# SUFFICIENT DIMENSION REDUCTION WITH MISSING DATA

---

A Dissertation
Submitted to
the Temple University Graduate Board

---

in Partial Fulfillment
of the Requirements for the Degree of
DOCTOR OF PHILOSOPHY

---

by
Qi Xia
August 2017

Examining Committee Members:

Yuexiao Dong, Main Advisor, Department of Statistical Science
Cheng Yong Tang, Co-Advisor, Department of Statistical Science
William W. S. Wei, Examining Chair, Department of Statistical Science
Xu Han, Department of Statistical Science
Yang Yang, School of Tourism and Hospitality Management

# ABSTRACT

SUFFICIENT DIMENSION REDUCTION WITH MISSING DATA

Qi Xia

DOCTOR OF PHILOSOPHY

Temple University, August 2017

Existing sufficient dimension reduction (SDR) methods typically consider cases with no missing data. The dissertation aims to propose methods to facilitate the SDR methods when the response can be missing. The first part of the dissertation focuses on the seminal sliced inverse regression (SIR) approach proposed by Li (1991). We show that missing responses generally affect the validity of the inverse regressions under the mechanism of missing at random. We then propose a simple and effective adjustment with inverse probability weighting that guarantees the validity of the SIR. Furthermore, a marginal coordinate test is introduced for this adjusted estimator. The proposed method share the simplicity of SIR and requires the linear conditional mean assumption. The second part of the dissertation proposes two new estimating equation procedures: the complete case estimating equation approach and the inverse probability weighted estimating equation approach. The two approaches are applied to a family of dimension reduction methods, which includes ordinary least squares, principal Hessian directions, and SIR. By solving the estimating equations, the two approaches are able to avoid the common assumptions in the SDR literature, the linear conditional mean assumption, and the con-

stant conditional variance assumption. For all the aforementioned methods, the asymptotic properties are established, and their superb finite sample performances are demonstrated through extensive numerical studies as well as a real data analysis.

In addition, existing estimators of the central mean space have uneven performances across different types of link functions. To address this limitation, a new hybrid SDR estimator is proposed that successfully recovers the central mean space for a wide range of link functions. Based on the new hybrid estimator, we further study the order determination procedure and the marginal coordinate test. The superior performance of the hybrid estimator over existing methods is demonstrated in simulation studies. Note that the proposed procedures dealing with the missing response at random can be simply adapted to this hybrid method.

# ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to all those who have been side by side with me, who supported and challenged me along the way. I am sincerely grateful to have my advisors, Dr. Yuexiao Dong and Dr. Cheng Yong Tang, who have been supporting my projects and giving such thoughtful feedback, who always aimed at moving me forward. Their inspiration and instructions have guided me along the journey and benefited to who I am and who I will be. I would like to thank my committees members, Dr. William W. S. Wei and Dr. Xu Han for their inspiring comments, their generosity and kindness. A special thanks to Dr. Yang Yang for bringing in expertise and insights as my external examiner. I also would like to thank my fellow doctoral students - those who have moved on, those in the quagmire, and those just beginning - for their support, feedback, and friendship.

To Xinyang Tang,

my amazing husband,

whose sacrificial care made it possible for me to complete the work.

And to my beloved parents,

Weiguo Xia and Shaoying Wang,

and parents in law,

Weiping Tang and Yuxian Li,

for their endless encouragement and patience.

# TABLE OF CONTENTS

# List of Figures

# List of Tables

# CHAPTER 1

# INTRODUCTION

As technologies advance dramatically, scientific data grows in both size and complexity. The high dimensionality of data leads to some challenges for statistical inference. One is that in regression analysis, the large amount of predictors makes it cumbersome to detect the relationship between the response variable and the collection of predictors. Another is that the more commonly encountered missingness in high-dimensional data can complicate and weaken the interpretation of statistical analysis.

Sufficient dimension reduction (SDR, Li, 1991; Cook, 1998) has attracted considerable interests in the analysis of high-dimensional data. It aims to identify a lower dimensional vector of linear combinations of the predictors, while maintain full regression information and impose no parametric models. However, the existing SDR methods typically consider cases with no missing data. Therefore, it is desirable to develop SDR methods under missing data framework. This book provides approaches for SDR methods dealing with missing data.

Throughout the book, we denote $\mathbb{R}$ as the set of all real numbers and $\mathbb{R}^p$ as

the $p$-dimensional Euclidean space. We present a scalar by a lower case letter $x$, a (random) vector by a bold lowercase letter $\mathbf{x}$, and a (random) matrix by a bold upper case letter $\mathbf{X}$. A random variable is denoted by an upper case letter X.

Following the introduction, we will show basic dimension reduction concepts as well as some classic and modern SDR approaches. After that, we will give a brief review for missing data analysis.

## 1.1 Sufficient Dimension Reduction

Sufficient dimension reduction approach assumes that the response variable relates to only a few linear combinations of the many predictors. Thus, even when all the predictors have explanatory effect, we can formulate the effect sufficiently as one or more linear combinations. The goal of dimension reduction is to identify these few linear combinations.

### 1.1.1 Central Space and Central Mean Space

Let $\mathbf{x} = (X_1, ..., X_p)^{\mathrm{T}} \in \mathbb{R}^p$ be a random vector representing the predictors, and let $Y$ be a random variable representing the univariate response. A main class of SDR problems concern the conditional distribution of Y given as $F(y \mid \mathbf{x})$, referred to as central space problem. If there exits a matrix $\mathbf{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d) \in \mathbb{R}^{p \times d}$ $(d < p)$ with the smallest possible $d$ such that the distribution of $Y$ conditional on $\mathbf{x}$ is the same as that conditional on $\mathbf{B}^{\mathrm{T}}\mathbf{x}$, written as,

$$F(y \mid \mathbf{x}) = F(y \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}), \text{ for all } y \in \mathbb{R}, \tag{1.1}$$

where $F(y \mid \mathbf{x}) = P(Y \leq y \mid \mathbf{x})$. Then, the $p$-dimensional covariates x can be replaced by the $d$-dimensional linear combinations $\mathbf{B}^{\mathrm{T}}\mathbf{x}$. Model (1.1) was mentioned but not explicitly explored in both Li (1991) and Cook (1998). Zeng and Zhu (2010) proved that (1.1) is equivalent to Li's (1991) dimension reduction model, $Y = f(\mathbf{B}^{\mathrm{T}}\mathbf{x}, \varepsilon)$, where $f(\cdot)$ is an unspecified link function, $\varepsilon$ is a random error independent of $\mathbf{x}$ and $\mathrm{E}(\varepsilon) = 0$; also it is equivalent to Cook's (1998) independence model, $Y \perp\!\!\!\perp \mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}$, where $\perp\!\!\!\perp$ denotes statistical independence. It states that given $\mathbf{B}^{\mathrm{T}}\mathbf{x}$, $Y$ and $\mathbf{x}$ are independent of each other. Note that for $\mathbf{B}$ satisfying (1.1), its multiplication to any non-singular matrix $\mathbf{A} \in \mathbb{R}^{q \times d}$, will still satisfy (1.1). In other words, if $F(y \mid \mathbf{x}) = F(y \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$, then $F(y \mid \mathbf{x}) = F(y \mid (\mathbf{BA})^{\mathrm{T}}\mathbf{x})$. Therefore, $\mathbf{B}$ is not identifiable. Instead of matrix $\mathbf{B}$ itself, the column space of $\mathbf{B}$ which satisfies (1.1) is the one that really drives the conditional independence relationship. Such a column space is called dimension reduction subspace (DRS; Cook, 1994). Aiming at finding a minimum dimension reduction subspace, Cook (1998) introduced the concept of central space as the intersection of all dimension reduction subspaces, and expressed by $\mathcal{S}_{Y|\mathbf{x}}$. The smallest subspace usually exists under the premise that the intersection of two dimension reduction subspaces is again a dimension reduction subspace (Yin et al., 2008), and is uniquely defined except for some degenerated cases (Cook, 2004). Central space thus becomes the main objective in dimension reduction inquiry with model (1.1). The central space dimension $d = \dim(\mathcal{S}_{Y|\mathbf{x}})$ is called structural dimension.

The central space problem concerns the conditional distribution of response given the predictors and provides a complete picture of their relation. In many situations, regression analysis is mostly interested in inferring about the

conditional mean of the response given the predictors. In this case, dimension reduction hinges on finding the matrix $\mathbf{B}$ satisfies a less restrictive assumption

$$\mathrm{E}(Y \mid \mathbf{x}) = \mathrm{E}(Y \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) \tag{1.2}$$

such that $\mathbf{B}^{\mathrm{T}}\mathbf{x}$ contains all the information about $Y$ that is available from $\mathrm{E}(Y \mid \mathbf{x})$. It is equivalent to the assumption that $Y \perp\!\!\!\perp \mathrm{E}(Y \mid \mathbf{x}) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}$. Similarly, $\mathbf{B}$ in model (1.2) is not unique and the column space of $\mathbf{B}$ satisfying (1.2) is of interest, defined as a mean dimension reduction subspace. The intersection of all dimension reduction subspaces is called the central mean space (Cook and Li, 2002), and expressed as $\mathcal{S}_{\mathrm{E}(Y|\mathbf{x})}$.

Generalizing the idea of the central mean space and central space, Yin and Cook (2002) introduced the central $k$-th moment subspace. Extending the conditional mean to the conditional variance, Zhu and Zhu (2009) defined the notion of central variance space.

Given standardized predictor $\mathbf{z} = \boldsymbol{\Sigma}^{-1/2}(\mathbf{x} - \boldsymbol{\mu})$, where $\boldsymbol{\mu} = \mathrm{E}(\mathbf{x})$ and $\boldsymbol{\Sigma} = \mathrm{Cov}(\mathbf{x})$, for $\boldsymbol{\eta}$ being a basis matrix of the corresponding dimension reduction subspaces, $Y \perp\!\!\!\perp \mathbf{z} \mid \boldsymbol{\eta}^{\mathrm{T}}\mathbf{z}$ if and only if $Y \perp\!\!\!\perp \mathbf{x} \mid (\boldsymbol{\Sigma}^{-1/2}\boldsymbol{\eta})^{\mathrm{T}}\mathbf{x}$. We will immediately have the relationships $\mathcal{S}_{Y|\mathbf{x}} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{Y|\mathbf{z}}$ and $\mathcal{S}_{\mathrm{E}(Y|\mathbf{x})} = \boldsymbol{\Sigma}^{-1/2}\mathcal{S}_{\mathrm{E}(Y|\mathbf{z})}$. This property is known as the invariance law (Cook, 1998). A practical implication is that standardizing the predictor does not change the nature for implementing dimension reduction, one can first estimate $\mathbf{z}$-scale central space $\mathcal{S}_{Y|\mathbf{z}}$ and then transform it back to the $\mathbf{x}$-scale $\mathcal{S}_{Y|\mathbf{x}}$.

### 1.1.2 Classic Sufficient Dimension Reduction Methods

As described in Section 1.1.1, the target of SDR is to estimate a space, which is the column space of $\mathbf{B}$ with the smallest dimension $d$. There are mainly two groups of classic methods for estimating the column space of $\mathbf{B}$, first-order methods and second-order methods. First-order methods include ordinary least squares (OLS; Li and Duan, 1989), sliced inverse regression (SIR; Li, 1991), parametric inverse regression (PIR; Bura and Cook, 2001b) and kernel inverse regression (KIR; Zhu and Fang, 1996; Ferre and Yao, 2005). These methods are also referred as inverse-regression based methods, and mostly require linear conditional mean (LCM) assumption on the predictor. The LCM condition assumes that

$$\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) \text{ is linear in } \mathbf{x}. \tag{1.3}$$

The geometric implication of the LCM assumption (1.3) is that the conditional expectation $\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$ coincides with $\mathbf{P}_{\boldsymbol{\Sigma}}\mathbf{x}$, represented as $\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = \mathbf{P}_{\boldsymbol{\Sigma}}\mathbf{x}$, where $\mathbf{P}_{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}\mathbf{B}(\mathbf{B}^{\mathrm{T}}\boldsymbol{\Sigma}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}$ is called the projection matrix. When LCM (1.3) holds for all possible $\mathbf{B}$, it indicates that the predictor has an elliptical distribution (Eaton, 1986). Under condition (1.3), we will introduce OLS and SIR methods.

#### Ordinary Least Squares

In the pioneering article, Li and Duan (1989) showed that if (1.3) holds and the structural dimension $d = \dim(\mathcal{S}_{Y|\mathbf{x}}) = 1$, then $\boldsymbol{\beta}_{\mathrm{OLS}} \in \mathcal{S}_{Y|\mathbf{x}}$, where $\boldsymbol{\beta}_{\mathrm{OLS}}$ denotes the ordinary least squares estimator and $\mathrm{Cov}(\mathbf{x}, Y)$ infers a subspace

of the column space of $\boldsymbol{\beta}_{\text{OLS}}$. Without loss of generality, the OLS estimator is simply $\boldsymbol{\Sigma}^{-1}\text{E}(\mathbf{x}Y)$ when $\text{E}(\mathbf{x}) = \mathbf{0}$. A brief proof (Li and Duan, 1989) is presented as following under model (1.1) and condition (1.3) given $\text{span}(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{x}}$.

$$\begin{aligned}\text{E}(\mathbf{x}Y) &= \text{E}(\text{E}(\mathbf{x}Y \mid \mathbf{x})) = \text{E}(\mathbf{x}\text{E}(Y \mid \mathbf{x})) \\ &= \text{E}(\mathbf{x}\text{E}(Y \mid \mathbf{B}^{\text{T}}\mathbf{x})) = \text{E}(\text{E}(\mathbf{x} \mid \mathbf{B}^{\text{T}}\mathbf{x})Y) = \mathbf{P}_{\boldsymbol{\Sigma}}\text{E}(\mathbf{x}Y).\end{aligned}$$

Thus,

$$\boldsymbol{\beta}_{\text{OLS}} = \boldsymbol{\Sigma}^{-1}\text{E}(\mathbf{x}Y) = \mathbf{B}(\mathbf{B}^{\text{T}}\boldsymbol{\Sigma}\mathbf{B})^{-1}\mathbf{B}^{\text{T}}\text{E}(\mathbf{x}Y) \Rightarrow \boldsymbol{\Sigma}^{-1}\text{E}(\mathbf{x}Y) \subseteq \mathcal{S}_{Y|\mathbf{x}}.$$

In addition, Cook and Li (2002) illustrated that OLS is indeed to estimate central mean space and $\boldsymbol{\beta}_{\text{OLS}} \subseteq \mathcal{S}_{\text{E}(Y|\mathbf{x})} \subseteq \mathcal{S}_{Y|\mathbf{x}}$. Therefore, OLS is more often regarded as a dimension reduction method targeting the central mean space $\mathcal{S}_{\text{E}(Y|\mathbf{x})}$. The limitations of OLS is that it can at most recover one direction of the central space. And when the link function between response and predictors is symmetric about 0 (such as U-shaped curve) OLS will fail to recover the central space, for instance, $\boldsymbol{\beta}_{\text{OLS}} = \mathbf{0}$ when $\mathbf{x}$ follows normal.

**Sliced Inverse Regression**

The main idea for inverse regression is reversing the response and the predictors. Sliced inverse regression method coined by Li (1991) is the most representative inverse-regression SDR method. Denote $\mathbf{M}_{\text{SIR}} = \boldsymbol{\Sigma}^{-1}\text{Cov}\{\text{E}(\mathbf{x} \mid Y)\}\boldsymbol{\Sigma}^{-1}$. It can be shown that $\text{span}(\mathbf{M}_{\text{SIR}}) \subseteq \mathcal{S}_{Y|\mathbf{x}}$ under condition (1.3), thus the eigenvectors associated with the $d$ largest eigenvalues of $\mathbf{M}_{\text{SIR}}$ are used

to recover $\mathcal{S}_{Y|\mathbf{x}}$. At the population level, let $\{J_1, \ldots, J_H\}$ be a partition of the sample space of $Y$ and denote $R_h = I(Y \in J_h)$, where $p_h = \mathrm{E}(R_h)$ and $\mathbf{u}_h = \mathrm{E}(\mathbf{x}R_h)$. We have

$$\mathbf{M}_{\mathrm{SIR}} = \boldsymbol{\Sigma}^{-1}\mathrm{Cov}\{\mathrm{E}(\mathbf{x} \mid Y)\}\boldsymbol{\Sigma}^{-1} = \boldsymbol{\Sigma}^{-1}\left(\sum_{h=1}^{H} p_h^{-1}\mathbf{u}_h\mathbf{u}_h^{\mathrm{T}}\right)\boldsymbol{\Sigma}^{-1},$$

where $p_h = \mathrm{E}(R_h)$ and $\mathbf{u}_h = \mathrm{E}(\mathbf{x}R_h)$. Then $\mathrm{span}(\mathbf{M}_{\mathrm{SIR}}) \subseteq \mathcal{S}_{Y|\mathbf{x}}$. Let $\{\mathbf{x}_i, Y_i\}$ $i = 1, \ldots, n$ be an i.i.d. sample. Calculate

$$\widehat{\mathbf{M}}_{\mathrm{SIR}} = \widehat{\boldsymbol{\Sigma}}^{-1}\left(\sum_{h=1}^{H} \widehat{p_h}^{-1}\widehat{\mathbf{u}}_h\widehat{\mathbf{u}}_h^{\mathrm{T}}\right)\widehat{\boldsymbol{\Sigma}}^{-1},$$

where, for example, $\widehat{\mathbf{u}}_h = n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})I(Y_i \in J_h)$ Let $\widehat{\boldsymbol{\beta}}_1, \cdots, \widehat{\boldsymbol{\beta}}_d$ be the eigenvectors corresponding to the $d$ leading eigenvalues of $\widehat{\mathbf{M}}_{\mathrm{SIR}}$. Then $\widehat{\mathbf{B}} = (\widehat{\boldsymbol{\beta}}_1, \cdots, \widehat{\boldsymbol{\beta}}_d)$.

Unlike OLS, SIR can recover multiple directions of central space. However, it has the same limitation when the link function between response and predictors is symmetric about 0, SIR will fail for $\mathrm{E}(\mathbf{x} \mid Y) = \mathbf{0}$. The subsequent second-order methods can resolve this issue.

Second-order methods include principal Hessian direction (PHD; Li, 1992; Cook and Li, 2004a), the sliced average variance estimator (Cook and Weisberg, 1991). These methods require not only (1.3), but also constant conditional variance (CCV):

$$\mathrm{Cov}(\mathbf{x} \,|\, \mathbf{B}^{\mathrm{T}}\mathbf{x}) \text{ is a nonrandom matrix .} \tag{1.4}$$

Combined with (1.3), (1.4) implies that $\mathrm{Cov}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = \mathbf{Q}_{\boldsymbol{\Sigma}}$, where $\mathbf{Q}_{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} - \mathbf{P}_{\boldsymbol{\Sigma}}\boldsymbol{\Sigma}\mathbf{P}_{\boldsymbol{\Sigma}}^{\mathrm{T}}$. When LCM(1.3) holds for all possible $\mathbf{B}$, it indicates that the predictor has an elliptical distribution (Eaton, 1986). When both LCM (1.3) and CCV (1.4) are held for all possible $\mathbf{B}$, it infers that the predictor follows multivariate normal distribution. Under condition (1.3) and (1.4), we will introduce SAVE and PHD methods.

**Sliced Average Variance Estimator**

To overcome the limitation of SIR, Cook and Weisberg (1991) proposed sliced average variance estimator (SAVE) to recover the central space by within slice variance. Denote

$$\mathbf{M}_{\mathrm{SAVE}} = \boldsymbol{\Sigma}^{-1}\mathrm{E}\left[\{\boldsymbol{\Sigma} - \mathrm{Var}(\mathbf{x} \mid Y)\}\boldsymbol{\Sigma}^{-1}\{\boldsymbol{\Sigma} - \mathrm{Var}(\mathbf{x} \mid Y)\}\right]\boldsymbol{\Sigma}^{-1}.$$

Under the two conditions (1.3) and (1.4), they demonstrated that $\mathrm{span}(\mathbf{M}_{\mathrm{SAVE}}) \subseteq \mathcal{S}_{Y|\mathbf{x}}$, consequently, the column space of $\mathbf{M}_{\mathrm{SAVE}}$ belongs to the central space $\mathcal{S}_{Y|\mathbf{x}}$. The estimation algorithm for $\mathbf{M}_{\mathrm{SAVE}}$ is similar to the procedure for SIR.

**Principal Hessian Directions**

Principal Hessian directions (PHD; Li, 1992) is a well-known second-order method. The idea derives from the observation that the Hessian matrix $H(\mathbf{x}) = \frac{\partial^2 \mathrm{E}(Y|\mathbf{x})}{\partial \mathbf{x} \partial \mathbf{x}^{\mathrm{T}}}$ is degenerated along any directions that are orthogonal to $\mathcal{S}_{Y|\mathbf{x}}$. Denote

$$\mathbf{M}_{\mathrm{PHD}} = \boldsymbol{\Sigma}^{-1}\mathrm{E}(Y\mathbf{x}\mathbf{x}^{\mathrm{T}})\boldsymbol{\Sigma}^{-1}\mathrm{E}(Y\mathbf{x}\mathbf{x}^{\mathrm{T}})\boldsymbol{\Sigma}^{-1}.$$

Li (1992) showed that span($\mathbf{M}_{\text{PHD}}$) $\subseteq \mathcal{S}_{Y|\mathbf{x}}$ and similar to OLS. Cook and Li (2002) pointed that span($\mathbf{M}_{\text{PHD}}$) $\subseteq \mathcal{S}_{\text{E}(Y|\mathbf{x})} \subseteq \mathcal{S}_{Y|\mathbf{x}}$. PHD overcomes the limitation that OLS can only recover one direction in the central mean space. And it also can handle symmetric model cases. It can give a more comprehensive estimate but in a trade off for requiring additional assumption (1.4). As Hessian matrices are important in studying multivariate nonlinear functions, when it comes to linear function, PHD will not work since $H(\mathbf{x}) = \mathbf{0}$.

### 1.1.3 Modern Sufficient Dimension Reduction Methods

The preceding classic sufficient dimension reduction methods rely on stringent conditions on the predictors. Because these conditions are too strong for some applications and not easy to check in practice, modern literature show more interests to relax the underlying assumptions.

#### Central Solution Space

Li and Dong (2009) generated the novel construction of central solution space to circumvent the linear conditional mean assumption and reformulate the commonly used first-order (inverse-regression based) methods, such as OLS, SIR, KIR and PIR. They synthesized the estimators into a common form and focused on defined inverse regression equation, for example $\text{E}(\mathbf{x} \mid Y) = \text{E}\{\text{E}(\mathbf{x} \mid \mathbf{B}^{\text{T}}\mathbf{x}) \,|\, Y\}$ in SIR. The span of the $\mathbf{B}$ that solved in the equation is called a solution subspace. And comparably, the intersection of the all such spaces will be called the central solution space $\mathcal{S}_{\text{CSS}}$. Also it is showed that $\mathcal{S}_{\text{CSS}} \subseteq \mathcal{S}_{Y|\mathbf{x}}$. Linearity assumption (1.3) is relaxed by requiring

$\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$ to be a polynomial function of $\mathbf{B}^{\mathrm{T}}\mathbf{x}$.

$$\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = \mathrm{E}\left[\mathbf{x}G^{\mathrm{T}}(\mathbf{B}^{\mathrm{T}}\mathbf{x})\{G^{\mathrm{T}}(\mathbf{B}^{\mathrm{T}}\mathbf{x})G^{\mathrm{T}}(\mathbf{B}^{\mathrm{T}}\mathbf{x})\}^{-1}G^{\mathrm{T}}(\mathbf{B}^{\mathrm{T}}\mathbf{x})\right], \tag{1.5}$$

where $G(\mathbf{B}^{\mathrm{T}}\mathbf{x}) = (f_1(\mathbf{B}^{\mathrm{T}}\mathbf{x}), ..., f_k(\mathbf{B}^{\mathrm{T}}\mathbf{x}))^{\mathrm{T}}$ and $f_1, ..., f_k$ are functions from $\mathbb{R}^q$ to $\mathbb{R}$. Dong and Li (2010) extended the central solution space idea to second methods such as SAVE and Directional Regression.

**Semiparametrics Approach**

LCM condition is eliminated in Li and Dong (2009) and Dong and Li (2010), while the constant variance condition (1.4) is still retained. Ma and Zhu (2012b) provided an innovative and completely different approach to sufficient dimension reduction through semiparametrics. Consequently, both conditions are released. More advancedly, the approach stimulate deeper and richer class estimators, obtaining the classic dimension reduction techniques as special cases in the class.

Their approach starts from deriving the complete family of influence functions via a geometric tool, semiparametrics, in Bickel et al. (1993) and Tsiatis (2006). Then a general class of estimating equations (1.6) is derived from the influence functions, that is for any functions $\mathbf{g}$ and $a$,

$$\mathrm{E}\big([\mathbf{g}(Y, \mathbf{B}^{\mathrm{T}}\mathbf{x}) - \mathrm{E}\{\mathbf{g}(Y, \mathbf{B}^{\mathrm{T}}\mathbf{x}) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}\}] \times [a(\mathbf{x}) - \mathrm{E}\{a(\mathbf{x}) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}\}]\big) = \mathbf{0}. \tag{1.6}$$

With the double robustness property of (1.6), root-$n$ consistent (Newey, 1990) estimators for central space can be obtained through sample version of (1.6) by nonparametrically estimating both $\mathrm{E}\{\mathbf{g}(Y, \mathbf{B}^{\mathrm{T}}\mathbf{x}) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}\}$ and $\mathrm{E}\{a(\mathbf{x}) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}\}$.

Ma and Zhu (2012b) further showed that the estimator $\widehat{\mathbf{B}}$ from the sample version estimating equation

$$\sum_{i=1}^{n}[\mathbf{g}(Y_i, \widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{x}_i) - \widehat{\mathrm{E}}\{\mathbf{g}_i(Y, \widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{x}_i) \mid \widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{x}_i\}] \times [a(\mathbf{x}_i) - \widehat{\mathrm{E}}\{a(\mathbf{x}_i) \mid \widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{x}_i\}] = \mathbf{0}$$

is distributed as

$$\sqrt{n}\mathbf{A}\mathrm{vec}(\widehat{\mathbf{B}} - \mathbf{B}) \to \mathcal{N}(\mathbf{0}, \mathbf{\Sigma_B}),$$

where $\mathrm{vec}(\mathbf{M})$ denotes the vector formed by concatenating the columns of $\mathbf{M}$ and

$$\mathbf{A} = \mathrm{E}\left\{\frac{\partial \mathrm{vec}\big([\mathbf{g}(Y, \mathbf{B}^{\mathrm{T}}\mathbf{x}) - \mathrm{E}\{\mathbf{g}(Y, \mathbf{B}^{\mathrm{T}}\mathbf{x}) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}\}] \times [a(\mathbf{x}) - \mathrm{E}\{a(\mathbf{x}) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}\}]\big)}{\partial \mathrm{vecl}(\mathbf{B})^{\mathrm{T}}}\right\}$$

$$\mathbf{\Sigma_B} = \mathrm{Cov}\{\mathrm{vec}([\mathbf{g}(Y, \mathbf{B}^{\mathrm{T}}\mathbf{x}) - \mathrm{E}\{\mathbf{g}(Y, \mathbf{B}^{\mathrm{T}}\mathbf{x}) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}\}] \times [a(\mathbf{x}) - \mathrm{E}\{a(\mathbf{x}) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}\}])\}.$$

And when with the sole choice of $\mathbf{g}(Y, \mathbf{B}^{\mathrm{T}}\mathbf{x}) = Y$, (1.6) becomes a general estimating equation to obtain estimators for central mean subspace. The connection of semiparametrics approach with classic SDR methods such as SIR, SAVE, OLS and PHD are showed.

With the novel semiparametrics techniques in dimension reduction, Ma and Zhu (2012a) discovered a paradoxical phenomenon that even when the linearity condition (1.3) holds in the underlying data, applying the condition will cause efficiency loss.

## 1.2 Missing Data

Missing data is prevalent in many disciplines. Statistical analysis with missing data has been an area of considerable interest since 1980s. Many statistical issues such as data with measurement error and two-phase sampling can be viewed as missing data problems. see Little and Rubin (2002), Tsiatis (2006) and Kim and Shao (2014) for comprehensive review. In this section, we will introduce mechanisms leading to missing data, commonly used methods for missing data and existing SDR methods with missing data.

### 1.2.1 Missing Mechanisms

Whether the fact that variables are missing is related to the underlying values of the variables leads to the concern of mechanisms of missing data. Rubin (1976) and colleagues (Little and Rubin, 2002) formalized the classification system for missing mechanisms: missing completely at random, missing at random and not missing at random. The mechanisms describe relationships between measured variables and the probability of missing data.

Let the random vector $\mathbf{z} = (Y, \mathbf{x}^{\mathrm{T}})^{\mathrm{T}}, \mathbf{x} \in \mathbb{R}^p$. The general notion of missing data is denoted by $(\mathbf{z}_1, \mathbf{z}_2, \delta)$, where $\mathbf{z}_1 \in \mathbb{R}^{p_1}$ represents the component that can be missing, $\mathbf{z}_2 \in \mathbb{R}^{p_2}$ is the part that is always observable $(p_1 + p_2 = p + 1)$, and $\delta$ is the missing indicator which takes value 1 when $\mathbf{z}_2$ is observed. A conceptual description of each mechanism in following is given as following.

#### Missing Completely at Random (MCAR)

If missingness does not depend on the values of the data $\mathbf{z}$, that is, if $\delta \perp\!\!\!\perp \mathbf{z}$, then the propensity for a data point to be missing is completely at

random (MCAR). And the missing data can be thought of a random subset of the complete data. For example, if a data entry clerk randomly misses some inputs, the missingness can be considered MCAR.

### Missing at Random (MAR)

An assumption less restrictive than MCAR is that missingness only depends on the other measured components ($\mathbf{z}_1$) but not on the components that are missing ($\mathbf{z}_2$), that is

$$\delta \perp\!\!\!\perp \mathbf{z}_1 \mid \mathbf{z}_2, \tag{1.7}$$

then the mechanism is defined as missing at random (MAR). For example, if a two-level test grades second level test for the participants only when he meets the cut-off for the first level test, then the missingness in the second-level test will be MAR.

### Not Missing at Random (NMAR)

If the propensity of missing data is systematically related to the hypothetical values that are missing, the mechanism is called not missing at random (NMAR). In other words, the NMAR mechanism describes data that are missing based on the would-be values of the missing scores. For example, suppose a questionnaire only asks respondents to fill in questions related to their hobbies, then the missingness for the unfilled questions can be NMAR.

### 1.2.2 Commonly Used Missing Data Methods

Missing data problem attracted extensive research. Among the existing approaches, complete-case analysis is the most commonly used in sufficient dimension reduction with missing data. It confines attention to cases that all the variables are present, thus removing all subjects with missingness. This approach is simple to apply, however, it causes the potential loss of precision and bias when the missingness mechanism is not MCAR. There are several methods for estimating parameters when data are MAR including score propensity methods, likelihood-based methods and imputation.

#### Propensity Scoring Methods

The propensity scoring methods utilizes the propensity score, termed by Rosenbaum and Rubin (1983), which is defined as $\pi = \pi(\mathbf{z}_2) = P(\delta = 1 \mid \mathbf{z}_2)$. Note that $\mathbf{z}_2$ is always observable and $\delta = 1$ indicates that $\mathbf{z}_1$ is observed. The most representative approach involving the propensity score is the inverse probability weighted (IPW) estimators first suggested by Horvitz and Thompson (1952), and later J. M. Robins and Zhao (1995). The basic intuition is that for any randomly chosen individual $\mathbf{z} = (\mathbf{z}_1^{\mathrm{T}}, \mathbf{z}_2^{\mathrm{T}})^{\mathrm{T}}$, the probability that $\mathbf{z}$ will have complete data is $\pi(\mathbf{z}_2)$. Therefore any $\mathbf{z}$ with complete data can be thought of as representing $\pi^{-1}(\mathbf{z}_2)$ individuals at random from the population, some of which may have missing data. For a size-n random sample $(\mathbf{z}_i, \delta_i) = \{(\mathbf{z}_{1i}^{\mathrm{T}}, \mathbf{z}_{2i}^{\mathrm{T}}), \delta_i\}$, the estimator for $\boldsymbol{\mu} = \mathrm{E}(\mathbf{z})$ can be suggested as

$$\hat{\boldsymbol{\mu}} = \frac{1}{n} \sum_{i=1}^{n} \frac{\delta_i \mathbf{z}_i}{\pi_i(\mathbf{z}_2)}.$$

It can be easily showed that $\hat{\boldsymbol{\mu}}$ is an unbiased estimator for $\mathrm{E}(\mathbf{z})$ under MAR assumption. However, the IPW estimator rely on the correctly specification for the propensity score, otherwise the estimator would be biased.

Scharfstein et al. (1999) first introduced the notion of double robust estimators, resulting in another prevailing propensity score method. Double robust estimators were also studies by Lipsitz et al. (1999), and Robins et al. (2000). Bang and Robins (2005) gives an excellent overview. The method suggests the estimator for $\boldsymbol{\mu} = \mathrm{E}(\mathbf{z})$ as

$$\hat{\boldsymbol{\mu}} = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i \mathbf{z}_i}{\pi_i(\mathbf{z}_2)} + \left(1 - \frac{\delta_i \mathbf{z}_i}{\pi_i(\mathbf{z}_2)}\right)\mathrm{E}(\mathbf{z}_i \mid \mathbf{z}_2)\right\}.$$

The estimator is also referred to as an augmented inverse probability weighted (AIPW) estimator. Although it needs specification for both $\pi(\mathbf{z}_2)$ and $\mathrm{E}(\mathbf{z} \mid \mathbf{z}_2)$, it offers double robustness property. The property can be showed in a sense that the estimator is unbiased if either $\pi(\mathbf{z}_2)$ or $\mathrm{E}(\mathbf{z} \mid \mathbf{z}_2)$ is correctly specified. In later chapters, we will propose approaches partially inspired by the scoring propensity methods.

### Imputation

The motivation of imputation is to provide completed data set so that the resulting point estimates are consistent among different analysts. Since some of the data points are missing, the natural strategy is to impute a value for such missing data and then estimating the parameter as if the imputed values were true values. For the same size-n random sample $(\mathbf{z}_i, \delta_i) = \{(\mathbf{z}_{1i}^{\mathrm{T}}, \mathbf{z}_{2i}^{\mathrm{T}}), \delta_i\}$, suppose that the parameter of interest is $\boldsymbol{\mu}_g = \mathrm{E}\{g(\mathbf{z})\}$. Then the imputed

estimator for $\boldsymbol{\mu}_g$ can be computed as

$$\hat{\boldsymbol{\mu}}_g = \frac{1}{n}\sum_{i=1}^{n}\{\delta_i g(\mathbf{z}_i) + (1-\delta_i)g(\mathbf{z}_i')\},$$

where $\mathbf{z}_i'$ is generated from the conditional distribution of $\mathbf{z}_i \mid \delta_i = 0$ or the average of multiple imputed values. Also see details in Rubin (1987), Schafer (1997), and Kim and Shao (2014).

**Likelihood-based Methods**

Maximizing the likelihood of the observed data is a commonly-used method for estimating unknown parameters of a model. The same principal holds when missing data occur, while the difficulty lies in specifying the likelihood of the observed data. As no plain solution exits to find an estimate that maximize the likelihood of the observed data when missing data occur, Dempster et al. (1977) proposed EM algorithm to solve the estimates. Excellent reference include the books by Little and Rubin (2002), Schafer (1997) and Kim and Shao (2014).

## 1.3 Existing Sufficient Dimension Reduction Methods Handling Missing Data

Li and Lu (2008) brought into new blood to SDR with missingness in data. Thereafter, the missing values in predictors and in response are treated separately in dimension reduction literature.

**Inverse Probability Weighted Estimators**

In the context of missing predictors, Li and Lu (2008) combined SIR with the augmented inverses probability weighted method (Robins et al., 1994) for dimension reduction problem. Partitioning the predictor vector $\mathbf{x} \in \mathbb{R}_p$ into $(\mathbf{x}_1^{\mathrm{T}}, \mathbf{x}_2^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{x}_1 \in \mathbb{R}^{p_1}$ are observed completely and $\mathbf{x}_1 \in \mathbb{R}^{p_2}$ has missing subjects, the MAR condition (1.7) is equivalent to $\delta \perp\!\!\!\perp \mathbf{x}_2 \mid (Y, \mathbf{x}_1)$. The challenge to estimate SIR matrix $\mathbf{M}_{\mathrm{SIR}} = \boldsymbol{\Sigma}^{-1}\mathrm{Cov}\{\mathrm{E}(\mathbf{x} \mid Y)\}\boldsymbol{\Sigma}^{-1}$ under missingness is to estimate $\boldsymbol{\Sigma}$ and $\mathrm{Cov}\{\mathrm{E}(\mathbf{x} \mid Y)\}$. Specifically, one needs to estimate the moments including $\mathrm{E}(\mathbf{x}_1)$, $\mathrm{E}(\mathbf{x}_1\mathbf{x}_1^{\mathrm{T}})$, $\mathrm{E}(\mathbf{x}_1 \mid Y)$, $\mathrm{E}(\mathbf{x}_2)$, $\mathrm{E}(\mathbf{x}_2\mathbf{x}_2^{\mathrm{T}})$, $\mathrm{E}(\mathbf{x}_2\mathbf{x}_1^{\mathrm{T}})$ and $\mathrm{E}(\mathbf{x}_2 \mid Y)$. Note that the first three terms can be estimated as usual, because they involve no missing observations. However, new consistent estimators need to be derived for the last four terms involving $\mathbf{x}_2$. Although evolving inverse probability weighted method can give unbiased estimators, as the method requires a parametric model for the missing propensity $\pi = P(\delta = 1 \mid Y, \mathbf{x}_1) = \pi(Y, \mathbf{x}_1)$, they proposed the augmented inverse probability weighted estimator for its embedded robustness property to avoid the harm of misspecification of $\pi$. Denote the resulting estimators as $\widehat{\boldsymbol{\Sigma}}$ and $\widehat{\boldsymbol{\Sigma}}_{\mathrm{E}(\mathbf{x}|Y)}$, they followed convention and spectral decomposition of $\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\boldsymbol{\Sigma}}_{\mathrm{E}(\mathbf{x}|Y)}$ to recover $\mathcal{S}_{Y|\mathbf{x}}$. However, inverting $\widehat{\boldsymbol{\Sigma}}$ may be problematic when predictors have colinearity, and estimating $\boldsymbol{\Sigma}^{-1}$ can be impossible when $n \leq p$. Dong and Zhu (2012) studied the inverse probability weighted method in estimating equation framework to avoid inverting the covariance matrix.

Dong and Zhu (2013) extended the inverse probability weighted procedure based on SAVE and DR and showed that the estimators can work well under either MAR or the MCAR mechanism.

**Nonparametric and Parametric Imputation**

Zhu et al. (2012) proposed alternative approach for SDR with missing predictors in an angle of semiparametric regression. Under a slightly different MAR assumption that $\delta \perp\!\!\!\perp \mathbf{x} \mid Y$, the challenge to estimate SIR matrix becomes to estimate the conditional moments $\mathrm{E}(X_k \mid Y)$ when $X_k$ has missingness and $\mathrm{E}(X_k X_l \mid Y)$ for $1 \leq k, l \leq p$ when either $X_k$ or $X_l$ has missingness. They imputed the missing values in $\mathrm{E}(X_k \mid Y)$ with $\mathrm{E}(\delta_k X_k \mid Y)/\mathrm{E}(\delta_k \mid Y)$ and $\mathrm{E}(X_k X_l \mid Y)$ with $\mathrm{E}(\delta_k \delta_l X_k X_l \mid Y)/\mathrm{E}(\delta_k \delta_l \mid Y)$ respectively. The conditional expectations evolving $\delta$ are estimated using standard nonparametric regressions. Correspondingly, the unconditional moments can be estimated through $\mathrm{E}\{\mathrm{E}(X_k X_l \mid Y)\} = \mathrm{E}(X_k X_l)$ technique.

With the partition idea and same MAR assumption in Li and Lu (2008), Zhu et al. (2012) proposed different estimators for the moments involve $\mathbf{x}_2$ such as $\mathrm{E}(\mathbf{x}_2)$ and $\mathrm{E}(\mathbf{x}_2 \mathbf{x}_2^{\mathrm{T}})$. For example, by stating that $\mathrm{E}(\mathbf{x}_{2,k}) = \mathrm{E}\{\mathrm{E}(\mathbf{x}_{2,k} \mid \mathbf{x}_1, Y)\} = \mathrm{E}\{\mathrm{E}(\mathbf{x}_{2,k}\delta_k \mid \mathbf{x}_1, Y)/\mathrm{E}(\delta_k \mid \mathbf{x}_1, Y)\}$, they estimated the quantities $\mathrm{E}(\mathbf{x}_{2,k}\delta_k \mid \mathbf{x}_1, Y)$ and $\mathrm{E}(\delta_k \mid \mathbf{x}_1, Y)$ nonparametricly when $p_2$ is small. However, parametric imputation was applied to address the cruse of dimensionality problem when $p_2$ is fairly large.

**Fusion-Refinement Procedure**

Being the first to target missing response problem in dimension reduction, Ding and Wang (2011) introduced a novel two-stage fusion-refinement (FR) procedure on SIR. MAR (1.7) condition is expressed as $\delta \perp\!\!\!\perp Y \mid \mathbf{x}$. By claiming $\mathcal{S}_{(Y,\delta)|\mathbf{x}} = \mathcal{S}_{\delta Y|\mathbf{x}}$, at the fusion stage, they recovered the central space $\mathcal{S}_{(Y,\delta)|\mathbf{x}}$ by $\mathbf{\Gamma}$ which is the basis of $\mathcal{S}_{\delta Y|\mathbf{x}}$. Then at the refinement stage, they built

a bridge $\mathcal{S}_{Y|\tilde{\mathbf{x}}}$ from $\mathcal{S}_{\delta Y|\mathbf{x}}$ to $\mathcal{S}_{Y|\mathbf{x}}$, where $\tilde{\mathbf{x}} = \boldsymbol{\Gamma}^{\mathrm{T}}\mathbf{x}$. By claiming that if $\mathbf{B}$ is a basis of $\mathcal{S}_{Y|\tilde{\mathbf{x}}}$, then $\boldsymbol{\Gamma}\mathbf{B}$ is a basis of $\mathcal{S}_{Y|\mathbf{x}}$, the refinement stage focuses on recovering $\mathcal{S}_{Y|\tilde{\mathbf{x}}}$. The estimation of $\mathcal{S}_{\delta Y|\tilde{\mathbf{x}}}$ is through probability mass function (pmf) imputation. Note that if $\mathcal{S}_{\delta|\mathbf{x}}$ is close to $\mathcal{S}_{Y|\mathbf{x}}$, the directions of $\mathcal{S}_{(Y,\delta)|\mathbf{x}}$ can provide a good estimation of $\mathcal{S}_{Y|\mathbf{x}}$. Under this situation, one can claim that $\mathcal{S}_{(Y,\delta)|\mathbf{x}} = \mathcal{S}_{\delta Y|\mathbf{x}} = \mathcal{S}_{\delta|\mathbf{x}} = \mathcal{S}_{Y|\mathbf{x}}$. Dong and Zhu (2013) employed this argument to extend the Fusion-Refinement procedure to SAVE and DR.

The rest of the book is organized as follows. In Chapter 2, we study the validity of complete case analysis for both forward and inverse regression approaches when response is missing at random. Focusing on the seminal sliced inverse regression (SIR) estimator, we propose a simple and effective adjustment with inverse probability weighting that guarantees the validity of the inverse regressions. In Chapter 3, we proposes two new estimating equation procedures to handle missing response in SDR but avoiding the common assumptions LCM and CCV. In Chapter 4, we propose a new hybrid SDR estimator that successfully recovers the central mean space for a wide range of link functions. The order determination procedure and the marginal coordinate test are studied based on the hybrid estimator. Note that extensive numerical results are provided in each Chapter to demonstrate the desirable performances of the proposed approaches. And the technical details are collected in Appendix. We conclude the dissertation with some discussions about the future work in Chapter 5.

# CHAPTER 2

# A NOTE ON INVERSE REGRESSIONS WHEN RESPONSES ARE MISSING AT RANDOM

Approaches for inverse regressions form a class of important methods for sufficient dimension reduction. Take SIR (Li, 1991) as an example, we show that when response variable $Y$ can be missing, inverse regressions with complete case analysis, encounter problems and adjustment is necessary. Assume $\mathrm{E}(\mathbf{x}) = \mathbf{0}$ and let $\mathbf{\Sigma}$ be the covariance matrix of $\mathbf{x}$ throughout this chapter.

## 2.1 Issues for Complete Case Analysis

Let $\delta = 1(0)$ if the response variable $Y$ is observed (missing). We consider the popular missing at random (Little and Rubin, 2002) for the data

missingness mechanism:

$$Y \perp\!\!\!\perp \delta \mid \mathbf{x}. \tag{2.1}$$

Due to MAR (2.1), we may define the propensity function as $P(\delta = 1 \mid Y, \mathbf{x}) = P(\delta = 1 \mid \mathbf{x}) = \pi(\mathbf{x})$.

A naive way handling missing response is the complete case analysis, which ignores the individual observations with missing items. We first note that the conventional regression approaches, also known as the forward regression approaches, are generally valid using the complete cases analysis. Examples of forward regressions include those methods based on linear models and generalized linear models. The key reason is that the conditional distribution of $Y$ given $\mathbf{x}$, whose probability mass function or probability density function is denoted by $f(Y \mid \mathbf{x})$, is the same as the conditional distribution of $Y$ given $\mathbf{x}$ and $\delta = 1$, denoted by $f(Y \mid \mathbf{x}, \delta = 1)$. To see that,

$$f(Y \mid \mathbf{x}, \delta = 1) = \frac{f(Y, \mathbf{x}, \delta = 1)}{f(\mathbf{x}, \delta = 1)} = \frac{P(\delta = 1 \mid Y, \mathbf{x}) f(Y, \mathbf{x})}{P(\delta = 1 \mid \mathbf{x}) f(\mathbf{x})} = f(Y \mid \mathbf{x}), \tag{2.2}$$

where the last equation is due to (2.1). The invariance of the conditional distributions in (2.2) implies that complete case analysis is valid with responses missing at random for forward regression approaches, because they essentially rely on the information from the conditional distribution of $Y$ given $\mathbf{x}$. In SDR, take the complete case based OLS estimator under linear model $Y = \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x} + \epsilon$ as a toy example. In this case, with $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, $\{\mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid \delta = 1)\}^{-1}\mathrm{E}(\mathbf{x}Y \mid \delta = 1)$ is a complete case estimator to recover $\mathcal{S}_{Y|\mathbf{x}}$. One can show that $\{\mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid \delta = 1)\}^{-1}\mathrm{E}(\mathbf{x}Y \mid \delta = 1) = \{\mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid \delta = 1)\}^{-1}(\mathbf{x}\mathbf{x}^{\mathrm{T}}\boldsymbol{\beta} \mid \delta = 1) = \boldsymbol{\beta}$. It follows that

$\{\mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid \delta = 1)\}^{-1}\mathrm{E}(\mathbf{x}Y \mid \delta = 1) \subseteq \mathcal{S}_{Y|\mathbf{x}}$.

However, ignoring observations with missing responses will generally lead to biased inverse regressions. With complete case analysis, the conditional distribution of $\mathbf{x}$ given $Y$ becomes

$$f(\mathbf{x} \mid Y, \delta = 1) = \frac{f(Y, \mathbf{x}, \delta = 1)}{f(Y, \delta = 1)} = \frac{P(\delta = 1 \mid Y, \mathbf{x}) f(Y, \mathbf{x})}{P(\delta = 1 \mid Y) f(Y)} = f(\mathbf{x} \mid Y) w(\mathbf{x}, Y), \tag{2.3}$$

with $w(\mathbf{x}, Y) = P(\delta = 1 \mid Y, \mathbf{x})/P(\delta = 1 \mid Y)$. In (2.3), $f(\mathbf{x} \mid Y) \neq f(\mathbf{x} \mid Y, \delta = 1)$ unless $w(\mathbf{x}, Y) = 1$. Therefore, the complete case based inverse regressions will possibly lead to bias. SIR, kernel inverse regression (KIR; Zhu and Fang, 1996), and SAVE all belong to the inverse regression family. Take the complete case based SIR as an example, we have

$$\begin{aligned} \mathrm{E}(\mathbf{x} \mid Y, \delta = 1) &= \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y, \mathbf{B}^{\mathrm{T}}\mathbf{x}, \delta = 1) \mid Y, \delta = 1\} \\ &= \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y, \mathbf{B}^{\mathrm{T}}\mathbf{x}, \delta = 1) \mid Y, \delta = 1\} \\ &= \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}, \delta = 1) \mid Y, \delta = 1\}. \end{aligned} \tag{2.4}$$

The derivation in (2.4) cannot step further unless the LCM condition (1.3) is true under the complete cases, namely $\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}, \delta = 1) = \mathbf{P}_{\mathbf{\Sigma}}\mathbf{x}$. Otherwise, $\mathrm{E}(\mathbf{x} \mid Y, \delta = 1) \neq \mathbf{P}_{\mathbf{\Sigma}}\mathrm{E}(\mathbf{x} \mid Y, \delta = 1)$ and $\mathrm{E}(\mathbf{x} \mid Y, \delta = 1)$ is biased to recover central space.

We simulate two toy examples to show the validity and impact of deleting missing observations with OLS and SIR, respectively. Suppose we generate 200 observations from model $Y = X_1 + \varepsilon$ where $\mathbf{x} = (X_1, X_2)^{\mathrm{T}}$ follows bivariate standard normal, and $\varepsilon \sim N(0, 0.2^2)$ is independent of $\mathbf{x}$. Note that in this

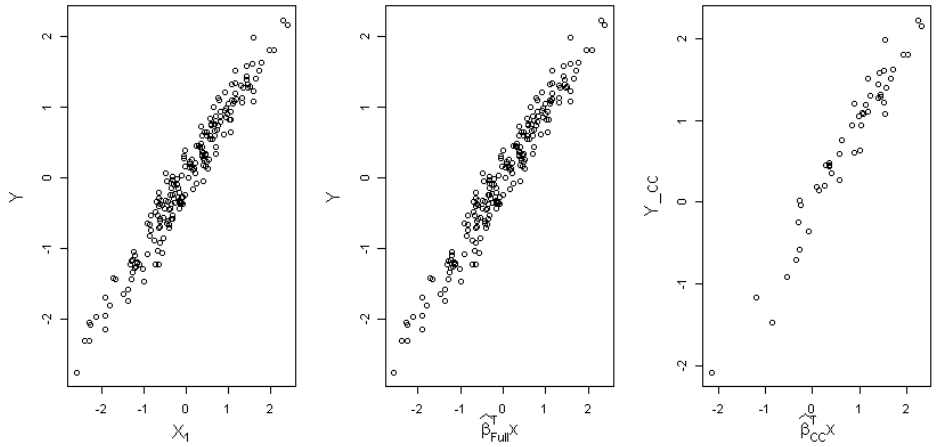
Table 2.1: Comparison of the estimators $\widehat{\boldsymbol{\beta}}_{Full}$ and $\widehat{\boldsymbol{\beta}}_{CC}$ for both OLS and SIR

| | OLS | | SIR | |
|---|---|---|---|---|
| $\boldsymbol{\beta}$ | $\widehat{\boldsymbol{\beta}}_{Full}$ | $\widehat{\boldsymbol{\beta}}_{CC}$ | $\widehat{\boldsymbol{\beta}}_{Full}$ | $\widehat{\boldsymbol{\beta}}_{CC}$ |
| 1.0000 | 0.9931 | 0.9687 | 0.9999 | 0.8076 |
| 0.0000 | 0.0068 | 0.0312 | 0.0238 | 0.5896 |

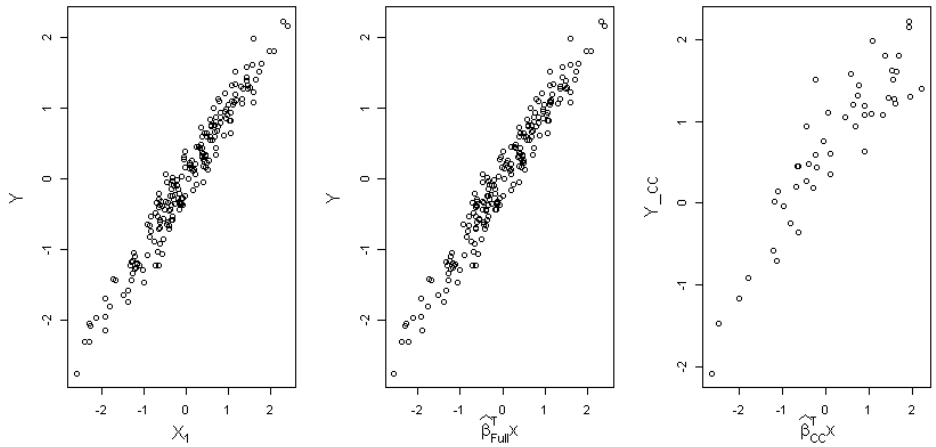
model $\boldsymbol{\beta} = (1, 0)$. By generating missingness with

$$P(\delta = 1 \mid \mathbf{x}) = \frac{\exp(-6 + (X_1 + 1)^2 + (X_2 - 1)^2)}{1 + \exp(-6 + (X_1 + 1)^2 + (X_2 - 1)^2)},$$

we have 49 complete cases. Denote $\widehat{\boldsymbol{\beta}}_{Full}$ and $\widehat{\boldsymbol{\beta}}_{CC}$ as the sample estimators for $\boldsymbol{\beta}$ using full data and complete data, respectively. The comparison of the OLS estimators with true $\boldsymbol{\beta}$ in Table 2.1 show the validity of the complete case analysis for OLS under the linear model. While $\widehat{\boldsymbol{\beta}}_{CC} = (0.8076, 0.5896)^{\mathrm{T}}$ is clearly a misleading direction and demonstrate the invalidity of the complete-case analysis for SIR. Figure 2.1a, Figure 2.1b provide the relationship between Y and $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$, $\widehat{\boldsymbol{\beta}}_{Full}^{\mathrm{T}}\mathbf{x}$, and $\widehat{\boldsymbol{\beta}}_{CC}^{\mathrm{T}}\mathbf{x}$ for OLS and SIR, respectively. It is expected that the third panel of Figure 2.1a shows obvious linear relationship between complete case $Y$ and $\widehat{\boldsymbol{\beta}}_{CC}^{\mathrm{T}}\mathbf{x}$. It behaves almost identical to $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ and $\widehat{\boldsymbol{\beta}}_{Full}^{\mathrm{T}}\mathbf{x}$ but with sparse observations. On the other hand, the result in the third panel of Figure 2.1b is much lousier due to the inaccurate estimator for $\widehat{\boldsymbol{\beta}}_{CC}$ with complete case based SIR.




(a)




(b)

Figure 2.1: Scatterplots of $Y$ versus $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$, $\widehat{\boldsymbol{\beta}}_{Full}^{\mathrm{T}}\mathbf{x}$, $\widehat{\boldsymbol{\beta}}_{CC}^{\mathrm{T}}\mathbf{x}$. (a). OLS estimators. (b). SIR estimators.

## 2.2 Adjustments for Inverse Regressions

Fortunately, adjustment can be developed for missing response in inverse regressions. We intend to develop an asymptotically unbiased estimator for $\mathrm{Cov}\{\mathrm{E}(\mathbf{x} \mid Y)\}$, adjusting for possibly missing $Y$. We illustrate the adjustment using SIR.

Let $\{J_1, \ldots, J_H\}$ be a measurable partition of the sample space of $Y$. Let $R_h = I(Y \in J_h)$ be the indicator function of $Y$ belonging to the $h$th slice. For $h = 1, \ldots, H$, let $p_h = \mathrm{E}\{\delta R_h/\pi(\mathbf{x})\}$, $\mathbf{u}_h = \mathrm{E}\{\delta \mathbf{x} R_h/\pi(\mathbf{x})\}$. The inverse probability weighting adjustment in $\mathbf{u}_h$ and $p_h$ is essential to show that $\mathrm{E}(\mathbf{x} \mid Y \in J_h) = \mathbf{u}_h/p_h$ under missing at random scheme (2.1). It follows that $\mathbf{u}_h = \mathrm{E}\{\mathrm{E}(\delta \mid \mathbf{x})\mathrm{E}(\mathbf{x}R_h \mid \mathbf{x})/\pi(\mathbf{x})\} = \mathrm{E}(\mathbf{x}R_h)$ and $p_h = \mathrm{E}\{\mathrm{E}(\delta \mid \mathbf{x})\mathrm{E}(R_h \mid \mathbf{x})/\pi(\mathbf{x})\} = \mathrm{E}\{\mathrm{E}(R_h \mid \mathbf{x})\} = \mathrm{E}(R_h)$. We refer to the adjusted approach as IPWSIR for the following. Define the IPWSIR kernel matrix as $\mathbf{M} = \boldsymbol{\Sigma}^{-1}\mathrm{Cov}\{\mathrm{E}(\mathbf{x} \mid Y)\}\boldsymbol{\Sigma}^{-1} = \sum_{h=1}^{H} \boldsymbol{\Sigma}^{-1} p_h^{-1} \mathbf{u}_h \mathbf{u}_h^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}$. We have the following result parallel to classic SIR.

**Theorem 2.1.** *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{x}) = \boldsymbol{\Sigma}$ *and all the moments involved exist. Under the LCM condition (1.3), we have span*$(\mathbf{M}) \subseteq \mathcal{S}_{Y|\mathbf{x}}$.

The result states that the column space of $\mathbf{M}$ can be used to recover the central space. The proof follows directly from the well-known property of kernel matrix of SIR thus is omitted.

Denote $\{\mathbf{x}_i, Y_i, \delta_i = 1\}_{i=1}^{n}$ be independent and identical copies of the data, and $Y_i$ is missing if $\delta_i = 1$. Let the kernel function $K_b(z) = b^{-1}K(z/b)$ for some symmetric probability density function $K(\cdot)$ and $b$ is a bandwidth which

can be estimated by leave-one-out cross-validation. Calculate

$$\widehat{\mathbf{M}} = \widehat{\mathbf{\Sigma}}^{-1} \left( \sum_{h=1}^{H} \widehat{p}_h^{-1} \widehat{\mathbf{u}}_h \widehat{\mathbf{u}}_h^{\mathrm{T}} \right) \widehat{\mathbf{\Sigma}}^{-1}, \tag{2.5}$$

where $\widehat{\mathbf{u}}_h = \frac{1}{n} \sum_{i=1}^{n} \delta_i R_{hi} (\mathbf{x}_i - \bar{\mathbf{x}}) / \widehat{\pi}(\mathbf{x}_i)$, $\widehat{p}_h = \frac{1}{n} \sum_{i=1}^{n} \delta_i R_{hi} / \widehat{\pi}(\mathbf{x}_i)$, and the sample estimator of $\pi(\mathbf{x}_i)$, $\widehat{\pi}(\mathbf{x}_i)$ can be estimated parametrically or nonparametrically. We have $\widehat{\mathbf{B}} = (\widehat{\boldsymbol{\beta}}_1, \cdots, \widehat{\boldsymbol{\beta}}_d)$, where $\widehat{\boldsymbol{\beta}}_1, \cdots, \widehat{\boldsymbol{\beta}}_d$ are the eigenvectors corresponding to the $d$ leading eigenvalues of $\widehat{\mathbf{M}}$.

We study the asymptotic property of $\widehat{\mathbf{M}}$ while estimating $\pi(\mathbf{x}_i)$ nonparametrically as

$$\widehat{\pi}(\mathbf{x}_i) = \frac{\sum_{j=1}^{n} K_b(\mathbf{x}_i - \mathbf{x}_j) \delta_j}{\sum_{j=1}^{n} K_b(\mathbf{x}_i - \mathbf{x}_j)}. \tag{2.6}$$

Let $\mathbf{a}_h = p_h^{-1/2} \mathbf{\Sigma}^{-1} \mathbf{u}_h$, $\mathbf{A} = (\mathbf{a}_1^{\mathrm{T}}, \ldots, \mathbf{a}_H^{\mathrm{T}})^{\mathrm{T}}$, and the corresponding sample estimators $\widehat{\mathbf{a}}_h = \widehat{p}_h^{-1/2} \widehat{\mathbf{\Sigma}}^{-1} \widehat{\mathbf{u}}_h$ and $\widehat{\mathbf{A}} = (\widehat{\mathbf{a}}_1^{\mathrm{T}}, \ldots, \widehat{\mathbf{a}}_H^{\mathrm{T}})^{\mathrm{T}}$. Then $\mathbf{M} = \mathbf{A}^{\mathrm{T}} \mathbf{A}$ and $\widehat{\mathbf{M}} = \widehat{\mathbf{A}}^{\mathrm{T}} \widehat{\mathbf{A}}$. For $h = 1, \ldots, H$, we have $\widehat{\mathbf{a}}_h - \mathbf{a}_h = (p_h^{-1/2})^* \mathbf{\Sigma}^{-1} \mathbf{u}_h + p_h^{-1/2} (\mathbf{\Sigma}^{-1})^* \mathbf{u}_h + p_h^{-1/2} \mathbf{\Sigma}^{-1} \mathbf{u}_h^* + o_p(n^{-1/2}) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\ell}_{h,i} + o_p(n^{-1/2})$. Then $\widehat{\mathbf{A}} - \mathbf{A} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{L}_i + o_p(n^{-1/2})$, where $\mathbf{L}_i = (\boldsymbol{\ell}_{1,i}^{\mathrm{T}}, \ldots, \boldsymbol{\ell}_{H,i}^{\mathrm{T}})^{\mathrm{T}}$. The definition for $(p_h^{-1/2})^*$, $(\mathbf{\Sigma}^{-1})^*$, $\mathbf{u}_h^*$ and $\boldsymbol{\ell}_{h,i}$ are illustrated in Appendix A. Note that $\mathrm{E}(\boldsymbol{\ell}_{h,i}) = \mathbf{0}$ and $\frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\ell}_{hi} = O_p(n^{-1/2})$. We have the next result shows the asymptotic distribution of $\mathrm{vec}(\widehat{\mathbf{A}})$, where $\mathrm{vec}(\mathbf{A})$ means the concatenation of the columns of $\mathbf{A}$.

**Theorem 2.2.** *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{x}) = \mathbf{\Sigma}$ *and all the moments involved exist. Suppose the LCM condition (1.3) holds. Let* $\mathbf{\Gamma} = \mathrm{E}\{\mathrm{vec}(\mathbf{L}) \mathrm{vec}^{T}(\mathbf{L})\}$.

*Then*

$$\sqrt{n}\left(\mathrm{vec}(\widehat{\mathbf{A}}) - \mathrm{vec}(\mathbf{A})\right) \to \mathcal{N}(\mathbf{0}, \mathbf{\Gamma}),$$

*where "$\to$" means converge in distribution.*

The technical details for Theorem 2.2 allow us to further show how the nonparametric estimation of $\pi(\mathbf{x})$ in (2.6) impacts the efficiency. To illustrate, we study sample estimators when true propensity score $\pi(\mathbf{x})$ is applied. Denote $\widetilde{\mathbf{u}}_h = \frac{1}{n}\sum_{i=1}^{n} \delta_i R_{hi}(\mathbf{x}_i - \bar{\mathbf{x}})/\pi(\mathbf{x}_i)$, $\widetilde{p}_h = \frac{1}{n}\sum_{i=1}^{n} \delta_i R_{hi}/\pi(\mathbf{x}_i)$, $\widetilde{\mathbf{a}}_h = \widetilde{p}_h^{-1/2}\widehat{\mathbf{\Sigma}}^{-1}\widetilde{\mathbf{u}}_h$, and $\widetilde{\mathbf{A}} = (\widetilde{\mathbf{a}}_1^{\mathrm{T}}, \ldots, \widetilde{\mathbf{a}}_H^{\mathrm{T}})^{\mathrm{T}}$. We conclude this section with the following proposition.

**Proposition 2.1.** *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, $\mathrm{Cov}(\mathbf{x}) = \mathbf{\Sigma}$ *and all the moments involved exist. Suppose the LCM condition (1.3) holds. Then* $\sqrt{n}\left(\mathrm{vec}(\widetilde{\mathbf{A}}) - \mathrm{vec}(\mathbf{A})\right) \to \mathcal{N}(\mathbf{0}, \mathbf{\Omega})$, *where "$\to$" means converge in distribution. In addition,* $(\mathbf{\Omega} - \mathbf{\Gamma})$ *is positive definite.*

Proposition 2.1 demonstrates a heuristic result that weighting by the inverse of a nonparametric estimate (2.6) of the propensity score $\widehat{\pi}(\mathbf{x})$, rather than the true propensity score $\pi(\mathbf{x})$, leads to a more efficient estimator. Its proof is included in Appendix A. This finding is not surprising as similar findings have been discussed in Hirano et al. (2003) when the propensity score is estimated noparametrically. In addition, Rosenbaum and Rubin (1983), Rubin and Thomas (1996), and J. M. Robins and Zhao (1995) show that using parametric estimates of the propensity score, rather than the true propensity score, can avoid some efficiency losses.

## 2.3 Marginal Coordinate Test

Cook (2004) introduced the SIR-based marginal coordinate hypothesis for model-free variable selection. Parallel to the SIR-based marginal coordinate tests studied in Cook (2004), we construct the IPWSIR-based marginal coordinate test. Model-free variable selection aims to select a subset of the predictors without assuming the functional form between $Y$ and $\mathbf{x}$, such that the selected predictors are sufficient to predict $F(Y \mid \mathbf{x})$, the conditional distribution of $Y$ given $\mathbf{x}$. Let $\mathcal{I} = \{1, \cdots, p\}$ be the full index set. To test whether the $k$th predictor is active or not for $k = 1, \cdots, p$, one can consider the hypothesis

$$\mathrm{H}_0^{[k]} : k \in \mathcal{A}^c \text{ versus } \mathrm{H}_a^{[k]} : k \in \mathcal{A}, \tag{2.7}$$

where the active set $\mathcal{A}$ and the the inactive set $\mathcal{A}^c$ are defined as

$$\mathcal{A} = \{k \in \mathcal{I} : F(Y \mid \mathbf{x}) \text{ functionally depends on } X_k\} \text{ and}$$
$$\mathcal{A}^c = \{k \in \mathcal{I} : F(Y \mid \mathbf{x}) \text{ does not functionally depend on } X_k\}.$$

Let $\mathbf{x}_{\mathcal{A}} = \{X_k : k \in \mathcal{A}\}$ denote the vector that contains all the active predictors. Then we have $Y \perp\!\!\!\perp \mathbf{x} \mid \mathbf{x}_{\mathcal{A}}$, where "$\perp\!\!\!\perp$" means independency. It is closely related to the concept of sufficient dimension reduction. Let $\mathbf{e}_k \in \mathbb{R}^p$, for $k = 1, \ldots, p$, where the $k$th element of $\mathbf{e}_k$ is 1 and all other elements are zero. For $\mathbf{B}$ being the basis for central space $\mathcal{S}_{Y|\mathbf{x}}$, we have $k \in \mathcal{A}^c$ if and only if $\mathbf{e}_k^{\mathrm{T}}\mathbf{B} = \mathbf{0}$. Thus testing hypothesis (2.7) is equivalent to testing

$$\mathrm{H}_0^{[k]} : \mathbf{e}_k^{\mathrm{T}}\mathbf{B} = \mathbf{0} \text{ versus } \mathrm{H}_a^{[k]} : \mathbf{e}_k^{\mathrm{T}}\mathbf{B} \neq \mathbf{0}. \tag{2.8}$$

Here $\mathbf{B} \in \mathbb{R}^{p\times d}$ can be replaced with any matrix that has the same column space as $\mathbf{B}$.

In Theorem 2.1, we have $\text{span}(\mathbf{M}) \subseteq \mathcal{S}_{Y|\mathbf{x}}$. With the additional coverage condition, $\text{span}\{\boldsymbol{\Sigma}^{-1}\mathbf{u}_h : h = 1,\ldots,H\} = \mathcal{S}_{Y|\mathbf{x}}$, we have $\text{span}(\mathbf{M}) = \mathcal{S}_{Y|\mathbf{x}}$. From the discussions following (2.8), we know $\mathbf{e}_k^{\mathrm{T}}\mathbf{M}\mathbf{e}_k = \mathbf{0}$ for $k \in \mathcal{A}^c$. Given an i.i.d. sample $\{\mathbf{x}_i, Y_i, \delta_i = 1\}_{i=1}^n$, we can calculate the sample estimator $\widehat{\mathbf{M}}$ as (2.5). The IPWSIR-based test statistic for (2.7) is then $n$ times the sample estimator of $\mathbf{e}_k^{\mathrm{T}}\mathbf{M}\mathbf{e}_k$, written as

$$T_k = n\sum_{h=1}^{H} \mathbf{e}_k^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{p_h}^{-1}\widehat{\mathbf{u}}_h\widehat{\mathbf{u}}_h^{\mathrm{T}}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{e}_k. \tag{2.9}$$

We have $T_k$ has an asymptotic distribution that is the sum of weighted $\chi^2(1)$ under $\mathrm{H}_0^{[k]}$ showed in Theorem 4.4.

**Theorem 2.3.** *Suppose* $\mathrm{E}(\mathbf{x}) = \mathbf{0}$*, and all the moments involved exist. Furthermore, suppose the LCM condition (1.3) holds and span*$\{\boldsymbol{\Sigma}^{-1}\mathbf{u}_h : h = 1,\ldots,H\} = \mathcal{S}_{Y|\mathbf{x}}$*. Then under* $\mathrm{H}_0^{[k]} : \mathbf{e}_k^T\mathbf{B} = \mathbf{0}$*,*

$$T_k \to \sum_{h=1}^{H}\omega_{k,h}\chi_h^2(1),$$

*where "*$\to$*" means converge in distribution,* $\omega_{k,1},\ldots,\omega_{k,H}$ *are the eigenvalues of* $\mathrm{Cov}(\mathbf{G}_k)$*, and* $\chi_h^2(1)$*'s are i.i.d.* $\chi^2(1)$ *random variables.*

The proof is demonstrated in Appendix A. Note that in practice, the unknown weights $\omega_{k,h}$'s can be replaced with their sample estimators.

## 2.4 Simulation

### 2.4.1 Accuracy

First we used synthetic data to demonstrate the effectiveness of the proposed inverse probability weighted SIR. Consider the models

$$\mathrm{I}: Y = \frac{\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}}{0.5 + (\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x} + 1.5)^2} + .2\varepsilon,$$
$$\mathrm{II}: Y = (\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x} + \boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x} + 3) + .2(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x})\varepsilon.$$

where $\varepsilon \sim \mathcal{N}(0,1)$. We generate $\mathbf{x} = (X_1, ..., X_{10})^{\mathrm{T}}$ from multivariate normal with mean zero and covariance matrix $(\sigma_{ij})_{p\times p}$ where $\sigma_{ij} = 0.5^{|i-j|}$. Let $\mathbf{B} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$ and the structural dimension $d = 2$, where $\boldsymbol{\beta}_1 = (.5, -.5, 0, 0, 0, 0, 0, 0, 0, 0)^{\mathrm{T}}$, and $\boldsymbol{\beta}_2 = (0, 0, 0, 0, 0, 0, 0, 0, .5, -.5)^{\mathrm{T}}$. Considering response $Y$ is missing, we define the MAR missingness schemes as

$$\pi(\mathbf{x}) = P(\delta = 1 \mid \mathbf{x}) = \frac{\exp(c_0 + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})}{1 + \exp(c_0 + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})}.$$

where $\boldsymbol{\alpha} = \boldsymbol{\beta}_1/||\boldsymbol{\beta}_1|| + (0, 0, 0, 0, 0, 0, 0, 0, 0, 1)^{\mathrm{T}}$. By letting $c_0 = -1, 0$ and 1, we get the corresponding missing proportion approximately as 30%, 50% and 70%. Each experiment is repeated 200 times with sample size $n = 400$. Slice the response $Y$ into $H = 4$ slices. Let $\widehat{\mathbf{B}}$ be the estimate for $\mathbf{B}$, the basis matrix of $\mathcal{S}_{Y|\mathbf{x}}$, and $\widehat{\mathbf{P}} = \widehat{\mathbf{B}}(\widehat{\mathbf{B}}^{\mathrm{T}}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^{\mathrm{T}}$ is the estimator for projection matrix $\mathbf{P}$. To evaluate the estimation accuracy, we display the boxplots of the Euclidean distance between $\widehat{\mathbf{B}}$ and $\mathbf{B}$, defined as $||\widehat{\mathbf{P}} - \mathbf{P}||^2$. Another popular criterion is also reported, that is the trace correlation coefficient, defined by $R^2(d) = \text{trace}(\widehat{\mathbf{P}}\mathbf{P}/d)$. A smaller distance and a correlation coefficient close to

1 indicate better performance. Denote Full for analysis with full data, CC for analysis with complete case, and IPW for analysis with IPW adjustment.

Based on 200 repetitions, we provide the boxplots of the Euclidean distances in Figure 2.2 and 2.3 for the two models, respectively. We see that for both models, although the performance of IPW estimator worsens with the increasing of missing proportion, it is consistently better than the complete case estimator. This is as expected because complete case estimator can be biased with inverse regression. Both CC and IPW estimators have worse performances than the oracle Full. We also report in Table 2.2 the mean trace correlation coefficient and its standard deviation. Note that while smaller Euclidean distance in Figure 2.2 and 2.3 means better estimation, larger value in Table 2.2 corresponds to more accurate estimator. The results confirm the findings in Figure 2.2 and 2.3.

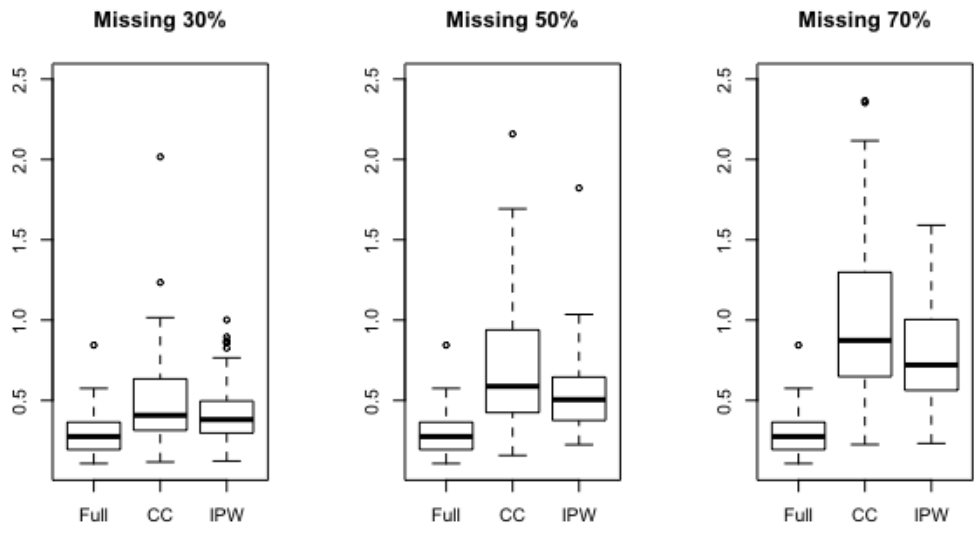



Figure 2.2: Boxplots of Euclidean distances for Model I with different missing proportions.

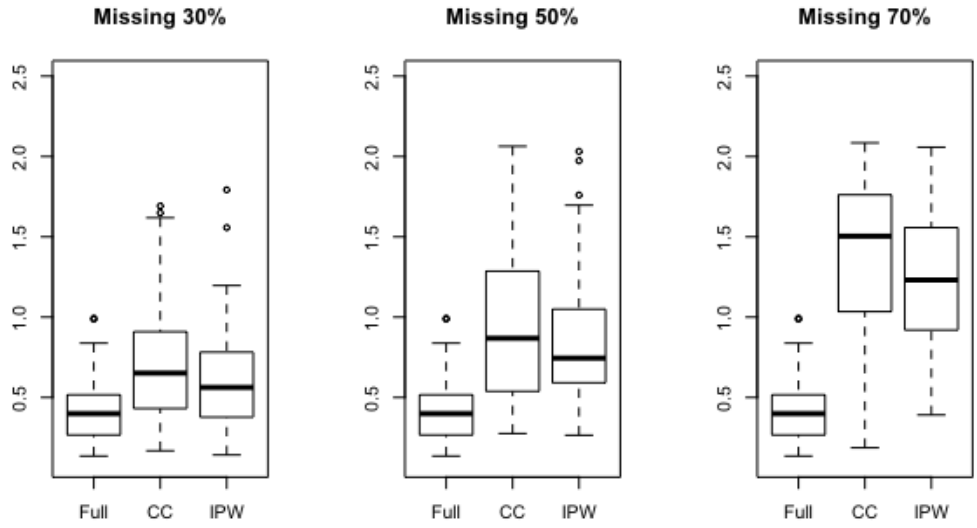



Figure 2.3: Boxplots of Euclidean distances for Model II with different missing proportions.

Table 2.2: Trace correlation Coefficient $R^2(d)$ for both models with different missing proportions

| Model | Missing | | Full | CC | IPW |
|---|---|---|---|---|---|
| I | 30% | ave | 0.927 | 0.877 | 0.896 |
| | | std | 0.031 | 0.067 | 0.046 |
| | 50% | ave | 0.927 | 0.827 | 0.863 |
| | | std | 0.031 | 0.095 | 0.061 |
| | 70% | ave | 0.927 | 0.750 | 0.799 |
| | | std | 0.031 | 0.124 | 0.076 |
| II | 30% | ave | 0.896 | 0.823 | 0.849 |
| | | std | 0.046 | 0.090 | 0.071 |
| | 50% | ave | 0.896 | 0.764 | 0.791 |
| | | std | 0.046 | 0.123 | 0.096 |
| | 70% | ave | 0.896 | 0.656 | 0.687 |
| | | std | 0.046 | 0.123 | 0.101 |

### 2.4.2 Marginal Coordinate Test

In addition, we use the same simulation setting to demonstrate the effectiveness of the marginal coordinate test via IPWSIR. We first approximate the asymptotic distribution of $T_k$ under $\mathrm{H}_0^{[k]} : \mathbf{e}_k^{\mathrm{T}}\mathbf{B} = \mathbf{0}$ as $\sum_{h=1}^{H}\omega_{k,h}\chi_h^2(1)$, where $\omega_{k,h}$'s are the eigenvalues of $\mathrm{Cov}(\mathbf{G}_k)$ for $h = 1, \ldots, H$. Let $\mathbf{W}_k = (\omega_{k,1}, \omega_{k,2}, \ldots, \omega_{k,H})^{\mathrm{T}}$, and $\mathbf{C}$ be an $m \times h$-dimensional matrix of i.i.d. $\chi^2(1)$ realizations. Then $\mathbf{C}\mathbf{W}_k$ is an $m$-dimensional vector of i.i.d. realizations of $\sum_{h=1}^{H}\omega_{k,h}\chi_h^2(1)$. The proportion of these $m$ realizations larger than $T_k$ is the approximated p-value for testing $\mathrm{H}_0^{[k]}$. We reject $\mathrm{H}_0^{[k]}$ if the approximated p-value is smaller than the nominal level $\alpha$. We set $m = 1000$ and test $\alpha = 0.01$, 0.05 and 0.10 respectively in all settings.

In Table 2.3 and 2.4, we report the frequencies that $\mathrm{H}_0^{[k]}$ is rejected for each predictor $X_k$ for Model I and II. Note that for the two models, the predictor dimension at $p = 10$, and the active predictors are $X_1$, $X_2$, $X_9$, and $X_{10}$. For both models, we consider sample sizes $n = 400$ and 800, and missing proportions varies with 30%, 50% and 70%. Note that the frequencies for predictors in the active set $\mathcal{A}$ are the estimated powers, and we want them to be close to 1. The boldfaced entries in Table 2.3 and 2.4 correspond to the estimated powers. On the other hand, the frequencies for predictors belonging to the inactive set $\mathcal{A}^c$ correspond to the estimated nominal levels, and it is ideal to have them close to the tested nominal levels.

For both models, the powers improve as sample size increases within each missing proportion. And not surprisingly, the performance deteriorates with the increasing of missing proportions. When $n = 800$, most of the powers for $X_1$ and $X_2$ become 1 even with 70% missing response. Interestingly, the pow-

Table 2.3: Marginal coordinate tests for Model I. Based on 1000 repetitions, frequencies of rejecting $H_0^{[k]}$ with different nominal tests and missing proportions are reported.

| Missing | n | $\alpha$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 30% | 400 | 0.01 | **1.00** | **1.00** | 0.02 | 0.01 | 0.01 | 0.01 | 0.03 | 0.05 | **0.70** | **0.77** |
| | | 0.05 | **1.00** | **1.00** | 0.08 | 0.03 | 0.07 | 0.07 | 0.02 | 0.07 | **0.88** | **0.88** |
| | | 0.10 | **1.00** | **1.00** | 0.13 | 0.09 | 0.08 | 0.12 | 0.06 | 0.12 | **0.94** | **0.94** |
| | 800 | 0.01 | **1.00** | **1.00** | 0.03 | 0.02 | 0.01 | 0.02 | 0.03 | 0.04 | **0.98** | **0.99** |
| | | 0.05 | **1.00** | **1.00** | 0.05 | 0.03 | 0.02 | 0.03 | 0.05 | 0.09 | **0.99** | **1.00** |
| | | 0.10 | **1.00** | **1.00** | 0.11 | 0.10 | 0.09 | 0.07 | 0.11 | 0.12 | **1.00** | **1.00** |
| 50% | 400 | 0.01 | **0.99** | **0.97** | 0.01 | 0.01 | 0.01 | 0.03 | 0.04 | 0.05 | **0.43** | **0.54** |
| | | 0.05 | **1.00** | **1.00** | 0.06 | 0.08 | 0.09 | 0.08 | 0.06 | 0.07 | **0.65** | **0.77** |
| | | 0.10 | **1.00** | **1.00** | 0.09 | 0.12 | 0.14 | 0.14 | 0.11 | 0.14 | **0.77** | **0.83** |
| | 800 | 0.01 | **1.00** | **1.00** | 0.01 | 0.01 | 0.01 | 0.03 | 0.04 | 0.05 | **0.86** | **0.85** |
| | | 0.05 | **1.00** | **1.00** | 0.03 | 0.05 | 0.03 | 0.04 | 0.06 | 0.09 | **0.97** | **0.98** |
| | | 0.10 | **1.00** | **1.00** | 0.04 | 0.13 | 0.11 | 0.13 | 0.12 | 0.13 | **0.97** | **0.98** |
| 70% | 400 | 0.01 | **0.42** | **0.41** | 0.01 | 0.01 | 0.00 | 0.03 | 0.03 | 0.07 | **0.20** | **0.49** |
| | | 0.05 | **0.83** | **0.74** | 0.06 | 0.12 | 0.12 | 0.11 | 0.09 | 0.19 | **0.44** | **0.76** |
| | | 0.10 | **0.97** | **0.94** | 0.14 | 0.17 | 0.20 | 0.19 | 0.21 | 0.26 | **0.56** | **0.84** |
| | 800 | 0.01 | **0.95** | **0.91** | 0.00 | 0.00 | 0.01 | 0.03 | 0.01 | 0.00 | **0.45** | **0.55** |
| | | 0.05 | **1.00** | **1.00** | 0.02 | 0.03 | 0.06 | 0.05 | 0.03 | 0.07 | **0.71** | **0.75** |
| | | 0.10 | **1.00** | **1.00** | 0.08 | 0.09 | 0.07 | 0.13 | 0.10 | 0.19 | **0.83** | **0.85** |

ers for $X_9$ and $X_{10}$ are relatively lower than the powers for $X_1$ and $X_2$. This phenomenon may be explained by the numerical findings in Li (1991) that the performance for SIR to estimate $\boldsymbol{\beta}_2$ may not be ideal. In their study, $\widehat{\boldsymbol{\beta}}_2$ performed worse than $\widehat{\boldsymbol{\beta}}_1$ in terms of accuracy and $\widehat{\boldsymbol{\beta}}_2$ is more sensitive to the noise level for both models. The imperfect estimation of $\boldsymbol{\beta}_2$ may negatively impact the selection of $X_9$ and $X_{10}$. In addition, when missing proportion is small, the estimated nominal levels are never too far away from the true nominal level with both sample sizes. With the increasing of missing proportions, larger sample size is required to show the desired estimated nominal levels.

Table 2.4: Marginal coordinate tests for Model II. Based on 1000 repetitions, frequencies of rejecting $H_0^{[k]}$ with different nominal tests and missing proportions are reported.

| Missing | n | $\alpha$ | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 30% | 400 | 0.01 | **1.00** | **1.00** | 0.01 | 0.02 | 0.02 | 0.02 | 0.04 | 0.00 | **0.54** | **0.67** |
| | | 0.05 | **1.00** | **1.00** | 0.06 | 0.05 | 0.06 | 0.07 | 0.06 | 0.02 | **0.74** | **0.81** |
| | | 0.10 | **1.00** | **1.00** | 0.11 | 0.12 | 0.13 | 0.10 | 0.10 | 0.08 | **0.82** | **0.91** |
| | 800 | 0.01 | **1.00** | **1.00** | 0.00 | 0.00 | 0.01 | 0.05 | 0.01 | 0.01 | **0.81** | **0.88** |
| | | 0.05 | **1.00** | **1.00** | 0.04 | 0.07 | 0.03 | 0.09 | 0.05 | 0.07 | **0.94** | **0.96** |
| | | 0.10 | **1.00** | **1.00** | 0.08 | 0.10 | 0.10 | 0.14 | 0.12 | 0.14 | **0.99** | **0.98** |
| 50% | 400 | 0.01 | **1.00** | **1.00** | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 | 0.02 | **0.39** | **0.60** |
| | | 0.05 | **1.00** | **1.00** | 0.04 | 0.06 | 0.05 | 0.08 | 0.06 | 0.07 | **0.66** | **0.80** |
| | | 0.10 | **1.00** | **1.00** | 0.08 | 0.14 | 0.13 | 0.11 | 0.10 | 0.12 | **0.73** | **0.85** |
| | 800 | 0.01 | **1.00** | **1.00** | 0.00 | 0.01 | 0.01 | 0.01 | 0.04 | 0.00 | **0.60** | **0.74** |
| | | 0.05 | **1.00** | **1.00** | 0.02 | 0.06 | 0.06 | 0.07 | 0.06 | 0.04 | **0.82** | **0.87** |
| | | 0.10 | **1.00** | **1.00** | 0.11 | 0.10 | 0.11 | 0.14 | 0.13 | 0.14 | **0.90** | **0.93** |
| 70% | 400 | 0.01 | **0.97** | **0.99** | 0.06 | 0.00 | 0.02 | 0.01 | 0.05 | 0.07 | **0.43** | **0.67** |
| | | 0.05 | **1.00** | **1.00** | 0.20 | 0.04 | 0.08 | 0.10 | 0.16 | 0.19 | **0.70** | **0.80** |
| | | 0.10 | **1.00** | **1.00** | 0.32 | 0.15 | 0.11 | 0.14 | 0.27 | 0.29 | **0.82** | **0.90** |
| | 800 | 0.01 | **1.00** | **0.99** | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.01 | **0.96** | **0.93** |
| | | 0.05 | **1.00** | **1.00** | 0.07 | 0.09 | 0.08 | 0.07 | 0.09 | 0.05 | **0.99** | **0.98** |
| | | 0.10 | **1.00** | **1.00** | 0.11 | 0.15 | 0.11 | 0.10 | 0.14 | 0.09 | **0.99** | **1.00** |

It needs to point out that for the current simulation study, $\omega_{k,h}$'s are the eigenvalues of $\mathrm{Cov}(\mathbf{G}_k)$, in which we apply the true propensity score $\pi(\mathbf{x})$ and $P(I(Y \in R_h) \mid \mathbf{x})$. The reported results show the validity of the asymptotic property of the proposed method in Theorem 2.2. Further study is needed for estimating $\widehat{\mathrm{Cov}}(\mathbf{G}_k)$, thus applying $\widehat{\omega}_{k,h}$'s to perform the marginal coordinate test.

# CHAPTER 3

# AN ESTIMATING EQUATION APPROACH FOR SDR WITH MISSING RESPONSE

Estimating approach based on estimating equations (EEs; Godambe, 1993) forms a general framework of statistical inference to accommodate a broad range of data structure and parameters, and is commonly used when data is not specified by a parametric model. Estimating equation approach has been substantially combined with Empirical Likelihood (EL) inference to solve missing data problem. Zhou et al. (2008), Wang and Chen (2009) proposed different reformulations for estimating equations with missing data and combined the EEs with EL theory. Tang and Qin (2012) introduced a semiparametric efficient EL approach for estimating equations with missing data. The similarity of the preceding approaches is that instead of imputing missing data directly, they impute the EEs by observed data. Chen et al. (2008) studied

the semiparametric efficiency bound for methods using estimating equations with missing data.

In sufficient dimension reduction literature, the employment of estimating equation approaches is considered in series of papers by Ma and Zhu (2012b,2012a); The semiparametric efficient estimators are in Ma and Zhu (2013a) for central space and Ma and Zhu (2013b) for central mean space. In this chapter, we will introduce an estimating equation approach for sufficient dimension reduction with missing data.

## 3.1 An Estimating Equation Approach

Let $\mathbf{B}_{p\times d}$ be the basis of central space/central mean space of interest in model (1.1). Define a vector of $r$ estimating functions as $\mathrm{g}(\mathbf{z};\mathbf{B}) = (g_1(\mathbf{z};\mathbf{B}), ..., g_r(\mathbf{z};\mathbf{B}))^{\mathrm{T}}$, $r \geq p$, where $\mathbf{z} = (Y, \mathbf{x}^{\mathrm{T}})^{\mathrm{T}}$. The estimating equation is specified by the general moment restrictions

$$\mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B})\} = \mathbf{0}.$$

Denote the missing data as $(\mathbf{z}_1, \mathbf{z}_2, \delta)$, where $\mathbf{z}_1 \in \mathbb{R}^{p_1}$ represents the component that can be missing, $\mathbf{z}_2 \in \mathbb{R}^{p_2}$ is the part that is always observable, and the random missing indicator be $\delta$ such that $\delta = 1(0)$ if $\mathbf{z}_1$ is observed(missing). To illustrate, the special case if $\mathbf{z}_1 = Y$ denotes the missing response case. Let $(\mathbf{z}_{1i}, \mathbf{z}_{2i}, \delta_i)$ be i.i.d observations, $i = 1, ...n$. Thus the missing at random assumption will be expressed as,

$$\delta_i \perp\!\!\!\perp \mathbf{z}_{1i} \mid \mathbf{z}_{2i}. \tag{3.1}$$

It is worth to mention that it is homogeneous missingness if $\mathbf{z}_{1i}$ corresponds

to the same component of $\mathbf{z}$ or all i, and heterogeneous if $\mathbf{z}_{1i}$ can be $i$ specific. We define the propensity function $\pi_i = \pi(\mathbf{z}_{2i}) = P(\delta_i = 1 \mid \mathbf{z}_{1i}, \mathbf{z}_{2i}) = P(\delta_i = 1 \mid \mathbf{z}_{2i})$.

Denote the estimating function with missing data as $\mathrm{g}(\mathbf{z}, \delta; \mathbf{B})$, thus the corresponding moment restriction becomes

$$\mathrm{E}\{\mathrm{g}(\mathbf{z}, \delta; \mathbf{B})\} = \mathbf{0}. \tag{3.2}$$

With only complete cases, one can write the estimating equation as

$$\mathrm{E}(\mathrm{g}(\mathbf{z}; \mathbf{B}) \mid \delta = 1) = \mathbf{0}, \tag{3.3}$$

which is referred to as CCEE for the following. The shortcoming of the complete case approach is the smaller sample size and possible bias. The case-wise validity of complete case estimating equation is discussed in Section 3.2 when response is missing at random. To address the possible issues in complete case approach, We further introduce two adjustments.

### 3.1.1 Inverse Probability Weighting Adjustment

Inverse Probability weighting (IPW) is a straight forward way to correct for bias, given $\mathrm{g}(\mathbf{z}; \mathbf{B})$, we define the inverse probability weighted estimating equation (IPWEE) as

$$\mathrm{E}\{\mathrm{g}_1(\mathbf{z}, \delta; \mathbf{B})\} = \mathrm{E}\left\{\frac{\delta}{\pi(\mathbf{z}_2)}\mathrm{g}(\mathbf{z}; \mathbf{B})\right\} = \mathbf{0}. \tag{3.4}$$

And the sample level inverse probability weighted estimating equation is given as

$$\sum_{i=1}^{n} \mathrm{g}_1(\mathbf{z}_i, \delta_i; \mathbf{B}) = \sum_{i=1}^{n} \frac{\mathrm{g}(\mathbf{z}_i; \mathbf{B})\delta_i}{\pi(\mathbf{z}_{2i})} = \mathbf{0}. \tag{3.5}$$

It can be shown that under MAR assumption(3.1), $\mathrm{g}_1(\mathbf{z}, \delta, \mathbf{B})$ is an unbiased estimating function for $\mathrm{g}(\mathbf{z}, \mathbf{B})$ by

$$\begin{aligned}
\mathrm{E}\left\{\frac{\delta}{\pi(\mathbf{z}_2)}\mathrm{g}(\mathbf{z}; \mathbf{B})\right\} &= \mathrm{E}\left[\mathrm{E}\left\{\frac{\delta}{\pi(\mathbf{z}_2)}\mathrm{g}(\mathbf{z}_1, \mathbf{z}_2; \mathbf{B}) \mid \mathbf{z}_2\right\}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left\{\frac{1}{\pi(\mathbf{z}_2)}\mathrm{g}(\mathbf{z}_1, \mathbf{z}_2; \mathbf{B}) \,|\, \mathbf{z}_2\right\} \mathrm{E}\{\delta \mid \mathbf{z}_2\}\right] \\
&= \mathrm{E}[\,\mathrm{E}\{\mathrm{g}(\mathbf{z}; \mathbf{B}) \,|\, \mathbf{z}_2\}\,] = \mathrm{E}\{\,\mathrm{g}(\mathbf{z}; \mathbf{B})\,\}.
\end{aligned}$$

Therefore, this approach will be more favorable than (3.3). The sacrifice is that we need to extra estimate the propensity function. When there is no substantial amount of missing data, and the covariate is not of high-dimensional, this is a viable approach.

### 3.1.2 Augmented Inverse Probability Weighting Adjustment

The IPW adjustment (3.4) achieves unbiased estimator, in this section, we will show another approach adjusting the bias, namely the augmented inverse probability weighted estimating equation (AIPWEE). We will show that this approach yields unbiased estimator and possesses doubly robust feature. The

formulation of AIPWEE is given as

$$\mathrm{E}\{\mathrm{g}_2(\mathbf{z},\delta;\mathbf{B})\} = \mathrm{E}\left[\frac{\delta \mathrm{g}(\mathbf{z};\mathbf{B})}{\pi(\mathbf{z}_2)} + \left(1 - \frac{\delta}{\pi(\mathbf{z}_2)}\right)\mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2\}\right] = \mathbf{0}. \quad (3.6)$$

And the sample level augmented inverse probability weighted estimating equation is given as

$$\sum_{i=1}^{n} \mathrm{g}_2(\mathbf{z}_i,\delta_i;\mathbf{B}) = \sum_{i=1}^{n}\left[\frac{\delta_i \mathrm{g}(\mathbf{z}_i;\mathbf{B})}{\pi_i(\mathbf{z}_i)} + \left(1 - \frac{\delta_i}{\pi_i(\mathbf{z}_i)}\right)\mathrm{E}\{\mathrm{g}(\mathbf{z}_i;\mathbf{B}) \mid \mathbf{z}_{2i}\}\right] = \mathbf{0}. \quad (3.7)$$

The unbiasedness of this approach is also simple to show,

$$\begin{aligned}
&\mathrm{E}\left[\frac{\delta}{\pi(\mathbf{z}_2)}\mathrm{g}(\mathbf{z};\mathbf{B}) + \left(1 - \frac{\delta}{\pi(\mathbf{z}_2)}\right)\mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2\}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left\{\frac{\delta}{\pi(\mathbf{z}_2)}\mathrm{g}(\mathbf{z};\mathbf{B}) + \left(1 - \frac{\delta}{\pi(\mathbf{z}_2)}\right)\mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2\}\right\} \,\middle|\, \mathbf{z}_2\right] \\
&= \mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B})\} + \mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B})\} - \mathrm{E}\left\{\frac{1}{\pi(\mathbf{z}_2)}\mathrm{E}(\delta \mid \mathbf{z}_2)\mathrm{E}(\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2)\right\} \\
&= \mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B})\} = \mathbf{0}.
\end{aligned}$$

Again, it remains to specify $\pi(.)$ and $\mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}\}$.

Let $\pi^*(\mathbf{z}_2)$ and $\mathrm{E}^*\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2\}$ denote misspecification of $\pi(\mathbf{z}_2)$ and $\mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid$

$\mathbf{z}_2\}$ respectively. Note that when we misspecify $\pi(\mathbf{z}_2)$, we have

$$\begin{aligned}
&\mathrm{E}\left[\frac{\delta}{\pi^*(\mathbf{z}_2)}\mathrm{g}(\mathbf{z};\mathbf{B}) + \left(1-\frac{\delta}{\pi^*(\mathbf{z}_2)}\right)\mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2\}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left\{\frac{\delta}{\pi^*(\mathbf{z}_2)}\mathrm{g}(\mathbf{z};\mathbf{B}) + \left(1-\frac{\delta}{\pi^*(\mathbf{z}_2)}\right)\mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2\}\right\} \,\middle|\, \mathbf{z}_2\right] \\
&= \mathrm{E}\left[\frac{\mathrm{E}(\delta \mid \mathbf{z}_2)}{\pi^*(\mathbf{z}_2)}\mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2\}\right] + \mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B})\} - \mathrm{E}\left[\frac{\mathrm{E}(\delta \mid \mathbf{z}_2)}{\pi^*(\mathbf{z}_2)}\mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2\}\right] \\
&= \mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B})\} = \mathbf{0}.
\end{aligned}$$

Similarly, when we misspecify $\mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2\}$, we have

$$\begin{aligned}
&\mathrm{E}\left[\frac{\delta}{\pi(\mathbf{z}_2)}\mathrm{g}(\mathbf{z};\mathbf{B}) + \left(1-\frac{\delta}{\pi(\mathbf{z}_2)}\right)\mathrm{E}^*\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2\}\right] \\
&= \mathrm{E}\left[\mathrm{E}\left\{\frac{\delta}{\pi(\mathbf{z}_2)}\mathrm{g}(\mathbf{z};\mathbf{B}) + \left(1-\frac{\delta}{\pi(\mathbf{z}_2)}\right)\mathrm{E}^*\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2\}\right\} \,\middle|\, \mathbf{z}_2\right] \\
&= \mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B})\} + \mathrm{E}[\mathrm{E}^*\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2\}] - \mathrm{E}\left[\frac{1}{\pi(\mathbf{z}_2)}\mathrm{E}(\delta \mid \mathbf{z}_2)\mathrm{E}^*\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2\}\right] \\
&= \mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B})\} = \mathbf{0}.
\end{aligned}$$

Thus the estimating equation (3.6) has a double robustness property. That is, if either the propensity function $\pi(.)$ or $\mathrm{E}\{\mathrm{g}(\mathbf{z};\mathbf{B}) \mid \mathbf{z}_2\}$ is correctly specified, the resulting estimator $\mathbf{B}$ is unbiased.

Ma and Zhu (2012b) derived a class of estimating equations from the semiparametrics tool for estimating the central space or central mean space. The general form is given as (1.6). We extend SDR estimating equation approaches to missing data framework. In the following sections, we will discuss in detail about the formulation of estimating equations for missing response problem with or without common SDR conditions. We define the estimating function

$\mathrm{g}(\mathbf{z};\mathbf{B})$ in (3.4) and (3.6) as

$$\mathrm{g}(\mathbf{z};\mathbf{B}) = \mathrm{g}(Y,\mathbf{x};\mathbf{B}) = \{l(Y) - \mathrm{E}(l(Y) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\} \times \{a(\mathbf{x}) - \mathrm{E}(a(\mathbf{x}) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\}, \tag{3.8}$$

for any functions $l(\cdot)$ and $a(\cdot)$. Here $\mathrm{g}(\mathbf{z};\mathbf{B})$ also enjoys a doubly robust property, such that the misspecification of either $\mathrm{E}\{l(Y) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}\}$ or $\mathrm{E}\{a(\mathbf{x}) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}\}$ will do no harm to give an unbiased estimating equation. By choosing specific $l(\cdot)$ and $a(\cdot)$, the approach can be connected to existing sufficient dimension reduction methods. When response is missing, we can plug in $\mathrm{g}(\mathbf{z};\mathbf{B})$ to (3.3), (3.4) and (3.6) for complete case, IPW and AIPW estimating equations, respectively. To illustrate, we will specify the estimating equations in terms of OLS and PHD and SIR approaches. The complete case estimating equation approaches are demonstrated in Section 3.2. And the IPW and AIPW estimating equation approaches are unified in Section 3.3.

## 3.2 Complete Case Estimating Equations

Ma and Zhu (2012b) developed the semiparametric approach for a family of SDR estimators. Specifically, misspecifying $\mathrm{E}\{l(Y) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}\} = \mathbf{0}$, $\mathrm{E}\{\mathrm{g}(\mathbf{x}, Y;\mathbf{B})\}$ (3.8) becomes

$$\mathrm{E}\left[l(Y)\{a(\mathbf{x}) - \mathrm{E}(a(\mathbf{x}) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\}\right] = \mathbf{0}. \tag{3.9}$$

Without loss of generality, we will assume $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, $\mathrm{E}(Y) = 0$ and $\mathrm{Cov}(\mathbf{x}) = \mathbf{I}_p$ throughout the chapter. Under the LCM condition (1.3), parametric assumption is given as

$$\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = \mathbf{P}\mathbf{x}, \tag{3.10}$$

where $\mathbf{P} = \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}$. The CCV condition(1.4) characterizes the variance-covariance matrix of $\mathbf{x}$ conditional on $\mathbf{B}^{\mathrm{T}}\mathbf{x}$ by assuming

$$\mathrm{Cov}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = \mathbf{Q}. \tag{3.11}$$

where $\mathbf{Q} = \mathbf{I_p} - \mathbf{P}$. Again, focusing on missing response at random, we have the MAR assumption (3.1) given as $\delta \perp\!\!\!\perp Y \mid \mathbf{x}$ and the propensity function $\pi(\mathbf{x}) = P(\delta = 1 \mid Y, \mathbf{x}) = P(\delta = 1 \mid \mathbf{x})$.

### 3.2.1 Sliced Inverse Regression

When specifying $l(Y) = \mathrm{E}(\mathbf{x} \mid Y)$ and $a(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}$, we connects (3.9) to SIR with the following estimating equation,

$$\mathrm{E}\left[\mathrm{E}(\mathbf{x} \mid Y)\{\mathbf{x} - \mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\}^{\mathrm{T}}\right] = \mathbf{0}. \tag{3.12}$$

The validity of (3.12) can be briefly justified as follows. First notice that

$$\mathrm{E}(\mathbf{x} \mid Y) = \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y, \mathbf{B}^{\mathrm{T}}\mathbf{x}) \mid Y\} = \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) \mid Y\}. \tag{3.13}$$

The first equality in (3.13) is from the law of iterated expectation, and the second equality is implied by the definition of the central space in (1.1). It follows that $\mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y)\mathbf{x}^{\mathrm{T}}\} = \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y)\mathrm{E}^{\mathrm{T}}(\mathbf{x} \mid Y)\} = \mathrm{E}[\mathrm{E}(\mathbf{x} \mid Y)\mathrm{E}^{\mathrm{T}}\{\mathrm{E}(\mathbf{x} \mid$

$\mathbf{B}^{\mathrm{T}}\mathbf{x}) \mid Y\}] = \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y)\mathrm{E}^{\mathrm{T}}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\}$, which implies (3.12). We see that the linear conditional mean assumption (3.10) is not required for the validity of (3.12).

**Remark 3.1.** *Let* $\mathbf{B}$ *be the basis of* $\mathcal{S}_{Y|\mathbf{x}}$ *and we have span*$(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{x}}$*. Here span*$(\mathbf{B})$ *denotes the column space of* $\mathbf{B}$*. When (3.10) is satisfied for* $\mathbf{B}$*,* $\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{T}\mathbf{x}) = \mathbf{P}\mathbf{x}$*, where* $\mathbf{P} = \mathbf{B}(\mathbf{B}^{T}\mathbf{B})^{-1}\mathbf{B}^{T}$*, thus (3.9) becomes* $\mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y)\mathbf{x}^{T}\}\mathbf{Q} = \mathbf{0}$*, where* $\mathbf{Q} = \{\mathbf{I}_p - \mathbf{P}\}$*. Note that* $\mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y)\mathbf{x}^{T}\} = \mathrm{Cov}\{\mathrm{E}(\mathbf{x} \mid Y)\}$*, and solving this equation for* $\mathbf{B}$ *becomes equivalent to finding the eigenvectors corresponding to the nonzero eigenvalues of* $\mathrm{Cov}\{\mathrm{E}(\mathbf{x} \mid Y)\}$*. In the dimension reduction literature, it is well-known that SIR estimates the central space from the eigenvalue decomposition of* $\mathrm{Cov}\{\mathrm{E}(\mathbf{x} \mid Y)\}$*. Thus solving the estimating equation (3.12) becomes equivalent to the classical SIR at the population level when the linear conditional mean assumption (4.1) is satisfied. Furthermore, (3.13) and (1.3) together lead to* $\mathrm{E}(\mathbf{x} \mid Y) = \mathbf{P}\mathrm{E}(\mathbf{x} \mid Y)$*. It follows that* $\mathrm{E}(\mathbf{x} \mid Y) \subseteq$ *span*$(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{x}}$*. This guarantees the validity of SIR when the linear conditional mean assumption (1.3) holds.*

When the response is subject to missingness, the original estimation equation (3.12) can not be evaluated directly at the sample level, which prompts us to consider the following modified estimating equation,

$$\mathrm{E}\left[\mathrm{E}(\mathbf{x} \mid Y, \delta = 1)\{\mathbf{x} - \mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}, \delta = 1)\}^{\mathrm{T}} \mid \delta = 1\right] = \mathbf{0}. \tag{3.14}$$

Equation (3.14) essentially replaces all the expectations and conditional expections in (3.12) with the expectations conditioning on an additional $\delta = 1$, which means the expectations and conditional expectations are now based

only on the complete cases. We refer to (3.14) as the complete case estimating equation of SIR. The verification of (3.14) is parallel to the justification of (3.12). Similar to (3.13), we have

$$\begin{aligned} \mathrm{E}(\mathbf{x} \mid Y, \delta = 1) &= \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y, \delta = 1, \mathbf{B}^{\mathrm{T}}\mathbf{x}) \mid Y, \delta = 1\} \\ &= \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}, \delta = 1) \mid Y, \delta = 1\}, \end{aligned} \tag{3.15}$$

where the first equality is from the law of iterated expectation, and the second equality is due to the fundamental SDR definition in (1.1). It follows that

$$\begin{aligned} &\mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y, \delta = 1)\mathbf{x}^{\mathrm{T}} \mid \delta = 1\} = \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y, \delta = 1)\mathrm{E}^{\mathrm{T}}(\mathbf{x} \mid Y, \delta = 1) \mid \delta = 1\} \\ &= \mathrm{E}[\mathrm{E}(\mathbf{x} \mid Y, \delta = 1)\mathrm{E}^{\mathrm{T}}\{\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}, \delta = 1) \mid Y, \delta = 1\} \mid \delta = 1] \\ &= \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y, \delta = 1)\mathrm{E}^{\mathrm{T}}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}, \delta = 1) \mid \delta = 1\}, \end{aligned}$$

which implies (3.14). The derivation above does not require the linear conditional mean assumption (3.10), nor does it make any assumptions about the missingness scheme.

**Remark 3.2.** *We have seen in Remark 3.1 that when the linear conditional mean assumption (3.10) holds, the estimating equation approach based on (3.12) becomes equivalent to the classical SIR. In the case with missing data, if we have*

$$\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{T}\mathbf{x}, \delta = 1) = \mathbf{P}\mathbf{x}, \tag{3.16}$$

*then (3.14) becomes* $\mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y, \delta = 1)\mathbf{x}^{T} \mid \delta = 1\}\mathbf{Q} = \mathbf{0}$. *Note, for example, (3.16) can be implied by (3.10) if* $Y$ *is missing completely at random. Solving*

*this equation for* $\mathbf{B}$ *becomes equivalent to finding the eigenvectors corresponding to the nonzero eigenvalues of* $\mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y, \delta = 1)\mathrm{E}^{T}(\mathbf{x} \mid Y, \delta = 1) \mid \delta = 1\}$*, which can be viewed as the complete-case version of the classical SIR. Furthermore, (3.15) and (3.16) together imply that* $\mathrm{E}(\mathbf{x} \mid Y, \delta = 1) = \mathbf{B}(\mathbf{B}^{T}\mathbf{B})^{-1}\mathbf{B}^{T}\mathrm{E}(\mathbf{x} \mid Y, \delta = 1)$*. It follows that* $\mathrm{E}(\mathbf{x} \mid Y, \delta = 1) \subseteq span(\mathbf{B}) = \mathcal{S}_{Y|\mathbf{x}}$*. However, (3.16) is generally not true even if (3.10) holds. Thus even with (3.10) being satisfied, complete case SIR could be biased. Similar bias issue has also been showed in Chapter 2 for inverse regressions. On the other hand, the estimating equation approach bypasses the requirement of (3.10) in the case of full data and the requirement of (3.16) in the case of missing data. Hence the original estimating equation (3.12) from Ma and Zhu (2012b) and our proposed complete case estimating equation (3.14) lead to unbiased estimators in the absence of the linear conditional mean assumption (3.10).*

### 3.2.2 Ordinary Least Squares

The OLS method (Li and Duan, 1989) takes $\mathrm{Cov}(\mathbf{x}, Y) = \mathrm{E}(\mathbf{x}Y)$ as a subspace of column space of $\mathbf{B}$. Regardless of LCM condition, it is always true that $\mathrm{E}(\mathbf{x}Y) = \mathrm{E}\{\mathrm{E}(\mathbf{x}Y) \mid \mathbf{x}\} = \mathrm{E}\{\mathbf{x}\mathrm{E}(Y \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\} = \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})Y\}$, the resulting estimating equation $\mathrm{E}\{\mathrm{g}(Y, \mathbf{x}; \mathbf{B})\} = \mathbf{0}$ is thus

$$\mathrm{E}[Y\{\mathbf{x} - \mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\}] = \mathbf{0}. \tag{3.17}$$

It is equivalent to (3.9) by letting $a(\mathbf{x}) = \mathbf{x}$ and $l(\mathbf{x}) = Y$. When the common assumed LCM condition (3.10) holds, the derivation continues with $\mathrm{E}(\mathbf{x}Y) = \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})Y\} = \mathrm{E}(\mathbf{P}\mathbf{x}Y)$, obviously, $\mathrm{E}\{\mathrm{g}(Y, \mathbf{x}; \mathbf{B})\}$ becomes $\mathrm{E}(\mathbf{Q}\mathbf{x}Y) = \mathbf{0}$,

where $\mathbf{Q} = \mathbf{I}_p - \mathbf{P}$. When the response is subject to missingness, we consider the following complete case estimating equation of OLS.

$$\mathrm{E}\left[Y\{\mathbf{x} - \mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}, \delta = 1)\} \mid \delta = 1\right] = \mathbf{0}. \tag{3.18}$$

The verification of (3.18) is parallel to the justification of (3.17). We have

$$\begin{aligned} \mathrm{E}(Y\mathbf{x} \mid \delta = 1) =& \mathrm{E}\{\mathrm{E}(Y\mathbf{x} \mid \mathbf{x}, \delta = 1) \mid \delta = 1\} \\ =& \mathrm{E}\{\mathbf{x}\mathrm{E}(Y \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}, \delta = 1) \mid \delta = 1\} \\ =& \mathrm{E}\{Y\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}, \delta = 1) \mid \delta = 1\}, \end{aligned} \tag{3.19}$$

which implies (3.18). The derivation above does not require the linear conditional mean assumption (1.3), nor does it make any assumptions about the missingness scheme.

**Remark 3.3.** *Similar to Remark 3.2, in the case of OLS, we have* $\mathrm{E}(Y\mathbf{x} \mid \delta = 1) = \mathrm{E}\{\mathrm{E}(Y\mathbf{x} \mid \mathbf{x}, \delta = 1) \mid \delta = 1\} = \mathrm{E}\{\mathbf{x}\mathrm{E}(Y \mid \mathbf{x}, \delta = 1) \mid \delta = 1\} = \mathrm{E}\{\mathbf{x}\mathrm{E}(Y \mid \mathbf{B}^{T}\mathbf{x}, \delta = 1) \mid \delta = 1\} = \mathrm{E}\{Y\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{T}\mathbf{x}, \delta = 1) \mid \delta = 1\}$. *The derivation cannot step further unless we have (3.16). With (3.16),* $\mathrm{E}(Y\mathbf{x} \mid \delta = 1) = \mathrm{E}(\mathbf{Q}\mathbf{x}Y \mid \delta = 1)$, *which can be viewed as the complete-case version of the classical OLS. However, (3.16) is generally not true even if (3.10) holds. Thus even with (3.10) being satisfied, classical OLS with only complete cases could be biased. The exemption is discussed in Chapter 1.1.2 when the link function between response and the predictor is linear, classical SIR is valid to recover central space with complete case analysis.*

### 3.2.3 Principal Hessian Directions

The PHD method (Li, 1992; Cook and Li, 2004a) recovers the central space via the eigenvectors associated with the $d$ eigenvalues of $\mathbf{M}_{\text{PHD}} = \text{E}(Y\mathbf{x}\mathbf{x}^T)$. In a similar fashion, one can derive the corresponding estimating equation $\text{E}\{\text{g}(Y, \mathbf{x}; \mathbf{B})\}$ directly by $\text{E}(Y\mathbf{x}\mathbf{x}^{\text{T}}) = \text{E}\{\text{E}(Y\mathbf{x}\mathbf{x}^{\text{T}}) \mid \mathbf{x}\} = \text{E}\{\text{E}(Y \mid \mathbf{B}^{\text{T}}\mathbf{x})\mathbf{x}\mathbf{x}^{\text{T}}\} = \text{E}\{Y\text{E}(\mathbf{x}\mathbf{x}^{\text{T}} \mid \mathbf{B}^{\text{T}}\mathbf{x})\}$. It follows that

$$\text{E}\{\text{g}(Y, \mathbf{x}; \mathbf{B})\} = \text{E}[Y\{\mathbf{x}\mathbf{x}^{\text{T}} - \text{E}(\mathbf{x}\mathbf{x}^{\text{T}} \mid \mathbf{B}^{\text{T}}\mathbf{x})\}], \tag{3.20}$$

which is the same as (3.8) but letting $a(\mathbf{x}) = \mathbf{x}\mathbf{x}^{\text{T}}$ and $l(Y) = Y$. As classic PHD method requires LCM (3.10) and CCV (3.11) conditions, we can continue to derive the following when incorporating both conditions,

$$\begin{aligned}\text{E}(Y\mathbf{x}\mathbf{x}^{\text{T}}) &= \text{E}\{Y\text{E}(\mathbf{x}\mathbf{x}^{\text{T}} \mid \mathbf{B}^{\text{T}}\mathbf{x})\} = \text{E}[Y\{\text{Var}(\mathbf{x} \mid \mathbf{B}^{\text{T}}\mathbf{x}) + \text{E}(\mathbf{x} \mid \mathbf{B}^{\text{T}}\mathbf{x})\text{E}(\mathbf{x}^{\text{T}} \mid \mathbf{B}^{\text{T}}\mathbf{x})\}] \\ &= \text{E}(Y)\mathbf{Q} + \mathbf{P}\text{E}(Y\mathbf{x}\mathbf{x}^{\text{T}})\mathbf{P} = \text{E}(Y\mathbf{P}\mathbf{x}\mathbf{x}^{\text{T}}\mathbf{P}),\end{aligned}$$

Thus the $\text{E}\{\text{g}(Y, \mathbf{x}; \mathbf{B})\}$ is simplified as $\text{E}\{Y(\mathbf{x}\mathbf{x}^{\text{T}} - \mathbf{P}\mathbf{x}\mathbf{x}^{\text{T}}\mathbf{P})\} = \mathbf{0}$.

When the response is subject to missingness, we consider the following complete case estimating equation of PHD.

$$\text{E}\left[Y\{\mathbf{x}\mathbf{x}^{\text{T}} - \text{E}(\mathbf{x}\mathbf{x}^{\text{T}} \mid \mathbf{B}^{\text{T}}\mathbf{x}, \delta = 1)\} \mid \delta = 1\right] = \mathbf{0}. \tag{3.21}$$

The verification of (3.21) is the same as the justification of the complete case estimating equation of OLS in (3.19) by replacing $\mathbf{x}$ with $\mathbf{x}\mathbf{x}^{\text{T}}$. Note that the derivation in (3.21) does not require the LCM assumption (3.10), nor does it make any assumptions about the missingness scheme.

**Remark 3.4.** *In order to relate the estimating equation (3.21) to complete case based classical PHD, we need to additionally introduce the following condition,*

$$\mathrm{Cov}(\mathbf{x} \mid \mathbf{B}^T\mathbf{x}, \delta = 1) = \mathbf{Q}, \tag{3.22}$$

*which can be viewed as the constant conditional variance assumption under complete cases. When both (3.16) and (3.22) are satisfied, we will have*

$$\mathrm{E}\{Y(\mathbf{x}\mathbf{x}^T - \mathbf{P}\mathbf{x}\mathbf{x}^T\mathbf{P}) \mid \delta = 1\} = \mathbf{0}, \tag{3.23}$$

*which can be viewed as the complete case based classical PHD. However, (3.16) and (3.22) are generally not true even if (3.10) and (3.11) hold. Thus even with (3.10) and (3.10) being satisfied, classical PHD with complete case analysis could be biased.*

## 3.3 Inverse Probability Weighted Estimating Equation

Although the complete case estimating equation approach leads to unbiased estimator, we lose efficiency by ignoring the information in the incomplete cases, especially when the missing proportion is large. To address this limitation, we continue to propose adjusted estimating equation approaches with both inverse probability and and augmented inverse probability weighting for SDR when response is missing at random. The adjusted estimating equations can be given by plugging $\mathrm{g}(\mathbf{z}; \mathbf{B}) = \mathrm{g}(Y, \mathbf{x}; \mathbf{B}) = l(Y)\{a(\mathbf{x}) - \mathrm{E}(a(\mathbf{x}) \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\}$ into (3.4) and (3.6) directly. Particularly, with appropriate sets of $a(\mathbf{x})$ and

$l(Y)$ defined the same as in Section 3.2, we have the corresponding adjusted estimating equations for SIR, OLS and PHD. To illustrate, the inverse probability weighted estimating equation for SIR is given as,

$$\mathrm{E}\left[\frac{\delta}{p(\mathbf{x})}\mathrm{E}(\mathbf{x} \mid Y)\{\mathbf{x} - \mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\}^{\mathrm{T}}\right] = \mathbf{0}. \tag{3.24}$$

To check the validity of (3.24), first note that

$$\mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y)\mathbf{x}^{\mathrm{T}}\delta p^{-1}(\mathbf{x})\} = \mathrm{E}[\mathrm{E}(\mathbf{x} \mid Y)\mathrm{E}^{\mathrm{T}}\{\mathbf{x}\delta p^{-1}(\mathbf{x}) \mid Y\}]. \tag{3.25}$$

On the other hand, the MAR assumption $Y \perp\!\!\!\perp \delta \mid \mathbf{x}$ implies that

$$\mathrm{E}\{\mathbf{x}\delta p^{-1}(\mathbf{x}) \mid Y\} = \mathrm{E}[\mathrm{E}\{\mathbf{x}\delta p^{-1}(\mathbf{x}) \mid \mathbf{x}, Y\} \mid Y] = \mathrm{E}(\mathbf{x} \mid Y). \tag{3.26}$$

Equations (3.25) and (3.26) lead to $\mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y)\mathbf{x}^{\mathrm{T}}\delta p^{-1}(\mathbf{x})\} = \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y)\mathbf{x}^{\mathrm{T}}\}$. Similarly, one can show $\mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y)\mathrm{E}^{\mathrm{T}}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\delta p^{-1}(\mathbf{x})\} = \mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y)\mathrm{E}^{\mathrm{T}}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\}$. Together with the original inverse regression estimating equation (3.12), (3.24) is guaranteed to hold. We refer to (3.24) as the inverse probability weighted estimating equation for SIR.

In addition, to access the double robustness property, the augmented inverse probability weighted estimating equation for SIR is given as follows,

$$\mathrm{E}\left[\left\{\frac{\delta}{\pi(\mathbf{x})}\mathrm{E}(\mathbf{x} \mid Y) + \left(1 - \frac{\delta}{\pi(\mathbf{x})}\right)\mathrm{E}(\mathrm{E}(\mathbf{x} \mid Y) \mid \mathbf{x})\right\}\{\mathbf{x} - \mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\}^{\mathrm{T}}\right] = \mathbf{0}. \tag{3.27}$$

**Remark 3.5.** *In the case when the common assumed LCM condition (3.10) holds,* $\{\mathbf{x} - \mathrm{E}(\mathbf{x} \mid \mathbf{B}^{T}\mathbf{x})\}^{T}$ *in equation (3.24) and (3.27) can be replaced with*

$\mathbf{x}^T\mathbf{Q}$, *where* $\mathbf{Q} = \mathbf{I}_p - \mathbf{P}$ *and* $\mathbf{P} = \mathbf{B}(\mathbf{B}^T\mathbf{B})^{-1}\mathbf{B}^T$. *The resulting estimating equations are valid only when LCM holds.*

Similarly, we obtain the IPWEE and AIPWEE for OLS and PHD. As IPWEE is part of AIPWEE, we only give IPWEE to avoid repetition.

$$\text{OLS: } \mathrm{E}\left[\left\{\frac{\delta Y}{\pi(\mathbf{x})} + \left(1 - \frac{\delta}{\pi(\mathbf{x})}\right)\mathrm{E}(Y \mid \mathbf{x})\right\}\{\mathbf{x} - \mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\}\right] = \mathbf{0}. \quad (3.28)$$

$$\text{PHD: } \mathrm{E}\left[\left\{\frac{\delta Y}{\pi(\mathbf{x})} + \left(1 - \frac{\delta}{\pi(\mathbf{x})}\right)\mathrm{E}(Y \mid \mathbf{x})\right\}\{\mathbf{x}\mathbf{x}^{\mathrm{T}} - \mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})\}\right] = \mathbf{0}. \quad (3.29)$$

**Remark 3.6.** *When the common assumed LCM assumption (3.10) holds,* $\mathbf{x} - \mathrm{E}(\mathbf{x} \mid \mathbf{B}^T\mathbf{x}) = \mathbf{Q}\mathbf{x}$ *in equation (3.28). And with additional CCV assumption (3.11),* $\mathbf{x}\mathbf{x}^T - \mathrm{E}(\mathbf{x}\mathbf{x}^T \mid \mathbf{B}^T\mathbf{x})$ *can be replaced with* $\mathbf{x}\mathbf{x}^T - \mathbf{P}\mathbf{x}\mathbf{x}^T\mathbf{P}$ *in equation (3.29). The resulting estimating equations are valid only when the required assumptions hold.*

## 3.4 Implementation

With the purpose of delivering the proposed approaches, we demonstrate the implementation details in practice. As exhibited in Section 3.2 and 3.3, the proposed approaches for SIR, OLS and PHD share similar components of estimating unobserved terms and solving estimating equations. For example, they require the derivations for the general conditional expectations $\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$, $\mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$, $\mathrm{E}(Y \mid \mathbf{x})$, $\mathrm{E}(Y \mid \mathbf{x})$, $\mathrm{E}\{\mathrm{E}(Y \mid \mathbf{x}) \mid \mathbf{x}\}$ and the propensity function $\pi(\mathbf{x})$. To ease the notation, we discuss the general $\mathbf{x}$ and $Y$ regardless of the subscript for observations. Let $K(\cdot)$ to denote a kernel function, and

$K_b(\cdot) = K(\cdot/b)/b$ for any bandwidth $h$.

Following Tang and Qin (2012), the propensity function is estimated nonparametrically by

$$\widehat{\pi}(\mathbf{x}) = \widehat{\mathrm{E}}(\delta \mid \mathbf{x}) = \frac{\sum_{j=1}^{n} \delta_j K_b(\mathbf{x} - \mathbf{x}_j)}{\sum_{i=1}^{n} K_b(\mathbf{x} - \mathbf{x}_j)}. \tag{3.30}$$

Note that $\pi(\mathbf{x})$ can also be estimated by other methods.

The conditional expectations $\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$ and $\mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$ can be estimated parametrically (Li and Dong, 2009) or nonparametrically (Ma and Zhu, 2012b). Li and Dong's parametric idea is expressed in (1.5), to estimate $\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$ as a function of $\mathbf{B}^{\mathrm{T}}\mathbf{x}$ but it does not solve the case for $\mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x})$. Therefore, Ma and Zhu's nonparametric estimation is adopted as following,

$$\widehat{\mathrm{E}}(\mathbf{x} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = \frac{\sum_{j=1}^{n} \mathbf{x}_j K_b(\mathbf{B}^{\mathrm{T}}\mathbf{x} - \mathbf{B}^{\mathrm{T}}\mathbf{x}_j)}{\sum_{i=1}^{n} K_b(\mathbf{B}^{\mathrm{T}}\mathbf{x} - \mathbf{B}^{\mathrm{T}}\mathbf{x}_j)}, \tag{3.31}$$

and

$$\widehat{\mathrm{E}}(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid \mathbf{B}^{\mathrm{T}}\mathbf{x}) = \frac{\sum_{j=1}^{n} \mathbf{x}_j \mathbf{x}_j^{\mathrm{T}} K_b(\mathbf{B}^{\mathrm{T}}\mathbf{x} - \mathbf{B}^{\mathrm{T}}\mathbf{x}_j)}{\sum_{i=1}^{n} K_b(\mathbf{B}^{\mathrm{T}}\mathbf{x} - \mathbf{B}^{\mathrm{T}}\mathbf{x}_j)}. \tag{3.32}$$

Inspired by Zhu et al. (2012), under MAR assumption,

$$\mathrm{E}(Y \mid \mathbf{x}) = \frac{\mathrm{E}(\delta Y \mid \mathbf{x})}{\mathrm{E}(\delta \mid \mathbf{x})}. \tag{3.33}$$

As both $\delta$ and $\delta Y$ are observable, the two quantities $\mathrm{E}(\delta Y \mid \mathbf{x})$ and $\mathrm{E}(\delta \mid \mathbf{x})$

can be estimated nonparametrically via kernel regression estimation,thus

$$\widehat{\mathrm{E}}(Y \mid \mathbf{x}) = \frac{\sum_{j=1}^{n} \delta_j Y_j K_b(\mathbf{x} - \mathbf{x}_j)}{\sum_{j=1}^{n} \delta_j K_b(\mathbf{x} - \mathbf{x}_j)}. \tag{3.34}$$

In Chapter 2, we introduced a sliced method for estimating $\mathrm{E}(\mathbf{x} \mid Y)$. While in this chapter, as all other conditional expectations are estimated nonparametrically, we propose a novel kernel estimator. Under missing at random scheme (3.1), it can be shown that $\mathrm{E}(\mathbf{x} \mid Y) = \mathrm{E}\{\mathbf{x}\delta p^{-1}(\mathbf{x}) \mid Y\}$, and $\mathrm{E}\{\delta p^{-1}(\mathbf{x}) \mid Y\} = \mathrm{E}[\mathrm{E}\{\delta p^{-1}(\mathbf{x}) \mid Y, \mathbf{x}\} \mid Y] = 1$, thus we have

$$\mathrm{E}(\mathbf{x} \mid Y) = \frac{\mathrm{E}\{\mathbf{x}\delta p^{-1}(\mathbf{x}) \mid Y\}}{\mathrm{E}\{\delta p^{-1}(\mathbf{x}) \mid Y\}}. \tag{3.35}$$

This motivates us to consider

$$\widehat{\mathrm{E}}(\mathbf{x} \mid Y) = \frac{\sum_{j=1}^{n} \mathbf{x}_j \delta_j \widehat{p}^{-1}(\mathbf{x}_j) K_b(Y - Y_j)}{\sum_{j=1}^{n} \widehat{p}^{-1}(\mathbf{x}_j) K_b(Y - Y_j)\delta_j}. \tag{3.36}$$

Similar to (3.33), we have

$$\mathrm{E}\{\mathrm{E}(\mathbf{x} \mid Y) \mid \mathbf{x}\} = \frac{\mathrm{E}\{\delta \mathrm{E}(\mathbf{x} \mid Y) \mid \mathbf{x}\}}{\mathrm{E}(\delta \mid \mathbf{x})}. \tag{3.37}$$

This prompts us to consider

$$\widehat{\mathrm{E}}\{\widehat{\mathrm{E}}(\mathbf{x} \mid Y) \mid \mathbf{x}\} = \frac{\sum_{j=1}^{n} \delta_j \widehat{\mathrm{E}}(\mathbf{x} \mid Y_j) K_b(\mathbf{x} - \mathbf{x}_j)}{\sum_{j=1}^{n} \delta_j K_b(\mathbf{x} - \mathbf{x}_j)}. \tag{3.38}$$

To achieve $\widehat{\mathbf{B}}$, we need to solve the sample estimating equations with numerical optimization. Let $\{(\mathbf{x}_i, Y_i, \delta_i), i = 1, \ldots, n\}$ be an i.i.d. sample. Without loss of generality, assume $\delta_i = 1$ for $i = 1, \ldots, n_1$ and $\delta_i = 0$ for

$i = n_1 + 1, \ldots, n$. We list the sample version of complete case estimating equation and AIPWEE for SIR in (3.39) and (3.40), respectively. Then other cases can be obtained similarly.

$$\sum_{i=1}^{n_1} \widehat{\mathrm{E}}(\mathbf{x} \mid Y_i, \delta = 1)\{\mathbf{x}_i - \widehat{\mathrm{E}}(\mathbf{x} \mid \widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{x}_i, \delta = 1)\}^{\mathrm{T}} = \mathbf{0}. \tag{3.39}$$

$$\sum_{i=1}^{n} \left[ \frac{\delta_i}{\widehat{\pi}(\mathbf{x}_i)} \widehat{\mathrm{E}}(\mathbf{x} \mid Y_i) + \left(1 - \frac{\delta_i}{\widehat{\pi}(\mathbf{x}_i)}\right) \widehat{\mathrm{E}}\{\widehat{\mathrm{E}}(\mathbf{x} \mid Y_i) \mid \mathbf{x}_i\} \right] \left\{ \mathbf{x}_i - \widehat{\mathrm{E}}(\mathbf{x} \mid \widehat{\mathbf{B}}^{\mathrm{T}}\mathbf{x}_i) \right\}^{\mathrm{T}} = \mathbf{0}. \tag{3.40}$$

**Remark 3.7.** *Note that for those complete case approaches, we only include the complete data for nonparametrically estimating the conditional expectations. Take* $\mathrm{E}(\mathbf{x} \mid Y_i, \delta = 1)$ *for* $i = 1, \ldots, n_1$ *as an example,*

$$\widehat{\mathrm{E}}(\mathbf{x} \mid Y_i, \delta = 1) = \frac{\sum_{j=1}^{n_1} K_h(Y_i - Y_j)\mathbf{x}_j}{\sum_{j=1}^{n_1} K_h(Y_i - Y_j)}. \tag{3.41}$$

*Then other complete case conditional expectations can be obtained correspondingly.*

To solve for $\widehat{\mathbf{B}}$, the practical implementation adopts Newton-Raphson procedure to minimize the Fronbenius norm of the estimating equations. The algorithm is showed below.

1. Pick an arbitrary starting value for $\mathbf{B}^{(0)}$.

2. At the $j$th iteration, nonparametrically estimate $\mathrm{E}(\mathbf{x} \mid \mathbf{B}^{(j)T}\mathbf{x})$, $\mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid \mathbf{B}^{(j)T}\mathbf{x})$, $\mathrm{E}(Y \mid \mathbf{x})$, $\mathrm{E}(Y \mid \mathbf{x})$, $\mathrm{E}\{\mathrm{E}(Y \mid \mathbf{x}) \mid \mathbf{x}\}$, and $\pi(\mathbf{x})$ by (3.31), (3.32), (3.34), (3.36), (3.38), and (3.30), respectively.

3. Let $\mathrm{g}(\mathbf{B}^{(j)})$ be the sample version of the estimating functions at the $j$th iteration, for example, the left-hand side of (3.40) for AIPW SIR without common conditions is given as

$$\mathrm{g}(\mathbf{B}^{(j)}) = \sum_{i=1}^{n} \left[ \frac{\delta_i}{\widehat{\pi}(\mathbf{x}_i)} \widehat{\mathrm{E}}(\mathbf{x} \mid Y_i) + \left( 1 - \frac{\delta_i}{\widehat{\pi}(\mathbf{x}_i)} \right) \widehat{\mathrm{E}}\{\widehat{\mathrm{E}}(\mathbf{x} \mid Y_i) \mid \mathbf{x}_i\} \right] \times \left\{ \mathbf{x}_i - \widehat{\mathrm{E}}(\mathbf{x} \mid \mathbf{B}^{(j)^{\mathrm{T}}} \mathbf{x}_i) \right\}^{\mathrm{T}}.$$

4. Update $\mathbf{B}^{(j)}$ using the Newton-Raphson algorithm which is a second-derivative-based gradient descent process. Form $\mathbf{B}^{(j)}$ to get $\mathbf{B}_{k+}^{(j)} = \mathbf{B}^{(j)} + \alpha \mathbf{e}_k$ and $\mathbf{B}_{k-}^{(j)} = \mathbf{B}^{(j)} - \alpha \mathbf{e}_k$ where $\mathbf{e}_k$ has the same size as $\mathbf{B}$ but has 1 in the $k$th entry while 0 elsewhere. Set the $k$th row of the first partial derivative $\partial\{\|\mathrm{g}(\mathbf{B})\|^2\}/\partial\{\mathrm{vec}(\mathbf{B})\}$ to be $\{\|\mathrm{g}(\mathbf{B}_{k+}^{(j)})\|^2 - \|\mathrm{g}(\mathbf{B}_{k-}^{(j)})\|^2\}/(2\alpha)$. Repeat for all the entries of $\mathbf{B}^{(j)}$. Then following similar procedures to get the second partial derivative $\partial^2\{\|\mathrm{g}(\mathbf{B})\|^2\}/\partial\{\mathrm{vec}(\mathbf{B})\}\partial\{\mathrm{vec}(\mathbf{B})\}^{\mathrm{T}}$. $\alpha$ is designed to be 0.001 in practice.

5. Update

$$\mathrm{vec}(\mathbf{B}^{(j+1)}) = \mathrm{vec}(\mathbf{B}^{(j)}) - \left[ \frac{\partial^2\{\|\mathrm{g}(\mathbf{B}^{(j)})\|^2\}}{\partial\{\mathrm{vec}(\mathbf{B}^{(j)})\}\partial\{\mathrm{vec}(\mathbf{B}^{(j)})\}^T} \right]^{-1} \times \frac{\partial\{\|\mathrm{g}(\mathbf{B}^{(j)})\|^2\}}{\partial\{\mathrm{vec}(\mathbf{B}^{(j)})\}}.$$

6. Repeat Steps 2 to 5 until $\|\mathrm{vec}(\mathbf{B}^{(j+1)}) - \mathrm{vec}(\mathbf{B}^{(j)})\| < \varepsilon$ where $\varepsilon$ is set to 0.0001 in practice.

## 3.5 Asymptotic Properties

In this chapter, we study the asymptotic properties for the estimators derived from complete case (3.3), inverse probability weighted (3.4) and augmented inverse probability weighted (3.6) estimating equations. We formulate a unified form for the embedded estimating function $\mathrm{g}(Y, \mathbf{x}; \mathbf{B})$ as

$$\mathrm{g}^\star = l^\star(Y)\{a^\star(\mathbf{x}) - \mathrm{E}(a^\star(\mathbf{x}) \,|\, \mathbf{B}^\mathrm{T}\mathbf{x})\},$$

for any functions $a^\star(\cdot)$ of $\mathbf{x}$ and $l^\star(\cdot)$ that satisfies $\mathrm{E}(l^\star \mid \mathbf{x}) = \mathrm{E}(l^\star \mid \mathbf{B}^\mathrm{T}\mathbf{x})$. And we define the general CCEE, IPWEE and AIPWEE in order to consider the special cases SIR, OLS and PHD simultaneously:

$$\begin{aligned}
\text{CCEE:} \quad & \mathrm{E}\left\{\mathrm{g}^\star \mid \delta = 1\right\} = \mathbf{0}. \\
\text{IPWEE:} \quad & \mathrm{E}\left\{\frac{\delta}{\pi(\mathbf{x})}\mathrm{g}^\star\right\} = \mathbf{0}. \\
\text{AIPWEE:} \quad & \mathrm{E}\left\{\frac{\delta \mathrm{g}^\star}{\pi(\mathbf{x})} + \left(1 - \frac{\delta}{\pi(\mathbf{x})}\right)\mathrm{E}(\mathrm{g}^\star \mid \mathbf{x})\right\} = \mathbf{0}.
\end{aligned}$$

Note that Specifying $l^\star(Y) = \mathrm{E}(\mathbf{x} \mid Y)$ and $a^\star(\mathbf{x}) = \mathbf{x}$ yields the SIR case; $l^\star(Y) = Y$ and $a^\star(\mathbf{x}) = \mathbf{x}$ yields the OLS case; $l^\star(Y) = Y$ and $a^\star(\mathbf{x}) = \mathbf{x}\mathbf{x}^\mathrm{T}$ yields the PHD case.

In order to study the properties of the space estimation via studying the properties in estimating the parameters in $\mathbf{B}$, we follow the particular parameterization in Ma and Zhu (2012a). We require the upper $d \times d$ sub-matrix of $\mathbf{B}$ to $\mathbf{I}_d$ to ensure identifiability of $\mathbf{B}$. Thus estimating $\mathcal{S}_{Y|\mathbf{x}}$ is equivalent to estimating the lower $(p-d) \times d$ sub-matrix in $\mathbf{B}$. We use $\mathrm{vecl}(\mathbf{B})$ to denote the length $(p-d)d$ vector formed by the concatenation of the columns in lower $(p-d) \times d$ sub-matrix in $\mathbf{B}$, and use $\mathrm{vec}(\mathbf{X})$ to denote the concatenation of the columns of any matrix $\mathbf{X}$.

As showed in Ma and Zhu (2012a), they performed a generalized method of moments (GMM) treatment to reduce the number of estimating equations to $(p-d)d$, and obtain the simplified sample version of $\mathrm{E}(g^\star) = \mathbf{0}$ as

$$\sum_{i=1}^{n}\sum_{j=1}^{k}[l_j(Y_i)\{a_j(\mathbf{x}_i) - \mathrm{E}(a_j(\mathbf{x}_i)\,|\,\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\}] = \mathbf{0}. \tag{3.42}$$

Note that for complete case, $n = n_1$. We will directly use (3.42) in our discussion for CCEE, IPWEE and AIPWEE. The Analysis of (3.42) for $k > 1$ can be easily obtained after thorough study for $k = 1$. Therefore, we ignore the subscript $j$ in the sequel. In the $k = 1$ case, given

$$\widehat{\mathrm{g}}_i^\star = l(Y_i)\{a(\mathbf{x}_i) - \widehat{\mathrm{E}}(a(\mathbf{x}_i)\,|\,\mathbf{B}^{\mathrm{T}}\mathbf{x}_i)\},$$

we obtain the estimator $\widehat{\mathbf{B}}_{CC}$ for CCEE, $\widehat{\mathbf{B}}_{IPW}$ for IPWEE, and $\widehat{\mathbf{B}}_{AIPW}$ for AIPWEE by solving the following quations respectively,

$$\begin{gathered}
\sum_{i=1}^{n_1}\widehat{\mathrm{g}}_i^\star = \mathbf{0}, \\
\sum_{i=1}^{n}\left(\frac{\delta_i}{\widehat{\pi}(\mathbf{x}_i)}\widehat{\mathrm{g}}_i^\star\right) = \mathbf{0}, \\
\sum_{i=1}^{n}\left\{\frac{\delta_i}{\widehat{\pi}(\mathbf{x}_i)}\widehat{\mathrm{g}}_i^\star + \left(1 - \frac{\delta_i}{\widehat{\pi}(\mathbf{x}_i)}\right)\widehat{\mathrm{E}}\left(\widehat{\mathrm{g}}_i^\star \mid \mathbf{x}_i\right)\right\} = \mathbf{0},
\end{gathered}$$

where $\widehat{\mathrm{E}}(\cdot \mid \mathbf{x}_i)$ is estimated nonparametrically as stated in Section 3.4. We will display the following important theorems to illustrate the asymptotic properties of the two estimators. The technical details are demonstrated in Appendix C.

### 3.5.1 Asymptotic Property for Complete Case Estimating Equation Approach

**Theorem 3.1.** *Under regularity conditions, as $n \to \infty$, the estimator $\widehat{\mathbf{B}}_{cc}$ satisfies*

$$\sqrt{n_1}\text{vecl}(\widehat{\mathbf{B}}_{CC} - \mathbf{B}) \to \mathcal{N}(\mathbf{0}, \mathbf{A}_1^{-1}\mathbf{\Gamma}\mathbf{A}_1^{-1}),$$

*where*

$$\mathbf{A}_1 = \text{E}\left\{ \frac{\partial \text{vec}(\text{g}_{cc})}{\partial \text{vecl}(\mathbf{B})^T} \,\middle|\, \delta = 1 \right\} \quad \textit{and} \quad \mathbf{\Gamma} = \text{E}[\text{vec}(\text{g}_{cc})\{\text{vec}(\text{g}_{cc})\}^T \mid \delta = 1]$$

*with*

$$\text{g}_{cc} = \{l(Y) - \text{E}(l(Y) \mid \mathbf{B}^T\mathbf{x}, \delta = 1)\}\{a(\mathbf{x}) - \text{E}(a(\mathbf{x}) \mid \mathbf{B}^T\mathbf{x}, \delta = 1)\}.$$

The regularity conditions are provided in Appendix. The proof of Theorem 3.1 has the same fashion as the proof of Theorem 1 in Ma and Zhu (2012a) with complete data rather than the full data, and is neglected.

### 3.5.2 Properties of Inverse Probability Weighted Estimating Equation approaches

**Theorem 3.2.** *Under regularity conditions, when $n \to \infty$, $\widehat{\mathbf{B}}_{IPW}$ satisfies*

$$n^{1/2}\text{vecl}(\widehat{\mathbf{B}}_{IPW} - \mathbf{B}) \to \mathcal{N}(\mathbf{0}, \mathbf{A}^{-1}\mathbf{\Gamma}_1\mathbf{A}^{-1}), \tag{3.43}$$

*where*

$$\mathbf{A} = \text{E}\left\{ \frac{\partial \text{vec}(\text{g})}{\partial \text{vecl}(\mathbf{B})^T} \right\} \quad \textit{and} \quad \mathbf{\Gamma_1} = \text{Cov}\left[ \text{vec}\left\{ \frac{\delta}{\pi(\mathbf{x})}\text{g} \right\} \right]$$

*with*

$$\text{g} = \{l(Y) - \text{E}(l(Y) \,|\, \mathbf{B}^T\mathbf{x})\}\{a(\mathbf{x}) - \text{E}(a(\mathbf{x}) \,|\, \mathbf{B}^T\mathbf{x})\}.$$

**Theorem 3.3.** *Under regularity conditions, when* $n \to \infty$, $\widehat{\mathbf{B}}_{AIPW}$ *satisfies*

$$n^{1/2}\text{vecl}(\widehat{\mathbf{B}}_{AIPW} - \mathbf{B}) \to \mathcal{N}\{\mathbf{0}, \mathbf{A}^{-1}\mathbf{\Gamma}_2\mathbf{A}^{-1}\} \tag{3.44}$$

*where*

$$\mathbf{A} = \text{E}\left\{\frac{\partial \text{vec(g)}}{\partial \text{vecl}(\mathbf{B})^T}\right\} \text{ and } \mathbf{\Gamma_2} = \text{Cov}\left[\text{vec}\left\{\frac{\delta}{\pi(\mathbf{x})}\text{g} + \left(1 - \frac{\delta}{\pi(\mathbf{x})}\right)\text{E}(\text{g} \mid \mathbf{x})\right\}\right],$$

*with*

$$\text{g} = \{l(Y) - \text{E}(l(Y) \mid \mathbf{B}^T\mathbf{x})\}\{a(\mathbf{x}) - \text{E}(a(\mathbf{x}) \mid \mathbf{B}^T\mathbf{x})\}.$$

**Theorem 3.4.** *Under the conditions in Theorems 3.2 and 3.3, when* $n \to \infty$

$$n\left[\text{Cov}\{\text{vecl}(\widehat{\mathbf{B}}_{IPW})\} - \text{Cov}\{\text{vecl}(\widehat{\mathbf{B}}_{AIPW})\}\right]$$

*is positive definite.*

Technical details for Theorem 3.2, 3.3 and 3.4 are shown in B.1. Combining the results in Theorems 3.1, 3.2, 3.3 and 3.4. We can claim that $\widehat{\mathbf{B}}_{CC}$, $\widehat{\mathbf{B}}_{IPW}$ and $\widehat{\mathbf{B}}_{IPW}$ are unbiased estimators for $\mathbf{B}$, the basis of $\mathcal{S}_{Y|\mathbf{x}}$. And compared to $\widehat{\mathbf{B}}_{IPW}$, $\widehat{\mathbf{B}}_{AIPW}$ is a more efficient estimator, thus the AIPWEE framework will be a recommended estimating equation approach to solve sufficient dimension reduction with missing data problems.

## 3.6 Numerical Results

In this section, the simulation studies are conducted to evaluate the performance of the approaches. Each experiment is repeated 200 times with 200 sample size. Throughout the simulation, Epanechnikov kernel is applied with the default bandwidth selector implemented in Matlab routine `ksdensity`, which is $\widehat{\sigma}(3n/4)^{-1/(d+4)}$ with $\hat{\sigma}$ being the robust estimation of the standard

deviation of $\mathbf{B}^{\mathrm{T}}\mathbf{x}$. To ensure the uniquely mapping the central space $\mathcal{S}_{Y|\mathbf{x}}$ to one basis matrix/vector $\mathbf{B}$, here we only consider basis matrix/vector of $\mathcal{S}_{Y|\mathbf{x}}$ has the form $\mathbf{B} = (\mathbf{I}_d, \mathbf{B}_l^{\mathrm{T}})^{\mathrm{T}}$, where $\mathbf{B}_l$ is of dimension $(p-d) \times d$. For predictor $\mathbf{x} = (X_1, ..., X_p)$, the dimension $p$ is chosen to be 10 and the following two cases are considered.

Case (i): We generate $\mathbf{x} = (X_1, ..., X_p)^{\mathrm{T}}$ from multivariate normal with mean zero and covariance matrix $(\sigma_{ij})_{p\times p}$ where $\sigma_{ij} = 0.5^{|i-j|}$. In this case, the predictors satisfy both LCM (3.10) condition and CCV (3.11) conditions.

Case (ii): We again generate $\mathbf{x} = (X_1, ..., X_p)^{\mathrm{T}}$ from multivariate normal with mean zero and covariance matrix $(\sigma_{ij})_{p\times p}$ where $\sigma_{ij} = 0.5^{|i-j|}$. Then we regenerate $X_3$ and $X_4$ from nonlinear models $X_3 = |X_1 + X_2| + |X_1|\epsilon_1$ and $X_4 = |X_1 + X_2|^2 + |X_2|\epsilon_2$. Here for $i = 1, 2$, $\epsilon_i$ is independent from $X_1$ and $X_2$ and follows standard normal distribution. We also regenerate $X_5$ and $X_6$ independently from two Bernoulli distributions with success probability $\exp(X_2)/\{1+\exp(X_2)\}$ and $\Phi(X_2)$ respectively, where $\Phi(\cdot)$ denotes the cumulative distribution function of standard normal. The linear conditional mean assumption (4.1) is violated in this case.

For both cases, we generate the response using the following models,

$$\text{I: } Y = \exp(\boldsymbol{\beta}_1^{\mathrm{T}}\mathbf{x}) + .5\varepsilon,$$

$$\text{II: } Y = (\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x})^2 + (\boldsymbol{\beta}_3^{\mathrm{T}}\mathbf{x})^2 + .5\varepsilon,$$

$$\text{III :} Y = (\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x})(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x} + \boldsymbol{\beta}_3^{\mathrm{T}}\mathbf{x} + 3) + .5(\boldsymbol{\beta}_2^{\mathrm{T}}\mathbf{x})\varepsilon,$$

where $\mathbf{B}$ is a $p$-dimensional vector and $\varepsilon \sim \mathcal{N}(0, 1)$ is the random error term. For Model I, $d = 1$, we let $\boldsymbol{\beta}_1 = (1, 1/3, 1/3, \ldots, 1/3)^{\mathrm{T}}$. For Model II and III, $d = 2$, we let $\boldsymbol{\beta}_2 = (1, 0, 1/\sqrt{8}, 1/\sqrt{8}..., 1/\sqrt{8})^{\mathrm{T}}$ and $\boldsymbol{\beta}_3 = (0, 1, 1/\sqrt{8}, -1/\sqrt{8}..., -1/\sqrt{8})^{\mathrm{T}}$.

Models I to III are chosen to test OLS, PHD and SIR, respectively with their estimating equation counterparts. Consider the case when $Y$ is missing, we denote missingness indicator $\delta = 1$ when $Y$ is observed and $\delta = 0$ otherwise. We define the MAR missingness schemes as

$$P(\delta = 1 \mid \mathbf{x}) = \frac{\exp(c_1 + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})}{1 + \exp(c_1 + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x})}.$$

The missing proportion of the response can be controlled by adjusting $c$. Setting $\boldsymbol{\alpha} = (1, 1, 1, ..., 1)_p^{\mathrm{T}}/\sqrt{p}$. Let $c_1 = 1, 0, -1$, we compare results for 30%, 50% and 70% missing proportions.

### 3.6.1 Simulation Study

Denote $\widehat{\mathbf{B}}$ as the estimator of $\mathbf{B}$. Two criteria are used to evaluate the accuracy of $\widehat{\mathbf{B}}$. Following Li and Dong (2009), we can measure the distance between $\widehat{\mathbf{B}}$ and $\mathbf{B}$ as $\Delta = ||\widehat{\mathbf{P}} - \mathbf{P}||^2$, where $\widehat{\mathbf{P}} = \widehat{\mathbf{B}}(\widehat{\mathbf{B}}^{\mathrm{T}}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^{\mathrm{T}}$, $\mathbf{P} = \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}$, and $||\cdot||$ denotes the matrix Frobenius norm. Smaller values of $\Delta$ imply better estimation. We also report $R^2 = \mathrm{trace}(\widehat{\mathbf{P}}\mathbf{P}/d)$, which is always between 0 and 1, with values close to 1 implying better estimation. We refer to $\Delta$ and $R^2$ as the Euclidean distance and the trace correlation coefficient, respectively.

For each case, we examine the performance with or without incorporating the common SDR conditions, LCM (3.10) and CCV (3.11). We denote Full, CC, EE1, EE2 to represent the estimates with full data analysis, CCEE, IPWEE analysis and AIPWEE analysis respectively. The subscripts 1 and 2 refers to analysis with and without the common conditions. The boxplots of the Euclidean distances regrading to Model I to III are reported in Figure 3.1, 3.2, and 3.3, respectively. In each subfigure, the left panel presents the

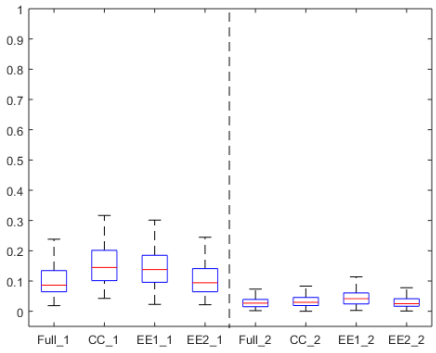
analysis incorporating the common conditions, while the right panel presents the analysis without the common conditions.

Focusing on case (i), within each panel, we can see that the two estimating equation estimators EE1 and EE2 show better performance than CC with smaller Euclidean distances in both panels. The worst performances of CC_1 for all models correspond to the possible bias caused by classical SDR methods under complete cases. Also it is noticed that the AIPWEE approach performs better than the IPWEE approach. The cross panel comparison with case (i) show that even when the common conditions hold for the model, avoiding them (right panels) result in better performance. This paradox coincides with the claim of efficiency loss caused by linearity condition in Ma and Zhu (2012a).

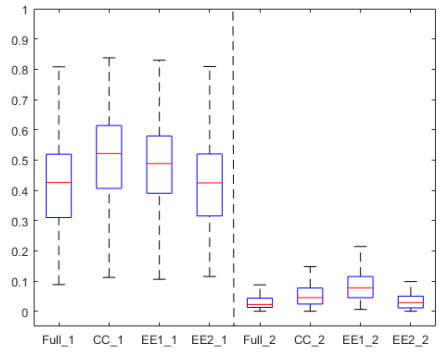
For case (ii), the cross panel comparison clearly demonstrate the invalidity of incorporating LCM and CCV conditions when they do not hold for the underlying predictor. And the observation confirms that the estimating equation estimators do not rely on the common conditions, which is not true for the left panel estimators. Again, within the right panels, we can see that the two estimating equation estimators EE1_2 and EE2_2 show better performance than CC_2 with smaller Euclidean distances. The exception is in model I, when missing proportion is relatively small, CC_2 gives slightly better results than EE1_2. Note that all estimators CC, EE1 and EE2 perform worse than the oracle Full estimator. Also it is noticed that the AIPWEE approach performs better than the IPWEE approach. Overall the good performance of CC_2 is reasonable as the estimators by solving the complete case estimating equations are unbiased. The same conclusions can be drawn by looking into the numbers for mean trace correlations displayed in Table 3.1, 3.2 and 3.3. Note that while

a smaller Euclidean distance in Figures means better estimation, larger value in the tables corresponds to more accurate estimators.

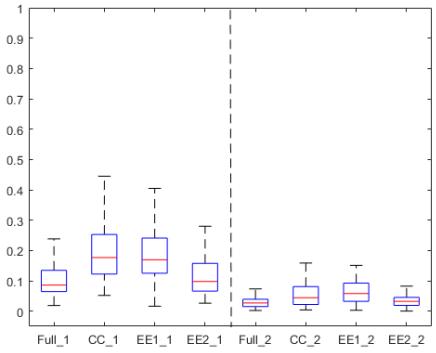
By varying $c_0 = -1, 0, 1$ to get different missing proportions, it reveals that the performance of CC_2 deteriorates with increasing missing proportion and may become significantly worse than Full_EE when the missing proportion is large. It is the same case for IPWEE estimators EE1_2, but slightly better than the CCEE estimators. On the other hand, the AIPWEE estimators EE2_1 under case (i), and EE2_2 for both cases are less sensitive to the missing proportion. Take EE2_2 as an example, in the challenging case of 70% missing response for case (i), the mean trace correlation is as high as $0.986, 0.970, 0.962$ for model I to III respectively; for case (ii), the mean trace correlation is as high as $0.989, 0.972, 0.940$ for model I to III with 70% missing proportion.



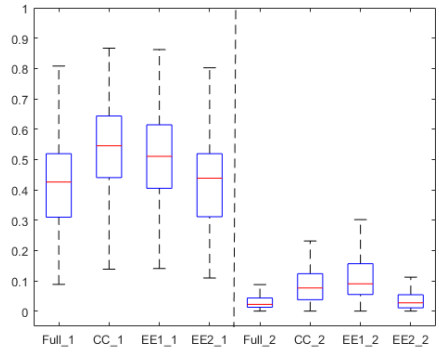

(a) Missing 30% Case (i)



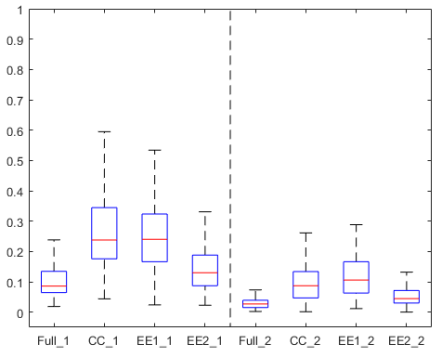

(b) Missing 30% Case (ii)



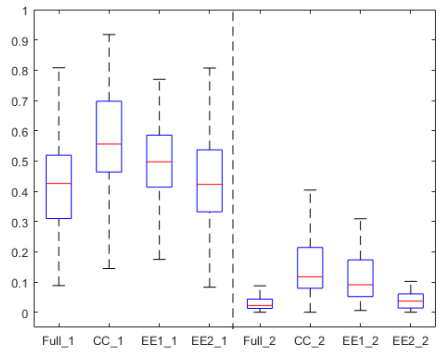

(c) Missing 50% Case (i)




(d) Missing 50% Case (ii)

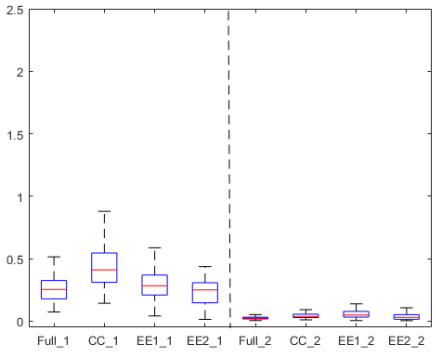



(e) Missing 70% Case (i)

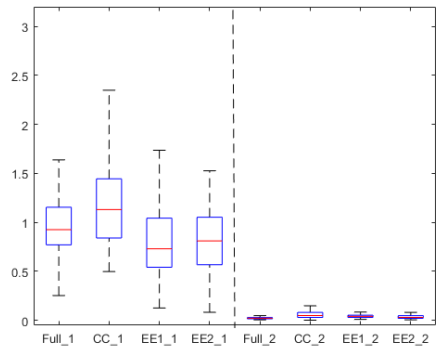



(f) Missing 70% Case (ii)

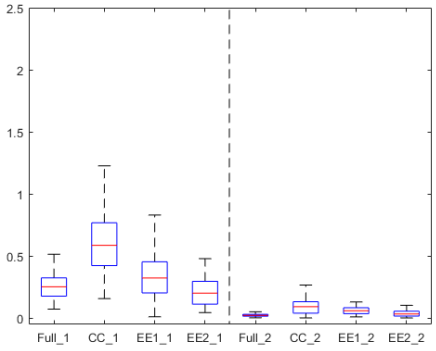
Figure 3.1: Boxplot of Euclidean distance for OLS estimating equations with different missing proportions (Model I)



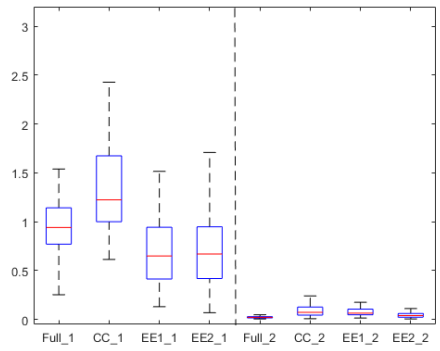

(a) Missing 30% Case (i)



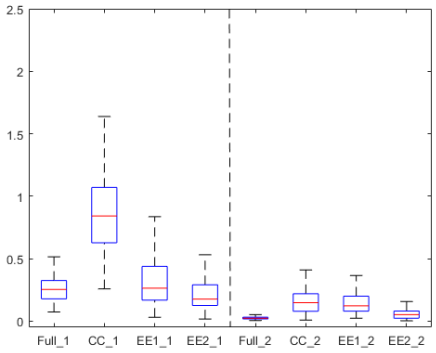

(b) Missing 30% Case (ii)



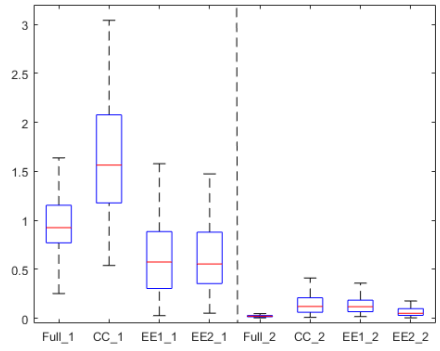

(c) Missing 50% Case (i)




(d) Missing 50% Case (ii)

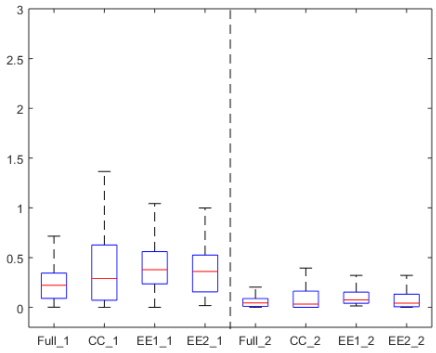



(e) Missing 70% Case (i)

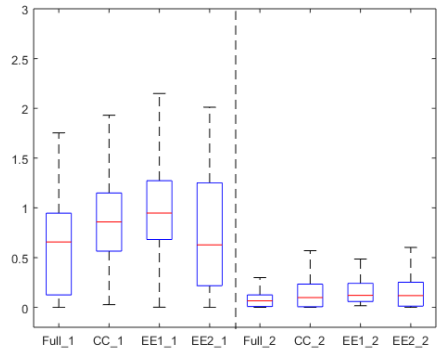



(f) Missing 70% Case (ii)

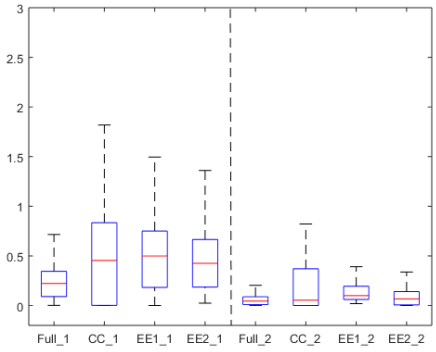
Figure 3.2: Boxplot of Euclidean distance for PHD estimating equations with different missing proportions (Model II)



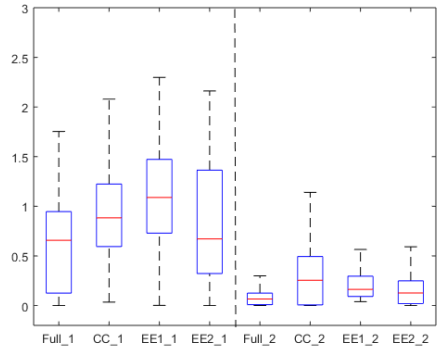

(a) Missing 30% Case (i)



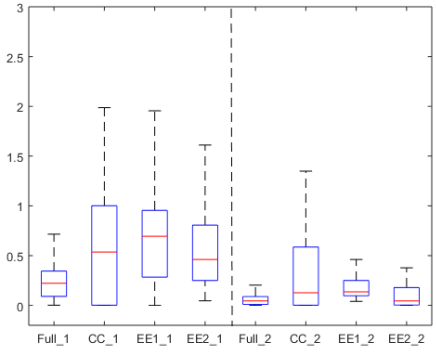

(b) Missing 30% Case (ii)



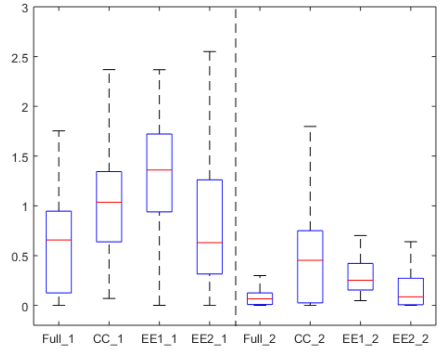

(c) Missing 50% Case (i)




(d) Missing 50% Case (ii)




(e) Missing 70% Case (i)




(f) Missing 70% Case (ii)

Figure 3.3: Boxplot of Euclidean distance for SIR estimating equations with different missing proportions (Model III)

Table 3.1: Mean and standard deviation of trace correlation coefficient $R^2$ for OLS estimating equations with different missing proportions (Model I).

| Case | Missing Proportion | | Full_1 | CC_1 | EE1_1 | EE2_1 | Full_2 | CC_2 | EE1_2 | EE2_2 |
|---|---|---|---|---|---|---|---|---|---|---|
| (i) | 30% | ave | 0.973 | 0.956 | 0.961 | 0.972 | 0.993 | 0.992 | 0.989 | 0.992 |
| | | std | 0.018 | 0.024 | 0.021 | 0.018 | 0.005 | 0.006 | 0.008 | 0.006 |
| | 50% | ave | 0.973 | 0.948 | 0.950 | 0.970 | 0.993 | 0.987 | 0.984 | 0.991 |
| | | std | 0.018 | 0.027 | 0.026 | 0.020 | 0.005 | 0.009 | 0.010 | 0.006 |
| | 70% | ave | 0.973 | 0.945 | 0.941 | 0.940 | 0.993 | 0.976 | 0.967 | 0.986 |
| | | std | 0.018 | 0.036 | 0.033 | 0.022 | 0.005 | 0.017 | 0.027 | 0.010 |
| (ii) | 30% | ave | 0.886 | 0.880 | 0.867 | 0.886 | 0.993 | 0.986 | 0.975 | 0.992 |
| | | std | 0.042 | 0.040 | 0.042 | 0.043 | 0.006 | 0.014 | 0.026 | 0.007 |
| | 50% | ave | 0.886 | 0.871 | 0.869 | 0.885 | 0.993 | 0.978 | 0.967 | 0.991 |
| | | std | 0.042 | 0.042 | 0.044 | 0.035 | 0.006 | 0.015 | 0.033 | 0.009 |
| | 70% | ave | 0.886 | 0.873 | 0.864 | 0.886 | 0.993 | 0.959 | 0.962 | 0.989 |
| | | std | 0.042 | 0.046 | 0.046 | 0.044 | 0.006 | 0.031 | 0.046 | 0.011 |

Table 3.2: Mean and standard deviation of trace correlation coefficient $R^2$ for PHD estimating equations with different missing proportions (Model II).

| Case | Missing Proportion | | Full_1 | CC_1 | EE1_1 | EE2_1 | Full_2 | CC_2 | EE1_2 | EE2_2 |
|---|---|---|---|---|---|---|---|---|---|---|
| (i) | 30% | ave | 0.934 | 0.892 | 0.928 | 0.939 | 0.994 | 0.989 | 0.984 | 0.987 |
| | | std | 0.029 | 0.041 | 0.029 | 0.030 | 0.004 | 0.006 | 0.016 | 0.010 |
| | 50% | ave | 0.934 | 0.843 | 0.915 | 0.945 | 0.994 | 0.983 | 0.976 | 0.982 |
| | | std | 0.029 | 0.081 | 0.047 | 0.034 | 0.004 | 0.012 | 0.018 | 0.012 |
| | 70% | ave | 0.934 | 0.776 | 0.915 | 0.943 | 0.994 | 0.962 | 0.957 | 0.970 |
| | | std | 0.029 | 0.099 | 0.061 | 0.046 | 0.004 | 0.026 | 0.035 | 0.021 |
| (ii) | 30% | ave | 0.747 | 0.694 | 0.798 | 0.792 | 0.994 | 0.989 | 0.984 | 0.987 |
| | | std | 0.105 | 0.123 | 0.099 | 0.097 | 0.003 | 0.005 | 0.014 | 0.009 |
| | 50% | ave | 0.747 | 0.654 | 0.807 | 0.807 | 0.994 | 0.980 | 0.978 | 0.982 |
| | | std | 0.105 | 0.136 | 0.111 | 0.114 | 0.003 | 0.016 | 0.017 | 0.011 |
| | 70% | ave | 0.747 | 0.599 | 0.830 | 0.830 | 0.994 | 0.966 | 0.965 | 0.972 |
| | | std | 0.105 | 0.140 | 0.114 | 0.120 | 0.003 | 0.020 | 0.024 | 0.017 |

Table 3.3: Mean and standard deviation of trace correlation coefficient $R^2$ for SIR estimating equations with different missing proportions (Model III).

| Case | Missing Proportion | | Full_1 | CC_1 | EE1_1 | EE2_1 | Full_2 | CC_2 | EE1_2 | EE2_2 |
|---|---|---|---|---|---|---|---|---|---|---|
| (i) | 30% | ave | 0.939 | 0.905 | 0.898 | 0.905 | 0.985 | 0.975 | 0.969 | 0.977 |
| | | std | 0.051 | 0.072 | 0.071 | 0.090 | 0.016 | 0.035 | 0.034 | 0.032 |
| | 50% | ave | 0.939 | 0.860 | 0.875 | 0.884 | 0.985 | 0.943 | 0.964 | 0.975 |
| | | std | 0.051 | 0.124 | 0.100 | 0.086 | 0.016 | 0.086 | 0.032 | 0.029 |
| | 70% | ave | 0.939 | 0.851 | 0.836 | 0.854 | 0.985 | 0.915 | 0.943 | 0.962 |
| | | std | 0.051 | 0.136 | 0.120 | 0.110 | 0.016 | 0.112 | 0.088 | 0.068 |
| (ii) | 30% | ave | 0.848 | 0.785 | 0.757 | 0.816 | 0.985 | 0.962 | 0.956 | 0.957 |
| | | std | 0.114 | 0.107 | 0.121 | 0.146 | 0.027 | 0.043 | 0.045 | 0.052 |
| | 50% | ave | 0.848 | 0.764 | 0.732 | 0.791 | 0.977 | 0.924 | 0.942 | 0.945 |
| | | std | 0.114 | 0.129 | 0.134 | 0.156 | 0.027 | 0.078 | 0.056 | 0.105 |
| | 70% | ave | 0.848 | 0.740 | 0.679 | 0.804 | 0.977 | 0.884 | 0.917 | 0.940 |
| | | std | 0.114 | 0.131 | 0.152 | 0.167 | 0.027 | 0.114 | 0.081 | 0.112 |

### 3.6.2 Real Data Application

We compare complete-case estimating equation and inverse probability weighted estimating equation approaches through a real data example. Consider the mussel data with $n = 82$ observations collected in an ecological study of New Zealand. The response is the mussel's muscle mass in g. The four predictors are the shell mass $S$ in g, the shell length $L$, the shell height $Ht$, and the shell width $W$ in mm. This data set is available in $R$ under the `dr` package. The data set has been studied in Bura and Cook (2001a) and Dong et al. (2015), and the structural dimension is estimated to be 1. The scatterplot matrix of the predictors is provided in the left panel of Figure 4. With $d = 1$, we first estimate $\boldsymbol{\beta}$ with the sample version of estimating equation (3.9) based on all 82 observations. The corresponding estimator is $\widehat{\boldsymbol{\beta}}_0 = (.385, 0.804, .442, .104)^{\mathrm{T}}$, which is the Full_2 estimator in the previous section. We refer to $\widehat{\boldsymbol{\beta}}_0^{\mathrm{T}}\mathbf{x}$. as the oracle predictor. The right panel of Figure 4 is the sufficient plot of the response versus the oracle predictor.

Denote $\delta$ as the missingness indicator of the response $Y$. In each repetition,

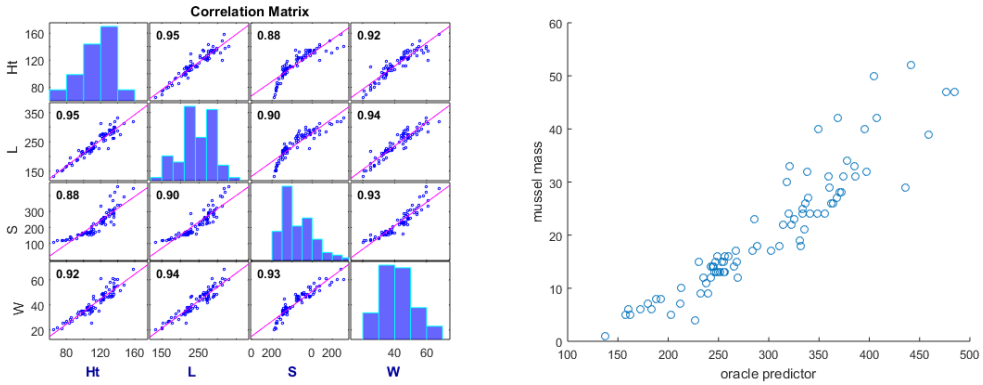



Figure 3.4: Left panel: the scatterplot matrix of the predictor. Right panel: the sufficient plot of muscle mass versus the oracle predictor $\widehat{\boldsymbol{\beta}}_0^{\mathrm{T}}\mathbf{x}$.

we generate $\delta_i$ from the Bernoulli distribution with success ($Y_i$ observed) probability $p(\mathbf{x}_i) = \{\exp(c_0 + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\} / \{1 + \exp(c_0 + \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{x}_i)\}$ for $i = 1, \ldots, 82$. Here we set $\boldsymbol{\alpha} = (0.4, 0.3, 0.2, 0.1)^{\mathrm{T}}$ and $c_0 = -1$. We treat $Y_i$ as observed if $\delta_i = 1$, and $Y_i$ is considered missing if $\delta_i = 0$. Based on $\{(\mathbf{x}_i, Y_i, \delta_i), i = 1, \ldots, 82\}$, we calculate the complete-case estimating equation estimator (the CC_2 estimator) and denote it as $\widehat{\boldsymbol{\beta}}_1$. We also get the IPWEE (EE1_2) and AIPWEE (EE2_2) estimators and denote them as $\widehat{\boldsymbol{\beta}}_2$ and $\widehat{\boldsymbol{\beta}}_3$. For $j = 1, 2, 3$, calculate $\Delta_j = ||\widehat{\mathbf{P}}_0 - \widehat{\mathbf{P}}_j||^2$, where $\widehat{\mathbf{P}}_0 = \widehat{\boldsymbol{\beta}}_0(\widehat{\boldsymbol{\beta}}_0^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_0)^{-1}\widehat{\boldsymbol{\beta}}_0^{\mathrm{T}}$ and $\widehat{\mathbf{P}}_j = \widehat{\boldsymbol{\beta}}_j(\widehat{\boldsymbol{\beta}}_j^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_j)^{-1}\widehat{\boldsymbol{\beta}}_j^{\mathrm{T}}$. Based on 200 repetitions, the average missing proportion is 72.5%, the average $\Delta_1$ value for the CCEE estimator $\widehat{\boldsymbol{\beta}}_1$ is 0.129, the average $\Delta_2$ value for the IPWEE estimator $\widehat{\boldsymbol{\beta}}_2$ is 0.072, and the average $\Delta_3$ value for the AIPWEE estimator $\widehat{\boldsymbol{\beta}}_3$ is 0.056. The result confirms our findings in the simulation section that EE1_2 and EE2_2 have better performance than CC_2 when the missing proportion is large.

# CHAPTER 4

# ON A NEW HYBRID ESTIMATOR FOR THE CENTRAL MEAN SPACE

As aforementioned, SIR and OLS can be categorized as first-order methods, which depend on linear functions of $\mathbf{x}$ such as $\mathrm{E}(\mathbf{x} \mid Y)$ and $\mathrm{E}(Y\mathbf{x})$. On the other hand, SAVE and PHD belong to the second-order methods, which involve both linear and quadratic functions for $\mathbf{x}$, such as $\mathrm{E}(\mathbf{x} \mid Y)$, $\mathrm{E}(\mathbf{x}\mathbf{x}^{\mathrm{T}} \mid Y)$, and $\mathrm{E}(Y\mathbf{x}\mathbf{x}^{\mathrm{T}})$. First-order methods do not work well when the link function between the response and the predictors is symmetric about the origin, while second-order methods are not efficient with linear link functions between $Y$ and $\mathbf{x}$. See, for example, Dong and Li (2010) and Dong (2016) for a nice summary. Due to the complementary nature of the first-order and the second-order methods, hybrid moment-based SDR estimators have been proposed in the literature. See, for example, Gannoun and Saracco (2003), Ye and

Weiss (2003), Zhu et al. (2007), Li and Wang (2007), Shanker and Prendergast (2011), and Yu et al. (2014). With the exception of Shanker and Prendergast (2011), the existing hybrid methods all focus on estimating the central space. Shanker and Prendergast (2011) proposed an iterative algorithm, which can be applied to estimators of the central space as well as estimators of the central mean space. To the best knowledge, there is no non-iterative hybrid estimator in the literature that exclusively focuses on estimating the central mean space.

The main contribution of this chapter (Xia and Dong (2017)) is to propose a new hybrid estimator of the central mean space, such that it works well for linear as well as symmetric link functions. The dimensionality of the central mean space is denoted by $d$, which is generally unknown in practice. The procedure to estimate $d$ is called order determination. The second contribution of the paper is to propose an order determination procedure based on the newly proposed hybrid estimator. Last but not least, following the idea of Cook (2004), the proposed hybrid method naturally leads to a new marginal coordinate test. Without assuming any parametric models, test is conducted for the contribution of an individual predictor to the regression mean in the presence of all the other predictors.

Without loss of generality, we assume $\mathrm{E}(Y) = 0$ and $\mathrm{E}(\mathbf{x}) = \mathbf{0}$ throughout this chapter. Let $\mathbf{\Sigma} = \mathrm{Var}(\mathbf{x})$ denote the covariance matrix of $\mathbf{x}$, and $\mathbf{B} \in \mathbb{R}^{p\times d}$ denotes the basis of the central mean space.

## 4.1 A New Hybrid Estimator for The Central Mean Space

### 4.1.1 Population Level Development

Before introducing the new hybrid estimator, we first review OLS and PHD as estimators of the central mean space. The OLS estimator can be expressed as $\boldsymbol{\beta}_{\text{OLS}} = \boldsymbol{\Sigma}^{-1}\text{E}(\mathbf{x}Y)$. The following condition was first proposed in Li and Duan (1989),

$$\text{E}(\mathbf{x} \mid \mathbf{B}^{\text{T}}\mathbf{x}) \text{ is linear in } \mathbf{B}^{\text{T}}\mathbf{x}. \tag{4.1}$$

We refer to (4.1) as the linear conditional mean (LCM) condition. Under the LCM condition, Li and Duan (1989) proved that $\boldsymbol{\beta}_{\text{OLS}} \in \mathcal{S}_{Y|\mathbf{x}}$. In addition to the LCM condition, the following condition is needed for PHD,

$$\text{Cov}(\mathbf{x} \mid \mathbf{B}^{\text{T}}\mathbf{x}) \text{ is a nonrandom matrix.} \tag{4.2}$$

We refer to (4.2) as the constant conditional variance (CCV) condition. Define matrix $\mathbf{M}_{\text{PHD}} = \boldsymbol{\Sigma}^{-1}\text{E}(Y\mathbf{x}\mathbf{x}^{\text{T}})$. Under both the LCM and the CCV conditions, Li (1992) proved that $\text{span}(\mathbf{M}_{\text{PHD}}) \subseteq \mathcal{S}_{Y|\mathbf{x}}$, where $\text{span}(\mathbf{M}_{\text{PHD}})$ denotes the column space of $\mathbf{M}_{\text{PHD}}$. Although OLS and PHD were initially proposed as estimators of the central space, Cook and Li (2002) revealed later that they are estimators of the central mean space. It is now well-known that $\boldsymbol{\beta}_{\text{OLS}} \in \mathcal{S}_{\text{E}(Y|\mathbf{x})} \subseteq \mathcal{S}_{Y|\mathbf{x}}$ with the LCM condition, and $\text{span}(\mathbf{M}_{\text{PHD}}) \subseteq \mathcal{S}_{\text{E}(Y|\mathbf{x})} \subseteq \mathcal{S}_{Y|\mathbf{x}}$ when both the LCM and the CCV conditions are satisfied.

As estimators of the central mean space, OLS and PHD have their limitations. OLS performs well with monotone link functions but fails with sym-

metric link functions. PHD, on the other hand, does not work well when the link function is linear. This motivates us to consider a hybrid estimator that simply concatenates $\boldsymbol{\beta}_{\mathrm{OLS}}$ and $\mathbf{M}_{\mathrm{PHD}}$. Specifically, we define $\mathbf{M} \in \mathbb{R}^{p\times(p+1)}$ as

$$\mathbf{M} = (\boldsymbol{\beta}_{\mathrm{OLS}}, \mathbf{M}_{\mathrm{PHD}}) = \boldsymbol{\Sigma}^{-1}\left(\mathrm{E}(Y\mathbf{x}), \mathrm{E}(Y\mathbf{x}\mathbf{x}^{\mathrm{T}})\right) = \boldsymbol{\Sigma}^{-1}\mathrm{E}\left(Y\mathbf{x}(\mathbf{x}^{*})^{\mathrm{T}}\right), \quad (4.3)$$

where $\mathbf{x}^{*} = (1, \mathbf{x}^{\mathrm{T}})^{\mathrm{T}}$. The next result states that the column space of $\mathbf{M}$ can be used to recover the central mean space. We omit its proof, which follows directly from the well-known properties of the OLS and the PHD estimators.

**Theorem 4.1.** *Suppose* $\mathrm{E}(Y) = 0$, $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, *and all the moments involved exist. Under the LCM condition (4.1) and the CCV condition (4.2), we have* $span(\mathbf{M}) \subseteq \mathcal{S}_{\mathrm{E}(Y|\mathbf{x})} \subseteq \mathcal{S}_{Y|\mathbf{x}}$.

### 4.1.2 Sample Estimator

Given an i.i.d sample $\{(\mathbf{x}_1, Y_1), \cdots, (\mathbf{x}_n, Y_n)\}$, the sample level algorithm becomes

1. Calculate $\bar{\mathbf{x}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$, $\bar{Y} = n^{-1}\sum_{i=1}^{n}Y_i$, and $\widehat{\boldsymbol{\Sigma}} = n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}$.

2. Center the data $\widetilde{\mathbf{x}}_i = \mathbf{x}_i - \bar{\mathbf{x}}$ and $\widetilde{Y}_i = Y_i - \bar{Y}$. Denote $\widetilde{\mathbf{x}}_i^{*} = (1, \widetilde{\mathbf{x}}_i^{\mathrm{T}})^{\mathrm{T}}$, $i = 1, \ldots, n$.

3. Calculate $\widehat{\mathbf{M}} = \widehat{\boldsymbol{\Sigma}}^{-1}n^{-1}\sum_{i=1}^{n}\widetilde{Y}_i\widetilde{\mathbf{x}}_i(\widetilde{\mathbf{x}}_i^{*})^{\mathrm{T}}$ and $\widehat{\mathbf{F}} = \widehat{\mathbf{M}}\widehat{\mathbf{M}}^{\mathrm{T}}$.

4. Perform eigenvalue decomposition of $\widehat{\mathbf{F}}$. Denote $\widehat{\mathbf{B}}_1, \ldots, \widehat{\mathbf{B}}_d$ as the eigenvectors corresponding to the $d$ largest eigenvalues of $\widehat{\mathbf{F}}$. Then $\widehat{\mathbf{B}} = (\widehat{\mathbf{B}}_1, \ldots, \widehat{\mathbf{B}}_d) \in \mathbb{R}^{p\times d}$ is the final estimator of $\mathbf{B}$.

In step 4 of the algorithm above, we assume $d$ is known. The procedure to determine $d$ will be discussed later in Section 3. Let $\mathbf{F} = \mathbf{M}\mathbf{M}^{\mathrm{T}}$ be the population counterpart of $\widehat{\mathbf{F}}$ in step 3. From the definition of $\mathbf{M}$ in (4.3), we see that $\mathbf{F} = \mathbf{F}_{\mathrm{OLS}} + \mathbf{F}_{\mathrm{PHD}}$, where $\mathbf{F}_{\mathrm{OLS}} = \boldsymbol{\beta}_{\mathrm{OLS}}\boldsymbol{\beta}_{\mathrm{OLS}}^{\mathrm{T}}$ and $\mathbf{F}_{\mathrm{PHD}} = \mathbf{M}_{\mathrm{PHD}}\mathbf{M}_{\mathrm{PHD}}^{\mathrm{T}}$. Thus our proposed hybrid estimator can be viewed as a convex combination of the OLS kernel matrix $\mathbf{F}_{\mathrm{OLS}}$ and the PHD kernel matrix $\mathbf{F}_{\mathrm{PHD}}$, where each component has equal weight. Convex combination methods, among others, have been studied in Ye and Weiss (2003) and Zhu et al. (2007). While these existing combination methods focus on estimating the central space, our estimator is proposed to recover the central mean space.

Let $\mathbf{A} = \boldsymbol{\Sigma}^{-1}Y\mathbf{x}(\mathbf{x}^*)^{\mathrm{T}} - \boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathrm{E}(Y\mathbf{x}(\mathbf{x}^*)^{\mathrm{T}})$. We conclude this section with the asymptotic distribution of $\mathrm{vec}(\widehat{\mathbf{M}})$, where $\mathrm{vec}(\widehat{\mathbf{M}})$ means the concatenation of the columns of $\widehat{\mathbf{M}}$.

**Theorem 4.2.** *Suppose* $\mathrm{E}(Y) = 0$, $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, *and all the moments involved exist. Let* $\boldsymbol{\Gamma} = \mathrm{E}\left(\mathrm{vec}(\mathbf{A})\mathrm{vec}^{T}(\mathbf{A})\right)$. *Then*

$$\sqrt{n}\left(\mathrm{vec}(\widehat{\mathbf{M}}) - \mathrm{vec}(\mathbf{M})\right) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Gamma}),$$

*where "*$\to$*" means converge in distribution.*

### 4.1.3 Hybrid Estimator When Response is Missing at Random

Recall the discussions in Chapter 2 for the validity of complete case OLS and PHD when response is missing at random. The conclusion can pass to the proposed hybrid estimator under the same missing scheme. In other words, it can be shown that $\boldsymbol{\Sigma}^{-1}\mathrm{E}(Y\mathbf{x}(\mathbf{x}^*) \mid \delta = 1)$ is invalid to recover central mean

space. Thus in this section, we propose a simple and valid inverse probability weighted hybrid estimator in a similar fashion as in Chapter 2. The IPW hybrid estimator is denoted as

$$\mathbf{M}^{\delta} = \mathbf{\Sigma}^{-1}\mathrm{E}\left\{\frac{\delta}{\pi(\mathbf{x})}Y\mathbf{x}(\mathbf{x}^*)\right\} \tag{4.4}$$

where $\delta$ is the missing indicator with value 1(0) representing observed (missing). And under the missing at random scheme, $\delta \perp\!\!\!\perp Y \mid \mathbf{x}$, thus the propensity score $\pi(\mathbf{x}) = P(\delta = 1 \mid \mathbf{x})$. It can be shown that $\mathbf{M}^{\delta} \subseteq \mathcal{S}_{\mathrm{E}(Y|\mathbf{x})}$ with the same conditions listed in Theorem 4.1.

Given an i.i.d sample $\{(\mathbf{x}_1, Y_1, \delta_1), \cdots, (\mathbf{x}_n, Y_n, \delta_n)\}$, the sample estimator of $\mathbf{M}^{\delta}$ becomes $\widehat{\mathbf{M}}^{\delta} = \widehat{\mathbf{\Sigma}}^{-1} n^{-1} \sum_{i=1}^{n} \delta_i \widehat{\pi}^{-1}(\mathbf{x}_i) \widetilde{Y}_i \widetilde{\mathbf{x}}_i (\widetilde{\mathbf{x}}_i^*)^{\mathrm{T}}$ and $\widehat{\mathbf{F}}^{\delta} = \widehat{\mathbf{M}}^{\delta}(\widehat{\mathbf{M}}^{\delta})^{\mathrm{T}}$, where $\widehat{\mathbf{\Sigma}}^{-1}$, $\widetilde{Y}_i$ and $\widetilde{\mathbf{x}}_i$ are calculated the same as in Section 4.1.2, and $\widehat{\pi}(\mathbf{x}_i)$ is estimated nonparametrically as in (3.30). Then perform eigenvalue decomposition of $\widehat{\mathbf{F}}^{\delta}$ to achieve the final estimator of $\mathbf{B}$.

## 4.2 Sequential Tests for Order Determination

In the sample level algorithm from Section 2, we have assumed that the dimension of the central mean space $d$ is known. In practice, the structural dimension $d$ is unknown and has to be estimated. The procedure of determining $d$ is called order determination. We follow the development of Li (1991), Cook and Li (2004b), and Li and Wang (2007), and develop a sequential test approach for order determination in this section.

Note that $d = \mathrm{rank}(\mathbf{F})$ for $\mathbf{F} = \mathbf{M}\mathbf{M}^{\mathrm{T}}$. We consider the following hypothe-

ses,

$$H_0^{(\ell)} : \text{rank}(\mathbf{F}) = \ell \text{ v.s. } H_a^{(\ell)} : \text{rank}(\mathbf{F}) > \ell \text{ for } \ell = 1, ..., p-1.$$

The rank of $\mathbf{F}$ is then estimated as the first $\ell$ for which $H_0^{(\ell)}$ is accepted. If $H_0^{(\ell)}$ is accepted for some $0 \leq \ell \leq p-1$, we have $\hat{d} = \text{argmin}\{\ell : H_0^{(\ell)} \text{ is accepted}\}$. If $H_0^{(\ell)}$ is rejected for all $0 \leq \ell \leq p-1$, we estimate $d$ by $\hat{d} = p$. Let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_p$ be the eigenvalues of $\widehat{\mathbf{F}} = \widehat{\mathbf{M}}\widehat{\mathbf{M}}^{\mathrm{T}}$. The test statistic for $H_0^{(\ell)}$ is

$$T_\ell = n \sum_{j=\ell+1}^{p} \hat{\lambda}_j, \tag{4.5}$$

and we reject $H_0^{(\ell)}$ for large values of $T_\ell$.

Following Li and Wang (2007), we study the asymptotic distribution of $T_\ell$ next. Under $H_0^{(\ell)}$, $\mathbf{M}$ has rank $\ell$. Denote the singular value decomposition of $\mathbf{M}$ under $H_0^{(\ell)}$ as

$$\mathbf{M} = \begin{pmatrix} \mathbf{U}_1 & \mathbf{U}_0 \end{pmatrix} \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V}_1^{\mathrm{T}} \\ \mathbf{V}_0^{\mathrm{T}} \end{pmatrix},$$

where $\mathbf{U}_1 \in \mathbb{R}^{p\times\ell}$ and $\mathbf{U}_0 \in \mathbb{R}^{p\times(p-\ell)}$ form the $p \times p$ othogonal matrix $(\mathbf{U}_1, \mathbf{U}_0)$, $\mathbf{V}_1 \in \mathbb{R}^{(p+1)\times\ell}$ and $\mathbf{V}_0 \in \mathbb{R}^{(p+1)\times(p+1-\ell)}$ form the $(p+1) \times (p+1)$ orthogonal matrix $(\mathbf{V}_1, \mathbf{V}_0)$, and $\mathbf{D}$ is an $\ell \times \ell$ diagonal matrix with positive diagonal elements. The asymptotic distribution of $T_\ell$ under $H_0^{(\ell)}$ is provided in the following theorem.

**Theorem 4.3.** *Suppose* $\mathrm{E}(Y) = 0$, $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, *and all the moments involved exist. Let* $\mathbf{H} = \mathbf{U}_0^T\mathbf{A}\mathbf{V}_0$, *and* $\mathbf{\Lambda} = \mathrm{E}\left(\text{vec}(\mathbf{H})\text{vec}^T(\mathbf{H})\right)$. *Then under* $H_0^{(\ell)}$, $T_\ell$ *converges in distribution to*

$$T_\ell \to \sum_{j=1}^{(p-l)(p+1-l)} \nu_j \chi_j^2(1),$$

*where "$\rightarrow$" means converge in distribution, $\nu_j$'s are the eigenvalues of* $\mathbf{\Lambda}$*, and* $\chi^2_j(1)$*'s are i.i.d.* $\chi^2(1)$ *random variables.*

The asymptotic variance $\mathbf{\Lambda}$ needs to be estimated in practice. Recall that $\widehat{\mathbf{\Sigma}} = n^{-1}\sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^{\mathrm{T}}$. Denote $\widehat{\mathbf{A}}_i = \widehat{\mathbf{\Sigma}}^{-1}\widetilde{Y}_i\widetilde{\mathbf{x}}_i(\widetilde{\mathbf{x}}_i^*)^{\mathrm{T}} - \widehat{\mathbf{\Sigma}}^{-1}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^T\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Sigma}}^*_{\mathrm{Yxx}}$, where $\widehat{\mathbf{\Sigma}}^*_{\mathrm{Yxx}} = n^{-1}\sum_{i=1}^{n}\widetilde{Y}_i\widetilde{\mathbf{x}}_i(\widetilde{\mathbf{x}}_i^*)^{\mathrm{T}}$. Let $\widehat{\mathbf{U}}_0$ and $\widehat{\mathbf{V}}_0$ be the sample estimators of $\mathbf{U}_0$ and $\mathbf{V}_0$, and let $\widehat{\mathbf{H}}_i = \widehat{\mathbf{U}}_0^{\mathrm{T}}\widehat{\mathbf{A}}_i\widehat{\mathbf{V}}_0$. Then the sample estimator of $\mathbf{\Lambda}$ becomes $\widehat{\mathbf{\Lambda}} = n^{-1}\sum_{i=1}^{n}\mathrm{vec}(\widehat{\mathbf{H}}_i)\mathrm{vec}^T(\widehat{\mathbf{H}}_i)$. The asymptotic distribution of $T_\ell$ under $H_0^{(\ell)}$ can then be approximated by $\sum_{j=1}^{(p-l)(p+1-l)}\hat{\nu}_j\chi^2_j(1)$, where $\hat{\nu}_j$'s are the eigenvalues of $\widehat{\mathbf{\Lambda}}$. Denote $(p-l)(p+1-l)$ by $s$. Let $\boldsymbol{\nu} = (\hat{\nu}_1, \ldots, \hat{\nu}_s)^{\mathrm{T}}$ and $\boldsymbol{W}$ be an $N \times s$ matrix of i.i.d. $\chi^2(1)$ realizations. Then the $N$ elements of $\boldsymbol{W\nu}$ become realizations of the approximate $T_\ell$ distribution, and the proportion of $T_\ell$ in (4.5) greater than these $N$ elements is the approximate $p$-value. We use $N = 500$ in our simulation studies.

## 4.3 Testing Predictor Effects

Without assuming any parametric models, the hybrid estimator provides a unique opportunity to test the contribution of an individual predictor to the regression mean in the presence of all the other predictors. Recall that $\mathbf{B} \in \mathbb{R}^{p\times d}$ denotes the basis of the central mean space. For $k = 1, ..., p$, define $\mathbf{e}_k \in \mathbb{R}^p$, where its $k$th element is one and all the other elements are zero. To test the effect of the $k$th predictor $X_k$, consider

$$H_0^{[k]} : \mathbf{e}_k^{\mathrm{T}}\mathbf{B} = \mathbf{0} \text{ v.s. } H_a^{[k]} : \mathbf{e}_k^{\mathrm{T}}\mathbf{B} \neq \mathbf{0}. \tag{4.6}$$

Let $\mathbf{x}_{-k} = (X_1, \ldots, X_{k-1}, X_{k+1}, \ldots, X_p)^{\mathrm{T}}$. It can be shown that $\mathbf{e}_k^{\mathrm{T}}\mathbf{B} = \mathbf{0}$ if and only if $Y \perp\!\!\!\perp \mathrm{E}(Y \mid \mathbf{x}) \mid \mathbf{x}_{-k}$. Thus $H_0^{[k]}$ in (4.6) implies that $X_k$ is not

contributing to the regression mean given all the other predictors. Hypotheses (4.6) is a special case of the marginal coordinate hypotheses. The SIR-based and the SAVE-based marginal coordinate tests are studied in Cook (2004) and Shao and Cook (2007), respectively. A hybrid of SIR-based and SAVE-based procedure is proposed in Zhou and Dong (2016). While these existing procedures focus on testing the predictor contribution to the central space, hypotheses (4.6) aims at testing the predictor contribution to the central mean space.

The proposed hybrid estimator $\mathbf{M}$ satisfies $\text{span}(\mathbf{M}) \subseteq \text{span}(\mathbf{B}) = \mathcal{S}_{\text{E}(Y|\mathbf{x})}$. Under $\text{H}_0^{[k]}$, we have $\mathbf{e}_k^{\text{T}}\mathbf{M}\mathbf{M}^{\text{T}}\mathbf{e}_k = \mathbf{e}_k^{\text{T}}\mathbf{F}\mathbf{e}_k = 0$. A natural test statistic for (4.6) becomes $G_k = n\mathbf{e}_k^{\text{T}}\widehat{\mathbf{F}}\mathbf{e}_k$, and we reject $\text{H}_0^{[k]}$ for large values of $G_k$. The asymptotic distribution of $G_k$ under $\text{H}_0^{[k]}$ is provided next.

**Theorem 4.4.** *Suppose* $\text{E}(Y) = 0$, $\text{E}(\mathbf{x}) = \mathbf{0}$, *and all the moments involved exist. Furthermore, suppose the LCM condition (1.3) and the CCV condition (1.4) hold. Let* $\mathbf{C}_k = \text{E}(\mathbf{A}^T\mathbf{e}_k\mathbf{e}_k^T\mathbf{A})$. *Then under* $\text{H}_0^{[k]}$,

$$G_k \to \sum_{h=1}^{p+1} \omega_{k,h}\chi_h^2(1),$$

*where "*$\to$*" means converge in distribution,* $\omega_{k,1}, \ldots, \omega_{k,p+1}$ *are the eigenvalues of* $\mathbf{C}_k$, *and* $\chi_h^2(1)$*'s are i.i.d.* $\chi^2(1)$ *random variables.*

Recall from Section 3 that $\widehat{\mathbf{A}}_i = \widehat{\mathbf{\Sigma}}^{-1}\widetilde{Y}_i\widetilde{\mathbf{x}}_i(\widetilde{\mathbf{x}}_i^*)^{\text{T}} - \widehat{\mathbf{\Sigma}}^{-1}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^{\text{T}}\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Sigma}}^*_{\text{Yxx}}$. Let $\widehat{\mathbf{C}}_k = n^{-1}\sum_{i=1}^{n}\widehat{\mathbf{A}}_i^{\text{T}}\mathbf{e}_k\mathbf{e}_k^{\text{T}}\widehat{\mathbf{A}}_i$. We can then approximate the asymptotic distribution of $G_k$ under $\text{H}_0^{[k]}$ as $\sum_{h=1}^{p+1}\hat{\omega}_{k,h}\chi_h^2(1)$, where $\hat{\omega}_{k,1}, \ldots, \hat{\omega}_{k,p+1}$ are the eigenvalues of $\widehat{\mathbf{C}}_k$. Following similar development in Section 3, we can get the approximate $p$-value for the marginal coordinate hypotheses (4.6).

## 4.4 Simulation Studies

Extensive simulation studies are carried out in this section to evaluate the performance of the proposed hybrid estimator. Consider the following models,

$$\begin{aligned}
\text{I} &: Y = (.5X_1 - .5X_p) + \text{sgn}(.5X_1 - .5X_p) + .1\varepsilon, \\
\text{II} &: Y = (.5X_1 - .5X_2)^2 + \sin(.5X_1 - .5X_2) + .1\varepsilon, \\
\text{III} &: Y = \sin(X_1) + \cos(X_p) + .1\varepsilon, \\
\text{IV} &: Y = (X_1 + X_p)^2 + 2(X_2 + X_3 + X_p) + .1\varepsilon.
\end{aligned}$$

In all four models, $\mathbf{x} = (X_1, ..., X_p)^{\text{T}}$ is generated from the multivariate normal distribution with zero mean vector and covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{p\times p}$, where $\sigma_{ij} = 0.5^{|i-j|}$ for $0 \le i, j \le p$. The error $\varepsilon \sim \mathcal{N}(0, 1)$ and $\varepsilon$ is independent of $\mathbf{x}$. The LCM condition (1.3) and the CCV condition (1.4) are satisfied due to the normality of $\mathbf{x}$. We fix the predictor dimension at $p = 10$ and consider sample sizes $n = 200$ and 800.

### 4.4.1 Accuracy

For estimator $\widehat{\mathbf{B}}$ and the basis $\mathbf{B}$, denote their corresponding projection matrices as $\mathbf{P}_{\widehat{\mathbf{B}}} = \widehat{\mathbf{B}}(\widehat{\mathbf{B}}^{\text{T}}\widehat{\mathbf{B}})^{-1}\widehat{\mathbf{B}}^{\text{T}}$ and $\mathbf{P}_{\mathbf{B}} = \mathbf{B}(\mathbf{B}^{\text{T}}\mathbf{B})^{-1}\mathbf{B}^{\text{T}}$. Then we measure the effectiveness of the estimator by $\Delta = \|\mathbf{P}_{\widehat{\mathbf{B}}} - \mathbf{P}_{\mathbf{B}}\|^2$, where $\|\cdot\|$ is the matrix Frobenius norm. Smaller values of $\delta$ indicate better estimation of the central mean space. Based on 200 repetitions, we report the mean and the standard deviation of $\Delta$ in Table 4.1. Three estimators of $\mathbf{B}$ are included for this comparison: the OLS, the PHD, and our proposed hybrid estimator. OLS always estimates $\mathbf{B}$ as a vector in $\mathbb{R}^p$, while PHD and the hybrid estimators use the true structural dimension $d$ in the eigenvalue decomposition step.

Table 4.1: Comparison for estimating $\mathbf{B}$. The mean and the standard deviation of $\Delta$ are reported based on 200 repetitions.

| | | OLS | | PHD | | HYB | |
|---|---|---|---|---|---|---|---|
| Model | n | mean | s.d. | mean | s.d. | mean | s.d. |
| I | 200 | 0.232 | 0.061 | 1.796 | 0.268 | 0.051 | 0.030 |
| | 800 | 0.065 | 0.018 | 1.763 | 0.252 | 0.004 | 0.002 |
| II | 200 | 0.485 | 0.171 | 0.292 | 0.208 | 0.108 | 0.073 |
| | 800 | 0.337 | 0.091 | 0.062 | 0.037 | 0.022 | 0.014 |
| III | 200 | 1.209 | 0.081 | 2.107 | 0.251 | 0.558 | 0.343 |
| | 800 | 1.161 | 0.038 | 1.898 | 0.241 | 0.118 | 0.052 |
| IV | 200 | 1.209 | 0.063 | 2.103 | 0.306 | 0.873 | 0.481 |
| | 800 | 1.180 | 0.035 | 1.883 | 0.253 | 0.159 | 0.083 |

Note that the linear component in model I favors OLS, which has smaller $\Delta$ values than PHD. On the other hand, the quadratic component in model II leads to smaller $\Delta$ values for PHD when compared with OLS. Although models I and II are suitable for OLS or PHD, the hybrid estimator can still improve over OLS and PHD in both models. While the first two models have $d = 1$, the next two models are more complex with $d = 2$. For $\mathbf{B} \in \mathbb{R}^{10\times 2}$ in models III and IV, the OLS estimator is clearly not ideal as $\boldsymbol{\beta}_{\text{OLS}} \in \mathbb{R}^p$ can recover at most one direction in the central mean space. Due to the sine function in model III and the linear link function in model IV, PHD also fails to recover $\mathbf{B}$ accurately. The hybrid estimator is again the best performer with the smallest $\Delta$ values for models III and IV. As $n$ increases from 200 to 800, all three methods improve, and the hybrid estimator maintains its advantage over OLS and PHD across all four models.

Consider the case when $Y$ is missing, we denote missingness indicator $\delta = 1$ when $Y$ is observed and $\delta = 0$ otherwise. We define the MAR missingness

schemes as

$$P(\delta = 1 \mid \mathbf{x}) = \frac{\exp(c_1 + \boldsymbol{\alpha}^T \mathbf{x})}{1 + \exp(c_1 + \boldsymbol{\alpha}^T \mathbf{x})}.$$

The missing proportion of the response can be controlled by adjusting $c$. Setting $\boldsymbol{\alpha} = (1, 1, 1, ..., 1)_p^{\mathrm{T}}/\sqrt{p}$. Let $c_1 = 1, 0, -1$, we compare results for Model I and II across different missing proportions as 30%, 50% and 70%. We measure the effectiveness of the estimator by the Euclidean distances $\Delta$ . Based on 200 repetitions, we report the mean and the standard deviation of $\Delta$ in Table 4.2. The OLS, PHD, and our proposed hybrid estimator of $\mathbf{B}$ are compared with Full data analysis, complete case analysis and inverse probability weighting adjustment. Full data analysis serves as the oracle estimator. for this comparison: the OLS, the PHD, and our proposed hybrid estimator. For Model I and II, all estimators estimate $\mathbf{B}$ as a vector in $\mathbb{R}^p$. It is as expected that the performance for the inverse probability weighted hybrid estimator are consistently better than the inverse probability weighted OLS and PHD estimators. In addition, the inverse probability weighted hybrid estimator performs better than the complete case estimators and worsens with the increasing of missing proportion.

### 4.4.2 Order Determination and Marginal Coordinate Test

To evaluate the effectiveness of the proposed order determination procedure, we report the frequencies of the estimated structural dimension $\hat{d}$ over 200 repetitions in Table 4.3. The nominal level for testing $H_0^{(\ell)} : d = \ell$ v.s. $H_a^{(\ell)} : d > \ell$ is set as 0.05, and the proposed procedure from Section 3 is com-

Table 4.2: Comparison for estimating $\mathbf{B}$ when response is missing at random. The mean and the standard deviation of $\Delta$ are reported based on 200 repetitions.

| Model | Missing | $\widehat{\boldsymbol{\beta}}$ | OLS | | PHD | | HYB | |
|---|---|---|---|---|---|---|---|---|
| | | | mean | s.d. | mean | s.d. | mean | s.d. |
| I | 30% | Full | 0.232 | 0.061 | 1.796 | 0.268 | 0.051 | 0.030 |
| | | CC | 0.547 | 0.111 | 1.230 | 0.483 | 0.248 | 0.192 |
| | | IPW | 0.539 | 0.113 | 1.118 | 0.483 | 0.257 | 0.158 |
| | 50% | Full | 0.232 | 0.061 | 1.796 | 0.268 | 0.051 | 0.030 |
| | | CC | 0.580 | 0.126 | 1.126 | 0.515 | 0.495 | 0.411 |
| | | IPW | 0.577 | 0.132 | 0.867 | 0.405 | 0.377 | 0.192 |
| | 70% | Full | 0.232 | 0.061 | 1.796 | 0.268 | 0.051 | 0.030 |
| | | CC | 0.603 | 0.160 | 1.213 | 0.520 | 0.815 | 0.528 |
| | | IPW | 0.616 | 0.173 | 0.812 | 0.393 | 0.554 | 0.279 |
| II | 30% | Full | 0.485 | 0.171 | 0.292 | 0.208 | 0.108 | 0.073 |
| | | CC | 0.532 | 0.185 | 0.401 | 0.287 | 0.202 | 0.138 |
| | | IPW | 0.543 | 0.190 | 0.401 | 0.281 | 0.174 | 0.167 |
| | 50% | Full | 0.485 | 0.171 | 0.292 | 0.208 | 0.108 | 0.073 |
| | | CC | 0.596 | 0.214 | 0.470 | 0.315 | 0.282 | 0.180 |
| | | IPW | 0.640 | 0.235 | 0.460 | 0.303 | 0.249 | 0.195 |
| | 70% | Full | 0.485 | 0.171 | 0.292 | 0.208 | 0.108 | 0.073 |
| | | CC | 0.733 | 0.268 | 0.648 | 0.417 | 0.453 | 0.342 |
| | | IPW | 0.822 | 0.328 | 0.596 | 0.385 | 0.440 | 0.294 |

pared with the PHD-based sequential test. Recall that $d = 1$ for models I and II, and $d = 2$ for models III and IV. We boldface the entries with the dominate frequency across $\hat{d} = 0$, 1, or 2 for easy reference. From the left panel of Table 4.3, we see that except for model II with $n = 800$, PHD fails to estimate the correct structural dimension $d$ with dominate frequency. Order determination with the hybrid method, on the other hand, can estimate the true structural dimension of single-index models I and II with frequency close to 1 ($n = 200$) or equal to 1 ($n = 800$). In the more challenging case of multi-index models III and IV, our proposal can still estimate the true $d$ with dominate frequency

Table 4.3: Comparison for order determination. The frequencies of the estimated structural dimension $\hat{d}$ are reported based on 200 replications.

| | | PHD | | | HYB | | |
|---|---|---|---|---|---|---|---|
| Model | n | $\hat{d}=0$ | $\hat{d}=1$ | $\hat{d}=2$ | $\hat{d}=0$ | $\hat{d}=1$ | $\hat{d}=2$ |
| I | 200 | **1.000** | 0.000 | 0.000 | 0.010 | **0.990** | 0.000 |
| | 800 | **1.000** | 0.000 | 0.000 | 0.000 | **1.000** | 0.000 |
| II | 200 | **0.735** | 0.265 | 0.000 | 0.020 | **0.980** | 0.000 |
| | 800 | 0.000 | **1.000** | 0.000 | 0.000 | **1.000** | 0.000 |
| III | 200 | **0.510** | 0.490 | 0.000 | 0.010 | **0.780** | 0.210 |
| | 800 | 0.000 | **1.000** | 0.000 | 0.000 | 0.000 | **1.000** |
| IV | 200 | **0.820** | 0.180 | 0.000 | 0.335 | **0.660** | 0.005 |
| | 800 | 0.000 | **1.000** | 0.000 | 0.000 | 0.055 | **0.945** |

when the sample size is large ($n = 800$).

The results of testing predictor contribution are summarized in Table 4.4. At 0.05 nominal level, we report the frequencies of rejecting $H_0^{[k]} : \mathbf{e}_k^{\mathrm{T}}\mathbf{B} = \mathbf{0}$ based on 200 replications. Note that rejecting $H_0^{[k]}$ is to conclude that the $k$th predictor $X_k$ contributes to the regression mean. For the ease of reference, we boldface all the frequencies that correspond to the contributing predictors, or when the corresponding $H_0^{[k]}$ is false. Ideally, we want the frequency to be close to 1 for predictors in the regression mean, and the frequency close to the nominal level for predictors not in the regression mean. The proposed procedure from Section 4 is compared with the PHD-based marginal coordinate test. The hybrid-based test works very well across all four models. Specifically, when $H_0^{[k]}$ is false, we reject the null with high frequencies for $n = 200$, and the power of our proposal to detect the contributing predictors becomes 1 with $n = 800$. When $H_0^{[k]}$ is true, the frequency of rejecting the null is close to the nominal level. The PHD-based marginal coordinate test works well for model II, where the quadratic link function favors PHD. While it can detect

Table 4.4: Comparison for testing predictor contribution. The frequencies of rejecting $H_0^{[k]} : \mathbf{e}_k^{\mathrm{T}}\mathbf{B} = \mathbf{0}$ are reported based on 200 replications.

| Model | n | Method | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ | $X_7$ | $X_8$ | $X_9$ | $X_{10}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I | 200 | PHD | **0.005** | 0.020 | 0.040 | 0.030 | 0.0200 | 0.020 | 0.035 | 0.060 | 0.030 | **0.010** |
| | | HYB | **1.000** | 0.020 | 0.040 | 0.030 | 0.020 | 0.020 | 0.035 | 0.055 | 0.030 | **1.000** |
| | 800 | PHD | **0.010** | 0.025 | 0.010 | 0.025 | 0.025 | 0.015 | 0.020 | 0.015 | 0.030 | **0.000** |
| | | HYB | **1.000** | 0.025 | 0.010 | 0.025 | 0.025 | 0.015 | 0.020 | 0.015 | 0.025 | **1.000** |
| II | 200 | PHD | **0.780** | **0.745** | 0.020 | 0.025 | 0.015 | 0.03 | 0.015 | 0.025 | 0.035 | 0.040 |
| | | HYB | **0.995** | **0.990** | 0.025 | 0.020 | 0.015 | 0.020 | 0.015 | 0.025 | 0.045 | 0.035 |
| | 800 | PHD | **1.000** | **1.000** | 0.025 | 0.015 | 0.030 | 0.025 | 0.040 | 0.005 | 0.030 | 0.025 |
| | | HYB | **1.000** | **1.000** | 0.035 | 0.015 | 0.030 | 0.025 | 0.030 | 0.010 | 0.030 | 0.025 |
| III | 200 | PHD | **0.005** | 0.020 | 0.050 | 0.030 | 0.010 | 0.030 | 0.035 | 0.035 | 0.015 | **1.000** |
| | | HYB | **1.000** | 0.025 | 0.050 | 0.030 | 0.010 | 0.030 | 0.035 | 0.035 | 0.015 | **1.000** |
| | 800 | PHD | **0.010** | 0.040 | 0.030 | 0.015 | 0.040 | 0.020 | 0.035 | 0.020 | 0.070 | **1.000** |
| | | HYB | **1.000** | 0.040 | 0.035 | 0.015 | 0.045 | 0.020 | 0.035 | 0.020 | 0.075 | **1.000** |
| IV | 200 | PHD | **0.920** | **0.015** | **0.020** | 0.025 | 0.020 | 0.025 | 0.015 | 0.015 | 0.015 | **0.855** |
| | | HYB | **0.920** | **0.405** | **0.310** | 0.025 | 0.020 | 0.020 | 0.015 | 0.015 | 0.015 | **0.980** |
| | 800 | PHD | **1.000** | **0.010** | **0.010** | 0.055 | 0.025 | 0.010 | 0.015 | 0.020 | 0.030 | **1.000** |
| | | HYB | **1.000** | **1.000** | **1.000** | 0.055 | 0.020 | 0.015 | 0.015 | 0.015 | 0.025 | **1.000** |

$X_{10}$ in model III as well as $X_1$ and $X_{10}$ in model IV, the PHD-based test can not detect the contributing predictors that appear in the linear component of model I and model IV or the sine component of model III.

## 4.5 Discussion

In this chapter, a new hybrid estimator for the central mean space is proposed. The method inherits the strengths of both the OLS component and the PHD component, and works well across a wide range of models. In addition, the sequential test approach is developed for order determination, and the marginal coordinate test procedure is studied for testing predictor contributions. Our proposal is different from existing hybrid estimators, most of which focus on recovering the central space. Our comprehensive treatment

of the hybrid estimator, together with the ease of computation, suggests that the new proposal could be a useful addition to the SDR literature. To improve the asymptotic efficiency of the hybrid central mean space estimator, the minimum discrepancy approach Cook and Ni (2005) for hybrid SDR estimators can be explored, which chooses the weights for the hybrid components adaptively. Yoo and Cook (2007) proposed a central mean space estimator for multivariate response. A hybrid procedure to estimate the central mean space with multivariate response is worth further investigation. In addition, parallel to CCEE, IPWEE and AIPWEE for OLS and PHD, the estimating equation approaches can be readily applied to the hybrid estimator.

# CHAPTER 5

# SUMMARY AND FUTURE RESEARCH

In current stage, we proposed a inverse probability weighting SIR and two novel estimating equation approaches,the complete-case estimating equation and the inverse probability weighted estimating equation, to solve missingness problem in sufficient dimension reduction. In the study of inverse probability weighting SIR, we differentiate missing issues in inverse regressions from forward regression. With the asymptotic study for proposed inverse probability weighting SIR estimator, we interestingly find that weighting with the nonparametrically estimated propensity score gains efficiency than using the true propensity score. In addition, the estimator allows us to further perform marginal coordinate test for model-free variable selection to detect the active predictor. The study of this proposal can be further explored along the following directions. First, the proposal can be generalized to other inverse regression family estimators such as SAVE and directional regression. Second, besides the inverse probability weighted adjustment to the estimat-

ing equation, augmented inverse probability weighted adjustment can also be implemented under a similar framework.

To the best of our knowledge, all the existing SDR methods in the literature that deal with missing data require linear conditional mean and sometimes conditional constant covariance assumptions, and our estimating equation proposals fill this important gap. In addition, the study delightfully finding that under the proposed complete case estimating equations framework, the resulting SDR estimators are still unbiased even when applying to inverse regression approaches such as SIR. The proposed framework has more to explore. The following is a list to prospect for future researches.

1. Besides the current nonparametric estimation, establishing a parametric estimation for the propensity function $\pi(\mathbf{x})$ to build a more complete asymptotic theory.

2. The structural dimension $d$ has been assumed as known for the estimating equation approaches. Ma and Zhang (2015) developed a novel procedure to determine $d$ in the estimating equation framework. The extension of Ma and Zhang (2015) in the case of missing data is worth further investigation.

3. Although the study in this book focus on special cases such as OLS, PHD and SIR, the proposals can be generalized to other SDR estimators such as SAVE, and directional regression to derive a rich class of estimators.

4. For now, we only discovered the missing response case, expanding it to the general missingness is of interest to tell a more thorough story.

5. In the discussion of complete case estimating equation, no missingness schemes are required to show its validity. However, IPWEE and AIPWEE rely on missing at random assumption. More missing mechanisms can be studied

under the framework, such as missing not at random.

Last but not least, the hybrid SDR estimator address the limitation that existing estimators of the central mean space have uneven performances across different types of link functions. Based on the new hybrid estimator, theories and numerical studies can be further implemented to combine the aforementioned procedures to deal with missing response.

# BIBLIOGRAPHY

Bang, H. and J. M. Robins (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics 61*, 962–972.

Bickel, P. J., C. A. J. Klaassen, Y. Ritov, and J. A. Wellner (1993). Efficient and adaptive estimation for semiparametric models. *Baltimore: The Johns Hopkins University Press*.

Bura, E. and R. D. Cook (2001a). Extending sir: the weighted chi-square test. *Journal of the American Statistical Association 96*, 996–1003.

Bura, E. and R. D. Cook (2001b). Extending sliced inverse regression: the weighted chi-squared test. *J. Amer. Statist. Assoc. 96*, 996–1003.

Chen, X., H. Hong, and A. Tarozzi (2008). Semiparametric efficiency in in gmm models with auxiliary data. *Ann. Statist. 36*, 808–843.

Cook, R. D. (1994). On the interpretation of regression plots. *J. Amer. Statist. Assoc. 89*, 177–189.

Cook, R. D. (1998). Regression graphics: Ideas for studying regressions through graphics. *New York: Wiley*, 50, 102–106.

Cook, R. D. (2004). Testing predictor contributions in sufficient dimension reduction. *Ann. Statist. 32*, 1062–1092.

Cook, R. D. and B. Li (2002). Dimension reduction for conditional mean in regression. *Ann. Statist. 40*, 455–474.

Cook, R. D. and B. Li (2004a). Determining the dimension of iterative hessian transformation. *Ann. Statist. 32*, 2501–31.

Cook, R. D. and B. Li (2004b). Determining the dimension of iterative hessian transformation. *Annals of Statistics 32*, 2501–2531.

Cook, R. D. and L. Ni (2005). Sufficient dimension reduction via inverse regression. *Statistica Sinica 100*, 410–428.

Cook, R. D. and S. Weisberg (1991). Discussion of sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc. 86*, 28–33.

Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the em algorithm. *J. Roy. Statist. Soc. Ser. B 39*, 1–38.

Ding, X. and Q. Wang (2011). Fusion-refinement procedure for dimension reduction with missing response at random. *J. Amer. Statist. Assoc. 108*, 1193–1207.

Dong, Y. (2016). A note on moment-based sufficient dimension reduction estimators. *Statistics and Its Interface*, 141–145.

Dong, Y. and B. Li (2010). Dimension reduction for non-elliptically distributed predictors: second-order moments. *Biometrika 97*, 279–294.

Dong, Y., Z. Yu, and L. P. Zhu (2015). Robust inverse regression for dimension reduction. *Journal of Multivariate Analysis 134*, 71–81.

Dong, Y. and L. P. Zhu (2012). A note on sliced inverse regression with missing predictors. *Statistical Analysis and Data Mining 5*, 128–138.

Dong, Y. and L. P. Zhu (2013). Direction estimation in single-index model with missing values. *Statistics and Its Interface 6*, 379–385.

Eaton, M. L. (1986). A characterization of spherical distributions. *J. Multivariate Anal. 34*, 439–446.

Ferre, L. and A. F. Yao (2005). Smooth function inverse regression. *Statist. Sinica 15*, 665–83.

Gannoun, A. and J. Saracco (2003). An asymptotic theory for sir$\alpha$ method. *Statistica Sinica 13*, 97–310.

Godambe, V. P. (1993). Estimating functions. *Oxford Univ. Press Oxford.*, 2–17.

Härdle, W. (1990). Applied nonparametric regression. *Cambridge: Cambridge University Press*.

Härdle, W. and E. Mammen (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist. 21*, 1926–1947.

Hirano, K., G. W. Imbens, and F. Chiaromonte (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica 71*, 1161–1189.

Horvitz, D. G. and D. J. Thompson (1952). A generalization of sampling without replacement from a finite universe. *J. Amer. Statist. Assoc. 47*, 663– 685.

J. M. Robins, A. R. and L. P. Zhao (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Amer. Statist. Assoc. 90*, 106–121.

Kim, J. K. and J. Shao (2014). Statistical methods for handling incomplete data. *CRC Press*.

Li, B., R. D. Cook, and F. Chiaromonte (2003). Dimension reduction for the conditional mean in regressions with categorical predictors. *Annals of Statistics 31*, 1636–1668.

Li, B. and Y. X. Dong (2009). Dimension reduction for non-elliptically distributed predictors. *Ann. Statist. 37*, 1272–1298.

Li, B. and S. Wang (2007). On directional regression for dimension reduction. *J. Amer. Statist. Assoc. 102*, 997–1008.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc. 86*, 316–327.

Li, K. C. (1992). On principal hessian directions for data visualization and dimension reduction: another application of steins lemma. *J. Amer. Statist. Assoc 87*, 1025–1039.

Li, K. C. and N. Duan (1989). Regression analysis under link violation. *Ann. Statist. 17*, 1009–1052.

Li, L. and W. Lu (2008). Sufficient dimension reduction with missing predictors. *J. Amer. Statist. Assoc. 103*, 882–831.

Lipsitz, S. R., J. G. Ibrahim, and L. P. Zhao (1999). A weighted estimating equation for missing covariate data with properties similar to maximum likelihood. *J. Amer. Statist. Assoc. 94*, 1147–1160.

Little, R. J. A. and D. B. Rubin (2002). Statistical analysis with missing data. *2nd Ed. Hoboken, NJ: Wiley*.

Ma, Y. and X. Zhang (2015). A validated information criterion to determine the structural dimension in dimension reduction models. *Biometrika 102*, 409–420.

Ma, Y. and L. P. Zhu (2012a). Efficiency loss caused by linearity condition in dimension reduction. *Biometrica 99*, 1–13.

Ma, Y. and L. P. Zhu (2012b). A semiparametric approach to dimension reduction. *J. Amer. Statist. Assoc 107(497)*, 168–179.

Ma, Y. and L. P. Zhu (2013a). Efficient estimation in sufficient dimension reduction. *Ann. Statist. 41*, 250–268.

Ma, Y. and L. P. Zhu (2013b). On estimation efficiency of the central mean subspace. *J. R. Statist. Soc. B 76*, 885–901.

Newey, W. K. (1990). Semiparametric efficiency bounds. *J. Appl. Econometrics 5*, 99–135.

Robins, J. M., A. Rotnitzky, and M. van der Laan (2000). Comment on "on profile likelihood". *J. Amer. Statist. Assoc. 95*, 431–435.

Robins, J. M., A. Rotnitzky, and L. Zhao (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc. 89*, 846–866.

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

Rubin, D. and N. Thomas (1996). Matching using estimated propensity scores: Relating theory to practice. *Biometrics 52*, 249–264.

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

Rubin, D. B. (1987). Multiple imputation for nonresponse in surveys. *Wiley, New York*.

Schafer, J. L. (1997). Analysis of incomplete multivariate data. *CRC Press*.

Scharfstein, D. O., A. Rotnitzky, and J. M. Robins (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models. *J. Amer. Statist. Assoc. 94*, 1135–1146.

Shanker, A. J. and L. A. Prendergast (2011). Iterative application of dimension reduction methods. *Electronic Journal of Statistics 5*, 1471–1494.

Shao, Y. and R. D. Cook (2007). Marginal tests with sliced average variance estimation. *Biometrika 94*, 285–296.

Tang, C. Y. and Y. S. Qin (2012). An efficient empirical likelihood approach for estimating equations with missing data. *Biometrika 99*, 1001–1007.

Tsiatis, A. A. (2006). Semiparametric theory and missing data. *New York: Springer*.

Wang, D. and S. X. Chen (2009). Empirical likelihood for estimating equations with missing values. *Ann. Statist. 37*, 490–517.

Xia, Q. and Y. Dong (2017). On a new hybrid estimator for the central mean space. *Journal of Systems Science and Complexity 30*, 111121.

Ye, Z. and R. E. Weiss (2003). Using the bootstrap to select one of a new class of dimension reduction methods. *Statistica Sinica 98*, 968–979.

Yin, X. and R. D. Cook (2002). Dimension reduction for the conditional $k$th moment in regression. *J. R. Stat. Soc. Ser. B 64*, 159–175.

Yin, X., B. Li, and R. D. Cook (2008). Successive direction extraction for estimating the central subspace in a multipleindex regression. *J. Multivar. Anal. 99*, 1733–1757.

Yoo, J. K. and R. D. Cook (2007). Optimal sufficient dimension reduction for the conditional mean in multivariate regression. *Biometrika 94*, 231–242.

Yu, Z., Y. Dong, and M. Huang (2014). General directional regression. *Journal of Multivariate Analysis 124*, 94–104.

Zeng, P. and Y. Zhu (2010). An integral transform method for estimating the central mean and central subspaces. *J. Multivariate Anal. 101*, 271–290.

Zhou, Y. and Y. Dong (2016). Model-free coordinate test and variable selection via directional regression. *Statistica Sinica 26*, 1159–1174.

Zhou, Y., A. T. K. Wan, and X. Wang (2008). Estimating equations inference with missing data. *J. Amer. Statist. Assoc. 103*, 1187–1199.

Zhu, L., M. O. M, and Y. Li (2007). On hybrid methods of inverse regression-based algorithms. *Computational Statistics and Data Analysis 51*, 2621–2635.

Zhu, L. P., T. Wang, and L. X. Zhu (2012). Sufficient dimension reduction in regressions with missing predictors. *Statist. Sinica 22*, 1611–1637.

Zhu, L. P. and L. X. Zhu (2009). Dimension-reduction for conditional variance in regressions. *Statist. Sinica 19*, 869–883.

Zhu, L. X. and K. T. Fang (1996). Asymptotics for the kernel estimates of sliced inverse regression. *Ann. Statist. 24*, 1053–1067.

# APPENDIX A

# TECHNICAL DETAILS FOR CHAPTER 2

*Proof of Theorem 2.2.* To study the asymptotic property of $\widehat{\mathbf{M}}$, we investigate $\widehat{\mathbf{a}}_h = \widehat{\mathbf{\Sigma}}^{-1}\widehat{p}_h^{-1/2}\widehat{\mathbf{u}}_h$, where $\widehat{\mathbf{u}}_h = \frac{1}{n}\sum_{i=1}^n \delta_i R_{hi}(\mathbf{x}_i - \bar{\mathbf{x}})/\widehat{\pi}(\mathbf{x}_i)$, and $\widehat{p}_h = \frac{1}{n}\sum_{i=1}^n \delta_i R_{hi}/\widehat{\pi}(\mathbf{x}_i)$. Let $f(\mathbf{x})$ be the probability density function of $\mathbf{x}$, $g(\mathbf{x}) = \pi(\mathbf{x})f(\mathbf{x})$, and $\widehat{g}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n K_b(\mathbf{x} - \mathbf{x}_i)\delta_i$ and $\widehat{f}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n K_b(\mathbf{x} - \mathbf{x}_i)$ are the sample estimators of $g(\mathbf{x})$ and $f(\mathbf{x})$ respectively. We have

$$\widehat{\pi}(\mathbf{x}_i) = \frac{\sum_{j=1}^n K_b(\mathbf{x}_i - \mathbf{x}_j)\delta_i}{\sum_{j=1}^n K_b(\mathbf{x}_i - \mathbf{x}_j)} = \frac{\widehat{g}(\mathbf{x}_i)}{\widehat{f}(\mathbf{x}_i)}.$$

It can be shown that

$$\frac{1}{\widehat{\pi}(\mathbf{x}_i)} - \frac{1}{\pi(\mathbf{x}_i)} = \frac{1}{\pi^2(\mathbf{x}_i)}\left[\frac{1}{f(\mathbf{x})}\left\{\widehat{g}(\mathbf{x}_i) - g(\mathbf{x}_i)\right\} - \frac{\pi(\mathbf{x}_i)}{f(\mathbf{x}_i)}\left\{\widehat{f}(\mathbf{x}_i) - f(\mathbf{x}_i)\right\}\right] + o_p(n^{-1/2}). \quad \text{(A.1)}$$

**Step 1** Show that

$$\widehat{p}_h - p_h = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(1-\frac{\delta_i}{\pi(\mathbf{x}_i)}\right)\mathrm{E}(R_h \mid \mathbf{x}_i) + \frac{\delta_i}{\pi(\mathbf{x}_i)}R_{hi}\right\} - p_h + o_p(n^{-1/2}) \tag{A.2}$$

Let $\widetilde{p}_h = \frac{1}{n}\sum_{i=1}^{n}\delta_i R_{hi}/\pi(\mathbf{x}_i)$, we have $\widehat{p}_h - p_h = \widehat{p}_h - \widetilde{p}_h + \widetilde{p}_h - p_h$. To show (A.2), we only need to study $\widehat{p}_h - \widetilde{p}_h$.

$$\begin{aligned}
\widehat{p}_h - \widetilde{p}_h =& \frac{1}{n}\sum_{i=1}^{n}\delta_i R_{hi}\left\{\frac{1}{\widehat{\pi}(\mathbf{x}_i)} - \frac{1}{\pi(\mathbf{x}_i)}\right\} \\
=& \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i R_{hi}}{\pi^2(\mathbf{x}_i)}\left[\frac{1}{f(\mathbf{x}_i)}\left\{\frac{1}{n}\sum_{j=1}^{n}K_b(\mathbf{x}_j-\mathbf{x}_i)\delta_j - g(\mathbf{x}_i)\right\}\right. \\
& \left. - \frac{\pi(\mathbf{x}_i)}{f(\mathbf{x}_i)}\left\{\frac{1}{n}\sum_{j=1}^{n}K_b(\mathbf{x}_j-\mathbf{x}_i) - f(\mathbf{x}_i)\right\}\right] + o_p(n^{-1/2}) \qquad \text{(A.3)} \\
=& S_1 - S_2 + o_p(n^{-1/2})
\end{aligned}$$

Note that

$$\begin{aligned}
S_1 =& \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i R_{hi}}{\pi^2(\mathbf{x}_i)f(\mathbf{x}_i)}\frac{1}{n}\sum_{j=1}^{n}K_b(\mathbf{x}_j-\mathbf{x}_i)\delta_j - p_h \\
=& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\frac{1}{2}\left\{\frac{\delta_i R_{hi}\delta_j}{\pi^2(\mathbf{x}_i)f(\mathbf{x}_i)}K_b(\mathbf{x}_j-\mathbf{x}_i) + \frac{\delta_i R_{hj}\delta_j}{\pi^2(\mathbf{x}_j)f(\mathbf{x}_j)}K_b(\mathbf{x}_j-\mathbf{x}_i)\right\} - p_h \\
=& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\{w(\mathbf{x}_i,\mathbf{x}_j) - p_h\} = \frac{2}{n}\sum_{i=1}^{n}\{w_1(\mathbf{x}_i) - p_h\} + o_p(n^{-1/2}) \qquad \text{(A.4)} \\
=& \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i R_{hi}}{\pi(\mathbf{x}_i)} + \frac{\delta_i}{\pi(\mathbf{x}_i)}\mathrm{E}(R_h \mid \mathbf{x}_i) - p_h\right\} + o_p(n^{-1/2})
\end{aligned}$$

where

$$w_1(\mathbf{x}_i) = \mathrm{E}_j\{w(\mathbf{x}_i, \mathbf{x}_j)\} = \mathrm{E}_j\left[\frac{1}{2}\left\{\frac{\delta_i R_{hi}\delta_j}{\pi^2(\mathbf{x}_i)f(\mathbf{x}_i)}K_b(\mathbf{x}_j - \mathbf{x}_i) + \frac{\delta_i R_{hj}\delta_j}{\pi^2(\mathbf{x}_j)f(\mathbf{x}_j)}K_b(\mathbf{x}_j - \mathbf{x}_i)\right\}\right], \tag{A.5}$$

and the last two steps of (A.5) are true because $\mathrm{E}\{w_1(\mathbf{x}_i)\} = \mathrm{E}(R_h) = p_h$,

$$\begin{aligned}
&\mathrm{E}_j\left\{\frac{\delta_i R_{hi}\delta_j}{\pi^2(\mathbf{x}_i)f(\mathbf{x}_i)}K_b(\mathbf{x}_j - \mathbf{x}_i)\right\} = \frac{\delta_i R_{hi}}{\pi^2(\mathbf{x}_i)f(\mathbf{x}_i)}\mathrm{E}_j\left\{\delta_j K_b(\mathbf{x}_j - \mathbf{x}_i)\right\} \\
&= \frac{\delta_i R_{hi}\delta_j}{\pi^2(\mathbf{x}_i)f(\mathbf{x}_i)}\mathrm{E}_j\left\{\pi(\mathbf{x}_j)K_b(\mathbf{x}_j - \mathbf{x}_i)\right\} \\
&= \frac{\delta_i R_{hi}}{\pi^2(\mathbf{x}_i)f(\mathbf{x}_i)}\frac{1}{b}\int \pi(\mathbf{x})f(\mathbf{x})K\left(\frac{\mathbf{x} - \mathbf{x}_i}{b}\right)dx \\
&= \frac{\delta_i R_{hi}}{\pi^2(\mathbf{x}_i)f(\mathbf{x}_i)}\pi(\mathbf{x}_i)f(\mathbf{x}_i) = \frac{\delta_i R_{hi}}{\pi(\mathbf{x}_i)} + o_p(n^{-1/2}),
\end{aligned} \tag{A.6}$$

and

$$\begin{aligned}
&\mathrm{E}_j\left\{\frac{\delta_i R_{hj}\delta_j}{\pi^2(\mathbf{x}_j)f(\mathbf{x}_j)}K_b(\mathbf{x}_j - \mathbf{x}_i)\right\} = \delta_i \mathrm{E}_j\left\{\frac{R_{hj}}{\pi(\mathbf{x}_j)f(\mathbf{x}_j)}K_b(\mathbf{x}_j - \mathbf{x}_i)\right\} \\
&= \frac{\delta_i}{b}\int\int \frac{R_h(y)}{\pi(\mathbf{x}_j)f(\mathbf{x}_j)}K\left(\frac{\mathbf{x} - \mathbf{x}_i}{b}\right)f(\mathbf{x}, y)d\mathbf{x}dy \\
&= \delta_i \int \frac{R_h(y)}{\pi(\mathbf{x}_i)}\frac{f(\mathbf{x}_i, y)}{f(\mathbf{x}_i)}dy = \frac{\delta_i}{\pi(\mathbf{x}_i)}\mathrm{E}(R_h \mid \mathbf{x}_i) + o_p(n^{-1/2}).
\end{aligned} \tag{A.7}$$

In addition, with the similar U-statistic technique applied in (A.5), we have

$$\begin{aligned} S_2 =& \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i R_{hi}}{\pi^2(\mathbf{x}_i)}\frac{\pi(\mathbf{x}_i)}{f(\mathbf{x}_i)}\frac{1}{n}\sum_{j=1}^{n}K_b(\mathbf{x}_j-\mathbf{x}_i)-p_h \\ =& \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{\delta_i R_{hi}}{\pi(\mathbf{x}_i)f(\mathbf{x}_i)}+\frac{\delta_j R_{hj}}{\pi(\mathbf{x}_j)f(\mathbf{x}_j)}\right\}K_b(\mathbf{x}_j-\mathbf{x}_i)-p_h \\ =& \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i R_{hi}}{\pi(\mathbf{x}_i)}+\mathrm{E}(R_h \mid \mathbf{x}_i)\right\}+o_p(n^{-1/2}). \end{aligned} \tag{A.8}$$

By combining (A.3), (A.5) and ($A$.8), we have

$$\widehat{p}_h-\widetilde{p}_h=\frac{1}{n}\sum_{i=1}^{n}\left(1-\frac{\delta_i}{\pi(\mathbf{x}_i)}\right)\mathrm{E}(R_h \mid \mathbf{x}_i)+o_p(n^{-1/2}). \tag{A.9}$$

and thus (A.2) is obtained.

**Step 2** Show that

$$\widehat{\mathbf{u}}_h-\mathbf{u}_h=\frac{1}{n}\sum_{i=1}^{n}\left\{\left(1-\frac{\delta_i}{\pi(\mathbf{x}_i)}\right)\mathbf{x}_i\mathrm{E}(R_{hi} \mid \mathbf{x}_i)+\frac{\delta_i\mathbf{x}_i R_h}{\pi(\mathbf{x}_i)}-\mathbf{x}_i p_h-\mathbf{u}_h\right\}. \tag{A.10}$$

*Proof.* Let $\widetilde{\mathbf{u}}_h=\frac{1}{n}\delta_i\mathbf{x}_i R_{hi}/\pi(\mathbf{x}_i)$. Note that $\widehat{\mathbf{u}}_h-\mathbf{u}_h=\widehat{\mathbf{u}}_h-\widetilde{\mathbf{u}}_h+\widetilde{\mathbf{u}}_h-\mathbf{u}_h$, we first write

$$\begin{aligned}\widetilde{\mathbf{u}}_h - \mathbf{u}_h =& \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i R_{hi}(\mathbf{x}_i - \bar{\mathbf{x}})}{\pi(\mathbf{x}_i)} - \mathrm{E}(\mathbf{x}R_h) \\ =& \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i R_{hi}\mathbf{x}_i}{\pi(\mathbf{x}_i)} - \mathrm{E}(\mathbf{x}R_h) - \frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}}\left\{\frac{\delta_i R_{hi}}{\pi(\mathbf{x}_i)} - p_h + p_h\right\} \\ =& \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\delta_i R_{hi}\mathbf{x}_i}{\pi(\mathbf{x}_i)} - \mathrm{E}(\mathbf{x}R_h) - \mathbf{x}_i p_h\right\} + o_p(n^{-1/2}).\end{aligned} \tag{A.11}$$

Then it remains to study $\widehat{\mathbf{u}}_h - \widetilde{\mathbf{u}}_h$.

$$\begin{aligned}\widehat{\mathbf{u}}_h - \widetilde{\mathbf{u}}_h =& \frac{1}{n}\sum_{i=1}^{n}\delta_i R_{hi}\mathbf{x}_i\left\{\frac{1}{\widehat{\pi}(\mathbf{x}_i)} - \frac{1}{\pi(\mathbf{x}_i)}\right\} \\ =& \frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i R_{hi}\mathbf{x}_i}{\pi^2(\mathbf{x}_i)}\left[\frac{1}{f(\mathbf{x}_i)}\left\{\frac{1}{n}\sum_{j=1}^{n}K_b(\mathbf{x}_j - \mathbf{x}_i)\delta_j - g(\mathbf{x}_i)\right\}\right. \\ & \left. - \frac{\pi(\mathbf{x}_i)}{f(\mathbf{x}_i)}\left\{\frac{1}{n}\sum_{j=1}^{n}K_b(\mathbf{x}_j - \mathbf{x}_i) - f(\mathbf{x}_i)\right\}\right] + o_p(n^{-1/2}).\end{aligned} \tag{A.12}$$

With a similar fashion as (A.5) to (A.8) in step 1, it can be shown that

$$\widehat{\mathbf{u}}_h - \widetilde{\mathbf{u}}_h = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{\delta_i}{\pi(\mathbf{x}_i)}\right)\mathbf{x}_i\mathrm{E}(R_h \mid \mathbf{x}_i) + o_p(n^{-1/2}). \tag{A.13}$$

□

Let $p_h^* = \widehat{p}_h - p_h$, $\mathbf{u}_h^* = \widehat{\mathbf{u}}_h - \mathbf{u}_h$, and $\boldsymbol{\Sigma}^* = \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma} = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \boldsymbol{\Sigma}$. We have $(p_h^{-1/2})^* = \widehat{p}_h^{-1/2} - p_h^{-1/2} = -\frac{1}{2}p_h^{-3/2}p_h^* + o_p(n^{-1/2})$, and $(\boldsymbol{\Sigma}^{-1})^* = \frac{1}{n}\widehat{\boldsymbol{\Sigma}}^{-1} - \boldsymbol{\Sigma}^{-1} = -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}^{-1}$, we have $\boldsymbol{\Sigma}^* = \sum_{i=1}^{n}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}} - \boldsymbol{\Sigma}$, $\boldsymbol{\Sigma}^{-1*} = -\boldsymbol{\Sigma}^{-1}\boldsymbol{\Sigma}^*\boldsymbol{\Sigma}^{-1}$. Note that $\widehat{\mathbf{a}}_h = \widehat{p}_h^{-1/2}\widehat{\boldsymbol{\Sigma}}^{-1}\widehat{\mathbf{u}}_h$, we have $\widehat{\mathbf{a}}_h - \mathbf{a}_h = (p_h^{-1/2})^*\boldsymbol{\Sigma}^{-1}\mathbf{u}_h + p_h^{-1/2}(\boldsymbol{\Sigma}^{-1})^*\mathbf{u}_h +$

$p_h^{-1/2}\boldsymbol{\Sigma}^{-1}\mathbf{u}_h^* + o_p(n^{-1/2}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\ell}_{h,i} + o_p(n^{-1/2})$, where

$$\begin{aligned}\boldsymbol{\ell}_{h,i} =& -\frac{1}{2}\left\{\left(1-\frac{\delta_i}{\pi(\mathbf{x}_i)}\right)\mathrm{E}(R_h \mid \mathbf{x}_i) + \frac{\delta_i R_{hi}}{\pi(\mathbf{x}_i)}\right\}\boldsymbol{\Sigma}^{-1}\mathbf{u}_h + \frac{1}{2}p_h^{-1/2}\boldsymbol{\Sigma}^{-1}\mathbf{u}_h \\ & - p_h^{-1/2}\boldsymbol{\Sigma}^{-1}\mathbf{x}_i\mathbf{x}_i^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} \\ & + p_h^{-1/2}\boldsymbol{\Sigma}^{-1}\left\{\left(1-\frac{\delta_i}{\pi(\mathbf{x}_i)}\right)\mathbf{x}_i\mathrm{E}(R_h \mid \mathbf{x}_i) + \frac{\delta_i\mathbf{x}_i R_{hi}}{\pi(\mathbf{x}_i)}\right\} - p_h^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{x}_i.\end{aligned} \tag{A.14}$$

Then $\widehat{\mathbf{A}} - \mathbf{A} = \frac{1}{n}\sum_{i=1}^n \mathbf{L}_i + o_p(n^{-1/2})$, where $\mathbf{L}_i = (\boldsymbol{\ell}_{1,i}^{\mathrm{T}}, \ldots, \boldsymbol{\ell}_{H,i}^{\mathrm{T}})^{\mathrm{T}}$. Note that $\mathrm{E}(\boldsymbol{\ell}_{h,i}) = \mathbf{0}$, $\frac{1}{n}\sum_{i=1}^n \boldsymbol{\ell}_{hi} = O_p(n^{-1/2})$, Theorem 2.2 is demonstrated, with $\mathbf{L} = (\boldsymbol{\ell}_1^{\mathrm{T}}, \ldots, \boldsymbol{\ell}_H^{\mathrm{T}})^{\mathrm{T}}$ and

$$\begin{aligned}\boldsymbol{\ell}_h =& -\frac{1}{2}\left\{\left(1-\frac{\delta}{\pi(\mathbf{x})}\right)\mathrm{E}(R_h \mid \mathbf{x}) + \frac{\delta R_h}{\pi(\mathbf{x})}\right\}\boldsymbol{\Sigma}^{-1}\mathbf{u}_h + \frac{1}{2}p_h^{-1/2}\boldsymbol{\Sigma}^{-1}\mathbf{u}_h \\ & - p_h^{-1/2}\boldsymbol{\Sigma}^{-1}\mathbf{x}\mathbf{x}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} \\ & + p_h^{-1/2}\boldsymbol{\Sigma}^{-1}\left\{\left(1-\frac{\delta}{\pi(\mathbf{x})}\right)\mathbf{x}\mathrm{E}(R_h \mid \mathbf{x}) + \frac{\delta\mathbf{x}R_h}{\pi(\mathbf{x})}\right\} - p_h^{1/2}\boldsymbol{\Sigma}^{-1}\mathbf{x}.\end{aligned} \tag{A.15}$$

□

*Proof of Theorem 4.4.* Let $\mathbf{t}_k = (t_{k,1}, \ldots, t_{k,H})^{\mathrm{T}}$ and $\widehat{\mathbf{t}}_k = (\widehat{t}_{k,1}, \ldots, \widehat{t}_{k,H})^{\mathrm{T}}$, where $t_{k,h} = \mathbf{e}_k^{\mathrm{T}}\mathbf{a}_h$ and $\widehat{t}_{k,h} = \mathbf{e}_k^{\mathrm{T}}\widehat{\mathbf{a}}_h$ for $h = 1, \ldots, H$. Under $\mathrm{H}_0^{[k]}$, the Frechet derivative of $t_{k,h}$ is defined as $g_{k,h} = p_h^{-1/2}\mathbf{e}_k^{\mathrm{T}}(\boldsymbol{\Sigma}^{-1})^*\mathbf{u}_h + p_h^{-1/2}\mathbf{e}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{u}_h^*$, where $(p_h^{-1/2})^*\mathbf{e}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{u}_h$ is omitted because $\mathbf{e}_k^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{u}_h = 0$ for $k \in \mathcal{A}^c$. It follows that $\sqrt{n}(\widehat{\mathbf{t}}_k - \mathbf{t}_k) \to \mathcal{N}(\mathbf{0}, \mathrm{Cov}(\mathbf{G_k}))$, where $\mathbf{G}_k = (g_{k,1}, \cdots, g_{k,H})^{\mathrm{T}}$. Because $\mathbf{t}_k = \mathbf{0}$ for $k \in \mathcal{A}^c$, We have $T_k = n\widehat{\mathbf{t}}_k^{\mathrm{T}}\widehat{\mathbf{t}}_k$ has an asymptotic distribution that is the sum of weighted $\chi^2(1)$ under $\mathrm{H}_0^{[k]}$. □

*Proof of Theorem 2.1.* Recall that $\widetilde{\mathbf{a}}_h = \widehat{\mathbf{\Sigma}}^{-1}\widetilde{p}_h^{-1/2}\widetilde{\mathbf{u}}_h$, and $\widehat{\mathbf{a}}_h = \widehat{\mathbf{\Sigma}}^{-1}\widehat{p}_h^{-1/2}\widehat{\mathbf{u}}_h$, we can apply the same procedures in the proof of Theorem 2.2 to find the covariance matrix $\mathbf{\Omega}$ by replacing $(p_h^{-1/2})^*$ with $-\frac{1}{2}p_h^{-3/2}(\widetilde{p}_h - p_h)$, and $\mathbf{u}_h^*$ with $\widetilde{\mathbf{u}}_h - \mathbf{u}_h$.

As $\mathbf{A} = (\mathbf{a}_1^{\mathrm{T}}, \ldots, \mathbf{a}_H^{\mathrm{T}})^{\mathrm{T}}$, to compare $\mathbf{\Gamma}$ and $\mathbf{\Omega}$, we only need to study $\mathrm{Cov}(\widehat{\mathbf{a}}_h)$ and $\mathrm{Cov}(\widetilde{\mathbf{a}}_h)$. In addition, $\widehat{\mathbf{\Sigma}}^{-1}$ can be ignored as it has the same effect for both counterparts. Let $\mathbf{c}_1 = \widehat{p}_h^{-1/2}\widehat{\mathbf{u}}_h - \widetilde{p}_h^{-1/2}\widetilde{\mathbf{u}}_h$ and $\mathbf{c}_2 = \widetilde{p}_h^{-1/2}\widetilde{\mathbf{u}}_h - p_h^{-1/2}\mathbf{u}_h$, we have $\widehat{p}_h^{-1/2}\widehat{\mathbf{u}}_h - p_h^{-1/2}\mathbf{u}_h = \mathbf{c}_1 + \mathbf{c}_2$, we have

$$\begin{aligned}
\mathbf{c}_1 &= (\widehat{p}_h^{-1/2} - \widetilde{p}_h^{-1/2})\widetilde{\mathbf{u}}_h + \widetilde{p}_h^{-1/2}(\widehat{\mathbf{u}}_h - \widetilde{\mathbf{u}}_h) + o_p(n^{-1/2}) \\
&= -\frac{1}{2}p_h^{-3/2}(\widehat{p}_h - \widetilde{p}_h)\mathbf{u}_h + p_h^{-1/2}(\widehat{\mathbf{u}}_h - \widetilde{\mathbf{u}}_h) + o_p(n^{-1/2}) \qquad \text{(A.16)} \\
&= \mathbf{b}_1 + \mathbf{b}_2 + o_p(n^{-1/2}) \\
\mathbf{c}_2 &= -\frac{1}{2}p_h^{-3/2}(\widetilde{p}_h - p_h)\mathbf{u}_h + p_h^{-1/2}(\widetilde{\mathbf{u}}_h - \mathbf{u}_h) + o_p(n^{-1/2}) \\
&= \mathbf{b}_3 + \mathbf{b}_4 + o_p(n^{-1/2}) \qquad \text{(A.17)}
\end{aligned}$$

Thus $\mathrm{Var}(\widehat{p}_h^{-1/2}\widehat{\mathbf{u}}_h - p_h^{-1/2}\mathbf{u}_h) = \mathrm{Var}(\mathbf{c}_1) + \mathrm{Var}(\mathbf{c}_2) + 2\mathrm{Cov}(\mathbf{c}_1, \mathbf{c}_2)$. The following steps (A.18) to (A.27) show that $2\mathrm{Cov}(\mathbf{c}_1, \mathbf{c}_2) = -2\mathrm{Var}(\mathbf{c}_1)$ and therefore $\mathrm{Var}(\widetilde{p}_h^{-1/2}\widetilde{\mathbf{u}}_h - p_h^{-1/2}\mathbf{u}_h) - \mathrm{Var}(\widehat{p}_h^{-1/2}\widehat{\mathbf{u}}_h - p_h^{-1/2}\mathbf{u}_h) = \mathrm{Var}(\mathbf{c}_1)$ is positive definite.

First, we study $\mathrm{Var}(\mathbf{c}_1)$ via (A.18) to (A.21).

$$\begin{aligned}
\mathrm{Var}(\mathbf{c}_1) =& \mathrm{Var}(\mathbf{b}_1) + \mathrm{Var}(\mathbf{b}_2) + 2\mathrm{Cov}(\mathbf{b}_1, \mathbf{b}_2) \\
=& \frac{1}{4}p_h^{-3}\mathbf{u}_h\mathrm{Var}(\widehat{p}_h - \widetilde{p}_h)\mathbf{u}_h^{\mathrm{T}} + p_h^{-1}\mathrm{Var}(\widehat{\mathbf{u}}_h - \widetilde{\mathbf{u}}_h) \qquad \text{(A.18)} \\
& - p_h^{-2}\mathrm{Cov}(\widehat{p}_h - \widetilde{p}_h, \widehat{\mathbf{u}}_h - \widetilde{\mathbf{u}}_h)\mathbf{u}_h,
\end{aligned}$$

where

$$\begin{aligned}\mathrm{Var}(\widehat{p}_h - \widetilde{p}_h) =& \mathrm{Var}\left\{\left(1 - \frac{\delta}{\pi(\mathbf{x})}\right)\mathrm{E}(R_h \mid \mathbf{x})\right\} = \mathrm{E}\left\{\left(1 - \frac{\delta}{\pi(\mathbf{x})}\right)^2 \mathrm{E}^2(R_h \mid \mathbf{x})\right\} \\ =& \mathrm{E}\left\{\left(\frac{1}{\pi(\mathbf{x})} - 1\right)\mathrm{E}^2(R_h \mid \mathbf{x})\right\}, \end{aligned} \tag{A.19}$$

$$\begin{aligned}\mathrm{Var}(\widehat{\mathbf{u}}_h - \widetilde{\mathbf{u}}_h) =& \mathrm{Var}\left\{\left(1 - \frac{\delta}{\pi(\mathbf{x})}\right)\mathbf{x}\mathrm{E}(R_h \mid \mathbf{x})\right\} \\ =& \mathrm{E}\left\{\left(1 - \frac{\delta}{\pi(\mathbf{x})}\right)^2 \mathbf{x}^2\mathrm{E}^2(R_h \mid \mathbf{x})\right\} \\ =& \mathrm{E}\left\{\left(\frac{1}{\pi(\mathbf{x})} - 1\right)\mathbf{x}^2\mathrm{E}^2(R_h \mid \mathbf{x})\right\}, \end{aligned} \tag{A.20}$$

and

$$\begin{aligned}\mathrm{Cov}(\widehat{p}_h - \widetilde{p}_h, \widehat{\mathbf{u}}_h - \widetilde{\mathbf{u}}_h) =& \mathrm{E}\left\{\left(1 - \frac{\delta}{\pi(\mathbf{x})}\right)^2 \mathbf{x}\mathrm{E}^2(R_h \mid \mathbf{x})\right\} \\ =& \mathrm{E}\left\{\left(\frac{1}{\pi(\mathbf{x})} - 1\right)\mathbf{x}\mathrm{E}^2(R_h \mid \mathbf{x})\right\}. \end{aligned} \tag{A.21}$$

Then we study $\mathrm{Cov}(\mathbf{c}_1, \mathbf{c}_2)$ from (A.22) to (A.27).

$$\begin{aligned}\mathrm{Cov}(\mathbf{c}_1, \mathbf{c}_2) =& \mathrm{Cov}(\mathbf{b}_1, \mathbf{b}_3) + \mathrm{Cov}(\mathbf{b}_2, \mathbf{b}_4) + \mathrm{Cov}(\mathbf{b}_1, \mathbf{b}_4) + \mathrm{Cov}(\mathbf{b}_2, \mathbf{b}_3) \\ =& \frac{1}{4}p_h^{-3}\mathbf{u}\mathrm{Cov}(\widehat{p}_h - \widetilde{p}_h, \widetilde{p}_h - p_h)\mathbf{u}^{\mathrm{T}} + p_h^{-1}\mathrm{Cov}(\widehat{\mathbf{u}}_h - \widetilde{\mathbf{u}}_h, \widetilde{\mathbf{u}}_h - \mathbf{u}) \\ & - \frac{1}{2}p_h^{-2}\left\{\mathrm{Cov}(\widehat{p}_h - \widetilde{p}_h, \widetilde{\mathbf{u}}_h - \mathbf{u}_h) + \mathrm{Cov}(\widetilde{p}_h - p_h, \widehat{\mathbf{u}}_h - \widetilde{\mathbf{u}}_h)\right\}\mathbf{u}_h, \end{aligned} \tag{A.22}$$

where

$$\begin{aligned}\mathrm{Cov}(\widehat{p}_h - \widetilde{p}_h, \widetilde{p}_h - p_h) =& \mathrm{E}\left\{\frac{\delta}{\pi(\mathbf{x})}\left(1 - \frac{\delta}{\pi(\mathbf{x})}\right) R_h \mathrm{E}(R_h \mid \mathbf{x})\right\} \\ =& \mathrm{E}\left\{\left(1 - \frac{1}{\pi(\mathbf{x})}\right) \mathrm{E}^2(R_h \mid \mathbf{x})\right\}, \end{aligned} \tag{A.23}$$

$$\begin{aligned}&\mathrm{Cov}(\widehat{\mathbf{u}}_h - \widetilde{\mathbf{u}}_h, \widetilde{\mathbf{u}}_h - \mathbf{u}_h) \\ &= \mathrm{E}\left\{\left(1 - \frac{\delta}{\pi(\mathbf{x})}\right) \mathbf{x}\mathrm{E}(R_h \mid \mathbf{x}) \left(\frac{\delta}{\pi(\mathbf{x})}\mathbf{x}R_h - \mathbf{x}p_h\right)\right\} \\ &= \mathrm{E}\left\{\left(1 - \frac{\delta}{\pi(\mathbf{x})}\right) \frac{\delta}{\pi(\mathbf{x})}\mathbf{x}^2 R_h \mathrm{E}(R_h \mid \mathbf{x})\right\} - p_h \mathrm{E}\left\{\left(1 - \frac{\delta}{\pi(\mathbf{x})}\right) \mathbf{x}\mathrm{E}(R_h \mid \mathbf{x})\mathbf{x}p_h\right\} \\ &= \mathrm{E}\left\{\mathrm{E}\left(\frac{\delta}{\pi(\mathbf{x})} - \frac{\delta}{\pi^2(\mathbf{x})} \,\middle|\, \mathbf{x}\right) \mathbf{x}^2 \mathrm{E}^2(R_h \mid \mathbf{x})\right\} - \mathbf{0} \\ &= \mathrm{E}\left\{\left(1 - \frac{1}{\pi(\mathbf{x})}\right) \mathbf{x}^2 \mathrm{E}^2(R_h \mid \mathbf{x})\right\}, \end{aligned} \tag{A.24}$$

and

$$\begin{aligned}\mathrm{Cov}(\widehat{p}_h - \widetilde{p}_h, \widetilde{\mathbf{u}}_h - \mathbf{u}_h) =& \mathrm{Cov}(\widetilde{p}_h - p_h, \widehat{\mathbf{u}}_h - \widetilde{\mathbf{u}}_h) \\ =& \mathrm{E}\left\{\frac{\delta}{\pi(\mathbf{x})}\left(1 - \frac{\delta}{\pi(\mathbf{x})}\right) \mathbf{x}R_h \mathrm{E}(R_h \mid \mathbf{x})\right\} \\ =& \mathrm{E}\left\{\mathrm{E}\left(\frac{\delta}{\pi(\mathbf{x})} - \frac{\delta}{\pi^2(\mathbf{x})} \,\middle|\, \mathbf{x}\right) \mathbf{x}\mathrm{E}^2(R_h \mid \mathbf{x})\right\} \\ =& \mathrm{E}\left\{\left(1 - \frac{1}{\pi(\mathbf{x})}\right) \mathbf{x}\mathrm{E}^2(R_h \mid \mathbf{x})\right\}. \end{aligned} \tag{A.25}$$

It follows that

$$2\mathrm{Cov}(\mathbf{c}_1, \mathbf{c}_2) = -2\mathrm{Var}(\mathbf{b}_1) - 2\mathrm{Var}(\mathbf{b}_2) - 4\mathrm{Cov}(\mathbf{b}_1, \mathbf{b}_2) = -2\mathrm{Var}(\mathbf{c}_1), \tag{A.26}$$

and thus

$$\mathrm{Var}(\widehat{p}_h^{-1/2}\widehat{\mathbf{u}}_h - p_h^{-1/2}\mathbf{u}_h) = \mathrm{Var}(\mathbf{c}_2) - \mathrm{Var}(\mathbf{c}_1). \tag{A.27}$$

□

# APPENDIX B

# TECHNICAL DETAILS FOR CHAPTER 3

## B.1 Technical Details

**Lemma B.1.** *Under Conditions (C1)-(C5), we have*

$$\max_{1\leq i\leq n}|\widehat{\pi}(\mathbf{x}_i)-\pi(\mathbf{x}_i)| = o_p(n^{-1/4})$$

Following the approach in Härdle (1990) and Härdle and Mammen (1993), one can show under certain conditions (will be listed) Lemma C.1 holds.

**Lemma B.2.** *Under Conditions (C1)-(C5), we have*

$$\sum_{i=1}^{n}\frac{\delta_i}{\pi_i(\mathbf{x})}\left[\mathrm{E}(l\mid\boldsymbol{\eta}_i)\{a(\mathbf{x}_i)-\mathrm{E}(a\mid\boldsymbol{\eta}_i)\}+\mathrm{E}(l\mid\boldsymbol{\eta}_i)\left\{\mathrm{E}(a\mid\boldsymbol{\eta})-\widehat{\mathrm{E}}(a\mid\boldsymbol{\eta}_i)\right\}\right]=o_p(n^{1/2})$$

The proof of Lemma C.2 is similar to the proof of Lemma A2 in Ma and Zhu (2012a).

**Lemma B.3.**

$$\frac{1}{n}\sum_{i=1}^{n}\frac{\delta_i}{\pi_i(\mathbf{x})}\epsilon_i\{\mathrm{E}(a \mid \boldsymbol{\eta}_i) - \widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i)\} = O_p\{h^m/n^{1/2} + h^{2m} + \log^2 n/(nh^d)\}$$

The proof of Lemma B.3 is similar to the proof of Lemma A1 in Ma and Zhu (2012a) and is neglected.

*Proof of Theorem 3.2.* The estimating equation which yields $\widetilde{\mathbf{B}}$ is given as

$$\sum_{i=1}^{n}\left\{\frac{\delta_i}{\widehat{\pi}_i(\mathbf{x})}l(Y_i)\{a(\mathbf{x}_i) - \widehat{\mathrm{E}}(a(\mathbf{x}_i) \mid \widetilde{\mathbf{B}}^T\mathbf{x}_i)\}\right\} = \mathbf{0} \tag{B.1}$$

Note that

$$\frac{1}{\widehat{\pi}_i(\mathbf{x})} = \frac{1}{\pi_i(\mathbf{x})} + \frac{\pi_i(\mathbf{x}) - \widehat{\pi}_i(\mathbf{x})}{\pi_i(\mathbf{x})\widehat{\pi}_i(\mathbf{x})}. \tag{B.2}$$

We have

$$\sum_{i=1}^{n}\left(\frac{\delta_i}{\widehat{\pi}(\mathbf{x}_i)}\widehat{\mathrm{g}}_i^{\star}\right) = \sum_{i=1}^{n}\left(\frac{\delta_i}{\pi_i(\mathbf{x})}\widehat{\mathrm{g}}_i^{\star}\right) + \sum_{i=1}^{n}\left(\frac{\pi_i(\mathbf{x}) - \widehat{\pi}_i(\mathbf{x})}{\pi_i(\mathbf{x})\widehat{\pi}_i(\mathbf{x})}\delta_i\widehat{\mathrm{g}}_i^{\star}\right). \tag{B.3}$$

It is clear that by Lemma C.1, the second term at the right hand side of (B.3)

is

$$\sum_{i=1}^{n}\left(\frac{\pi_i(\mathbf{x})-\widehat{\pi}_i(\mathbf{x})}{\pi_i(\mathbf{x})\widehat{\pi}_i(\mathbf{x})}\delta_i\widehat{\mathrm{g}}_i^{\star}\right)=\sum_{i=1}^{n}\frac{\pi_i(\mathbf{x})-\widehat{\pi}_i(\mathbf{x})}{\pi_i(\mathbf{x})\widehat{\pi}_i(\mathbf{x})}\delta_i l(Y_i)\{a(\mathbf{x}_i)-\widehat{\mathrm{E}}(a(\mathbf{x}_i)\,|\,\widetilde{\mathbf{B}}^T\mathbf{x}_i)\}$$
$$=\sum_{i=1}^{n}\frac{\pi_i(\mathbf{x})-\widehat{\pi}_i(\mathbf{x})}{\pi_i(\mathbf{x})\widehat{\pi}_i(\mathbf{x})}\delta_i l(Y_i)\{a(\mathbf{x}_i)-\mathrm{E}(a(\mathbf{x}_i)\,|\,\mathbf{B}^T\mathbf{x}_i)\}$$
$$+\sum_{i=1}^{n}\frac{\pi_i(\mathbf{x})-\widehat{\pi}_i(\mathbf{x})}{\pi_i(\mathbf{x})\widehat{\pi}_i(\mathbf{x})}\delta_i l(Y_i)\{\mathrm{E}(a(\mathbf{x}_i)\,|\,\mathbf{B}^T\mathbf{x}_i)-\widehat{\mathrm{E}}(a(\mathbf{x}_i)\,|\,\mathbf{B}^T\mathbf{x}_i)\} \tag{B.4}$$
$$+\sum_{i=1}^{n}\frac{\pi_i(\mathbf{x})-\widehat{\pi}_i(\mathbf{x})}{\pi_i(\mathbf{x})\widehat{\pi}_i(\mathbf{x})}\delta_i l(Y_i)\{\widehat{\mathrm{E}}(a(\mathbf{x}_i)\,|\,\mathbf{B}^T\mathbf{x}_i)-\widehat{\mathrm{E}}(a(\mathbf{x}_i)\,|\,\widetilde{\mathbf{B}}^T\mathbf{x}_i)\}$$
$$=n*(o_p(n^{-1/2})+o_p(n^{-1/2})+o_p(n^{-1/2}))=o_p(n^{1/2})$$

Thus the estimating equation (B.1) becomes

$$\sum_{i=1}^{n}\left[\frac{\delta_i}{\pi_i(\mathbf{x})}l(Y_i)\{a(\mathbf{x}_i)-\widehat{\mathrm{E}}(a(\mathbf{x}_i)\,|\,\widetilde{\mathbf{B}}^T\mathbf{x}_i)\}\right]=\mathbf{0} \tag{B.5}$$

Denoting $\boldsymbol{\eta}_i=\mathbf{B}^{\mathrm{T}}\mathbf{x}_i$, $\widetilde{\boldsymbol{\eta}}_i=\widetilde{\mathbf{B}}^T\mathbf{x}_i$, $\mathrm{E}(a(\mathbf{x}_i)\,|\,\boldsymbol{\eta}_i)=\mathrm{E}(a\mid\boldsymbol{\eta})$ and $\mathrm{E}(l(Y_i)\,|\,\boldsymbol{\eta}_i)=\mathrm{E}(l\mid\boldsymbol{\eta}_i)$, we rewrite (B.5) to obtain

$$-\sum_{i=1}^{n}\left[\frac{\delta_i}{\pi_i(\mathbf{x})}\mathrm{E}(l\mid\boldsymbol{\eta}_i)\{a(\mathbf{x}_i)-\widehat{\mathrm{E}}(a\mid\widetilde{\boldsymbol{\eta}}_i)\}\right]$$
$$=\sum_{i=1}^{n}\frac{\delta_i}{\pi_i(\mathbf{x})}\{l(Y_i)-\mathrm{E}(l\mid\boldsymbol{\eta}_i)\}\{a(\mathbf{x}_i)-\widehat{\mathrm{E}}(a\mid\widetilde{\boldsymbol{\eta}}_i)\}+o_p(n^{1/2}). \tag{B.6}$$

We first study the left hand side of (B.6).

$$
\begin{aligned}
&-\sum_{i=1}^{n}\left[\frac{\delta_i}{\pi_i(\mathbf{x})}\mathrm{E}(l \mid \boldsymbol{\eta}_i)\{a(\mathbf{x}_i)-\widehat{\mathrm{E}}(a \mid \widetilde{\boldsymbol{\eta}}_i)\}\right] \\
&\quad=-\sum_{i=1}^{n}\frac{\delta_i}{\pi_i(\mathbf{x})}\mathrm{E}(l \mid \boldsymbol{\eta}_i)\{a(\mathbf{x}_i)-\mathrm{E}(a \mid \boldsymbol{\eta}_i)\} \\
&\qquad-\sum_{i=1}^{n}\frac{\delta_i}{\pi_i(\mathbf{x})}\mathrm{E}(l \mid \boldsymbol{\eta}_i)\{\mathrm{E}(a \mid \boldsymbol{\eta}_i)-\widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i)\} \\
&\qquad-\sum_{i=1}^{n}\frac{\delta_i}{\pi_i(\mathbf{x})}\mathrm{E}(l \mid \boldsymbol{\eta}_i)\{\widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i)-\widehat{\mathrm{E}}(a \mid \widetilde{\boldsymbol{\eta}}_i)\} \\
&\quad=-(U_{n1}+U_{n2}+S_n)
\end{aligned}
$$

Lemma C.2 gives that $U_{n1}+U_{n2}=o_p(n^{1/2})$. We continue to study $S_n$. Denoting $\otimes$ as Kronecker product, i.e. $\mathbf{A}\otimes\mathbf{B}=(a_{ij}\mathbf{B})$ for any matrices $\mathbf{A}$ and $\mathbf{B}$. Using Taylor's expansion, the weak law of large numbers and the fact that $\partial\mathrm{vec}(fg^T)/\partial\mathbf{x}^T = g\otimes\partial f/\partial\mathbf{x}^T+\partial g/\partial\mathbf{x}^T\otimes f$ and $\mathrm{E}(l \mid \mathbf{x})=\mathrm{E}(l \mid \boldsymbol{\eta})$, we vectorize $S_n$ to obtain the following,

$$
\begin{aligned}
&- \operatorname{vec}\left[\sum_{i=1}^{n} \frac{\delta_i}{\pi_i(\mathbf{x})} \mathrm{E}(l \mid \boldsymbol{\eta}_i)\{\widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i) - \widehat{\mathrm{E}}(a \mid \widetilde{\boldsymbol{\eta}}_i)\}\right] \\
&= \sum_{i=1}^{n} \operatorname{vec}\left[\frac{\delta_i}{\pi_i(\mathbf{x})} \mathrm{E}(l \mid \boldsymbol{\eta}_i)\widehat{\mathrm{E}}(a \mid \widetilde{\boldsymbol{\eta}}_i) - \frac{\delta_i}{\pi_i(\mathbf{x})} \mathrm{E}(l \mid \boldsymbol{\eta}_i)\widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i)\right] \\
&= \sum_{i=1}^{n} \left[\frac{\delta_i}{\pi_i(\mathbf{x})} \frac{\partial \operatorname{vec}\{\mathrm{E}(l \mid \boldsymbol{\eta}_i)\widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i)\}}{\partial \operatorname{vecl}(\mathbf{B})^{\mathrm{T}}}\right] \operatorname{vecl}(\widetilde{\mathbf{B}} - \mathbf{B}) + o_p(n^{1/2}) \\
&= \sum_{i=1}^{n} \left\{\frac{\delta_i}{\pi_i(\mathbf{x})} \frac{\partial \widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i)}{\partial \operatorname{vecl}(\mathbf{B})^{T}} \otimes \mathrm{E}(l \mid \boldsymbol{\eta}_i)\right\} \operatorname{vecl}(\widetilde{\mathbf{B}} - \mathbf{B}) + o_p(n^{1/2}) \\
&= n\mathrm{E}\left\{\frac{\delta}{\pi(\mathbf{x})} \frac{\partial \mathrm{E}(a^{\mathrm{T}} \mid \boldsymbol{\eta})}{\partial \operatorname{vecl}(\mathbf{B})^{T}} \otimes \mathrm{E}(l \mid \boldsymbol{\eta})\right\} \operatorname{vecl}(\widetilde{\mathbf{B}} - \mathbf{B}) + o_p(n^{1/2}) \\
&= n\mathrm{E}\left[\mathrm{E}\left\{\frac{\delta}{\pi(\mathbf{x})} \frac{\partial \mathrm{E}(a^{\mathrm{T}} \mid \boldsymbol{\eta})}{\partial \operatorname{vecl}(\mathbf{B})^{T}} \otimes \mathrm{E}(l \mid \boldsymbol{\eta}) \;\middle|\; \mathbf{x}\right\}\right] \operatorname{vecl}(\widetilde{\mathbf{B}} - \mathbf{B}) + o_p(n^{1/2}) \\
&= n\mathrm{E}\left[\left\{\frac{\mathrm{E}(\delta \mid \mathbf{x})}{\pi(\mathbf{x})}\right\} \mathrm{E}\left\{\frac{\partial \mathrm{E}(a^{\mathrm{T}} \mid \boldsymbol{\eta})}{\partial \operatorname{vecl}(\mathbf{B})^{T}} \otimes \mathrm{E}(l \mid \boldsymbol{\eta}) \;\middle|\; \mathbf{x}\right\}\right] \operatorname{vecl}(\widetilde{\mathbf{B}} - \mathbf{B}) + o_p(n^{1/2}) \\
&= n\mathrm{E}\left\{\frac{\partial \mathrm{E}(a^{\mathrm{T}} \mid \boldsymbol{\eta})}{\partial \operatorname{vecl}(\mathbf{B})^{T}} \otimes \mathrm{E}(l \mid \boldsymbol{\eta})\right\} \operatorname{vecl}(\widetilde{\mathbf{B}} - \mathbf{B}) + o_p(n^{1/2}) \\
&= -n\mathbf{A}\operatorname{vecl}(\widetilde{\mathbf{B}} - \mathbf{B}) + o_p(n^{1/2}) \qquad \text{(B.7)}
\end{aligned}
$$

where

$$
\mathbf{A} = \mathrm{E}\left\{\frac{\partial \operatorname{vec}\left[\{g(Y) - \mathrm{E}(l \mid \boldsymbol{\eta})\}\{a(\mathbf{x}) - \mathrm{E}(a \mid \boldsymbol{\eta})\}\right]}{\partial \operatorname{vecl}(\mathbf{B})^{T}}\right\}
$$

Let $a(\mathbf{x}, \boldsymbol{\eta}) = a(\mathbf{x}) - \mathrm{E}(a \mid \boldsymbol{\eta})$ and $l(Y, \boldsymbol{\eta}) = l(Y) - \mathrm{E}(l \mid \boldsymbol{\eta})$, we can show that

the last equality of (B.7) holds by the following derivation.

$$
\begin{aligned}
\mathbf{A} =& \mathrm{E}\left\{\frac{\partial \mathrm{vec}\left[\{l(Y)-\mathrm{E}(l \mid \boldsymbol{\eta})\}\{a(\mathbf{x})-\mathrm{E}(a \mid \boldsymbol{\eta})\}\right]}{\partial \mathrm{vecl}(\mathbf{B})^{T}}\right\} \\
=& \mathrm{E}\left\{a^{T}(\mathbf{x}, \boldsymbol{\eta}) \otimes \frac{\partial l(Y, \boldsymbol{\eta})}{\partial \mathrm{vecl}(\mathbf{B})^{\mathrm{T}}}\right\}+\mathrm{E}\left\{\frac{\partial a^{T}(\mathbf{x}, \boldsymbol{\eta})}{\partial \mathrm{vecl}(\mathbf{B})^{\mathrm{T}}} \otimes l(Y, \boldsymbol{\eta})\right\} \\
=&-\mathrm{E}\left\{a^{T}(\mathbf{x}, \boldsymbol{\eta}) \otimes \frac{\partial \mathrm{E}(l \mid \boldsymbol{\eta})}{\mathrm{vecl}(\mathbf{B})^{\mathrm{T}}}\right\}-\mathrm{E}\left\{\frac{\partial \mathrm{E}(a \mid \boldsymbol{\eta})^{T}}{\mathrm{vecl}(\mathbf{B})^{\mathrm{T}}} \otimes l(Y, \boldsymbol{\eta})\right\} \\
=&-\mathrm{E}\left\{\frac{\partial \mathrm{vec}\{\mathrm{E}(l \mid \boldsymbol{\eta}) a(\mathbf{x}, \boldsymbol{\eta})\}}{\partial \mathrm{vecl}(\mathbf{B})^{\mathrm{T}}}\right\}-\mathrm{E}\left\{\frac{\partial \mathrm{E}(a^{\mathrm{T}} \mid \boldsymbol{\eta})}{\mathrm{vecl}(\mathbf{B})^{\mathrm{T}}} \otimes \mathrm{E}(l \mid \boldsymbol{\eta})\right\} \\
&-\mathrm{E}\left\{\frac{\partial \mathrm{E}(a^{\mathrm{T}} \mid \boldsymbol{\eta})}{\mathrm{vecl}(\mathbf{B})^{\mathrm{T}}} \otimes l(Y)\right\}+\mathrm{E}\left\{\frac{\partial \mathrm{E}(a^{\mathrm{T}} \mid \boldsymbol{\eta})}{\mathrm{vecl}(\mathbf{B})^{\mathrm{T}}} \otimes \mathrm{E}(l \mid \boldsymbol{\eta})\right\} \\
=&-\mathrm{E}\left\{\frac{\partial \mathrm{E}(a^{\mathrm{T}} \mid \boldsymbol{\eta})}{\mathrm{vecl}(\mathbf{B})^{\mathrm{T}}} \otimes l(Y)\right\} \\
=&-\mathrm{E}\left\{\frac{\partial \mathrm{E}(a^{\mathrm{T}} \mid \boldsymbol{\eta})}{\partial \mathrm{vecl}(\mathbf{B})^{T}} \otimes \mathrm{E}(l \mid \boldsymbol{\eta})\right\},
\end{aligned}
$$

where the last two equalities hold because $\mathrm{E}\left[\mathrm{E}(l \mid \boldsymbol{\eta})\{a(\mathbf{x})-\mathrm{E}(a \mid \boldsymbol{\eta})\}\right]=\mathbf{0}$ and $\mathrm{E}\{l(Y)-\mathrm{E}(l \mid \boldsymbol{\eta})\}=\mathbf{0}$, hence

$$
\begin{gathered}
\mathrm{E}\left\{\frac{\partial \mathrm{vec}[\mathrm{E}(l \mid \boldsymbol{\eta}) a(\mathbf{x}, \boldsymbol{\eta})]}{\partial \mathrm{vecl}(\mathbf{B})^{\mathrm{T}}}\right\}=\mathbf{0} \\
\mathrm{E}\left\{\frac{\partial \mathrm{vecE}(a^{\mathrm{T}} \mid \boldsymbol{\eta})}{\partial \mathrm{vecl}(\mathbf{B})^{T}} \otimes\{l(Y)-\mathrm{E}(l \mid \boldsymbol{\eta})\}\right\}=\mathbf{0}
\end{gathered}
$$

Next we study the right hand side of (B.6).

$$
\begin{aligned}
&\sum_{i=1}^{n} \frac{\delta_i}{\pi_i(\mathbf{x})}\{l(Y_i) - \mathrm{E}(l \mid \boldsymbol{\eta}_i)\}\{a(\mathbf{x}_i) - \widehat{\mathrm{E}}(a \mid \widetilde{\boldsymbol{\eta}}_i)\} + o_p(n^{1/2}) \\
&= \sum_{i=1}^{n} \frac{\delta_i}{\pi_i(\mathbf{x})}\{l(Y_i) - \mathrm{E}(l \mid \boldsymbol{\eta}_i)\}\{a(\mathbf{x}_i) - \mathrm{E}(a \mid \boldsymbol{\eta}_i)\} \\
&\quad + \sum_{i=1}^{n} \frac{\delta_i}{\pi_i(\mathbf{x})}\{l(Y_i) - \mathrm{E}(l \mid \boldsymbol{\eta}_i)\}\{\mathrm{E}(a \mid \boldsymbol{\eta}_i) - \widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i)\} \\
&\quad + \sum_{i=1}^{n} \frac{\delta_i}{\pi_i(\mathbf{x})}\{l(Y_i) - \mathrm{E}(l \mid \boldsymbol{\eta}_i)\}\{\widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i) - \widehat{\mathrm{E}}(a \mid \widetilde{\boldsymbol{\eta}}_i)\} + o_p(n^{1/2}) \\
&= R_n + T_n + W_n + o_p(n^{1/2})
\end{aligned}
$$

Because $\mathrm{E}\{l(Y) - \mathrm{E}(l \mid \boldsymbol{\eta})\} = \mathbf{0}$, envoking Lemma B.3 results that $T_n = O_p\{n^{1/2}h^m + nh^{2m} + \log^2 n/h^d\}$, which is $o_p(n^{1/2})$ when $nh^{2d} \to \infty$ and $nh^{4m} \to \infty$. We vectorize $W_n$ by Taylor expansion,

$$
\begin{aligned}
&\mathrm{vec}\left[\sum_{i=1}^{n} \frac{\delta_i}{\pi_i(\mathbf{x})}\{l(Y_i) - \mathrm{E}(l \mid \boldsymbol{\eta}_i)\}\{\widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i) - \widehat{\mathrm{E}}(a \mid \widetilde{\boldsymbol{\eta}}_i)\}\right] \\
&= \sum_{i=1}^{n} \mathrm{vec}\left[\frac{\delta_i}{\pi_i(\mathbf{x})}\{l(Y_i) - \mathrm{E}(l \mid \boldsymbol{\eta}_i)\}\widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i) - \frac{\delta_i}{\pi_i(\mathbf{x})}\{l(Y_i) - \mathrm{E}(l \mid \boldsymbol{\eta}_i)\}\widehat{\mathrm{E}}(a \mid \widetilde{\boldsymbol{\eta}}_i)\}\right] \\
&= -\sum_{i=1}^{n} \left\{\frac{\delta_i}{\pi_i(\mathbf{x})} \frac{\partial \mathrm{vec}[\{l(Y_i) - \mathrm{E}(l \mid \boldsymbol{\eta}_i)\}\widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i)]}{\partial \mathrm{vecl}(\mathbf{B})^{\mathrm{T}}}\right\} \mathrm{vecl}(\widetilde{\mathbf{B}} - \mathbf{B}) + o_p(n^{1/2}) \\
&= -\sum_{i=1}^{n} \left[\frac{\delta_i}{\pi_i(\mathbf{x})} \frac{\partial \widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i)}{\partial \mathrm{vecl}(\mathbf{B})^{\mathrm{T}}} \otimes \{l(Y_i) - \mathrm{E}(l \mid \boldsymbol{\eta}_i)\}\right] \mathrm{vecl}(\widetilde{\mathbf{B}} - \mathbf{B}) + o_p(n^{1/2}) \\
&= -n\mathrm{E}\left[\frac{\delta}{\pi(\mathbf{x})} \frac{\partial \mathrm{E}(a \mid \boldsymbol{\eta})}{\partial \mathrm{vecl}(\mathbf{B})^{\mathrm{T}}} \otimes \{l(Y) - \mathrm{E}(l \mid \boldsymbol{\eta})\}\right] \mathrm{vecl}(\widetilde{\mathbf{B}} - \mathbf{B}) + o_p(n^{1/2}) \\
&= o_p(n^{1/2}),
\end{aligned}
$$

where the last equality holds because $\mathrm{E}\{l(Y) - \mathrm{E}(l \mid \boldsymbol{\eta})\} = \mathbf{0}$. Vectorizing $R_n$ completes the vectorization of the right hand side of (B.6). To summarize the

results concerning (B.6), we obtain

$$-n\mathbf{A}\text{vecl}(\widetilde{\mathbf{B}}-\mathbf{B})=\sum_{i=1}^{n}\text{vec}\left[\frac{\delta_i}{\pi_i(\mathbf{x})}\{l(Y_i)-\text{E}(l\mid\boldsymbol{\eta}_i)\}\{a(\mathbf{x}_i)-\text{E}(a\mid\boldsymbol{\eta}_i)\}\right]+o_p(n^{1/2}) \tag{B.8}$$

The proof of Theorem 3.2 is completed by accessing the asymptotic property of $\widetilde{\mathbf{B}}$ from (B.8).

□

*Proof of Theorem 3.3.* We redefine $\widehat{\text{g}}_i^\star = l(Y_i)\{a(\mathbf{x}_i)-\widehat{\text{E}}(a\mid\widehat{\boldsymbol{\eta}}_i)\}$, and denote $\widehat{m}(\widehat{\text{g}}_i^\star)=\widehat{\text{E}}(\widehat{\text{g}}_i^\star\mid\mathbf{x}_i)$ to show the following.

$$\begin{aligned}
&\sum_{i=1}^{n}\widehat{m}\left[l(Y_i)\{a(\mathbf{x}_i)-\widehat{\text{E}}(a\mid\widehat{\boldsymbol{\eta}}_i)\}\right]=\sum_{i=1}^{n}\widehat{m}\left[g(Y_i)\{a(\mathbf{x}_i)-\text{E}(a\mid\boldsymbol{\eta}_i)\}\right]\\
&\quad+\sum_{i=1}^{n}\widehat{m}\left[l(Y_i)\{a(\mathbf{x}_i)-\widehat{\text{E}}(a\mid\widehat{\boldsymbol{\eta}}_i)\}-l(Y_i)\{a(\mathbf{x}_i)-\widehat{\text{E}}(a\mid\boldsymbol{\eta}_i)\}\right]\\
&\quad+\sum_{i=1}^{n}\widehat{m}\left[l(Y_i)\{a(\mathbf{x}_i)-\widehat{\text{E}}(a\mid\boldsymbol{\eta}_i)\}-l(Y_i)\{a(\mathbf{x}_i)-\text{E}(a\mid\boldsymbol{\eta}_i)\}\right]\\
&=\sum_{i=1}^{n}\widehat{m}\left(l(Y_i)\{a(\mathbf{x}_i)-\text{E}(a\mid\boldsymbol{\eta}_i)\}\right)+M_{1n}+M_{2n}\\
&=\sum_{i=1}^{n}\widehat{m}\left[l(Y_i)\{a(\mathbf{x}_i)-\text{E}(a\mid\boldsymbol{\eta}_i)\}\right]+o_p(n^{1/2})\\
&=M_{3n}+\sum_{i=1}^{n}m\left[l(Y_i)\{a(\mathbf{x}_i)-\text{E}(a\mid\boldsymbol{\eta}_i)\}\right]+o_p(n^{1/2})\\
&=\sum_{i=1}^{n}m\left[l(Y_i)\{a(\mathbf{x}_i)-\text{E}(a\mid\boldsymbol{\eta}_i)\}\right]+o_p(n^{1/2})
\end{aligned}$$

where

$$M_{1n} = \sum_{i=1}^{n} \widehat{m} \left[ l(Y_i)\{a(\mathbf{x}_i) - \widehat{\mathrm{E}}(a \mid \widehat{\boldsymbol{\eta}}_i)\} - l(Y_i)\{a(\mathbf{x}_i) - \widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i)\} \right]$$

$$M_{2n} = \sum_{i=1}^{n} \widehat{m} \left[ l(Y_i)\{a(\mathbf{x}_i) - \widehat{\mathrm{E}}(a \mid \boldsymbol{\eta}_i)\} - l(Y_i)\{a(\mathbf{x}_i) - \mathrm{E}(a \mid \boldsymbol{\eta}_i)\} \right]$$

$$M_{3n} = \sum_{i=1}^{n} \{\widehat{m} \left[ l(Y_i)\{a(\mathbf{x}_i) - \mathrm{E}(a \mid \boldsymbol{\eta}_i)\} \right] - m \left[ l(Y_i)\{a(\mathbf{x}_i) - \mathrm{E}(a \mid \boldsymbol{\eta}_i)\} \right]\}$$

We will state lemmas to show that $M_{in}$s, $i = 1, 2, 3$, are of order $o_p(n^{1/2})$. Thus by further applying $(B.2)$ and Lemma C.1, the estimating equation which yields $\widehat{\mathbf{B}}$ is thus given as

$$\sum_{i=1}^{n} \left\{ \frac{\delta_i}{\pi_i(\mathbf{x})} l(Y_i)\{a(\mathbf{x}_i) - \widehat{\mathrm{E}}(a \mid \widehat{\boldsymbol{\eta}}_i)\} + \left(1 - \frac{\delta_i}{\pi_i(\mathbf{x})}\right) m \left( l(Y_i)\{a(\mathbf{x}_i) - \mathrm{E}(a \mid \boldsymbol{\eta}_i)\} \right) \right\} + o_p(n^{1/2}) = \mathbf{0} \tag{B.9}$$

We rewrite (B.9) to obtain

$$\begin{aligned} &- \sum_{i=1}^{n} \left\{ \frac{\delta_i}{\pi_i(\mathbf{x})} \mathrm{E}(l \mid \boldsymbol{\eta}_i)\{a(\mathbf{x}_i) - \widehat{\mathrm{E}}(a \mid \widehat{\boldsymbol{\eta}}_i) \right\} \\ &= \sum_{i=1}^{n} \left\{ \frac{\delta_i}{\pi_i(\mathbf{x})} \{l(Y_i) - \mathrm{E}(l \mid \boldsymbol{\eta}_i)\}\{a(\mathbf{x}_i) - \widehat{\mathrm{E}}(a \mid \widehat{\boldsymbol{\eta}}_i) \right\} \\ &\quad + \left(1 - \frac{\delta_i}{\pi_i(\mathbf{x})}\right) \mathrm{E}\left[ l(Y_i)\{a(\mathbf{x}_i) - \mathrm{E}(a \mid \boldsymbol{\eta}_i)\} \mid \mathbf{x} \right] + o_p(n^{1/2}) \end{aligned} \tag{B.10}$$

By the proof of Theorem 3.2, we get the vectorized left hand side of (B.10) is $-n\mathbf{A}\mathrm{vecl}(\widehat{\mathbf{B}} - \mathbf{B})$. Also we can show that

$$\begin{aligned}
&\mathrm{E}\left[l(Y_i)\{a(\mathbf{x}_i)-\mathrm{E}(a \mid \boldsymbol{\eta}_i)\} \mid \mathbf{x}\right] = \mathrm{E}\left[\{l(Y_i)-\mathrm{E}(l \mid \boldsymbol{\eta}_i)\}\{a(\mathbf{x}_i)-\mathrm{E}(a \mid \boldsymbol{\eta}_i)\} \mid \mathbf{x}\right] \\
&\quad + \mathrm{E}\left[\mathrm{E}(l \mid \boldsymbol{\eta}_i)\{a(\mathbf{x}_i)-\mathrm{E}(a \mid \boldsymbol{\eta}_i)\} \mid \mathbf{x}\right] \qquad \text{(B.11)} \\
&= \mathrm{E}\left[\{l(Y_i)-\mathrm{E}(l \mid \boldsymbol{\eta}_i)\}\{a(\mathbf{x}_i)-\mathrm{E}(a \mid \boldsymbol{\eta}_i)\} \mid \mathbf{x}\right]
\end{aligned}$$

The last equality holds for $\mathrm{E}\left[\mathrm{E}(l \mid \boldsymbol{\eta})\{a(\mathbf{x})-\mathrm{E}(a \mid \boldsymbol{\eta})\}\right] = \mathbf{0}$. Therefor the proof of Theorem 3.3 is completed by combining (B.11) and then vectorizing the right hand side of (B.10) and access the asymptotic property for $\widehat{\mathbf{B}}$.

□

*Proof of Theorem 3.4.* From Theorem 3.2 and 3.3, we can easily obtain that

$$\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2 = -\mathrm{Cov}\left[\mathrm{vec}\left\{\left(1-\frac{\delta}{\pi(\mathbf{x})}\right)\mathrm{E}(\mathrm{g} \mid \mathbf{x})\right\}\right] - 2\mathrm{E}\left[\mathrm{vec}\left\{\frac{\delta}{\pi(\mathbf{x})}\left(1-\frac{\delta}{\pi(\mathbf{x})}\right)\mathrm{gE}(\mathrm{g}^T \mid \mathbf{x})\right\}\right]$$

Similar to the proof of Lemma 1 in Tang and Qin (2012),

$$\mathrm{Cov}\left[\mathrm{vec}\left\{\left(1-\frac{\delta}{\pi(\mathbf{x})}\right)\mathrm{E}(\mathrm{g} \mid \mathbf{x})\right\}\right] = \mathrm{E}\left[\mathrm{vec}\left\{\left(\frac{1}{\pi(\mathbf{x})}-1\right)\mathrm{E}(\mathrm{g} \mid \mathbf{x})\mathrm{E}(\mathrm{g}^T \mid \mathbf{x})\right\}\right]$$

and

$$\mathrm{E}\left[\mathrm{vec}\left\{\frac{\delta}{\pi(\mathbf{x})}\left(1-\frac{\delta}{\pi(\mathbf{x})}\right)\mathrm{gE}(\mathrm{g}^T \mid \mathbf{x})\right\}\right] = \mathrm{E}\left[\mathrm{vec}\left\{\left(1-\frac{1}{\pi(\mathbf{x})}\right)\mathrm{E}(\mathrm{g} \mid \mathbf{x})\mathrm{E}(\mathrm{g}^T \mid \mathbf{x})\right\}\right]$$

Thus

$$\boldsymbol{\Gamma}_1 - \boldsymbol{\Gamma}_2 = \mathrm{E}\left[\mathrm{vec}\left\{\left(\frac{1}{\pi(\mathbf{x})}-1\right)\mathrm{E}(\mathrm{g} \mid \mathbf{x})\mathrm{E}(\mathrm{g}^T \mid \mathbf{x})\right\}\right]$$

which is positive definite. Hence

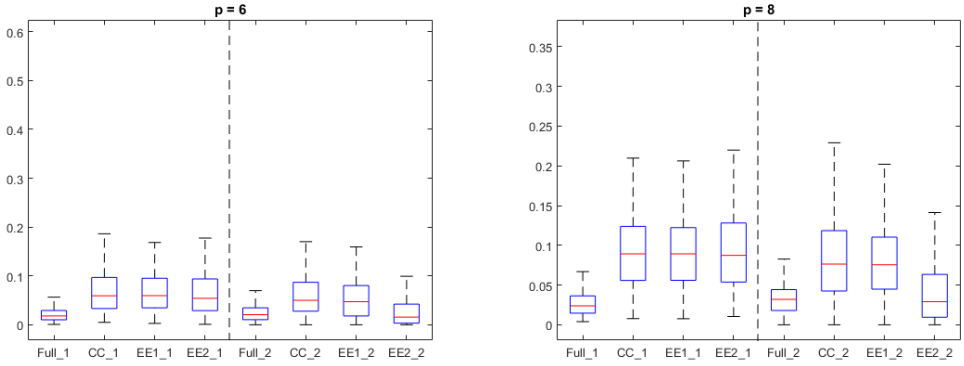
$$n\left[\text{Cov}\{\text{vecl}(\widetilde{\mathbf{B}})\} - \text{Cov}\{\text{vecl}(\widehat{\mathbf{B}})\}\right] = \mathbf{A}^{-1}(\mathbf{\Gamma}_1 - \mathbf{\Gamma}_2)\mathbf{A}^{-1}$$

is positive definite.

□

## B.2 More Simulation Results

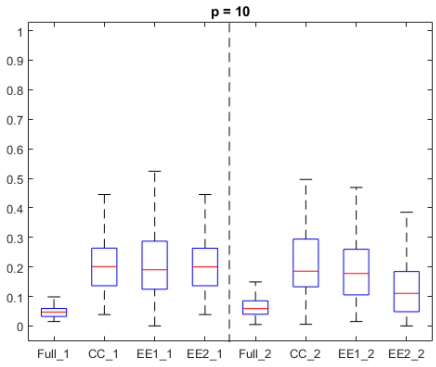
With the same simulation setting in Chapter 3, we have more numerical results for OLS and PHD estimating equations for the following two models. Model (I): $Y = \mathbf{B}^{\mathrm{T}}\mathbf{x} + \varepsilon_1$ for OLS method; $Y = (\mathbf{B}^{\mathrm{T}}\mathbf{x})^2 + \varepsilon_2$ for PHD method; Model (II): $Y = \exp(\mathbf{B}^{\mathrm{T}}\mathbf{x}) + \varepsilon_3$, where $\mathbf{B}$ is a $p$-dimensional vector and $\varepsilon_i$'s $\sim \mathcal{N}(0,1)$ is the random error term. As we fix $q = 1$, we let $\mathbf{B} = (\beta_1, ..., \beta_p)^T$, where $\beta_1 = 1$ and $(\beta_2, ..., \beta_p)^T = (1, ..., 1)^T/\sqrt{p-1}$. The dimension $p$ is chosen to be 6, 8, 10 and 12 respectively. The boxplots of Euclidean distance are showed in Figure B.1 to B.8. In addition, the canonical correlation and Euclidean distance are reported in Table B.1 to B.8. The numerical results match the finding in Section 3.5



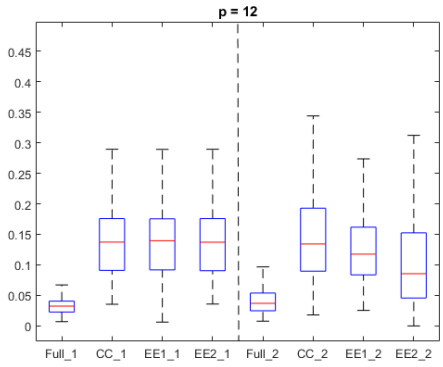

(a) $p = 6$




(b) $p = 8$

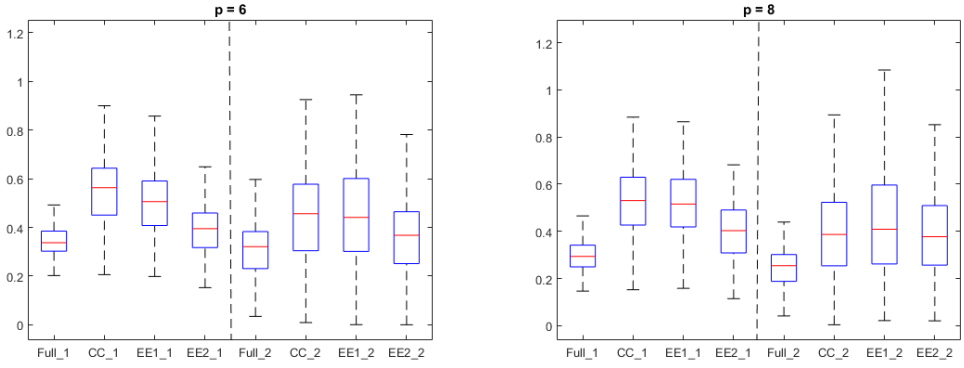



(c) $p = 10$




(d) $p = 12$

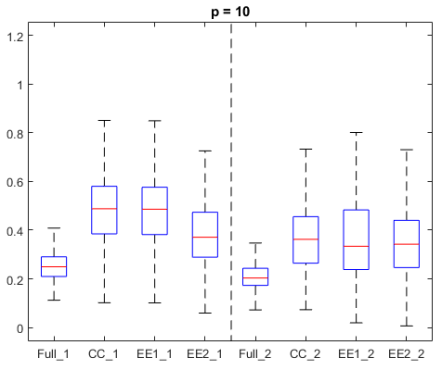
Figure B.1: Boxplot of Euclidean distance for OLS Case (i)-Model (I)



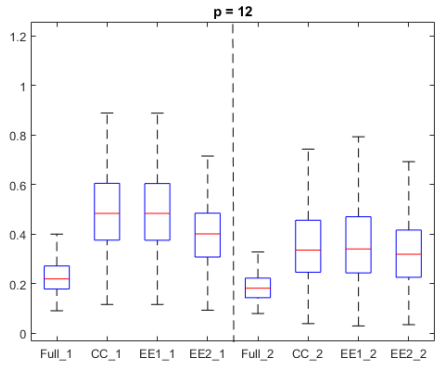

(a) $p = 6$




(b) $p = 8$

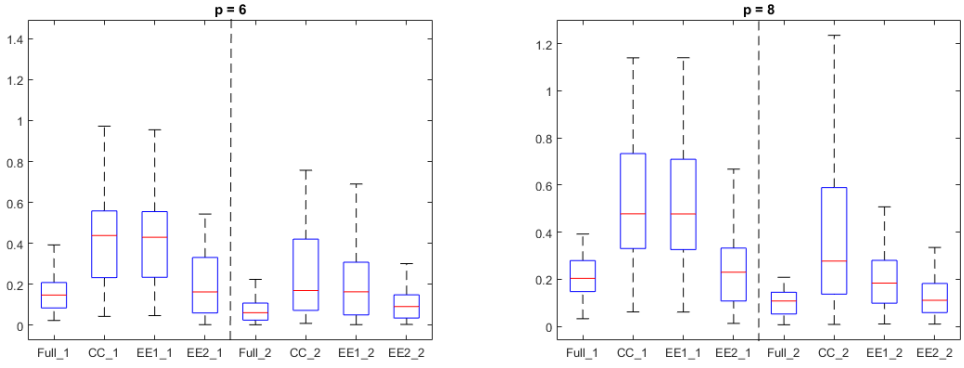



(c) $p = 10$




(d) $p = 12$

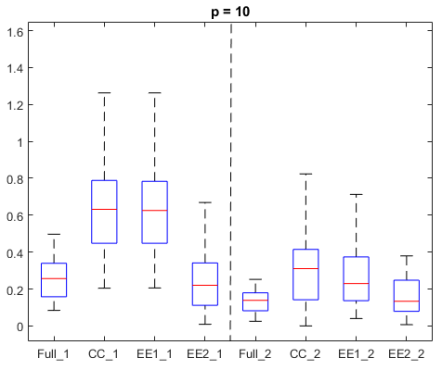
Figure B.2: Boxplot of Euclidean distance for OLS Case (ii)-Model (I)



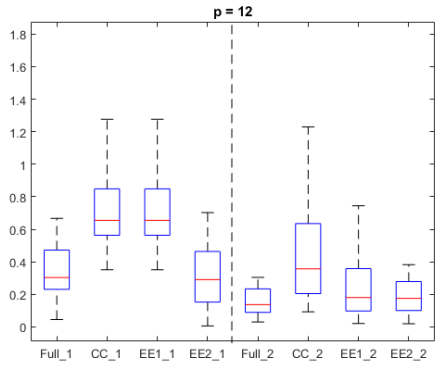

(a) $p = 6$




(b) $p = 8$

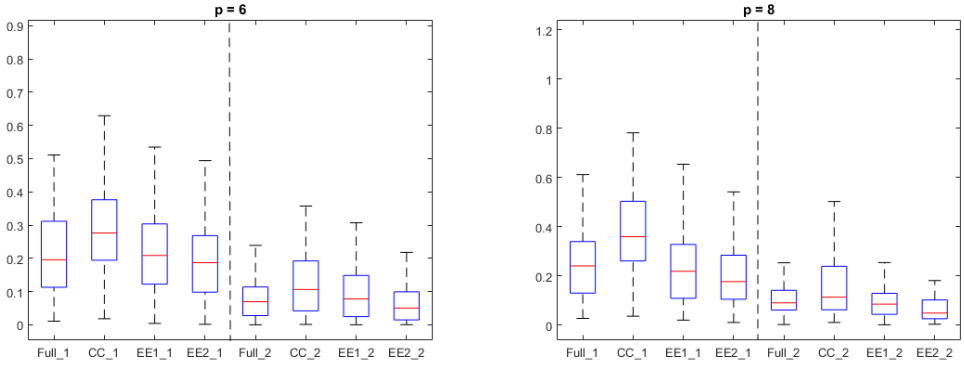



(c) $p = 10$




(d) $p = 12$

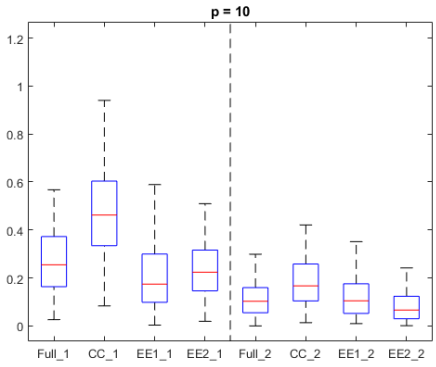
Figure B.3: Boxplot of Euclidean distance for PHD Case (i)-Model (I)



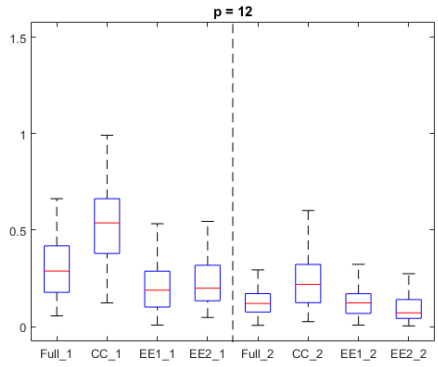

(a) $p = 6$




(b) $p = 8$




(c) $p = 10$




(d) $p = 12$

Figure B.4: Boxplot of Euclidean distance for PHD Case (ii)-Model (I)

| **B** | p=6 | | p=8 | | p=10 | | p=12 | |
|---|---|---|---|---|---|---|---|---|
| | Corr | Dis | Corr | Dis | Corr | Dis | Corr | Dis |
| Full_1 | 0.995 | 0.021 | 0.993 | 0.027 | 0.988 | 0.049 | 0.992 | 0.033 |
| CC_1 | 0.980 | 0.072 | 0.975 | 0.093 | 0.950 | 0.218 | 0.964 | 0.143 |
| EE1_1 | 0.982 | 0.070 | 0.976 | 0.093 | 0.942 | 0.224 | 0.963 | 0.143 |
| EE2_1 | 0.981 | 0.075 | 0.975 | 0.097 | 0.944 | 0.217 | 0.963 | 0.143 |
| Full_2 | 0.994 | 0.024 | 0.992 | 0.033 | 0.984 | 0.065 | 0.990 | 0.041 |
| CC_2 | 0.984 | 0.062 | 0.977 | 0.089 | 0.940 | 0.229 | 0.963 | 0.146 |
| EE1_2 | 0.986 | 0.057 | 0.979 | 0.084 | 0.951 | 0.192 | 0.967 | 0.129 |
| EE2_2 | 0.992 | 0.032 | 0.988 | 0.049 | 0.967 | 0.128 | 0.972 | 0.109 |

Table B.1: Correlation and Euclidean distance for OLS Case (i)-Model (I)

| **B** | p=6 | | p=8 | | p=10 | | p=12 | |
|---|---|---|---|---|---|---|---|---|
| | Corr | Dis | Corr | Dis | Corr | Dis | Corr | Dis |
| Full_1 | 0.911 | 0.340 | 0.923 | 0.296 | 0.935 | 0.253 | 0.942 | 0.225 |
| CC_1 | 0.863 | 0.555 | 0.863 | 0.525 | 0.874 | 0.477 | 0.874 | 0.488 |
| EE1_1 | 0.864 | 0.503 | 0.860 | 0.516 | 0.872 | 0.476 | 0.868 | 0.487 |
| EE2_1 | 0.896 | 0.393 | 0.894 | 0.401 | 0.899 | 0.380 | 0.892 | 0.406 |
| Full_2 | 0.918 | 0.313 | 0.935 | 0.251 | 0.947 | 0.207 | 0.952 | 0.188 |
| CC_2 | 0.877 | 0.453 | 0.892 | 0.403 | 0.901 | 0.372 | 0.902 | 0.369 |
| EE1_2 | 0.878 | 0.451 | 0.907 | 0.436 | 0.904 | 0.362 | 0.904 | 0.362 |
| EE2_2 | 0.902 | 0.368 | 0.896 | 0.390 | 0.909 | 0.346 | 0.910 | 0.341 |

Table B.2: Correlation and Euclidean distance for OLS Case (ii)-Model (I)

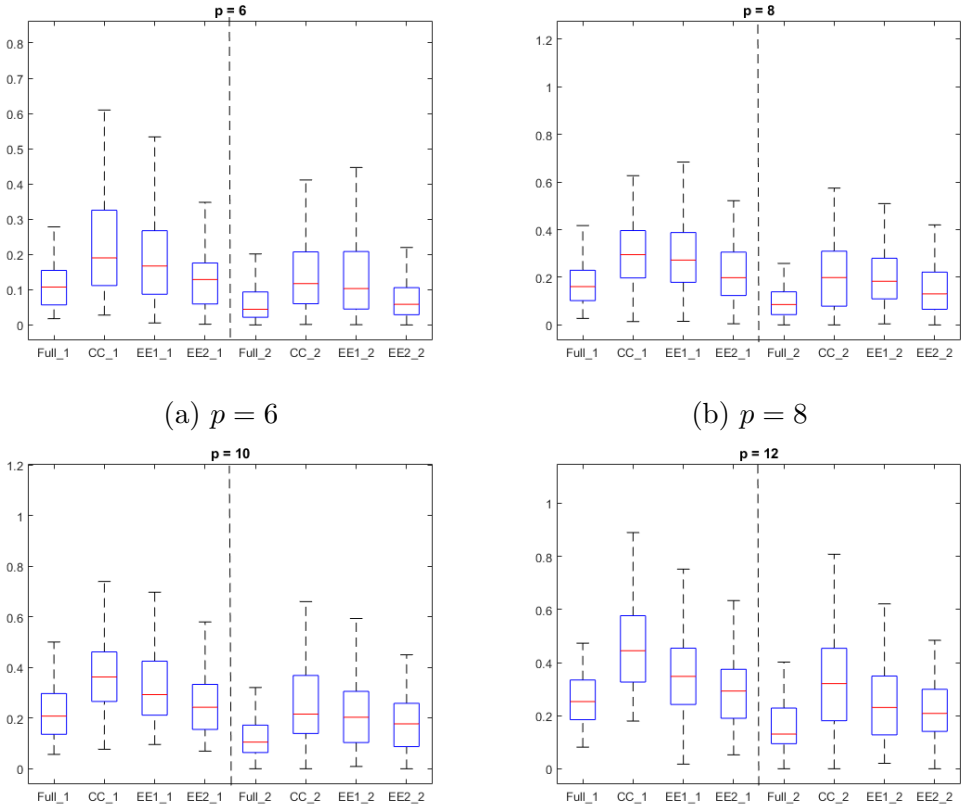
| **B** | p=6 | | p=8 | | p=10 | | p=12 | |
|---|---|---|---|---|---|---|---|---|
| | Corr | Dis | Corr | Dis | Corr | Dis | Corr | Dis |
| Full_1 | 0.957 | 0.166 | 0.940 | 0.229 | 0.927 | 0.277 | 0.910 | 0.340 |
| CC_1 | 0.885 | 0.454 | 0.865 | 0.535 | 0.836 | 0.639 | 0.804 | 0.753 |
| EE1_1 | 0.877 | 0.445 | 0.854 | 0.529 | 0.821 | 0.638 | 0.781 | 0.753 |
| EE2_1 | 0.934 | 0.247 | 0.927 | 0.275 | 0.928 | 0.270 | 0.908 | 0.336 |
| Full_2 | 0.980 | 0.080 | 0.970 | 0.119 | 0.957 | 0.165 | 0.955 | 0.172 |
| CC_2 | 0.925 | 0.273 | 0.894 | 0.381 | 0.910 | 0.334 | 0.879 | 0.439 |
| EE1_2 | 0.936 | 0.238 | 0.947 | 0.202 | 0.922 | 0.292 | 0.936 | 0.242 |
| EE2_2 | 0.964 | 0.137 | 0.958 | 0.159 | 0.957 | 0.166 | 0.948 | 0.201 |

Table B.3: Correlation and Euclidean distance for PHD Case (i)-Model (I)

| **B** | p=6 | | p=8 | | p=10 | | p=12 | |
|---|---|---|---|---|---|---|---|---|
| | Corr | Dis | Corr | Dis | Corr | Dis | Corr | Dis |
| Full_1 | 0.944 | 0.217 | 0.932 | 0.260 | 0.924 | 0.288 | 0.917 | 0.313 |
| CC_1 | 0.943 | 0.301 | 0.919 | 0.389 | 0.893 | 0.472 | 0.877 | 0.534 |
| EE1_1 | 0.941 | 0.226 | 0.940 | 0.232 | 0.946 | 0.206 | 0.946 | 0.209 |
| EE2_1 | 0.950 | 0.193 | 0.948 | 0.202 | 0.939 | 0.234 | 0.941 | 0.228 |
| Full_2 | 0.979 | 0.083 | 0.972 | 0.109 | 0.971 | 0.113 | 0.965 | 0.138 |
| CC_2 | 0.963 | 0.142 | 0.956 | 0.169 | 0.947 | 0.203 | 0.931 | 0.261 |
| EE1_2 | 0.972 | 0.108 | 0.973 | 0.104 | 0.968 | 0.125 | 0.962 | 0.146 |
| EE2_2 | 0.978 | 0.086 | 0.981 | 0.075 | 0.978 | 0.086 | 0.974 | 0.100 |

Table B.4: Correlation and Euclidean distance for PHD Case (ii)-Model (I)

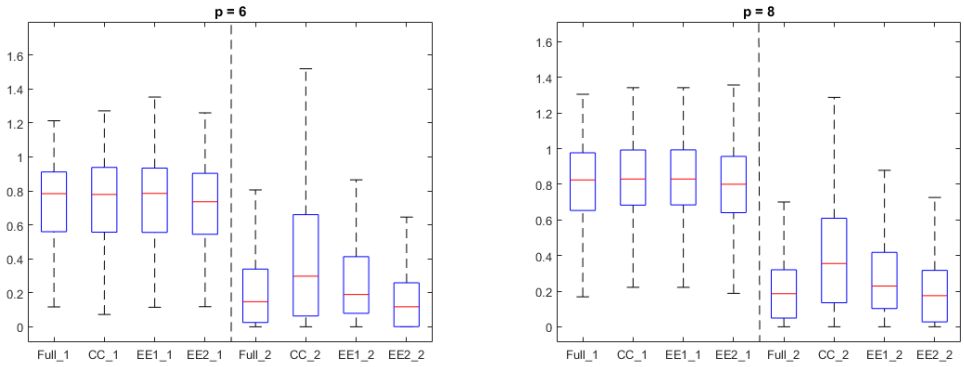
**Figures and Tables for Model (II)**



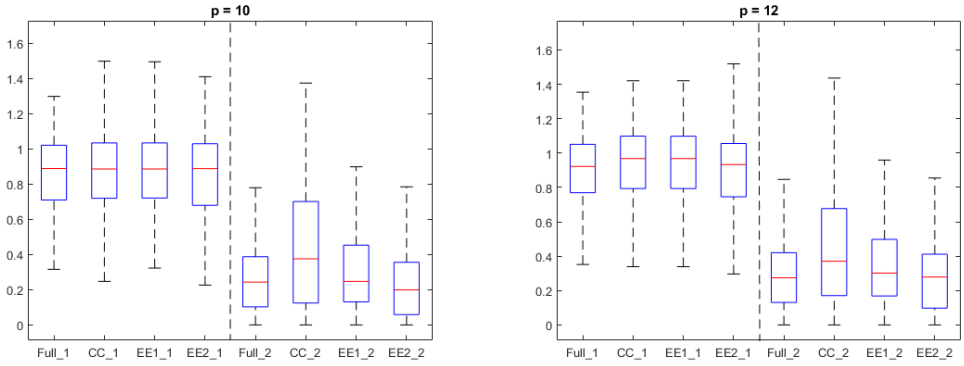

(a) $p = 6$ (b) $p = 8$

(c) $p = 10$ (d) $p = 12$

Figure B.5: Boxplot of Euclidean distance for OLS Case (i)-Model (II)




(a) $p = 6$ (b) $p = 8$

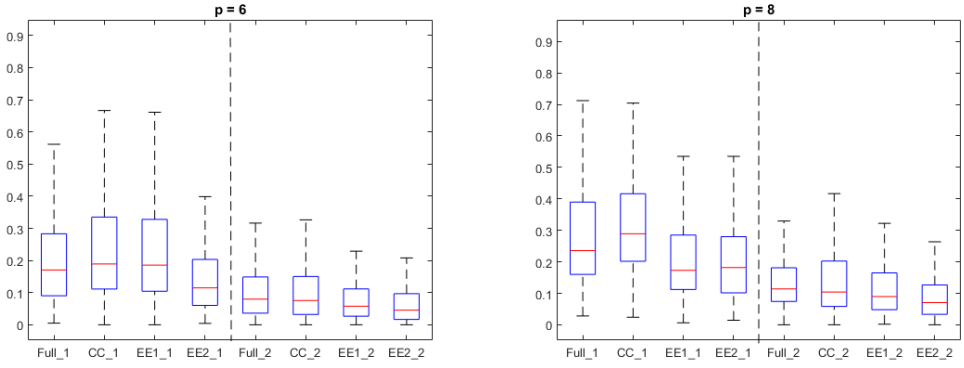



(c) $p = 10$ (d) $p = 12$

Figure B.6: Boxplot of Euclidean distance for OLS Case (ii)-Model (II)

| **B** | p=6 | | p=8 | | p=10 | | p=12 | |
|---|---|---|---|---|---|---|---|---|
| | Corr | Dis | Corr | Dis | Corr | Dis | Corr | Dis |
| Full_1 | 0.967 | 0.127 | 0.954 | 0.180 | 0.944 | 0.218 | 0.931 | 0.266 |
| CC_1 | 0.938 | 0.235 | 0.923 | 0.313 | 0.907 | 0.379 | 0.887 | 0.461 |
| EE1_1 | 0.948 | 0.200 | 0.923 | 0.291 | 0.914 | 0.327 | 0.902 | 0.369 |
| EE2_1 | 0.963 | 0.143 | 0.942 | 0.222 | 0.932 | 0.260 | 0.920 | 0.305 |
| Full_2 | 0.984 | 0.065 | 0.975 | 0.099 | 0.967 | 0.129 | 0.958 | 0.164 |
| CC_2 | 0.961 | 0.150 | 0.941 | 0.225 | 0.929 | 0.269 | 0.906 | 0.350 |
| EE1_2 | 0.963 | 0.144 | 0.944 | 0.215 | 0.938 | 0.236 | 0.935 | 0.248 |
| EE2_2 | 0.980 | 0.080 | 0.959 | 0.160 | 0.954 | 0.179 | 0.940 | 0.229 |

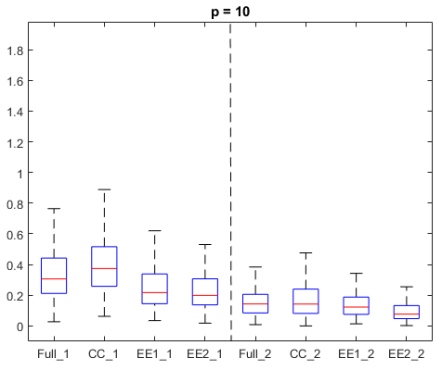
Table B.5: Correlation and Euclidean distance for OLS Case (i)-Model (II)



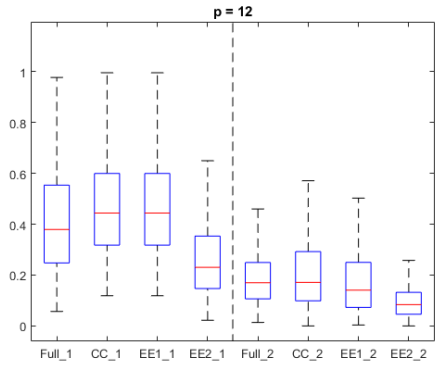

(a) $p = 6$




(b) $p = 8$

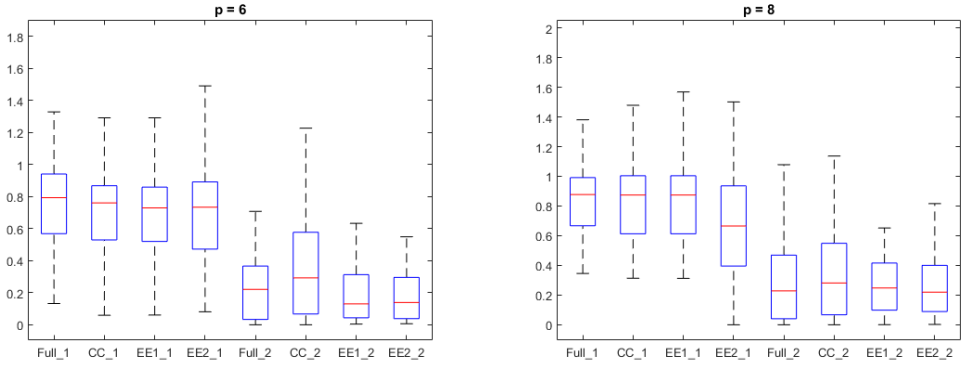



(c) $p = 10$




(d) $p = 12$

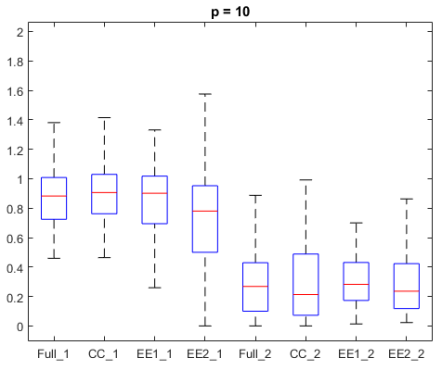
Figure B.7: Boxplot of Euclidean distance for PHD Case (i)-Model (II)



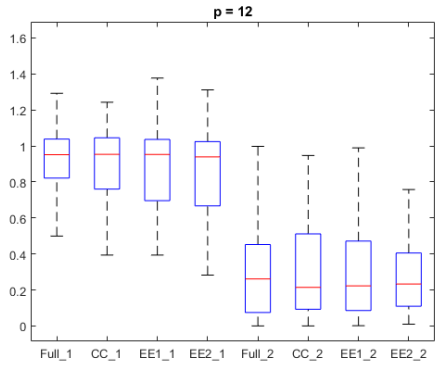

(a) $p = 6$




(b) $p = 8$




(c) $p = 10$




(d) $p = 12$

Figure B.8: Boxplot of Euclidean distance for PHD Case (ii)-Model (II)

| **B** | p=6 | | p=8 | | p=10 | | p=12 | |
|---|---|---|---|---|---|---|---|---|
| | Corr | Dis | Corr | Dis | Corr | Dis | Corr | Dis |
| Full_1 | 0.788 | 0.741 | 0.768 | 0.809 | 0.749 | 0.867 | 0.739 | 0.897 |
| CC_1 | 0.804 | 0.741 | 0.792 | 0.813 | 0.778 | 0.862 | 0.757 | 0.927 |
| EE1_1 | 0.787 | 0.745 | 0.765 | 0.814 | 0.750 | 0.862 | 0.727 | 0.929 |
| EE2_1 | 0.863 | 0.720 | 0.777 | 0.780 | 0.755 | 0.846 | 0.737 | 0.899 |
| Full_2 | 0.937 | 0.234 | 0.941 | 0.224 | 0.927 | 0.275 | 0.916 | 0.310 |
| CC_2 | 0.890 | 0.393 | 0.884 | 0.416 | 0.879 | 0.435 | 0.879 | 0.433 |
| EE1_2 | 0.924 | 0.277 | 0.918 | 0.304 | 0.913 | 0.321 | 0.907 | 0.346 |
| EE2_2 | 0.954 | 0.175 | 0.945 | 0.208 | 0.934 | 0.246 | 0.926 | 0.279 |

Table B.6: Correlation and Euclidean distance for OLS Case (ii)-Model (II)

| **B** | p=6 | | p=8 | | p=10 | | p=12 | |
|---|---|---|---|---|---|---|---|---|
| Full_1 | 0.946 | 0.206 | 0.926 | 0.280 | 0.908 | 0.345 | 0.8898 | 0.410 |
| CC_1 | 0.932 | 0.236 | 0.907 | 0.323 | 0.885 | 0.415 | 0.867 | 0.475 |
| EE1_1 | 0.963 | 0.144 | 0.946 | 0.207 | 0.934 | 0.252 | 0.871 | 0.475 |
| EE2_1 | 0.971 | 0.155 | 0.945 | 0.213 | 0.937 | 0.241 | 0.929 | 0.271 |
| Full_2 | 0.975 | 0.091 | 0.966 | 0.131 | 0.959 | 0.161 | 0.949 | 0.196 |
| CC_2 | 0.972 | 0.110 | 0.962 | 0.148 | 0.955 | 0.175 | 0.944 | 0.216 |
| EE1_2 | 0.980 | 0.078 | 0.970 | 0.116 | 0.963 | 0.146 | 0.956 | 0.170 |
| EE2_2 | 0.984 | 0.063 | 0.977 | 0.091 | 0.974 | 0.103 | 0.975 | 0.098 |

Table B.7: Correlation and Euclidean distance for PHD Case (i)-Model (II)

| **B** | p=6 | | p=8 | | p=10 | | p=12 | |
|---|---|---|---|---|---|---|---|---|
| Full_1 | 0.7739 | 0.7776 | 0.7527 | 0.8494 | 0.7449 | 0.8757 | 0.7286 | 0.9288 |
| CC_1 | 0.7975 | 0.7266 | 0.7826 | 0.8472 | 0.7554 | 0.91 | 0.7752 | 0.9007 |
| EE1_1 | 0.7938 | 0.7179 | 0.7396 | 0.8674 | 0.7343 | 0.8876 | 0.7387 | 0.8973 |
| EE2_1 | 0.7894 | 0.7234 | 0.7885 | 0.7103 | 0.7638 | 0.7898 | 0.7446 | 0.8761 |
| Full_2 | 0.9137 | 0.3068 | 0.9046 | 0.3388 | 0.9116 | 0.3159 | 0.9131 | 0.3167 |
| CC_2 | 0.8972 | 0.3665 | 0.8946 | 0.3654 | 0.9175 | 0.3037 | 0.9078 | 0.3395 |
| EE1_2 | 0.9382 | 0.2274 | 0.9203 | 0.295 | 0.9174 | 0.3114 | 0.921 | 0.2937 |
| EE2_2 | 0.9267 | 0.2586 | 0.9209 | 0.2861 | 0.9231 | 0.2879 | 0.9243 | 0.2832 |

Table B.8: Correlation and Euclidean distance for PHD Case (ii) -Model (II)

# APPENDIX C

# TECHNICAL DETAILS FOR CHAPTER 4

**Proof of Theorem 4.2**. Let $\mathbf{M} = \mathbf{\Sigma}^{-1}\mathbf{\Sigma}^{*}_{\mathrm{Yxx}}$, where $\mathbf{\Sigma}^{*}_{\mathrm{Yxx}} = \mathrm{E}\left(Y\mathbf{x}(\mathbf{x}^{*})^{T}\right)$. Similarly, denote $\widehat{\mathbf{M}} = \widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Sigma}}^{*}_{\mathrm{Yxx}}$, where $\widehat{\mathbf{\Sigma}} = n^{-1}\sum_{i=1}^{n}\widetilde{\mathbf{x}}_i\widetilde{\mathbf{x}}_i^{T}$ and $\widehat{\mathbf{\Sigma}}^{*}_{\mathrm{Yxx}} = n^{-1}\sum_{i=1}^{n}\widetilde{Y}_i\widetilde{\mathbf{x}}_i(\widetilde{\mathbf{x}}_i^{*})^{T}$. From Li et al. (2003), we have

$$\widehat{\mathbf{\Sigma}}^{-1} - \mathbf{\Sigma}^{-1} = -\mathbf{\Sigma}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i\mathbf{x}_i^{T} - \mathbf{\Sigma})\right)\mathbf{\Sigma}^{-1} + O_p(n^{-1}) \qquad \text{(C.1)}$$

Because $\mathrm{E}(Y) = 0$ and $\mathrm{E}(\mathbf{x}) = \mathbf{0}$, it can be shown that

$$\widehat{\mathbf{\Sigma}}^{*}_{\mathrm{Yxx}} - \mathbf{\Sigma}^{*}_{\mathrm{Yxx}} = \frac{1}{n}\sum_{i=1}^{n}\left(Y_i\mathbf{x}_i(\mathbf{x}_i^{*})^{T} - \mathbf{\Sigma}^{*}_{\mathrm{Yxx}}\right) + O_p(n^{-1}) \qquad \text{(C.2)}$$

By combining (C.1) and (C.2), we have

$$\widehat{\mathbf{M}} - \mathbf{M} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{A}_i + O_p(n^{-1}),$$

where

$\mathbf{A}_i = \mathbf{\Sigma}^{-1} Y_i \mathbf{x}_i (\mathbf{x}_i^*)^T - \mathbf{\Sigma}^{-1} \mathbf{x}_i \mathbf{x}_i^T \mathbf{\Sigma}^{-1} \mathbf{\Sigma}^*_{\text{Yxx}}$. It follows that $\sqrt{n}(\text{vec}(\widehat{\mathbf{M}}) - \text{vec}(\mathbf{M})) \to \mathcal{N}(\mathbf{0}, \mathbf{\Gamma})$, where $\mathbf{\Gamma} = \text{E}\left(\text{vec}(\mathbf{A})\text{vec}^T(\mathbf{A})\right)$.

**Proof of Theorem 4.3**. The proof is similar to Theorem 5 of Li and Wang (2007), and thus omitted.

**Proof of Theorem 4.4**. The proof is similar to Theorem 1 of Zhou and Dong (2016), and thus omitted.

**Lemma C.1.** *Under Conditions (C1)-(C5), we have*

$$\max_{1 \le i \le n} |\widehat{\pi}(\mathbf{x}_i) - \pi(\mathbf{x}_i)| = o_p(n^{-1/4})$$

Following the approach in Härdle (1990) and Härdle and Mammen (1993), one can show under certain conditions (will be listed) Lemma C.1 holds.

**Lemma C.2.** *Under Conditions (C1)-(C5), we have*

$$\sum_{i=1}^{n} \frac{\delta_i}{\pi_i(\mathbf{x})} \left[ \text{E}(l \mid \boldsymbol{\eta}_i)\{a(\mathbf{x}_i) - \text{E}(a \mid \boldsymbol{\eta}_i)\} + \text{E}(l \mid \boldsymbol{\eta}_i) \left\{ \text{E}(a \mid \boldsymbol{\eta}) - \widehat{\text{E}}(a \mid \boldsymbol{\eta}_i) \right\} \right] = o_p(n^{1/2})$$

The proof of Lemma C.2 is similar to the proof of Lemma A2 in Ma and Zhu (2012a).