.18) must satisfy

$$\frac{1}{n}\sum_{i=1}^{n} \{\delta_i \hat{w}(\mathbf{x}_i) - 1\}\xi_{\mathbf{x}_i} + \frac{n_0}{n}\hat{h} = 0, \tag{3.24}$$

where $\xi_{\mathbf{x}_i}$ is the reproducing kernel Hilbert space evaluator function. That is, for any function $f \in \mathcal{H}$, we have $\langle f, \xi_{\mathbf{x}_i}\rangle_{\mathcal{H}} = f(\mathbf{x}_i)$. Take the inner product in $\mathcal{H}$ of (3.24) with $\hat{m}$, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n} \{\delta_i \hat{w}(\mathbf{x}_i) - 1\}\hat{m}(\mathbf{x}_i) &= -\tau\frac{n_0}{n}\langle \hat{h}, \hat{m}\rangle_{\mathcal{H}} \\ &= -\tau\frac{n_0}{n}\{\langle h, m\rangle_{\mathcal{H}} + o_p(1)\} \\ &= \tau\frac{n_0}{n} \times O_p(1),\end{aligned}$$

where the second inequality holds due to the convegence of $\hat{m}$ and $\hat{h}$, the last equality holds due to Cauchy–Schwarz inequality. Thus, as long as $\tau = o(n^{-1/2})$, we have $R_n = o_p(n^{-1/2})$. Further, by Lemma 2 and the derivation in Theorem 1, $\mathbf{m}(\mathbf{x}_i) - \hat{\mathbf{m}}(\mathbf{x}_i) = O_p\left(\lambda^{1/2} + n^{-1/2}\{\gamma(\lambda)\}^{1/2}\right) = o_p(1)$, where the last equality is implied by the regularity condition for λ. Thus, we have $S_n = n^{-1}\sum_{i=1}^{n}\{\delta_i\hat{w}(\mathbf{x}_i) - 1\}\{m(\mathbf{x}_i) - \hat{m}(\mathbf{x}_i)\} = o_p(n^{-1/2})$. Therefore, we have established

$$\hat{\theta}_{PS} = \frac{1}{n}\sum_{i=1}^{n} m(\mathbf{x}_i) + \frac{1}{n}\sum_{i=1}^{n}\delta_i w(\mathbf{x}_i)e_i + o_p(n^{-1/2}).$$

□

CHAPTER 4. PROPENSITY SCORE ESTIMATION USING DENSITY RATIO MODEL UNDER ITEM NONRESPONSE

Hengfang Wang and Jae Kwang Kim

Iowa State University

Modified from a manuscript to be submitted to *Journal of the American Statistical Association*

4.1 Abstract

Missing data is frequently encountered in practice. Propensity score estimation is a popular tool for handling such missingness. The propensity score is often developed using the model for the response probability which can be subject to model misspecification. In this paper, we consider an alternative approach of estimating the inverse of the propensity scores using density ratio function. By partitioning the sample into two groups based on the response status of the elements, we can apply the density ratio function estimation method and obtain the inverse propensity scores for nonresponse adjustment. Density ratio estimation can be obtained by applying the so-called maximum entropy method which uses the Kullback-Leibler divergence measure under calibration constraints. By including the covariates for the outcome regression models only into the density ratio model, we can achieve efficient propensity score estimation. The proposed method can be extended to soft calibration in the kernel-based functional space. We further extend the proposed approach to the multivariate missing case. Some limited simulation studies are presented to compare with the existing methods.

4.2 Introduction

Missing data is frequently encountered in practice. The missingness can occur when the observational unit does not comply with the measurement or when the measurement tool does

not work for some mechanical reason or by mistake. In some cases, missingness is planned to reduce the cost or to reduce the response burden. Causal inference can be viewed as a missing data problem. Making valid statistical inference in spite of the missing data is a fundamental problem in statistics. (Kim and Shao, 2013; Little and Rubin, 2019).

Propensity score (PS) approach is a popular approach to handling the missing data problem using inverse weighting. The propensity score is often developed using the model for the response probability. In principle, regression models for binary response, e.g., logistic regression, can be utilized to model the response probability given the observed auxiliary information. An inverse probability weighting estimator can then be constructed to get an unbiased estimation of the target parameter. However, correct specification of the propensity score model can be challenging and we often do not have a good understanding of the response mechanism to specify the propensity model correctly. Furthermore, the final estimation can be unstable when some propensity scores are close to zero. The variance and bias of the resultant estimator are likely to be amplified by the nature of such 'inverse' fashion.

The existing methods for propensity score estimation are either based on maximum likelihood method (Rosenbaum and Rubin, 1983; Robins et al., 1994; Tan, 2006) or calibration method (Folsom, 1991; Tan, 2010; Graham et al., 2012; Hainmueller, 2012; Imai and Ratkovic, 2014; Kim and Haziza, 2014; Vermeulen and Vansteelandt, 2015; Chan et al., 2016; Tan, 2020) with some penalization in the calibration equation. The calibration method gives a doubly robust flavour, but the choice of the objective function for calibration estimation is not fully agreed.

In this chapter, we propose an alternative framework for inverse propensity weighting without modeling the response mechanism directly. By using density ratio representation of the inverse propensity scores, we now estimate the density ratio function directly. By including the covariates for the outcome regression models only into the density ratio model, we can achieve efficient propensity score estimation. The proposed method provides a unified framework for developing calibrated propensity score estimation.

Under the new framework for propensity score method employing the density ratio representation, we propose a two-step procedure for propensity score estimation. In the first step, we select important features from the outcome model and form a finite dimensional set of basis functions for modeling the log density ratio function. In the second step, we estimate the density ratio function using maximum entropy method (Nguyen et al., 2010) with the basis functions selected from the first step.

The first step is important in reducing the variance which might be caused by including nuisance variables into DRE. That is, the PS estimator can be inefficient when there are more auxiliary variables than we actually need (Shimodaira, 2000; Shortreed and Ertefaie, 2017). By including only important covariates for the outcome model to the log density model, we can obtain efficient PS estimation. We consider two main approaches to do the dimension reduction, one is variable selection and the other is sufficient dimension reduction (SDR). Given the observed study variable and the corresponding auxiliary variables, penalization methods (Tibshirani, 2011) can be implemented to select important covariates. For SDR, we wish to find sparse transformation of covariates such that the study variable and observed indicator are independent under such transformation(Fukumizu et al., 2004; Stojanov et al., 2019). We present some asymptotic theory of the resulting PS estimator. Variance estimation of the PS estimator can be implemented using either linearization method or bootstrap.

Furthermore, using the density ratio framework, the PS estimation can be easily extended to handle multivariate missing data. In multivariate missing patterns, we can partition the sample into multiple groups and apply the DRE method to obtain the inverse propensity scores. The proposed method under multivariate missing data setup can be used to combine information from multiple sources.

4.3 Basic Setup

Suppose that the parameter θ of interest can be written as a solution to $\mathbb{E}\{U(\theta; \mathbf{X}, Y)\} = 0$, where Y is the study variable that is subject to missingness and $\mathbf{X}$ is the auxiliary variable that is

always observed. Suppose we have an n *i.i.d.* realization of $(\mathbf{X}, Y)$, denoted as $\{(\mathbf{x}_i, y_i) : i = 1, \ldots, n\}$. Since Y is subject to missingness, the actual dataset we usually have is $\{(\mathbf{x}_i, \delta_i y_i, \delta_i) : i = 1, \cdots, n\}$ where δ_i is the response indicator variable defined as

$$\delta_i = \begin{cases} 1 & \text{if } y_i \text{ is observed} \\ 0 & \text{otherwise.} \end{cases}$$

We assume that the response mechanism is missing at random (MAR) in the sense of Rubin (1976). That is,

$$Y \perp \delta \mid \mathbf{X}. \tag{4.1}$$

To introduce the density ratio function, we use $f(\cdot)$ to denote generic density functions. In particular, let $f(\mathbf{x})$ to denote the density $f(\mathbf{X} = \mathbf{x})$. Further, let $f_j(\mathbf{x})$ denote the conditional density of $f(\mathbf{X} = \mathbf{x} \mid \delta = j)$ for $j = 0, 1$. Using this notation, we define the following density ratio function:

$$r(\mathbf{x}) = \frac{f_0(\mathbf{x})}{f_1(\mathbf{x})}.$$

By Bayes theorem, we obtain

$$\frac{\mathbb{P}(\delta = 0 \mid \mathbf{X})}{\mathbb{P}(\delta = 1 \mid \mathbf{X})} = \frac{\mathbb{P}(\delta = 0)}{\mathbb{P}(\delta = 1)} \times r(\mathbf{X}).$$

Assuming $c = \mathbb{P}(\delta = 0)/\mathbb{P}(\delta = 1)$ is known, we can express

$$\frac{1}{\mathbb{P}(\delta = 1 \mid \mathbf{X})} = 1 + c \cdot r(\mathbf{X})$$

and use

$$w_i = \frac{1}{\mathbb{P}(\delta_i = 1 \mid \mathbf{x}_i)} = 1 + c \cdot r(\mathbf{x}_i) \tag{4.2}$$

as the propensity score weight for unit i with $\delta_i = 1$. So, our PS weighting problem reduces to density ratio estimation. Note that we can easily estimate c by $\hat{c} = n_0/n_1$, $n_1 = \sum_{i=1}^{n} \delta_i$ and $n_0 = n - n_1$. The PS estimator of θ can be defined as the solution to $\hat{U}_{PS}(\theta) = 0$, where

$$\hat{U}_{PS}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \delta_i \left\{1 + \hat{c} \cdot r(\mathbf{x}_i)\right\} U(\theta; \mathbf{x}_i, y_i) \tag{4.3}$$

Traditionally, the PS estimator is justified under the response mechanism, which is based on the assumption that $\mathbb{P}(\delta_i = 1 \mid \mathbf{X}_i)$ is correctly specified. Note that the unbiasedness of $\hat{U}_{PS}(\theta)$ in (4.3) can also be justified without relying on the response mechanism. Under MAR in (4.1), we have

$$\begin{aligned}
\mathbb{E}\{r(\mathbf{X})U(\theta;\mathbf{X},Y) \mid \delta = 1\} &= \mathbb{E}\left\{\frac{f_0(\mathbf{X})}{f_1(\mathbf{X})}U(\theta;\mathbf{X},Y) \mid \delta = 1\right\} \\
&= \int \frac{f_0(\mathbf{x},y)}{f_1(\mathbf{x},y)}U(\theta;\mathbf{x},y)f_1(\mathbf{x},y)d\mu(\mathbf{x},y) \\
&= \int U(\theta;\mathbf{x},y)f_0(\mathbf{x},y)d\mu(\mathbf{x},y) \\
&= \mathbb{E}\left\{U(\theta;\mathbf{X},Y) \mid \delta = 0\right\}, \qquad (4.4)
\end{aligned}$$

where $\mu(\cdot)$ is the dominating measure of $(\mathbf{X}, Y)$. Thus, we can obtain

$$\mathbb{E}\{\hat{U}_{PS}(\theta) \mid \boldsymbol{\delta}\} = \mathbb{E}\{\hat{U}_n(\theta) \mid \boldsymbol{\delta}\}, \qquad (4.5)$$

where

$$\hat{U}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} U(\theta;\mathbf{x}_i,y_i) = \frac{1}{n}\sum_{i=1}^{n}\{\delta_i U(\theta;\mathbf{x}_i,y_i) + (1-\delta_i)U(\theta;\mathbf{x}_i,y_i)\}$$

and $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_n)^{\mathrm{T}}$. Since $\hat{U}_n(\theta)$ is an unbiased estimating function, $\hat{U}_{PS}(\theta)$ is also unbiased by (4.5). Note that the reference distribution in (4.5) is the joint distribution of $(\mathbf{X}, Y)$ conditional on δ. This is the reverse of the classical approach to PS estimation which uses the response mechanism, the conditional distribution of δ given $(\mathbf{X}, Y)$. Thus, the proposed density ratio approach to PS estimation is a totally different framework.

4.4 Proposed Method

4.4.1 Density Ratio Model

To motivate the proposed method, let $\mathbf{b}(\mathbf{x})$ be an integrable function of $\mathbf{x}$ and consider the density ratio function using the density function of $\mathbf{b}(\mathbf{x})$. That is,

$$\tilde{r}(\mathbf{x}) = \frac{f(\mathbf{b}(\mathbf{x}) \mid \delta = 0)}{f(\mathbf{b}(\mathbf{x}) \mid \delta = 1)}. \qquad (4.6)$$

We can use $\tilde{r}(\mathbf{x})$ to construct a propensity score estimator of $\theta = \mathbb{E}(Y)$ as follows:

$$\hat{\theta}_{PS2} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \left\{1 + \frac{n_0}{n_1}\tilde{r}(\mathbf{x}_i)\right\} y_i, \tag{4.7}$$

where $\tilde{r}(\mathbf{x})$ is the density ratio function using the density function of $\mathbf{b}(\mathbf{x})$. The density ratio function is derived from the reduced propensity model $\tilde{\pi}(\mathbf{x}) = \mathbb{P}(\delta = 1 \mid \mathbf{b}(\mathbf{x}))$. That is, similarly to (4.2), we can obtain

$$\{\tilde{\pi}(\mathbf{x})\}^{-1} = 1 + c \cdot \tilde{r}(\mathbf{x}). \tag{4.8}$$

Thus, if the MAR condition holds conditional on $\mathbf{b}(\mathbf{X})$, that is,

$$Y \perp \delta \mid \mathbf{b}(\mathbf{X}), \tag{4.9}$$

then $\hat{\theta}_{PS2}$ in (4.7) is also unbiased under the response probability model using (4.8).

Condition (4.9) can be called reduced MAR as the MAR condition holds for given a summary statistic $\mathbf{b}(\mathbf{x})$. Rosenbaum and Rubin (1983) called $\mathbf{b}(\mathbf{x})$ in (4.9) as a balancing score and discuss a propensity score method using a parametric model for $\tilde{\pi}(\mathbf{x}) = \mathbb{P}(\delta = 1 \mid \mathbf{b}(\mathbf{x}))$. Apparently, under MAR in (4.1), a sufficient condition for (4.9) is

$$\mathbb{P}(\delta = 1 \mid \mathbf{x}) = \mathbb{P}(\delta = 1 \mid \mathbf{b}(\mathbf{x})).$$

The following lemma provides another sufficient condition for (4.9).

Lemma 4.1. *If MAR condition in (4.1) holds and the reduced model for y holds such that*

$$f(y \mid \mathbf{x}) = f(y \mid \mathbf{b}(\mathbf{x})), \tag{4.10}$$

then (4.9) holds.

As long as the reduced outcome model in (4.10) is true, $\hat{\theta}_{PS2}$ is unbiased. Further, the following lemma gives another insight on $\tilde{r}(\mathbf{x})$.

Lemma 4.2. *Let $\tilde{r}(\mathbf{x})$ be the density ratio function in (4.6). If (4.9) holds, then we obtain*

$$\tilde{r}(\mathbf{x}) = \mathbb{E}\left\{r(\mathbf{x}) \mid \mathbf{b}(\mathbf{x}), y, \delta = 1\right\}, \tag{4.11}$$

where $r(\mathbf{x}) = f_0(\mathbf{x})/f_1(\mathbf{x})$ is the true density ratio function.

Note that result (4.11) implies that

$$\mathbb{E}\{\hat{\theta}_{PS1} \mid \mathbf{b}(\mathbf{X}), Y, \boldsymbol{\delta}\} = \hat{\theta}_{PS2},$$

where $\hat{\theta}_{PS1} = n^{-1}\sum_{i=1}^{n}\delta_i\{1+\hat{c}\cdot r(\mathbf{x}_i)\}y_i$. Because $\hat{\theta}_{PS2}$ is based on $\tilde{r}(\mathbf{x})$, the smoothed density ratio function, it is called the smoothed propensity score estimator. Therefore, we can establish the following theorem.

Proposition 1. *Under (4.9), we obtain*

$$Var(\hat{\theta}_{PS1}) \geq Var(\hat{\theta}_{PS2}). \tag{4.12}$$

To estimate the smoothed density ratio function from the sample, we introduce the maximum entropy method in the following subsection.

4.4.2 Maximum Entropy Estimation

We have seen that the propensity score estimation problem reduces to the density ratio estimation problem. Density ratio estimation (DRE), the problem of estimating the ratio of two density functions for two different populations, is a fundamental problem in machine learning (Sugiyama et al., 2012). By partitioning the sample into two groups based on the response status, we can apply the DRE method and thus obtain the inverse propensity scores. One important method of DRE is so called the maximum entropy method which minimizes the Kullback-Leibler (KL) divergence (or negative entropy) subject to the normalization constraint (Nguyen et al., 2010).

To explain the proposed DRE method, suppose that we have 2 probability distributions $\mathbb{P}_0, \mathbb{P}_1$, with $\mathbb{P}_0$ absolutely continuous with respect to $\mathbb{P}_1$. For simplicity, we also assume that $\mathbb{P}_k$ are absolutely continuous with respect to Lebesgue measure μ, with density f_k with support $\mathcal{X} \subset \mathbb{R}^p$, for $k = 0, 1$. The KL divergence between $\mathbb{P}_0$ and $\mathbb{P}_1$ is defined by

$$D_{KL}(\mathbb{P}_0, \mathbb{P}_1) = \int \log\left(\frac{d\mathbb{P}_0}{d\mathbb{P}_1}\right) d\mathbb{P}_0 = \int \log\left(\frac{f_0}{f_1}\right) f_0 d\mu.$$

We are interested in estimating density ratio functions $r = f_0/f_1$ from the sample.

By Nguyen et al. (2010), we have a variational representation for the KL divergence, i.e.,

$$D_{KL}(\mathbb{P}_0, \mathbb{P}_1) = \sup_{r>0} \left\{ \int r \log(r)\, d\mathbb{P}_1 - \int r d\mathbb{P}_1 + 1 \right\},$$

so that the density ratio function can be understood as the maximizer of

$$Q(r) = \int r \log(r)\, f_1 d\mu - \int r f_1 d\mu. \tag{4.13}$$

The sample version objective function is

$$\hat{Q}(r) = \frac{1}{n_1} \sum_{i=1}^{n} \delta_i r(\mathbf{x}_i) \left[\log\{r(\mathbf{x}_i)\} - 1 \right]. \tag{4.14}$$

The maximizer of $\hat{Q}(r)$ is an M-estimator of the density ratio function r.

To find the constraints for M-estimation of $\tilde{r}$, note that, by the definition of $\tilde{r}(\mathbf{x})$,

$$\begin{aligned} \mathbb{E}\{\tilde{r}(\mathbf{X})\mathbf{b}(\mathbf{X}) \mid \delta = 1\} &= \int \tilde{r}(\mathbf{x})\mathbf{b}(\mathbf{x}) f(\mathbf{b}(\mathbf{x}) \mid \delta = 1) d\mu \\ &= \int \mathbf{b}(\mathbf{x}) f(\mathbf{b}(\mathbf{x}) \mid \delta = 0) d\mu \\ &= \mathbb{E}\{\mathbf{b}(\mathbf{X}) \mid \delta = 0\}. \end{aligned} \tag{4.15}$$

Thus, from the integral equation in (4.15), we can construct an estimating equation for the density ratio function as

$$\frac{1}{n_1} \sum_{i=1}^{n} \delta_i \tilde{r}(\mathbf{x}_i)\mathbf{b}(\mathbf{x}_i) = \frac{1}{n_0} \sum_{i=1}^{n} (1 - \delta_i)\mathbf{b}(\mathbf{x}_i), \tag{4.16}$$

which is a finite-sample version of the integral equation in (4.15). Note that constraint (4.16) is equivalent to

$$\sum_{i=1}^{n} \delta_i \left\{ 1 + \frac{n_0}{n_1} \cdot \tilde{r}(\mathbf{x}_i) \right\} \mathbf{b}(\mathbf{x}_i) = \sum_{i=1}^{n} \mathbf{b}(\mathbf{x}_i), \tag{4.17}$$

which is called the calibration property (Deville and Särndal, 1992) or the covariate-balancing property (Imai and Ratkovic, 2014). The choice of the control function $\mathbf{b}(\mathbf{x})$ will be discussed in Section 3.3.

We propose to obtain an estimator of $\tilde{r}(\mathbf{x})$ by solving the optimization problem using $\hat{Q}(r)$ as the objective function with constraint (4.17). Using Lagrange multiplier method, we maximize

$$\hat{Q}_\lambda(r) = \frac{1}{n_1}\sum_{i=1}^{n}\delta_i r_i\{\log(r_i)-1\} + \boldsymbol{\lambda}'\left[\frac{1}{n_1}\sum_{i=1}^{n}\delta_i r_i \mathbf{b}(\mathbf{x}_i) - \frac{1}{n_0}\sum_{i=1}^{n}(1-\delta_i)\mathbf{b}(\mathbf{x}_i)\right].$$

Note that

$$\frac{\partial}{\partial r_i}\hat{Q}_\lambda(r) = \frac{1}{n_1}\delta_i\log(r_i) + \boldsymbol{\lambda}'\frac{1}{n_1}\delta_i\mathbf{b}(\mathbf{x}_i).$$

The solution $\partial\hat{Q}_\lambda(r)/\partial r_i = 0$ gives the form of the solution to the constrained optimization problem. Thus, we obtain that the solution satisfies the log-linear model for the density ratio function:

$$\log\{r(\mathbf{x};\boldsymbol{\phi})\} = \mathbf{b}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\phi} \tag{4.18}$$

for some $\boldsymbol{\phi}\in\mathbb{R}^{l+1}$. Roughly speaking, the calibration property means $r(\mathbf{x})\in\mathcal{H}^{\perp}$, where $\mathcal{H}^{\perp}$ is the orthogonal complement space of $\mathcal{H} = \mathrm{span}\{\mathbf{b}(\mathbf{x})\}$. Thus, the solution $\hat{r}(\mathbf{x})$ that is obtained by maximizing $\hat{Q}(r)$ in (4.14) subject to the calibration constraint in (4.17) can be viewed as a low-dimensional projection of $r(\mathbf{x})$ onto the space $\mathcal{H}^{\perp}$.

Based on the fact

$$\int r\{\log(r)-1\}p_1 d\mu = \int \log(r)p_0 d\mu - \int r p_1 d\mu,$$

the objective function (4.14) can be simplified as

$$\hat{Q}(r) = \frac{1}{n_0}\sum_{i=1}^{n}(1-\delta_i)\log(r_i) - \frac{1}{n_1}\sum_{i=1}^{n}\delta_i r_i. \tag{4.19}$$

Combining (4.18) and (4.19), the optimization problem reduces to maximizing

$$\hat{Q}(\boldsymbol{\phi}) = \frac{1}{n_0}\sum_{i=1}^{n}(1-\delta_i)\{\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\phi}\} - \frac{1}{n_1}\sum_{i=1}^{n}\delta_i\exp\{\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\phi}\}. \tag{4.20}$$

The maximizer of $\hat{Q}(\boldsymbol{\phi})$ satisfies $\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}) = \mathbf{0}$, where

$$\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}) = \frac{1}{n_1}\sum_{i=1}^{n}\delta_i\exp\{\mathbf{b}(\mathbf{x}_i)^{\mathrm{T}}\boldsymbol{\phi}\}\mathbf{b}(\mathbf{x}_i) - \frac{1}{n_0}\sum_{i=1}^{n}(1-\delta_i)\mathbf{b}(\mathbf{x}_i). \tag{4.21}$$

Note that the estimating equation using (4.21) leads to the covariate balancing property in (4.17).

4.4.3 Asymptotic Properties

We now develop the main asymptotic properties of the PS estimator

$$\hat{\theta}_{PS} = \frac{1}{n}\sum_{i=1}^{n} \delta_i \left\{1 + \frac{n_0}{n_1}\hat{r}(\mathbf{x}_i)\right\} y_i, \tag{4.22}$$

where $\hat{r}(\mathbf{x})$ is the maximum entropy estimator of the density ratio function under model (4.18). That is, $\hat{r}(\mathbf{x}) = \exp\{\mathbf{b}^{\mathrm{T}}(\mathbf{x})\hat{\boldsymbol{\phi}}\}$ and $\hat{\boldsymbol{\phi}}$ is the solution to the estimating equations using $\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi})$ in (4.21). Let $\boldsymbol{\phi}^{\star}$ be the maximizer of $Q(r)$ in (4.13) in the parametric class in (4.18) such that $\exp\{\mathbf{b}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\phi}^{\star}\} = \tilde{r}(\mathbf{x})$. Note that

$$\begin{aligned}
\mathbb{E}\{\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^{\star}) \mid \boldsymbol{\delta}\} &= \mathbb{E}\left\{\frac{1}{n_1}\sum_{i=1}^{n}\delta_i\tilde{r}(\mathbf{x}_i)\mathbf{b}(\mathbf{x}_i) - \frac{1}{n_0}\sum_{i=1}^{n}(1-\delta_i)\mathbf{b}(\mathbf{x}_i) \mid \boldsymbol{\delta}\right\} \\
&= \mathbb{E}\{\tilde{r}(\mathbf{x})\mathbf{b}(\mathbf{x}) \mid \delta = 1\} - \mathbb{E}\{\mathbf{b}(\mathbf{x}) \mid \delta = 0\} \\
&= \mathbb{E}\{\mathbf{b}(\mathbf{x}) \mid \delta = 0\} - \mathbb{E}\{\mathbf{b}(\mathbf{x}) \mid \delta = 0\} = \mathbf{0},
\end{aligned}$$

where the third equality follows from (4.4). Note that the reference distribution is the conditional distribution of $\mathbf{X}$ given δ. The unbiasedness of $\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^{\star})$ can also be derived under the response probability model associated with (4.18). That is, under the response model

$$\mathbb{P}(\delta = 1 \mid \mathbf{x}) = \frac{\exp\{\mathbf{b}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\phi}^{\star}\}}{1 + \exp\{\mathbf{b}^{\mathrm{T}}(\mathbf{x})\boldsymbol{\phi}^{\star}\}} := \tilde{\pi}(\mathbf{x}), \tag{4.23}$$

then we can also obtain $\mathbb{E}\{\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^{\star})\} = \mathbf{0}$.

Thus, as long as the sufficient conditions for $\mathbb{E}\{\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^{\star})\} = \mathbf{0}$ are satisfied, we can establish the weak consistency of $\hat{\boldsymbol{\phi}}$ and apply the standard Taylor linearization to obtain the following Theorem. The regularity conditions and the proof are presented in the Appendix.

Theorem 4.1. *Under the regularity conditions in the Appendix, we have*

$$\hat{\theta}_{PS} = \frac{1}{n}\sum_{i=1}^{n} d(\mathbf{x}_i, y_i, \delta_i; \boldsymbol{\phi}^{\star}) + o_p(n^{-1/2}), \tag{4.24}$$

where

$$d(\mathbf{x}_i, y_i, \delta_i; \boldsymbol{\phi}) = \mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\tilde{\boldsymbol{\beta}} + \delta_i\{1 + (n_0/n_1)\tilde{r}(\mathbf{x}_i; \boldsymbol{\phi})\}(y_i - \mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\tilde{\boldsymbol{\beta}}),$$

$\tilde{r}(\mathbf{x}_i;\boldsymbol{\phi}) = \exp\{\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\phi}\}$ *and* $\tilde{\boldsymbol{\beta}}$ *is the probability limit of the solution to*

$$\sum_{i=1}^{n} \delta_i \tilde{r}(\mathbf{x}_i;\boldsymbol{\phi})\mathbf{b}(\mathbf{x}_i)\left\{y_i - \mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\beta}\right\} = \mathbf{0}. \tag{4.25}$$

By Theorem 4.1, under the response probability model satisfying (4.23), we can establish

$$\sqrt{n}\left(\hat{\theta}_{PS} - \theta_0\right) \xrightarrow{\mathcal{L}} N(\mathbf{0}, V_r), \tag{4.26}$$

where

$$V_r = \mathbb{V}(Y) + c \cdot \mathbb{E}(\tilde{r}(\mathbf{X};\boldsymbol{\phi}^\star)\{Y - \mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\tilde{\boldsymbol{\beta}}\}^2).$$

Instead of assuming (4.23), one can obtain (4.26) using the outcome regression model. If $\mathbb{E}(Y \mid \mathbf{x})$ satisfies

$$\mathbb{E}(Y \mid \mathbf{x}) \in \mathrm{span}\{\mathbf{b}(\mathbf{x})\}, \tag{4.27}$$

then we have $\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\tilde{\boldsymbol{\beta}} = \mathbb{E}(Y \mid \mathbf{x}_i)$. In this case, we have the following results.

Corollary 4.1. *Suppose that the assumptions for Theorem 4.1 hold. If* $\mathbf{b}(\mathbf{x})$ *in (4.18) satisfies (4.27), we can also establish the asymptotic normality in (4.26) where*

$$V_r = \mathbb{V}(Y) + c \cdot \mathbb{E}[\tilde{r}(\mathbf{X};\boldsymbol{\phi}^\star)\mathbb{V}(Y \mid \mathbf{X})]. \tag{4.28}$$

Remark 4.1. *We have presented two different sufficient conditions for (4.26). One is the response probability model in (4.23). The other is the outcome regression model in (4.27). Either one of the two models is sufficient to establish the asymptotic unbiasedness of the proposed PS estimator in (4.22). Thus, the proposed PS estimator can be understood as a doubly robust estimator (Bang and Robins, 2005; Tsiatis, 2007; Cao et al., 2009; Han and Wang, 2013). As long as either one of the two models: the outcome regression model such as (4.27) or the response propensity model (4.23), the resulting estimator satisfies (4.26). However, the outcome regression model assumption in (4.27) is more useful in developing an efficient PS estimator. Note that we*

can express V_r *in (4.28) as*

$$\begin{aligned} V_r &= \mathbb{V}\{\mathbb{E}(Y \mid \mathbf{X})\} + \mathbb{E}[\{\tilde{\pi}(\mathbf{X})\}^{-1}\mathbb{V}(Y \mid \mathbf{X})] \\ &= \mathbb{V}\{\mathbb{E}(Y \mid \mathbf{b}(\mathbf{X}))\} + \mathbb{E}[\{\mathbb{E}(\pi(\mathbf{X}) \mid \mathbf{b}(\mathbf{X}))\}^{-1}\mathbb{V}(Y \mid \mathbf{b}(\mathbf{X}))] \\ &\leq \mathbb{V}\{\mathbb{E}(Y \mid \mathbf{b}(\mathbf{X}))\} + \mathbb{E}\left(\mathbb{E}\left[\{\pi(\mathbf{X})\}^{-1} \mid \mathbf{b}(\mathbf{X})\right]\mathbb{V}(Y \mid \mathbf{b}(\mathbf{X}))\right) \\ &= \mathbb{V}\{\mathbb{E}(Y \mid \mathbf{X})\} + \mathbb{E}\left(\{\pi(\mathbf{X})\}^{-1}\mathbb{V}(Y \mid \mathbf{X})\right) \end{aligned}$$

where the last inequality is based on Jensen's inequality applied to the concave function $g(x) = 1/x$ *for* $x \in (0, 1)$*. This is essentially an alternative proof for (4.12).*

Remark 4.2. *Condition (4.27) is the first moment version of the reduced model*

$$f(y \mid \mathbf{x}) = f(y \mid \mathbf{b}(\mathbf{x})). \tag{4.29}$$

Recall that, by Lemma 4.1, condition (4.29) implies the balancing score assumption (4.9)*, which in turn implies that the efficiency of the smoothed PS estimator using* $\tilde{r}(\mathbf{x})$ *in (4.6). The constrained optimization problem discussed in the maximum entropy method is a computational tool for implementing the smoothed PS estimator. If the space* $\mathcal{H} = span\{\mathbf{b}(\mathbf{x})\}$ *is large enough, then (4.27) is likely to be satisfied and result (4.11) will hold. However, if* $\mathcal{H}$ *is too large, then we can find* $\mathcal{H}_0 \subset \mathcal{H}$ *such that* $\mathbb{E}(Y \mid \mathbf{x}) \in \mathcal{H}_0$*. In this case, we can construct a smoothed density ratio function using the basis functions in* $\mathcal{H}_0$ *only and obtain a more efficient PS estimator. Therefore, including unnecessary constraints in the calibration equation will increase the variance. This is consistent with the empirical findings of Brookhart et al. (2006) and Shortreed and Ertefaie (2017). We will discuss this result further in Section 4.5.*

By (4.24), the PS estimator is approximated by the sample mean of $d_i = d(\mathbf{x}_i, y_i, \delta_i; \boldsymbol{\phi}^\star)$. The function $d(\mathbf{x}_i, y_i, \delta_i; \boldsymbol{\phi}^\star)$ is referred to as an influence function of $\hat{\theta}_{PS}$. The phrase influence function used by Hampel (1974) and is motivated by the fact that to the first order $d_i = d(\mathbf{x}_i, y_i, \delta_i; \boldsymbol{\phi}^\star)$ is the influence of observation $(\mathbf{x}_i, y_i, \delta_i)$ on the estimator $\hat{\theta}_{PS}$. One direct result of Theorem 4.1 is that the variance estimation of $\hat{\theta}_{PS}$ can be constructed by a standard

linearization method. In particular, let

$$\hat{d}_i = \mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\hat{\boldsymbol{\beta}} + \delta_i\{1 + (n_0/n_1)\tilde{r}(\mathbf{x}_i;\hat{\boldsymbol{\phi}})\}\{y_i - \mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\hat{\boldsymbol{\beta}}\},$$

where $\hat{\boldsymbol{\beta}}$ is the solution to (4.25) evaluated at $\boldsymbol{\phi} = \hat{\boldsymbol{\phi}}$. Then, the variance estimator can be written as

$$\widehat{\mathrm{Var}}(\hat{\theta}_{PS}) = \widehat{\mathrm{Var}}(\bar{d}_n) = \frac{1}{n}S^2_{\hat{d}},$$

where

$$S^2_{\hat{d}} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\hat{d}_i - \bar{\hat{d}}_n\right)^2,$$

where $\bar{\hat{d}}_n = 1/n\sum_{i=1}^{n}\hat{d}_i$.

4.5 Dimension Reduction

4.5.1 Introduction

In Section 4.4, we have seen that the log-linear DR model (4.18) is enough to obtain the PS estimation. The consistency of the PS estimator of $\theta_0 = \mathbb{E}(Y)$ depends on whether $\mathbb{E}(Y \mid \mathbf{x})$ lies in the linear space $\mathcal{H} = \mathrm{span}\{\mathbf{b}(\mathbf{x})\}$ generated by the basis functions in the log-linear DR model. If the linear space $\mathcal{H}$ is large enough to satisfy $\mathbb{E}(Y \mid \mathbf{x}) \in \mathcal{H}$, the consistency can be established. However, as pointed out in Remark 2, including other $\mathbf{x}$-variables outside the outcome model into $\mathcal{H}$ may lead to efficiency loss.

To explain the idea further, we assume that $\mathbf{b}(\mathbf{x}) = \mathbf{x}_{\mathcal{M}}$, where $\mathcal{M}$ is an index set for a subset of $\mathbf{x}$. The following lemma presents an interesting results, which is a natural corollary from Lemma 4.1.

Lemma 4.3. *If MAR condition in (4.1) holds and the reduced model for y holds such that*

$$f(y \mid \mathbf{x}) = f(y \mid \mathbf{x}_{\mathcal{M}}) \tag{4.30}$$

for $\mathbf{x} = (\mathbf{x}_{\mathcal{M}}, \mathbf{x}_{\mathcal{M}^c})$, then we can obtain MAR given $\mathbf{x}_{\mathcal{M}}$. That is,

$$Y \perp \delta \mid \mathbf{X}_{\mathcal{M}}. \tag{4.31}$$

Note that (4.31) is a special case of the reduced MAR in (4.9) using $\mathbf{b}(\mathbf{x}) = \mathbf{x}_{\mathcal{M}}$ as the balancing score function. In the spirit of Proposition 1, we can see that the PS estimator using $\mathbf{b}(\mathbf{x}) = \mathbf{x}_{\mathcal{M}}$ in (4.18) is more efficient than the PS estimator using $\mathbf{b}(\mathbf{x}) = \mathbf{x}$ in (4.18). Therefore, it is better to apply a model selection procedure to select the important variables (i.e., $\mathbf{x}_{\mathcal{M}}$ here) which satisfies (4.27).

4.5.2 Variable Selection Method

We utilize two-stage estimation strategy to complete DRE in propensity score approach,

- Step 1: Use penalized regression method to select the basis functions for the regression of y on $\mathbf{x}$.
- Step 2: Use the basis functions in Step 1 to construct the log-linear density ratio model in (4.18). Apply the proposed maximum entropy method to obtain $\hat{r}(\mathbf{x}) = \tilde{r}(\mathbf{x}; \hat{\boldsymbol{\phi}}) = \exp\{\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\hat{\boldsymbol{\phi}}\}$.

If we use the penalized regression method satisfying the oracle property, such as the SCAD penalty (Fan and Li, 2001), we can safely ignore the uncertainty due to model selection.

In the first stage, we adapt the penalized estimating equations (Johnson et al., 2008) to do the variable selection. To be more general, we utilize an Z-estimator as a working model and we denote the corresponding score function as $\mathbf{U}(\boldsymbol{\alpha})$. For example, $\mathbf{U}(\boldsymbol{\alpha})$ can be written as

$$\mathbf{U}(\boldsymbol{\alpha}) = \frac{2}{n_1}\sum_{i=1}^{n} \delta_i \mathbf{x}_i(\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\alpha} - y_i) \tag{4.32}$$

in traditional least squares estimation, where $\boldsymbol{\alpha} \in \mathbb{R}^{d+1}$. The penalized estimating equations can be written as

$$\mathbf{U}^P(\boldsymbol{\alpha}) = \mathbf{U}(\boldsymbol{\alpha}) - q_\lambda(|\boldsymbol{\alpha}|)\mathrm{sgn}(\boldsymbol{\alpha}), \tag{4.33}$$

where $q_\lambda(|\boldsymbol{\alpha}|) = (q_\lambda(|\alpha_0|), \ldots, (q_\lambda(|\alpha_d|))^{\mathrm{T}}$, $q_\lambda(\cdot)$ is a continuous function and $q_\lambda(|\boldsymbol{\alpha}|)\mathrm{sgn}(\boldsymbol{\alpha})$ is an elementwise product between two vectors. Further, let $p_\lambda(x) = \int q_\lambda(x)dx$. In M-estimation

framework, $p_\lambda(x)$ usually performs as penalization functions. Various penalization functions can chosen and the one we specify here is the smoothly clipped absolute deviation function (SCAD) (Fan and Li, 2001). In particular,

$$q_\lambda(\alpha) = \lambda \left\{ \mathbb{I}(|\alpha| < \lambda) + \frac{(a\lambda - |\alpha|)_+}{(a-1)\lambda} \mathbb{I}(|\alpha| \geq \lambda) \right\}, \tag{4.34}$$

where $(x)_+ = \max\{x, 0\}$, $\mathbb{I}$ is the indicator function and a is constant specified as 3.7 in Fan and Li (2001). Further, we let $\hat{\mathcal{M}}$ to denote the variable index set selected after SCAD procedure. Here we use a working model to select the variables. By Assumption 4.7, 4.8 in the Appendix and Theorem 1 in Johnson et al. (2008), we have

$$\begin{aligned} \mathbb{P}(\hat{\alpha}_j \neq 0) \to 1, \text{ for } j \in \mathcal{M}; \\ \mathbb{P}(\hat{\alpha}_j = 0) \to 1, \text{ for } j \in \mathcal{M}^c, \end{aligned} \tag{4.35}$$

which constructs the model selection consistency. After the first stage, we now obtain the important variable set $\hat{\mathcal{M}}$.

In the second stage, we use the selected variables to perform the maximum entropy method for the density ratio estimation. That is, we maximize

$$\hat{Q}(\boldsymbol{\phi}) = \frac{1}{n_0} \sum_{i=1}^{n} (1 - \delta_i)(\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\phi}) - \frac{1}{n_1} \sum_{i=1}^{n} \delta_i \exp\{\mathbf{x}_i^{\mathrm{T}} \boldsymbol{\phi}\}$$

subject to $\phi_k = 0$ for all $k \in \hat{\mathcal{M}}^c$. The resulting PS estimator is then computed by (4.22) with $\hat{r}(\mathbf{x}) = \exp(\mathbf{x}^{\mathrm{T}} \hat{\boldsymbol{\phi}})$, where $\hat{\phi}_k = 0$ for all $k \in \hat{\mathcal{M}}^c$.

Corollary 4.2. *Suppose that the assumptions for Theorem 4.1 hold. Also, the additional assumptions listed in the Appendix hold. If $\mathbf{x}_\mathcal{M}$ satisfies $\mathbb{E}(Y \mid \mathbf{x}) = \mathbb{E}(Y \mid \mathbf{x}_\mathcal{M})$, with probability goes to 1,*

$$\sqrt{n} \left(\hat{\theta}_{PS} - \theta_0 \right) \xrightarrow{\mathcal{L}} N(\mathbf{0}, V_r), \tag{4.36}$$

where

$$V_r = \mathbb{V}(Y) + \mathbb{E}[(n_0/n_1) r(\mathbf{X}_\mathcal{M}) \mathbb{V}(Y \mid \mathbf{X}_\mathcal{M})]. \tag{4.37}$$

Remark 4.3. *Note that we apply a two-stage procedure to perform the estimation. The asymptotic results in (4.36) is based on the model selection consistency in (4.35), whose probability goes to 1. That is, if we define $\mathcal{D}_n = \{\mathcal{M} = \hat{\mathcal{M}}\}$ where $\hat{\mathcal{M}}$ is obtained from the first stage procedure, the linearization in (4.36) is conditional on $\mathcal{D}_n$. By (4.35), we obtain $P(\mathcal{D}_n) \to 1$ and the limiting distribution of $T_n \equiv \sqrt{n}(\hat{\theta}_{PS} - \theta)$ given $\hat{\mathcal{M}}$, denoted by $\mathcal{L}(T_n \mid \hat{\mathcal{M}})$, is asymptotically equivalent to $\mathcal{L}(T_n \mid \mathcal{M})$. See also Theorem 1 of Yang et al. (2020) for a similar argument.*

4.5.3 Sufficient Subspace Construction

Apart from variable selection, another popular approach to reduce covariates is the sufficient dimension reduction (SDR) (Li, 1991; Cook, 1994, 2009). Sufficient dimension reduction finds a subspace $\mathcal{H}_0$ with minimal dimension such that

$$Y \perp \mathbf{X} \mid \mathcal{P}_{\mathcal{H}_0}\mathbf{X} \tag{4.38}$$

holds, where $\mathcal{P}_{\mathcal{H}_0}$ is the projection operator to $\mathcal{H}_0$. In particular, our goal is to find $\mathbf{W} \in \mathbb{R}^{l\times d}$ with

$$\mathbf{b}(\mathbf{x}) = \mathbf{W}\mathbf{x} \tag{4.39}$$

such that $\mathbf{b}(\mathbf{x})$ spans $\mathcal{H}_0$ satisfying (4.38), where $l < d$ and $d = \dim(\mathbf{x})$. Once $\mathbf{b}(\mathbf{x})$ in (4.39) is chosen, we can apply the maximum entropy method using model (4.18).

To find W in a function space, we employ the method in Stojanov et al. (2019). For a positive semi-definite kernel function k, let $\mathcal{H}_k$ denote the induced reproducing kernel Hilbert space (RKHS). In particular, there exists a feature map $\psi : \mathbb{R}^d \to \mathcal{H}_k$, which maps the previous covariates to an abstract space $\mathcal{H}_k$. On the other hand, the kernel t may induce a map $\mathbb{R} \to \mathcal{H}_t$ for the response variable. Further, let u induces a map $\mathbb{R}^l \to \mathcal{H}_u$ for sufficient dimension reduction variable $\mathbf{b}(\mathbf{x})$. For $\xi \in \mathcal{H}_u$, $\zeta \in \mathcal{H}_t$, the cross-covariance operator from $\mathcal{H}_u$ to $\mathcal{H}_t$ is defined as

$$\langle \zeta, \mathcal{C}_{Y,\mathbf{b}(\mathbf{X})}\xi\rangle_{\mathcal{H}_t} \equiv \mathbb{E}_{Y,\mathbf{b}(\mathbf{X})}\left\{\xi(\mathbf{b}(\mathbf{X}))\zeta(Y)\right\} - \mathbb{E}_{\mathbf{b}(\mathbf{X})}\left\{\xi(\mathbf{b}(\mathbf{X}))\right\}\mathbb{E}_Y\left\{\zeta(Y)\right\}.$$

In addition, the conditional covariance operator can be defined as

$$\mathcal{C}_{YY|\mathbf{b}(\mathbf{X})} \equiv \mathcal{C}_{YY} - \mathcal{C}_{Y,\mathbf{b}(\mathbf{X})}\mathcal{C}^{-1}_{\mathbf{b}(\mathbf{X}),\mathbf{b}(\mathbf{X})}\mathcal{C}_{\mathbf{b}(\mathbf{X}),Y}.$$

Intuitively, the above operator can depict the conditional indepedence between Y and $\mathbf{X}$ given $\mathbf{b}(\mathbf{X})$ (Fukumizu et al., 2004), i.e.,

$$Y \perp \mathbf{X} \mid \mathbf{b}(\mathbf{X}) \Longleftrightarrow \mathcal{C}_{YY|\mathbf{b}(\mathbf{X})} = \mathcal{C}_{YY|\mathbf{X}}.$$

Also, Theorem 7 in Fukumizu et al. (2004) shows that $\mathcal{C}_{YY|\mathbf{b}(\mathbf{X})} \geq \mathcal{C}_{YY|\mathbf{X}}$. Thus, the SDR problem can be transformed into the following optimization problem on Grassmann manifold for the data with $\delta_i = 1$:

$$\begin{aligned} \arg\min_{\mathbf{W}} \quad & \text{Trace}\left\{\hat{\mathcal{C}}_{YY|\mathbf{b}(\mathbf{X})}\right\} \\ s.t. \quad & \mathbf{W}\mathbf{W}^{\mathrm{T}} = \mathbf{I}. \end{aligned} \tag{4.40}$$

For convenience, we use the Euclidean distance as the inner product for each space. The computational details for the aforementioned optimization problem can are presented in the Appendix.

Remark 4.4. *In the two-step procedure for estimation of the PS model, we have two sources uncertainty. The first one comes from the step one, which is essentially estimating W to form the new basis functions. The second step is the parameter estimation given the basis functions. Now, to fully account for the uncertainty associated with the two-step procedure, we can use the following bootstrap methods mimicking the two-step procedure for estimation.*

1. *Use the bootstrap sample to apply the same dimension reduction technique treating the estimated dimension (l) is fixed. After that, we obtain a bootstrap version of the reduced basis functions.*

2. *Use the bootstrap basis functions to apply the same estimation procedure to obtain the bootstrap replicate of the parameter estimate and its density ratio estimates.*

3. *Finally, the bootstrap replicate of the PS estimator can be obtained using the bootstrap sample and the bootstrap replicate of the parameter estimates for the propensity scores.*

4.6 Multivariate Missing Data

The proposed method in Section 4.5 is based on the assumption that $\theta = E(Y)$ is the parameter of interest for a single study variable Y. If there are multiple parameters of interest and only a single set of propensity weights is used, (4.27) is no longer applicable. We now consider the case of multivariate study variables, denoted by $Y_1, \cdots, Y_p$, and they are subject to missingness. There are 2^p possible missing patterns with p study variables. Let $T \leq 2^p$ be the realized number of different missing patterns in the sample. Thus, the sample is partitioned into T disjoint subsets with the same missing patterns.

Let S_t be the t-th subset of the sample from this partition. We assume that S_1 consists of elements with complete response. Without loss of generality, we may define $\delta_{i,t} = 1$ if $i \in S_t$ and $\delta_{i,t} = 0$ otherwise. Let $f_t(\mathbf{x}, \mathbf{y})$ be the density function $f(\mathbf{x}, \mathbf{y} \mid \delta_t = 1)$ and define

$$r_t(\mathbf{x}, \mathbf{y}) = \frac{f_t(\mathbf{x}, \mathbf{y})}{f_1(\mathbf{x}, \mathbf{y})}$$

be the density ratio function that we are interested in estimating. We further assume that

$$\log\{r_t(\mathbf{x}, \mathbf{y})\} = \phi_{t0} + \boldsymbol{\phi}_{t1}^{\mathrm{T}}\mathbf{x} + \boldsymbol{\phi}_{t2}^{\mathrm{T}}\mathbf{y}_{obs(t)}, \tag{4.41}$$

where $\mathbf{y}_{obs(t)}$ is the observed part of $\mathbf{y}$ in S_t.

We can apply the maximum entropy method to estimate the parameters in (4.41). That is, the sample-version objective function for estimating r_t is

$$\hat{Q}_t(r_t) = \frac{1}{n_t}\sum_{i \in S_t} \log\{r_t(\mathbf{x}_i, \mathbf{y}_i)\} - \frac{1}{n_1}\sum_{i \in S_1} r_t(\mathbf{x}_i, \mathbf{y}_i).$$

Under model (4.41),

$$\hat{Q}_t(\boldsymbol{\phi}_t) \;=\; \frac{1}{n_t}\sum_{i \in S_t}\{\phi_{t0} + \boldsymbol{\phi}_{t1}^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\phi}_{t2}^{\mathrm{T}}\mathbf{y}_{i,obs(t)}\} - \frac{1}{n_1}\sum_{i \in S_1}\exp\{\phi_{t0} + \boldsymbol{\phi}_{t1}^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\phi}_{t2}^{\mathrm{T}}\mathbf{y}_{i,obs(t)}\}.$$

The estimating equation maximizing $\hat{Q}_t(\boldsymbol{\phi}_t)$ is obtained by $\hat{\mathbf{U}}_t(\boldsymbol{\phi}_t) = \partial\hat{Q}_t(\boldsymbol{\phi}_t)/\partial\boldsymbol{\phi}_t$. Thus, we have only to solve

$$\hat{\mathbf{U}}_t(\boldsymbol{\phi}_t) \equiv \sum_{i \in S_1} w_i(\boldsymbol{\phi}_t)\,(\mathbf{z}_i - \bar{\mathbf{z}}_t) = \mathbf{0}, \tag{4.42}$$

where $\mathbf{z}_i = (\mathbf{x}_i^{\mathrm{T}}, \mathbf{y}_{i,obs(t)}^{\mathrm{T}})^{\mathrm{T}}$ and

$$w_i(\boldsymbol{\phi}_t) = \frac{\exp(\boldsymbol{\phi}_{t1}^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\phi}_{t2}^{\mathrm{T}}\mathbf{y}_{i,obs(t)})}{\sum_{i\in S_1}\exp(\boldsymbol{\phi}_{t1}^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\phi}_{t2}^{\mathrm{T}}\mathbf{y}_{i,obs(t)})} \tag{4.43}$$

and $\bar{\mathbf{z}}_t = n_t^{-1}\sum_{i\in S_t}\mathbf{z}_i$. Also, the intercept term ϕ_{t0} is determined to satisfy

$$\frac{1}{n_1}\sum_{i\in S_1}\exp\left(\phi_{t0} + \boldsymbol{\phi}_{t1}^{\mathrm{T}}\mathbf{x}_i + \boldsymbol{\phi}_{t2}^{\mathrm{T}}\mathbf{y}_{i,obs(t)}\right) = 1. \tag{4.44}$$

Now, to estimate $\boldsymbol{\theta}$ defined through $\mathbb{E}\{\mathbf{U}(\boldsymbol{\theta};\mathbf{x},\mathbf{y})\} = \mathbf{0}$, one can use

$$\sum_{i=1}^{n}\mathbf{U}(\boldsymbol{\theta};\mathbf{x}_i,\mathbf{y}_i) = \mathbf{0}$$

as an estimating equation under complete response. Under the above multivariate missingness, the PS estimator can be obtained by solving

$$\sum_{i\in S_1} w_i\mathbf{U}(\boldsymbol{\theta};\mathbf{x}_i,\mathbf{y}_i) = \mathbf{0}$$

where

$$w_i = 1 + \sum_{t=2}^{T}\frac{n_t}{n_1}r(\mathbf{x}_i,\mathbf{y}_{i,obs(t)};\hat{\boldsymbol{\phi}}_t). \tag{4.45}$$

Let $\mathbf{z}_{i,t} = (1,\mathbf{x}_i^{\mathrm{T}},\mathbf{y}_{i,obs(t)}^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\phi}_t = (\phi_{t0},\boldsymbol{\phi}_{t1}^{\mathrm{T}},\boldsymbol{\phi}_{t2}^{\mathrm{T}})^{\mathrm{T}}$. As a result, the density ratio for missing pattern t can be simplified as $r_t(\mathbf{z}_{i,t};\boldsymbol{\phi}_t)$, for $t = 2,\ldots,T$. Define $\delta_{i,t}$ as the indicator function for unit i and missing pattern t. Let the true parameter of interest be $\boldsymbol{\theta}_0$ and the true parameter for density ratio be $\boldsymbol{\phi}_0 = (\boldsymbol{\phi}_2^{\mathrm{T}},\ldots,\boldsymbol{\phi}_T^{\mathrm{T}})^{\mathrm{T}}$. Then we have the following theorem

Theorem 4.2. *Under the regularity conditions, for multivariate missing case, we have the asymptotic expandsion*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} d(\mathbf{x}_i,\mathbf{y}_i,\delta_i;\boldsymbol{\phi}_0) + o_p(1),$$

where

$$\begin{aligned} d(\mathbf{x}_i,\mathbf{y}_i,\delta_i;\boldsymbol{\phi}_0) = &-\left[\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{U}(\boldsymbol{\theta}_0;\mathbf{x},\mathbf{y})\right\}\right]^{-1} \\ &\times\left[\delta_{i,1}\mathbf{U}_i + \sum_{t=2}^{T}\delta_{i,t}\tilde{\boldsymbol{\beta}}_t\mathbf{z}_{i,t} + \delta_{i,1}\sum_{t=2}^{T}\frac{n_t}{n_1}r_t(\mathbf{z}_{i,t};\boldsymbol{\phi}_t)\left\{\mathbf{U}_i - \tilde{\boldsymbol{\beta}}_t\mathbf{z}_{i,t}\right\}\right] + o_p(1), \end{aligned}$$

where $\mathbf{U}_i = \mathbf{U}(\boldsymbol{\theta}; \mathbf{x}_i, \mathbf{y}_i)$ *and* $\tilde{\boldsymbol{\beta}}_t$ *is the probability limit to the solution of*

$$\sum_{i=1}^{n} \delta_{i,1} r_t(\mathbf{z}_{i,t}; \boldsymbol{\phi}_t) \left\{ \mathbf{U}(\boldsymbol{\theta}_0; \mathbf{x}_i, \mathbf{y}_i) - \boldsymbol{\beta}_t \mathbf{z}_{i,t} \right\} \mathbf{z}_{i,t}^{\mathrm{T}} = \mathbf{0}.$$

The above theorem depicts the asymptotic behavior of our proposed estimators in multivariate missing case.

Table 4.1: Missing Pattern Example

	x	y_1	y_2	y_3
S_1	✓	✓	✓	✓
S_2	✓	✓		✓
S_3	✓	✓	✓	
S_4	✓	✓		

Example 4.1. *For illustration, suppose that we have a missing data pattern in Table 4.1. The parameter of interest is defined through* $\mathbb{E}\{\mathbf{U}(\theta \mid \mathbf{X}, Y_1, Y_2, Y_3)\} = 0$. *For example,* θ *can be the regression coefficients for the regression of* y_3 *on* $(\mathbf{x}, y_1, y_2)$. *We have bivariate missingness, with the response indicator functions* δ_k *for* y_k, $k = 1, 2, 3$. *In this particular example,* y_1 *is completely observed which is often the case with longitudinal survey data.*

We are interested in finding the propensity weights for subset S_1 *such that*

$$\sum_{i \in S_1} w_i \mathbf{U}(\boldsymbol{\theta}; \mathbf{x}_i, y_{i1}, y_{i2}, y_{i3}) = 0$$

is approximately unbiased for $\boldsymbol{\theta}$. *We also wish to include all the partial observations in* S_2, S_3, S_4 *into the propensity weights.*

To do this, we consider three density ratio function

$$r_t(\mathbf{x}, \mathbf{y}_{obs(t)}) = \frac{f_t(\mathbf{x}, \mathbf{y}_{obs(t)})}{f_1(\mathbf{x}, \mathbf{y}_{obs(t)})}$$

where $\mathbf{y}_{obs(t)}$ *is the observed part of* $\mathbf{y} = (y_1, y_2, y_3)^{\mathrm{T}}$ *with missing pattern* S_t*, for* $t = 2, 3, 4$*. We assume a log-linear model for* $r_t(\mathbf{x}, \mathbf{y}_{obs(t)})$*. Thus, we assume*

$$\begin{aligned}
\log r_2(\mathbf{x}, \mathbf{y}_{obs(2)}) &= \phi_{20} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\phi}_{2x} + \phi_{21}y_1 + \phi_{23}y_3 \\
\log r_3(\mathbf{x}, \mathbf{y}_{obs(3)}) &= \phi_{30} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\phi}_{3x} + \phi_{31}y_1 + \phi_{32}y_2 \\
\log r_4(\mathbf{x}, \mathbf{y}_{obs(4)}) &= \phi_{40} + \mathbf{x}^{\mathrm{T}}\boldsymbol{\phi}_{4x} + \phi_{41}y_1
\end{aligned}$$

so that MAR holds. We can use (4.42) and (4.44) to estimate the model parameters and then use (4.45) to obtain the propensity weights for the final estimation.

4.7 Simulation Study

4.7.1 Simulation for MAR

A limited simulation study is performed to compare the PS methods. The setup for the simulation study employed a 2×2 factorial structure with two factors. The first factor is the outcome regression (OR) model that generates the sample. The second factor is the response mechanism (RM). We generate δ and $\mathbf{x} = (x_1, x_2, x_3, x_4)^{\mathrm{T}}$ based on the RM first. We have

- RM1 (Logsitic model):

$$\begin{aligned}
x_{ik} &\sim N(2, 1), \text{for } k = 1, \ldots, 4, \\
\delta_i &\sim \text{Ber}(p_i), \\
\text{logit}(p_i) &= 1 - x_{i1} + 0.5x_{i2} + 0.5x_{i3} - 0.25x_{i4}.
\end{aligned}$$

- RM2(Gaussian mixture model):

$$\begin{aligned}
\delta_i &\sim \text{Bern}(0.6) \\
x_{ik} &\sim N(2, 1), \text{for } k = 1, 2, 3, \\
x_{i4} &\sim \begin{cases} N(3, 1), \text{if } \delta_i = 1 \\ N(1, 1), \text{otherwise.} \end{cases}
\end{aligned}$$

Once x and δ are generated, we can generate y from OR1 and OR2. That is, we generate y from

- OR1: $y_i = 1 + x_{i1} + x_{i2} + x_{i3} + x_{i4} + e_i$.
- OR2: $y_i = 1 + 0.5x_{i1}x_{i2} + 0.5x_{i3}^2x_{i4}^2 + e_i$.

Here, $e_i \sim N(0, 1)$. Further, we compare four other estimators:

- Maximum likelihood estimator (MLE) with Bernoulli distribution with parameter $\text{logit}(p_i) = \mathbf{x}_i^{\mathrm{T}}\zeta$.
- Covariate balancing propensity score method (CBPS) from Imai and Ratkovic (2014) using calibration variable $(1, x_1, x_2, x_3, x_4)^{\mathrm{T}}$.
- Improved covariate balancing propensity score method (iCBPS) from (Fan et al., 2016) using calibration variable $(1, x_1, x_2, x_3, x_4)^{\mathrm{T}}$.
- Entropy balancing method from Hainmueller (2012) using calibration variable $(1, x_1, x_2, x_3, x_4)^{\mathrm{T}}$.

We use the sample size $N = 5000$ with 5000 Monte Carlo samples. The results are presented in Table 4.2, where DR is our proposed method. When we use $(1, x_1, x_2, x_3, x_4)$ as the calibration variable, OR1 matches with the working outcome model and RM1 matches with the working response model. Hainmueller (2012) method also shows good performances when OR1 or RM1 is true (doubly robust), but when both models fail (i.e. OR2RM2 setup), the performance is really poor. Our proposed method is also doubly robust and it performs reasonably well even when both models fail.

4.7.2 MAR under High-dimensional Case

We further test our proposed method in high-dimension case. We test another 2×2 factoral structure simulations. We first generate $\mathbf{x} = (x_1, \ldots, x_{100})$ with each $x_k \sim N(2, 1)$, and generate y from

Table 4.2: Relative bias (R.B.), standard error (S.E.) and root meam square error (RMSE) for the model with 4 covariates

Model	Method	R.B. (%)	S.E. ($\times 10^2$)	RMSE ($\times 10^2$)
OR1RM1	DR	0.00	3.52	3.52
	Hainmueller	0.00	3.46	3.46
	CBPS	-0.24	6.85	7.17
	iCBPS	0.02	3.58	3.58
	MLE	-0.01	7.02	7.02
OR2RM1	DR	0.00	5.45	5.45
	Hainmueller	0.18	5.31	5.49
	CBPS	-0.22	7.53	7.73
	iCBPS	-0.07	5.92	5.95
	MLE	-0.02	7.66	7.66
OR1RM2	DR	-0.00	5.39	5.39
	Hainmueller	0.00	4.07	4.07
	CBPS	-0.67	22.50	23.34
	iCBPS	0.12	5.75	5.85
	MLE	-0.07	28.61	28.62
OR2RM2	DR	-0.36	9.34	9.87
	Hainmueller	-5.42	7.00	48.72
	CBPS	-0.61	23.58	24.19
	iCBPS	0.26	10.24	10.49
	MLE	-0.07	33.08	33.08

1. OR3: $y_i = 1 + x_{i1} + x_{i2} + x_{i3} + e_i$.

2. OR4: $y_i = 1 + 0.5x_{i1}x_{i2} + 0.5x_{i1}x_{i3} + e_i$.

Here, $e_i \sim N(0, 1)$. We consider two different response mechanism,

- RM3:

$$\delta_i \sim \text{Ber}(p_i),$$
$$\text{logit}(p_i) = 1 - x_{i1} + 0.5x_{i4} + 0.5x_{i5} - 0.25x_{i10}.$$

- RM4:

$$p_i = \begin{cases} 0.95 & \text{if } x_{i5} \geq 2, \\ 0.25 & \text{if } x_{i5} < 2. \end{cases}$$

In this setup, the response mechanism does not follow a logistic regression model. But, the MAR assumption still holds. We set the sample size as $N = 1000$ and response rate around 60%. We conduct the simulations for 1000 Monte Carlo samples. The simulation results for different estimator behaviors are presented in Table 4.3. The variance estimation results of our proposed estimators are presented in Table 4.4. Note that we apply the linearized variance estimation procedure to get the valid variance estimation by Corollary 4.1. We also test different variable cases in the density ratio model. In particular, we use 'DR_Full' to denote that all variables are considered. 'DR_VS' denotes the case that variables in the model are selected based the procedure described in Subsection 4.5.2. We use 'DR_SDR' to denote the sufficient dimension reduction method results.

From Table 4.4, the variance estimation relative biases dimension reduction approach for all scenarios are all under 8%, which verifies our proposed variance estimation procedure. Due to nuisance variable presentation, the variance estimation procedure for our proposed method with all variables won't work, which reflects the necessity of dimension reduction. As for point estimation , our proposed method DR_SDR and DR_VS perform best in the sense of RMSE. Aside from this, Hainmueller's method is also competitive, followed by DR_Full and iCBPS.

4.7.3 Simulation for Multivariate Missing Case

Suppose we have the data structure in Table 4.1. Now let $\mathbf{x}_i = (x_{1i}, x_{2i}, x_{3i})^{\mathrm{T}}$. Further, we use θ_2 and θ_3 to denote the mean of Y_2 and Y_3 correspondingly. For model D, we consider continuous variables. Specifically, we have

$$\begin{aligned} y_{i1} &= 1 + x_{i1} + x_{i1}^2 + x_{i2} + x_{i3} + \varepsilon_{i1}, \\ y_{i2} &= 2 + 2x_{i1} + 2x_{i2} + x_{i2}^2 + x_{i3} + \varepsilon_{i2}, \\ y_{i3} &= 1 + 0.5y_{i1} + 0.5y_{i2} + \varepsilon_{i3}, \end{aligned}$$

where $\varepsilon_{i,1}, \varepsilon_{i,2}, \varepsilon_{i,3} \overset{i.i.d}{\sim} N(0,1)$ and

$$\begin{aligned} \delta_{i2}, \delta_{i3} &\overset{i.i.d}{\sim} \mathrm{Ber}(p_i), \\ \mathrm{logit}(p_i) &= 0.35 - x_{i1} + x_{i2} - 0.25x_{i3} + 0.25y_{i1}. \end{aligned}$$

For comparison, we can consider three methods additionally. One is naive CC method, which uses only full respondents. That is, use the set of units with $\delta_i = 1$ for estimating θ, where $\delta_i = \delta_{i1}\delta_{i2}\delta_{i3}$. The other method is a simple propensity score method using a logistic regression model of δ_i on $(\mathbf{x}_i, y_{1i})$. In fact, we can apply the maximum entropy method for parameter estimation. That is, we are estimating the density ratio function based on $\delta_i = 1$ and $\delta_i = 0$. We use θ_2 to denote the mean of Y_2, θ_3 to denote the mean of Y_3 and θ_4 to denote the quantity $\mathbb{P}(Y_2 \leq Y_3)$. The corresponding results are presented in Table 4.5. In Table 4.5, *CC_Sample* method denotes the CC method, *Entropy_Detail* denotes the method illustrated in Section 4.6, *Entropy_Rough* denotes entropy methods described in this paragraph, *Logistic* to denote the aforementioned logistic regression. As we can see, the method *Entropy_Detail* behaves best among them for all three parameters in the sense of RMSE.

Table 4.3: Relative bias (R.B.), standard error (S.E.) and root meam square error (RMSE) for seven methods under 4 models

Model	Method	R.B. (%)	S.E. ($\times 10^2$)	RMSE ($\times 10^2$)
OR3RM3	DR_Full	-0.01	7.74	7.74
	DR_VS	0.00	7.13	7.13
	DR_SDR	0.06	6.72	6.73
	Hainmueller	0.00	7.35	7.35
	CBPS	-2.98	12.81	24.50
	iCBPS	-0.80	7.29	9.18
	MLE	0.22	27.13	27.17
OR3RM4	DR_Full	-0.02	8.01	8.01
	DR_VS	0.02	7.11	7.11
	DR_SDR	-0.01	6.97	6.98
	Hainmueller	-0.01	7.49	7.49
	CBPS	-1.92	12.70	18.51
	iCBPS	-0.02	7.49	7.49
	MLE	37.65	283.65	387.21
OR4RM3	DR_Full	0.01	9.97	9.97
	DR_VS	0.04	9.41	9.42
	DR_SDR	-0.08	8.98	8.98
	Hainmueller	0.01	9.60	9.60
	CBPS	-4.86	13.16	27.62
	iCBPS	-2.33	9.81	15.21
	MLE	0.40	29.65	29.72
OR4RM4	DR_Full	-0.01	10.61	10.61
	DR_VS	0.00	9.46	9.46
	DR_SDR	-0.03	9.47	9.47
	Hainmueller	0.00	9.96	9.96
	CBPS	-1.90	13.86	16.81
	iCBPS	-0.04	10.10	10.10
	MLE	38.40	225.38	296.06

Table 4.4: Monte Carlo variance (M.C.V.), variance esimation (V.E.) and variance estiamtion relative bias (V.E.R.B.) for 4 models.

Model	Method	M.C.V.($\times 10^3$)	V.E. ($\times 10^3$)	V.E.R.B. (%)
OR3RM3	DR_Full	5.99	4.67	-21.97
	DR_VS	5.09	4.90	-3.60
	DR_SDR	4.51	4.48	-0.72
OR3RM4	DR_Full	6.41	5.10	-20.43
	DR_VS	5.05	4.66	-7.72
	DR_SDR	4.86	4.81	-1.18
OR4RM3	DR_Full	9.95	8.27	-16.80
	DR_VS	8.86	8.84	-0.24
	DR_SDR	8.06	7.83	-2.77
OR4RM4	DR_Full	11.26	9.00	-20.00
	DR_VS	8.95	8.50	-5.04
	DR_SDR	8.97	8.62	-3.87

Table 4.5: Relative bias (R.B), standard error (S.E.) and root meam square error (RMSE) for 4 methods under multivariate missing simulation setup.

Parameter	Method	R.B (%)	S.E. ($\times 10^2$)	RMSE ($\times 10^2$)
θ_2	CC_Sample	20.95	8.16	63.38
	Entropy_Detail	-0.83	7.21	7.63
	Entropy_Rough	-2.44	7.88	10.77
	Logistic	-3.80	8.24	14.07
θ_3	CC_Sample	24.02	6.87	84.38
	Entropy_Detail	-0.31	6.30	6.39
	Entropy_Rough	-1.03	6.65	7.56
	Logistic	-2.01	8.08	10.71
θ_4	CC_Sample	5.75	0.99	3.61
	Entropy_Detail	0.40	1.83	1.85
	Entropy_Rough	1.55	1.65	1.90
	Logistic	-2.10	1.40	1.89

4.8 Real Data Application

We apply our methods mention in Section 4.6 for the multivariate missing case to Beijing pollution dataset (Zhang et al., 2017). As an illustrative example, we take the 5 stations' hourly data in January, 2016. The station contains Dongsi, Guanyuan, Nongzhanguan, Tiantan and Wanliu. There are 774 hourly records for each location in that month, there are 12 records with missing data and we exclude them and keep others as full sample. For each record, we fully oberseved weather conditions like temperature ($°C$), air pressure (hPa), dew point ($°C$) and cumulative wind speed (m/s). On the other hand, we artificially create missingness for the pollutant SO_2 ($\mu g/m^3$) and $PM_{2.5}$ ($\mu g/m^3$). Let $(Y_{ij,1}, Y_{ij,2})$ be the measurement of SO_2 and $PM_{2.5}$, respectively, for location i in j-th hour. Further, let $\mathbf{x}_{ij}$ be the covariate (temperature, air pressure, dew point, cumulative wind speed) for the location i and hour j. We preprocess all the weather variables by standarization, after which each covariate has zero mean and unit variance. Further, we use the log transformation for $PM_{2.5}$ due to its skewness. Then, we are interested in the following linear model

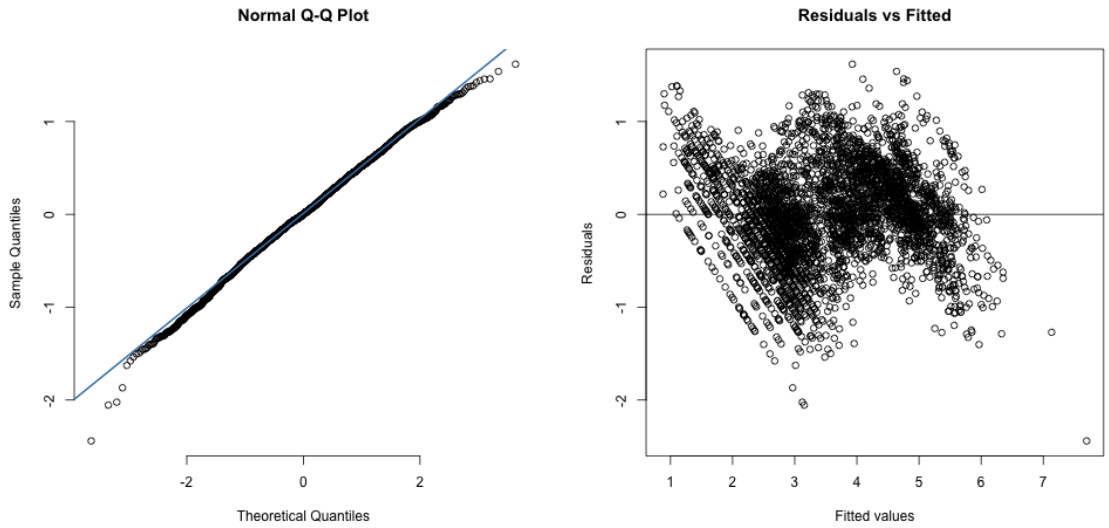
$$\log(y_{ij,2}) = \mu_i + \mathbf{x}_{ij}^{\mathrm{T}}\boldsymbol{\beta}_1 + y_{ij,1}\beta_2 + e_{ij},$$

where $e_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$ for unknown σ^2 and μ_i is the fixed effect for each station. Here, $\mathbf{x}_{ij}$ is always observed, but $(y_{ij,1}, y_{ij,2})^{\mathrm{T}}$ is subject to missingness. Specifically, we are interested in estiamting the parameter $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \beta_{13}, \beta_{14})^{\mathrm{T}}$ and β_2. The diagnostic plots for the above linear regression using full sample is presented in Figure 4.2.

We mannually create missingness with the following missing machanism:

$$\begin{aligned}
\delta_{ij,1} \sim Ber(p_{ij,1}), \delta_{ij,2} \sim Ber(p_{ij,2}),\\
\text{logit}(p_{ij,1}) = 0.5 + 0.1x_{ij,1} + 0.1x_{ij,2} + 0.1x_{ij,3} + 0.1x_{ij,4},\\
\text{logit}(p_{ij,2}) = 0.5 + 0.1x_{ij,1} + 0.1x_{ij,2} + 0.1x_{ij,3} + 0.1x_{ij,4} + 0.3y_{ij,2},
\end{aligned}$$

where the marginal missing rates for SO_2 and $PM_{2.5}$ are around 0.6. The regression parameters can be estimated by solving the weighted least square equations where the weights can be



(a) Q-Q plot (b) Residual plot

Figure 4.2: Diagnostic plots for the regression $\log(y_{ij,2})$ given $y_{ij,1}$ and $\mathbf{x}_{ij}$: (a) for Q-Q plot and (b) for residual plot.

estimated with the approach proposed in Section 4.6. In particular, we compare CC method, and the proposed PS estimation methods and logistic regression. We run 1000 Monte Carlo simulations to get the results, which are presented in Table 4.6.

From Table 4.6, we can see that our proposed method Entropy_Detail has relatively low relative bias for parameters compared with other methods. Although CC method has relatively low Monte Carlo standard error, it essentially suffers severe bias problem for β_{11}. In the sense of RMSE, Entropy_Detail also perfoms well.

Table 4.6: Relative bias (R.B.), Monte Carlo standard error (S.E.) and root meam square error (RMSE) for 4 methods for Beijing pollution dataset.

Criteria	Parameter	CC	Entropy_Detail	Entropy_Rough	Logistic Regression
R.B. ($\times 100\%$)	β_{11}	-104.14	-15.96	-74.21	-69.37
	β_{12}	-10.82	2.49	-8.50	-6.04
	β_{13}	-1.03	-5.32	-11.91	-12.50
	β_{14}	19.84	-4.52	-4.20	-4.57
	β_2	-23.05	-2.98	-12.04	-11.44
S.E. ($\times 10^3$)	β_{11}	14.20	26.05	28.22	36.20
	β_{12}	12.59	25.25	23.23	26.73
	β_{13}	12.75	24.14	19.44	19.79
	β_{14}	19.82	38.87	45.86	54.05
	β_2	0.60	1.06	0.85	0.99
RMSE ($\times 10^2$)	β_{11}	29.82	17.12	26.20	26.22
	β_{12}	14.00	15.96	16.14	16.74
	β_{13}	11.78	19.28	25.72	26.32
	β_{14}	20.56	19.95	21.57	23.40
	β_2	8.16	3.70	5.97	5.86

4.9 References

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61(4):962–973.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Stürmer, T. (2006). Variable selection for propensity score models. *American Journal of Epidemiology*, 163(12):1149–1156.

Cao, W., Tsiatis, A. A., and Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96(3):723–734.

Chan, K. C. G., Yam, S. C. P., and Zhang, Z. (2016). Globally efficient non-parametric inference of average treatment effects by empirical balancing calibration weighting. *Journal of the Royal Statistical Society: Series B (Statistical methodology)*, 78(3):673.

Cook, R. D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, 89(425):177–189.

Cook, R. D. (2009). *Regression graphics: Ideas for studying regressions through graphics*, volume 482. John Wiley & Sons.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87:376–382.

Fan, J., Imai, K., Liu, H., Ning, Y., and Yang, X. (2016). Improving covariate balancing propensity score: A doubly robust and efficient approach.

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

Folsom, R. E. (1991). Exponential and logistic weight adjustments for sampling and nonresponse error reduction. In *Proceedings of the American Statistical Association, Social Statistics Section*, volume 197201.

Fukumizu, K., Bach, F. R., and Jordan, M. I. (2004). Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. *Journal of Machine Learning Research*, 5(Jan):73–99.

Graham, B. S., de Xavier Pinto, C. C., and Egel, D. (2012). Inverse probability tilting for moment condition models with missing data. *The Review of Economic Studies*, 79(3):1053–1079.

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political analysis*, pages 25–46.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Han, P. and Wang, L. (2013). Estimation with missing data: beyond double robustness. *Biometrika*, 100(2):417–430.

Huang, W., Gallivan, K. A., and Absil, P.-A. (2015). A broyden class of quasi-newton methods for riemannian optimization. *SIAM Journal on Optimization*, 25(3):1660–1685.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, pages 243–263.

Johnson, B. A., Lin, D., and Zeng, D. (2008). Penalized estimating functions and variable selection in semiparametric regression models. *Journal of the American Statistical Association*, 103:672–680.

Kim, J. K. and Haziza, D. (2014). Doubly robust inference with missing data in survey sampling. *Statistica Sinica*, 24(1):375–394.

Kim, J. K. and Shao, J. (2013). *Statistical methods for handling incomplete data*. CRC press.

Li, K.-C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. (2010). Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861.

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

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Shimodaira, H. (2000). Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244.

Shortreed, S. M. and Ertefaie, A. (2017). Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*, 73:1111–1122.

Stojanov, P., Gong, M., Carbonell, J. G., and Zhang, K. (2019). Low-dimensional density ratio estimation for covariate shift correction. *Proceedings of machine learning research*, 89:3449.

Sugiyama, M., Suzuko, T., and Kanamori, T. (2012). *Density ratio estimation in machine learning*. Cambridge University Press, New York.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101(476):1619–1637.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97(3):661–682.

Tan, Z. (2020). Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data. *Biometrika*, 107(1):137–158.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Vermeulen, K. and Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110(511):1024–1036.

Yang, S., Kim, J., and Song, R. (2020). Doubly robust inference when combining probability and non-probability samples with high dimensional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82:445–465.

Zhang, S., Guo, B., Dong, A., He, J., Xu, Z., and Chen, S. X. (2017). Cautionary tales on air-quality improvement in beijing. In *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, volume 473, page 20170457. The Royal Society.

4.10 Appendix: Technical Details

4.10.1 Proof of Lemma 4.1

We wish to show that, under (4.1) and (4.10), we have

$$\mathbb{P}(Y \in A \mid \mathbf{b}(\mathbf{x}), \delta) = \mathbb{P}(Y \in A \mid \mathbf{b}(\mathbf{x})),$$

for any measurable set A. Now, for any $\mathbf{b}_0 = \mathbf{b}(\mathbf{x}_0)$, where $\mathbf{x}_0 \in \mathcal{H}$.

$$\begin{aligned}
\mathbb{P}(Y \in A \mid \mathbf{b}(\mathbf{X}) = \mathbf{b}_0, \delta) &= \frac{\int_A \int_{\mathbf{b}^{-1}(\mathbf{b}_0)} f(y|\mathbf{x}, \delta) f(\delta|\mathbf{x}) f(\mathbf{x}) d\mu(\mathbf{x}, y)}{\int_{\mathcal{Y}} \int_{\mathbf{b}^{-1}(\mathbf{b}_0)} f(y|\mathbf{x}, \delta) f(\delta|\mathbf{x}) f(\mathbf{x}) d\mu(\mathbf{x}, y)} \\
&\overset{(i)}{=} \frac{\int_A \int_{\mathbf{b}^{-1}(\mathbf{b}_0)} f(y|\mathbf{x}) f(\delta|\mathbf{x}) f(\mathbf{x}) d\mu(\mathbf{x}, y)}{\int_{\mathcal{Y}} \int_{\mathbf{b}^{-1}(\mathbf{b}_0)} f(y|\mathbf{x}) f(\delta|\mathbf{x}) f(\mathbf{x}) d\mu(\mathbf{x}, y)} \\
&\overset{(ii)}{=} \frac{\int_A \int_{\mathbf{b}^{-1}(\mathbf{b}_0)} f(y|\mathbf{b}(\mathbf{x})) f(\delta|\mathbf{x}) f(\mathbf{x}) d\mu(\mathbf{x}, y)}{\int_{\mathbf{b}^{-1}(\mathbf{b}_0)} f(\delta|\mathbf{x}) f(\mathbf{x}) d\mu(\mathbf{x})} \\
&= \frac{\int_{\mathbf{b}^{-1}(\mathbf{b}_0)} \mathbb{P}(Y \in A|\mathbf{b}(\mathbf{X}) = \mathbf{b}_0) f(\delta|\mathbf{x}) f(\mathbf{x}) d\mu(\mathbf{x})}{\int_{\mathbf{b}^{-1}(\mathbf{b}_0)} f(\delta|\mathbf{x}) f(\mathbf{x}) d\mu(\mathbf{x})} \\
&\overset{(iii)}{=} \mathbb{P}(Y \in A \mid \mathbf{b}(\mathbf{X}) = \mathbf{b}_0),
\end{aligned}$$

where (i) follows from (4.1), (ii) equality follows from (4.10) and (iii) follows from the fact that $\mathbb{P}(Y \in A \mid \mathbf{b}(\mathbf{X}) = \mathbf{b}_0)$ is a constant on the set $\mathbf{b}^{-1}(\mathbf{b}_0) = \{\mathbf{x} \in \mathcal{H} : \mathbf{b}(\mathbf{x}) = \mathbf{b}_0\}$.

4.10.2 Proof of Lemma 4.2

Note that

$$\begin{aligned}
& (n_0/n_1)\mathbb{E}\{r(\mathbf{x}) \mid y, \mathbf{b}(\mathbf{x}), \delta = 1\} \\
= \; & \mathbb{E}\left\{\frac{\mathbb{P}(\delta = 0 \mid \mathbf{x})}{\mathbb{P}(\delta = 1 \mid \mathbf{x})} \mid y, \mathbf{b}(\mathbf{x}), \delta = 1\right\} \\
= \; & \mathbb{E}\left\{\frac{\mathbb{P}(\delta = 0 \mid \mathbf{x}, \mathbf{b}(\mathbf{x}), y)}{\mathbb{P}(\delta = 1 \mid \mathbf{x}, \mathbf{b}(\mathbf{x}), y)} \mid y, \mathbf{b}(\mathbf{x}), \delta = 1\right\},
\end{aligned}$$

where the last equality holds by the MAR condition in (4.1). Also,

$$\begin{aligned}
& \mathbb{E}\left\{\frac{\mathbb{P}(\delta = 0 \mid \mathbf{x}, \mathbf{b}(\mathbf{x}), y)}{\mathbb{P}(\delta = 1 \mid \mathbf{x}, \mathbf{b}(\mathbf{x}), y)} \mid y, \mathbf{b}(\mathbf{x}), \delta = 1\right\} \\
= \; & \int \frac{f(\mathbf{x} \mid y, \mathbf{b}(\mathbf{x}), \delta = 0)\mathbb{P}(\delta = 0 \mid y, \mathbf{b}(\mathbf{x}))}{f(\mathbf{x} \mid y, \mathbf{b}(\mathbf{x}), \delta = 1)\mathbb{P}(\delta = 1 \mid y, \mathbf{b}(\mathbf{x}))} f(\mathbf{x} \mid y, \mathbf{b}(\mathbf{x}), \delta = 1) d\mu(\mathbf{x}) \\
= \; & \frac{\mathbb{P}(\delta = 0 \mid y, \mathbf{b}(\mathbf{x}))}{\mathbb{P}(\delta = 1 \mid y, \mathbf{b}(\mathbf{x}))} \int \frac{f(\mathbf{x} \mid y, \mathbf{b}(\mathbf{x}), \delta = 0)}{f(\mathbf{x} \mid y, \mathbf{b}(\mathbf{x}), \delta = 1)} f(\mathbf{x} \mid y, \mathbf{b}(\mathbf{x}), \delta = 1) d\mu(\mathbf{x}) \\
= \; & \frac{\mathbb{P}(\delta = 0 \mid y, \mathbf{b}(\mathbf{x}))}{\mathbb{P}(\delta = 1 \mid y, \mathbf{b}(\mathbf{x}))} \\
= \; & \frac{\mathbb{P}(\delta = 0 \mid \mathbf{b}(\mathbf{x}))}{\mathbb{P}(\delta = 1 \mid \mathbf{b}(\mathbf{x}))} \quad \text{(by (4.9))} \\
= \; & (n_0/n_1)\tilde{r}(\mathbf{x}).
\end{aligned}$$

Thus, Lemma 4.2 is established.

4.10.3 Regularity Conditions and Proof of Theorem 4.1

We have the following regularity conditions for Theorem 4.1.

Assumption 4.1. *Assume that there exists constants C_1 and C_2 such that*

$$0 < C_1 \leq \lambda_{min}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{b}(\mathbf{x}_i)\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\right) \leq \lambda_{max}\left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{b}(\mathbf{x}_i)\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\right) \leq C_2 < \infty,$$

where λ_{min} and λ_{max} are the smallest and largest eigenvalues of a specific matrix.

Assumption 4.2. *Assume the parameter space $\mathcal{G}$ for $\boldsymbol{\phi}$ is compact and the set*

$$\left\{\boldsymbol{\phi} \neq \mathbf{0} : \mathbf{b}(\mathbf{x}_i)^{\mathrm{T}}\boldsymbol{\phi} \leq 0 \text{ if } \delta_i = 1 \text{ for } i = 1, \ldots, n \text{ and } \mathbb{E}\left\{(1-\delta)\mathbf{b}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\phi}\right\} \geq 0\right\}$$

is empty.

Assumption 4.3. *The matrix* $\mathbb{E}[\partial/\partial\boldsymbol{\phi}\{\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi})\}|_{\boldsymbol{\phi}=\boldsymbol{\phi}^\star}]$ *defined in* (4.21) *is nonsingular.*

The detailed proof for Theorem 4.1 is presented as follows.

Proof. By Assumption 4.2 and Proposition S1 in Tan (2020), $\hat{Q}(\boldsymbol{\phi})$ is concave in $\boldsymbol{\phi}$ and is strictly concave and bounded from above, therefore has a unique maximizer $\hat{\boldsymbol{\phi}}$. By the weak law of large numbers, we have $\hat{Q}(\boldsymbol{\phi}) \xrightarrow{P} Q(\boldsymbol{\phi})$. Further, as $\mathcal{G}$ is compact, we have the uniform convergence of $\hat{Q}(\boldsymbol{\phi})$ to $Q(\boldsymbol{\phi})$. By Theorem 5.7 in Van der Vaart (2000), we have

$$\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^\star = o_p(1).$$

Then, by mean value theorem, we have

$$\hat{\mathbf{U}}_{DR}(\hat{\boldsymbol{\phi}}) - \hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^\star) = \frac{\partial}{\partial\boldsymbol{\phi}}\hat{\mathbf{U}}_{DR}(\tilde{\boldsymbol{\phi}})\left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^\star\right), \tag{4.46}$$

where $\tilde{\boldsymbol{\phi}}$ is a point between $\boldsymbol{\phi}^\star$ and $\hat{\boldsymbol{\phi}}$. Apparently, $\partial\hat{\mathbf{U}}_{DR}/\partial\boldsymbol{\phi}$ and $\mathbb{E}[\partial\hat{\mathbf{U}}_{DR}/\partial\boldsymbol{\phi}]$ are continuous within the set $\mathcal{G}$. Therefore, using the similar technique as above, we can arrive at

$$\frac{\partial}{\partial\boldsymbol{\phi}}\hat{\mathbf{U}}_{DR}(\tilde{\boldsymbol{\phi}}) = \mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\phi}}\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^\star)\right\} + o_p(1). \tag{4.47}$$

Combine (4.46), (4.47) and the fact that $\hat{\mathbf{U}}_{DR}(\hat{\boldsymbol{\phi}}) = \mathbf{0}$, we have

$$-\sqrt{n}\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^\star) = \sqrt{n}\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\phi}}\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^\star)\right\}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^\star) + o_p\left(\sqrt{n}\left\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^\star\right\|\right). \tag{4.48}$$

Then, by Cauchy-Schwarz inequality, we have

$$\begin{aligned}
\sqrt{n}\left\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^\star\right\| \leq& \left\|\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\phi}}\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^\star)\right\}^{-1}\right\| \left\|\sqrt{n}\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\phi}}\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^\star)\right\}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^\star)\right\| \\
=& \left\|\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\phi}}\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^\star)\right\}^{-1}\right\| \left\|\sqrt{n}\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^\star) + o_p\left(\sqrt{n}\left\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^\star\right\|\right)\right\| \\
=& \mathcal{O}_p(1) + o_p\left(\sqrt{n}\left\|\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^\star\right\|\right),
\end{aligned}$$

which implies the root-n convergence of $\hat{\boldsymbol{\phi}}$. Therefore, (4.48) can be written as

$$\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}^\star = -\left[\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\phi}'}\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^\star)\right\}\right]^{-1}\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^\star) + o_p(n^{-1/2}), \tag{4.49}$$

where

$$\begin{aligned}\frac{n_0}{n}\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}) =& \frac{n_0}{n}\left[\frac{1}{n_1}\sum_{i=1}^{n}\delta_i \exp\{\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\phi}\} - \frac{1}{n_0}\sum_{i=1}^{n}(1-\delta_i)\right]\mathbf{b}(\mathbf{x}_i)\\ =& \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\delta_i n_0}{n_1}\exp\{\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i)\boldsymbol{\phi}\} - (1-\delta_i)\right]\mathbf{b}(\mathbf{x}_i)\\ =& \frac{1}{n}\sum_{i=1}^{n}(\delta_i - 1)\,\mathbf{b}(\mathbf{x}_i) + \frac{1}{n}\cdot\frac{n_0}{n_1}\sum_{i=1}^{n}\delta_i\tilde{r}(\mathbf{x}_i;\boldsymbol{\phi})\mathbf{b}(\mathbf{x}_i)\\ =& \frac{1}{n}\sum_{i=1}^{n}\left[\delta_i\left\{1+\frac{n_0}{n_1}\tilde{r}(\mathbf{x}_i;\boldsymbol{\phi})\right\} - 1\right]\mathbf{b}(\mathbf{x}_i).\end{aligned}$$

By Taylor expansion and similar technique we used above, we have

$$\begin{aligned}\hat{\theta}_{PS}(\hat{\boldsymbol{\phi}}) =& \hat{\theta}_{PS}(\boldsymbol{\phi}^\star) + \mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\phi}}\hat{\theta}_{PS}(\boldsymbol{\phi}^\star)\right\}\left(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi}^\star\right) + o_p\left(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi}^\star\right) + \mathcal{O}\left(\left\|\hat{\boldsymbol{\phi}}-\boldsymbol{\phi}^\star\right\|_2^2\right)\\ =& \hat{\theta}_{PS}(\boldsymbol{\phi}^\star) + \mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\phi}}\hat{\theta}_{PS}(\boldsymbol{\phi}^\star)\right\}\left(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi}^\star\right) + o_p(n^{-1/2})\\ \overset{(i)}{=}& \hat{\theta}_{PS}(\boldsymbol{\phi}^\star) - \mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\phi}}\hat{\theta}_{PS}(\boldsymbol{\phi}^\star)\right\}\left[\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\phi}'}\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^\star)\right\}\right]^{-1}\hat{\mathbf{U}}_{DR}(\boldsymbol{\phi}^\star) + o_p(n^{-1/2})\\ =& \bar{y}_\pi + (\bar{\mathbf{b}}_n - \bar{\mathbf{b}}_\pi)^{\mathrm{T}}\tilde{\boldsymbol{\beta}} + o_p(n^{-1/2}),\end{aligned}$$

where

$$(\bar{\mathbf{b}}_\pi^{\mathrm{T}}, \bar{y}_\pi) = \frac{1}{n}\sum_{i=1}^{n}\delta_i\left\{1+\frac{n_0}{n_1}\tilde{r}(\mathbf{x}_i;\boldsymbol{\phi}^\star)\right\}(\mathbf{b}^{\mathrm{T}}(\mathbf{x}_i), y_i)$$

and $\bar{\mathbf{b}}_n = n^{-1}\sum_{i=1}^{n}\mathbf{b}(\mathbf{x}_i)$, and equality (i) follows from (4.49), which completes the proof of (4.24).

□

4.10.4 Regularity conditions and Proof of Theorem 4.2

We have the following regularity conditions for Theorem 4.2.

Assumption 4.4. *Assume that there exists constants $C_{1,t}$ and $C_{2,t}$ such that*

$$0 < C_{1,t} \le \lambda_{min}\left(\frac{1}{n_t}\sum_{i=1}^{n}\delta_{i,t}\mathbf{z}_{i,t}\mathbf{z}_{i,t}^{\mathrm{T}}\right) \le \lambda_{max}\left(\frac{1}{n_t}\sum_{i=1}^{n}\delta_{i,t}\mathbf{z}_{i,t}\mathbf{z}_{i,t}^{\mathrm{T}}\right) \le C_{2,t} < \infty,$$

where λ_{min} and λ_{max} are the smallest and largest eigenvalues of a specific matrix, for $t = 2, \ldots, T$.

Assumption 4.5. *Assume the parameter space $\mathcal{G}_t$ for $\boldsymbol{\phi}_t$ is compact and the set*

$$\left\{\boldsymbol{\phi} \neq \mathbf{0} : \mathbf{z}_{i,t}^{\mathrm{T}}\boldsymbol{\phi}_t \leq 0 \text{ if } \delta_{i,1} = 1 \text{ for } i = 1, \ldots, n \text{ and } \mathbb{E}\left\{\delta_{i,t}\mathbf{z}_{i,t}^{\mathrm{T}}\boldsymbol{\phi}_t\right\} \geq 0\right\}$$

is empty, for $t = 2, \ldots, T$.

Assumption 4.6. *The matrix $\mathbb{E}[\partial/\partial\boldsymbol{\phi}_t\{\hat{\mathbf{U}}_{DR,t}(\boldsymbol{\phi}_t)\}]$ is nonsingular, where*

$$\hat{\mathbf{U}}_{DR,t}(\boldsymbol{\phi}_t) = \frac{1}{n_1}\sum_{i=1}^{n}\left\{\delta_{i,1}\exp(\mathbf{z}_{i,t}^{\mathrm{T}}\boldsymbol{\phi}_t) - \delta_{i,t}\right\}\mathbf{z}_{i,t},$$

for $t = 2, \ldots, T$.

Proof. We present the main steps for expansion of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$, the details are similar to that in Theorem 4.1. First of all, by mean value theorem, there exists $\tilde{\boldsymbol{\theta}}$ between $\boldsymbol{\theta}_0$ and $\hat{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\phi}}$ between $\boldsymbol{\phi}_0$ and $\hat{\boldsymbol{\phi}}$ such that

$$\begin{aligned}
\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) &= -\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\delta_{i,1}\mathbf{U}(\tilde{\boldsymbol{\theta}}; \mathbf{x}_i, \mathbf{y}_i)w_i(\hat{\boldsymbol{\phi}})\right]^{-1} \\
&\quad \times \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_{i,1}\mathbf{U}(\boldsymbol{\theta}_0, \mathbf{x}_i, \mathbf{y}_i)w_i(\hat{\boldsymbol{\phi}})\right\} \\
&= -\left[\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\delta_{i,1}\mathbf{U}(\tilde{\boldsymbol{\theta}}; \mathbf{x}_i, \mathbf{y}_i)w_i(\hat{\boldsymbol{\phi}})\right]^{-1} \\
&\quad \left[\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta_{i,1}\mathbf{U}(\boldsymbol{\theta}_0, \mathbf{x}_i, \mathbf{y}_i)w_i(\boldsymbol{\phi}_0) + \frac{1}{n}\sum_{i=1}^{n}\delta_{i,1}\left\{\frac{\partial}{\partial\boldsymbol{\phi}}w_i(\tilde{\boldsymbol{\phi}})\right\}\left\{\sqrt{n}\left(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0\right)\right\}\right].
\end{aligned} \tag{4.50}$$

Note that

$$w_i(\boldsymbol{\phi}_0) = 1 + \sum_{t=2}^{T}\frac{n_t}{n_1}r_t(\mathbf{z}_{i,t}; \boldsymbol{\phi}_t),$$

we then have

$$\frac{\partial}{\partial\boldsymbol{\phi}}w_i(\boldsymbol{\phi}_0) = (\frac{n_2}{n_1}r_2(\mathbf{z}_{i,2}; \boldsymbol{\phi}_2)\mathbf{z}_{i,2}^{\mathrm{T}}, \ldots, \frac{n_T}{n_1}r_T(\mathbf{z}_{i,T}; \boldsymbol{\phi}_T)\mathbf{z}_{i,T}^{\mathrm{T}}). \tag{4.51}$$

Meanwhile, we have

$$\sqrt{n}\left(\hat{\boldsymbol{\phi}}-\boldsymbol{\phi}_0\right)=\sqrt{n}\begin{bmatrix}-\left[\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\phi}_2}\hat{\mathbf{U}}_{DR,2}(\boldsymbol{\phi}_2)\right\}\right]^{-1} & \cdots & \mathbf{0}\\ \vdots & \ddots & \vdots\\ \mathbf{0} & \cdots & -\left[\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\phi}_T}\hat{\mathbf{U}}_{DR,T}(\boldsymbol{\phi}_T)\right\}\right]^{-1}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{U}}_{DR,2}(\boldsymbol{\phi}_2)\\ \vdots\\ \hat{\mathbf{U}}_{DR,T}(\boldsymbol{\phi}_T)\end{bmatrix}+o_p(1). \tag{4.52}$$

Further, we have

$$\frac{n_t}{n}\hat{\mathbf{U}}_{DR,t}(\boldsymbol{\phi}_t)=\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_{i,1}\frac{n_t}{n_1}r_t(\mathbf{z}_{i,t};\boldsymbol{\phi}_t)-\delta_{i,t}\right\}\mathbf{z}_{i,t}, \tag{4.53}$$

and

$$\frac{n_t}{n}\frac{\partial}{\partial\boldsymbol{\phi}_t}\hat{\mathbf{U}}_{DR,t}(\boldsymbol{\phi}_t)=\frac{1}{n}\sum_{i=1}^{n}\left\{\delta_{i,1}\frac{n_t}{n_1}r_t(\mathbf{z}_{i,t};\boldsymbol{\phi}_t)\right\}\mathbf{z}_{i,t}\mathbf{z}_{i,t}^{\mathrm{T}}, \tag{4.54}$$

for $t=2,\ldots,T$.

Now, plug (4.51), (4.52), (4.53) and (4.54) into (4.50), we have

$$\begin{aligned}\sqrt{n}\left(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0\right)=&-\left[\mathbb{E}\left\{\delta_1 w(\boldsymbol{\phi}_0)\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{U}(\boldsymbol{\theta}_0;\mathbf{x},\mathbf{y})\right\}\right]^{-1}\\ &\times\frac{1}{\sqrt{n}}\left[\sum_{i=1}^{n}\delta_{i,1}\mathbf{U}(\boldsymbol{\theta}_0;\mathbf{x}_i,\mathbf{y}_i)w_i(\boldsymbol{\phi}_0)-\left\{\delta_{i,1}\sum_{t=2}^{T}\frac{n_t}{n_1}r_t(\mathbf{z}_{i,t};\boldsymbol{\phi}_t)\tilde{\boldsymbol{\beta}}_t\mathbf{z}_{i,t}\right\}+\sum_{t=2}^{T}\delta_{i,t}\tilde{\boldsymbol{\beta}}_t\mathbf{z}_{i,t}\right]+o_p(1),\end{aligned} \tag{4.55}$$

$$\begin{aligned}=&-\left[\mathbb{E}\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{U}(\boldsymbol{\theta}_0;\mathbf{x},\mathbf{y})\right\}\right]^{-1}\\ &\times\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\delta_{i,1}\mathbf{U}_i+\sum_{t=2}^{T}\delta_{i,t}\tilde{\boldsymbol{\beta}}_t\mathbf{z}_{i,t}+\delta_{i,1}\sum_{t=2}^{T}\frac{n_t}{n_1}r_t(\mathbf{z}_{i,t};\boldsymbol{\phi}_t)\left\{\mathbf{U}_i-\tilde{\boldsymbol{\beta}}_t\mathbf{z}_{i,t}\right\}\right]+o_p(1),\end{aligned}$$

where $\mathbf{U}_i=\mathbf{U}(\boldsymbol{\theta}_0;\mathbf{x}_i,\mathbf{y}_i)$ and $\tilde{\boldsymbol{\beta}}_t$ is the probability limit to the solution of

$$\sum_{i=1}^{n}\delta_{i,1}r_t(\mathbf{z}_{i,t};\boldsymbol{\phi}_t)\left\{\mathbf{U}(\boldsymbol{\theta}_0;\mathbf{x}_i,\mathbf{y}_i)-\boldsymbol{\beta}_t\mathbf{z}_{i,t}\right\}\mathbf{z}_{i,t}^{\mathrm{T}}=\mathbf{0}.$$

□

4.10.5 Proof of Lemma 4.3

We have only to prove that

$$f(y\mid\mathbf{x}_{\mathcal{M}},\delta)=f(y\mid\mathbf{x}_{\mathcal{M}}).$$

Now, using Bayes formula,

$$\begin{aligned} f(y \mid \mathbf{x}_{\mathcal{M}}, \delta) &= \frac{\int f(y \mid \mathbf{x}, \delta)\mathbb{P}(\delta \mid \mathbf{x}) f(\mathbf{x}_{\mathcal{M}^c} \mid \mathbf{x}_{\mathcal{M}}) f(\mathbf{x}_{\mathcal{M}}) d\mathbf{x}_{\mathcal{M}^c}}{\int\int f(y \mid \mathbf{x}, \delta)\mathbb{P}(\delta \mid \mathbf{x}) f(\mathbf{x}_{\mathcal{M}^c} \mid \mathbf{x}_{\mathcal{M}}) f(\mathbf{x}_{\mathcal{M}}) d\mathbf{x}_{\mathcal{M}^c} dy} \\ &= \frac{\int f(y \mid \mathbf{x}_{\mathcal{M}})\mathbb{P}(\delta \mid \mathbf{x}) f(\mathbf{x}_{\mathcal{M}^c} \mid \mathbf{x}_{\mathcal{M}}) f(\mathbf{x}_{\mathcal{M}}) d\mathbf{x}_{\mathcal{M}^c}}{\int\int f(y \mid \mathbf{x}_{\mathcal{M}})\mathbb{P}(\delta \mid \mathbf{x}) f(\mathbf{x}_{\mathcal{M}^c} \mid \mathbf{x}_{\mathcal{M}}) f(\mathbf{x}_{\mathcal{M}}) d\mathbf{x}_{\mathcal{M}^c} dy} \\ &= \frac{f(y \mid \mathbf{x}_{\mathcal{M}}) \int \mathbb{P}(\delta \mid \mathbf{x}) f(\mathbf{x}_{\mathcal{M}^c} \mid \mathbf{x}_{\mathcal{M}}) d\mathbf{x}_{\mathcal{M}^c}}{\int f(y \mid \mathbf{x}_{\mathcal{M}}) \int \mathbb{P}(\delta \mid \mathbf{x}) f(\mathbf{x}_{\mathcal{M}^c} \mid \mathbf{x}_{\mathcal{M}}) d\mathbf{x}_{\mathcal{M}^c} dy} \\ &= \frac{f(y \mid \mathbf{x}_{\mathcal{M}})\mathbb{P}(\delta \mid \mathbf{x}_{\mathcal{M}})}{\int f(y \mid \mathbf{x}_{\mathcal{M}})\mathbb{P}(\delta \mid \mathbf{x}_{\mathcal{M}}) dy} = f(y \mid \mathbf{x}_{\mathcal{M}}), \end{aligned}$$

where the second equality follows by MAR assumption and the reduced model assumption.

4.10.6 Proof for Corollary 4.1

Apparently, we only need to compute the asymptotic variance V_r. Without loss of generality, the (4.27) is equivalent to that there exists a $\boldsymbol{\beta}^\star \in \mathbb{R}^{l+1}$, such that $\mathbb{E}(Y \mid \mathbf{X}) = \mathbf{b}^{\mathrm{T}}(\mathbf{X})\boldsymbol{\beta}^\star$. From Theorem 4.1, we can directly calculate the variance of the influence function

$$d(\mathbf{X}, Y, \delta; \boldsymbol{\phi}^\star) = \mathbb{E}\left(Y \mid \mathbf{x}\right) + \delta\{1 + (n_0/n_1)\tilde{r}(\mathbf{X}; \boldsymbol{\phi}^\star)\}\{Y - \mathbb{E}\left(Y \mid \mathbf{x}\right)\}.$$

That is,

$$V_r = \mathbb{V}[\mathbb{E}\{d(\mathbf{X}, Y, \delta; \boldsymbol{\phi}^\star) \mid \boldsymbol{\delta}, \mathbf{X}\}] + \mathbb{E}[\mathbb{V}\{d(\mathbf{X}, Y, \delta; \boldsymbol{\phi}^\star) \mid \boldsymbol{\delta}, \mathbf{X}\}]$$

where the conditional expectation $\mathbb{E}(\cdot \mid \mathbf{x}, \boldsymbol{\delta})$ is with respect to Y conditional on $\mathbf{x}$ and δ. In particular, we have

$$\mathbb{V}[\mathbb{E}\{d(\mathbf{X}, Y, \delta; \boldsymbol{\phi}^\star) \mid \mathbf{X}, \boldsymbol{\delta}\}] = \mathbb{V}\{\mathbb{E}(Y \mid \mathbf{X})\}, \tag{4.56}$$

by MAR. Further, note that

$$\tilde{\pi}(\mathbf{x}) \equiv \mathbb{P}(\delta = 1 \mid \mathbf{b}(\mathbf{x})) = \left\{1 + \frac{n_0}{n_1}\tilde{r}(\mathbf{X}; \boldsymbol{\phi}^\star)\right\}^{-1},$$

we have

$$\begin{aligned} \mathbb{E}[\mathbb{V}\{d(\mathbf{X}, Y, \delta; \boldsymbol{\phi}^\star) \mid \mathbf{X}, \boldsymbol{\delta}\}] &= \mathbb{E}(\mathbb{V}[\delta_i\{\tilde{\pi}(\mathbf{x})\}^{-1}\{Y - \mathbb{E}(Y \mid \mathbf{X})\} \mid \mathbf{X}, \boldsymbol{\delta}]) \\ &= \mathbb{E}\left[\delta_i\{\tilde{\pi}(\mathbf{x})\}^{-2}\mathbb{V}(Y \mid \mathbf{X})\right] \\ &= \mathbb{E}\left[\{\tilde{\pi}(\mathbf{x})\}^{-1}\mathbb{V}(Y \mid \mathbf{X})\right]. \end{aligned} \tag{4.57}$$

Combine (4.56) and (4.57), we arrive the conclusion in Corollary 4.1.

4.10.7 Regularity Conditions for Corollary 4.2

Assumption 4.7. *For any constant M, there exists non-singular matrix $\mathbf{D}$ such that*

$$\sup_{|\boldsymbol{\alpha}-\boldsymbol{\alpha}_0|\leq Mn^{-1/2}} \left| n^{-1/2}\mathbf{U}(\boldsymbol{\alpha}) - n^{-1/2}\mathbf{U}(\boldsymbol{\alpha}_0) - n^{1/2}\mathbf{D}(\boldsymbol{\alpha}-\boldsymbol{\alpha}_0) \right| = o_p(1).$$

Additionally, $n^{-1/2}\mathbf{U}(\boldsymbol{\alpha}_0) \xrightarrow{\mathcal{L}} N(\mathbf{0}, \mathbf{F})$ for a positive definite matrix $\mathbf{F}$.

Assumption 4.8. *The tuning parameter λ in (4.33) satisfies*

$$\lambda \to 0, \sqrt{n}\lambda \to \infty$$

as $n \to \infty$.

4.10.8 Computational Details for SDR

Define the sample operator for a matrix as $S_{\Omega,M} : \mathbb{R}^{n\times d} \to \mathbb{R}^{n_1\times d}$. That is, all rows in a $n \times d$ matrix with $\delta_i = 1$ form as the image, where i is the index for rows. Naturally, we have $S_{\Omega,M}(\mathbf{X}_n) \in \mathbb{R}^{n_1\times d}$, where $\mathbf{X}_n = [\mathbf{x}_1^{\mathrm{T}}; \ldots; \mathbf{x}_n^{\mathrm{T}}]$. Further, we define the normalized sample operator for a matrix as $\tilde{S}_{\Omega,M}(\mathbf{M}) = S_{\Omega,M}(\mathbf{M}) - \{n_1^{-1}\mathbf{1}_{n_1}^{\mathrm{T}} S_{\Omega,M}(\mathbf{M})\} \otimes \mathbf{1}_d$ which ensures the matrix $\tilde{S}_{\Omega,M}(\mathbf{M})$ is a column mean zero $n_1 \times d$ matrix for $\mathbf{M} \in \mathbb{R}^{n\times d}$. Similarly, let the sample operator for a vector be $S_{\boldsymbol{\Omega},V} : \mathbb{R}^n \to \mathbb{R}^{n_1}$. We have the observed vector of $\mathbf{y}$ as $S_{\Omega,V}(\mathbf{y}) \in \mathbb{R}^{n_1}$. Define the normalized sample operator for vector as $\tilde{S}_{\Omega,V}(\mathbf{v}) = S_{\Omega,V}(\mathbf{v}) - n_1^{-1}\{\mathbf{1}_{n_1}^{\mathrm{T}} S_{\Omega,V}(\mathbf{v})\}\mathbf{1}_{n_1}$ for a vector $\mathbf{v} \in \mathbb{R}^n$. As a result, we have $\hat{\mathcal{C}}_{YY} = n_1^{-1}\tilde{S}_{\Omega,V}^{\mathrm{T}}(\mathbf{y})\tilde{S}_{\Omega,V}(\mathbf{y})$, $\hat{\mathcal{C}}_{\mathbf{b}(\mathbf{X}),Y} = n_1^{-1}\mathbf{W}\tilde{S}_{\Omega,M}^{\mathrm{T}}(\mathbf{X}_n)\tilde{S}_{\Omega,V}(\mathbf{y})$, $\hat{\mathcal{C}}_{\mathbf{b}(\mathbf{X}),\mathbf{b}(\mathbf{X})} = n_1^{-1}\mathbf{W}\tilde{S}_{\Omega,M}^{\mathrm{T}}(\mathbf{X}_n)\tilde{S}_{\Omega,M}(\mathbf{X}_n)\mathbf{W}^{\mathrm{T}}$. Also, in case of singularity of matrix during iteration algorithm, the objective function in (4.40) can be written as

$$\begin{aligned}\mathrm{Trace}\Bigg[&\frac{1}{n_1}\tilde{S}_{\Omega,V}^{\mathrm{T}}(\mathbf{y})\tilde{S}_{\Omega,V}(\mathbf{y}) - \frac{1}{n_1^2}\tilde{S}_{\Omega,V}^{\mathrm{T}}(\mathbf{y})\tilde{S}_{\Omega,M}(\mathbf{X}_n)\mathbf{W}^{\mathrm{T}} \\ &\times \left\{\mathbf{W}\tilde{S}_{\Omega,M}^{\mathrm{T}}(\mathbf{X}_n)\tilde{S}_{\Omega,M}(\mathbf{X}_n)\mathbf{W}^{\mathrm{T}} + \epsilon n_1 \mathbf{I}_{n_1}\right\}^{-1} \mathbf{W}\tilde{S}_{\Omega,M}^{\mathrm{T}}(\mathbf{X}_n)\tilde{S}_{\Omega,V}(\mathbf{y})\Bigg],\end{aligned}$$

where $\epsilon = 0.01$. We solve the optimization problem (4.40) with the limited-memory Riemannian Broyden–Fletcher–Goldfarb–Shanno (LRBFGS) algorithms (Huang et al., 2015).

Another important aspect for SDR is the choice of l. For a given l, let the minimizer of (4.40) be $\hat{\mathbf{W}}_{(l)}$. Like the idea in Subsection 4.5.2, we employ a working model by assuming a linear model for Y given $\hat{\mathbf{W}}_{(l)}\mathbf{b}(\mathbf{X})$. The optimal $\hat{l}$ can be chosen to minimize the following Bayesian information criterion

$$BIC(l) = n_1 \ln(\hat{\sigma}^2_{(l)}) + (l+2)\ln(n_1),$$

where $\hat{\sigma}^2_{(l)}$ is the estimated error variance for linear model $Y \mid \hat{\mathbf{W}}_{(l)}\mathbf{b}(\mathbf{X})$.

CHAPTER 5. GENERAL CONCLUSION

In this dissertation, we consider the problems related to survey sampling and missing data. For survey sampling, we propose a kernel-based functional calibration method to estimate the population mean with fully observed auxiliary information. The root-n consistency of our proposed estimator is studied. Further, under regularity conditions, our proposed the proposed calibration estimator attains the Godambe-Joshi lower bound asymptotically. As for missing data problem, we first proposed imputation and propensity score methods using kernel ridge regression to estimate the population mean, which is more robust than parametric approach. The root-n consistency of the proposed estimators is shown and valid variance estimators are proposed. Further, we propose a new framework with density ratio estimation to handle missing data problem based on propensity approach. The asymptotic property of our proposed estimators is studied. We further extend our proposed method to multivariate missing scenario.