# A simple and efficient estimation method for models with nonignorable missing data

Chunrong Ali
Oliver Linton
Zheng Zhang

# A Simple and Efficient Estimation Method for Models with Nonignorable Missing Data

Chunrong Ai

*Department of Economics, University of Florida*
*tsinghua@ufl.edu*


Oliver Linton

*Faculty of Economics, University of Cambridge*
*obl20@cam.ac.uk*


Zheng Zhang

*Institute of Statistics and Big Data, Renmin University of China*
*zhengzhang@ruc.edu.cn*

January 10, 2018

## Abstract

This paper proposes a simple and efficient estimation procedure for the model with non-ignorable missing data studied by Morikawa and Kim (2016). Their semiparametrically efficient estimator requires explicit nonparametric estimation and so suffers from the curse of dimensionality and requires a bandwidth selection. We propose an estimation method based on the Generalized Method of Moments (hereafter GMM). Our method is consistent and asymptotically normal regardless of the number of moments chosen. Furthermore, if the number of moments increases appropriately our estimator can achieve the semiparametric efficiency bound derived in Morikawa and Kim (2016), but under weaker regularity conditions. Moreover, our proposed estimator and its consistent covariance matrix are easily

1

computed with the widely available GMM package. We propose two data-based methods for selection of the number of moments. A small scale simulation study reveals that the proposed estimation indeed out-performs the existing alternatives in finite samples.

**Keywords**: Nonignorable nonresponse; Generalized method of moments; Semiparametric efficiency.

# 1   Introduction

Missing data is common in many fields of applications. One way to deal with the missing data problem is to delete observations containing missing data. In doing so we may produce biased estimates and erroneous conclusions, depending on the data missing mechanism. If data are missing completely at random, standard estimation and inference procedures are still consistent when the missing data observations are ignored, see Heitjan and Basu (1996), Little (1988) among others. If data are missing at random (MAR) in the sense that the propensity of missingness depends only on the observed covariates, consistent estimation can still be obtained through covariate balancing, see Rubin (1976a,b), Little and Rubin (1989), Robins and Rotnitzky (1995), Robins et al. (1995), Bang and Robins (2005), Qin and Zhang (2007), Chen et al. (2008), Tan (2010), Rotnitzky et al. (2012), Little and Rubin (2014) among others. In many applications, data are missing not at random (MNAR). For example, the income question in sample surveys is often not answered by people at the top end of the distribution, that is, their response frequency depends on an outcome variable that is often the key focus. An investigator is examining the effect of sleep on pain by calling subjects daily to ask them about last night's sleep and their pain today. Patients who are experiencing severe pain are more likely to not come to the phone leaving the data missing for that particular day; again this would violate the MAR assumption. From political science, roll-call votes, which measure legislatures ideological positions, are subject to non-ignorable nonresponse because, unsurprisingly, politicians behave strategically. In the MNAR case, the parameter of interest may not even be identified (e.g., Robins et al. (1997)), let alone be consistently estimated. To be more specific, let $T \in \{0,1\}$ denote the binary random variable indicating the missing status of the outcome variable $Y$: $Y$ is observed if $T$ takes the value one and $Y$ is not observed if $T$ takes the value zero. Let $\boldsymbol{X}$ denote a vector of explanatory variables, let $\pi(\boldsymbol{x}, y) = \mathbb{P}(T = 1 | \boldsymbol{X} = \boldsymbol{x}, Y = y)$ denote the propensity score function and let $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$ denote the conditional density function of $Y$ given $\boldsymbol{X}$. Robins et al. (1997) shows that if both the propensity score function and the conditional density function are completely unknown, the joint distribution of $(T, Y)$ given $\boldsymbol{X}$ is not point identifiable. In this case, a necessary identification

condition is the parameterization of either the propensity score function or the conditional density function. Molenberghs and Kenward (2007) proposes the parameterization of both the propensity score function and the conditional density function as an identification strategy, while Sverchkov (2008) and Riddles et al. (2016) parameterize the propensity score function and only a component of the conditional density function: $f_{Y|\boldsymbol{X},T}(y|\boldsymbol{x}, T = 1)$.

If the joint distribution is not the parameter of interest, the identification strategy above can be modified. For example, if the parameter of interest is the conditional density of $Y$ given $\boldsymbol{X}$ (i.e., $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$), parameterization of the propensity score function is not needed. However, parameterization of $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$ in this case is not sufficient for identification due to missing data. Tang et al. (2003) suggests parameterization of the marginal density $f_{\boldsymbol{X}}(\boldsymbol{x})$ as well, while Zhao and Shao (2015) imposes an exclusion retriction. In both studies, $f_{Y|\boldsymbol{X}}(y|\boldsymbol{x})$ is identified and consistently estimated.

We consider estimation of the parameter $\theta_0 = \mathbb{E}[U(\boldsymbol{X}, Y)]$, where $U(\cdot)$ is any known function. We suppose that the propensity score $\pi$ is parameterized but do not restrict the conditional density function of the outcome variable. In earlier work in this framework, either the coefficients in the propensity score function are known or consistently estimated from an external sample (Kim and Yu (2011)) or an exclusion restriction is imposed (Wang et al. (2014) and Shao and Wang (2016)). Morikawa and Kim (2016) study the efficient estimation of $\theta_0$. They derive the efficient score function (and hence the semiparametric efficiency bound) for $\theta_0$ in this model. They propose to estimate the efficient score function by estimating $f_{Y|\boldsymbol{X},T}(y|, \boldsymbol{x}, 1)$ by a working parametric model (MK1) or by kernel nonparametric estimation (MK2). Their approach MK1 is not efficient unless the working parametric model is correct, although it is consistent. Their method MK2 suffers from the curse of dimensionality (their smoothness conditions depend on the dimensionality of the covariates through their conditions C14) and the bandwidth selection problem (about which they give no guidance).

We study the same estimation problem as in Morikawa and Kim (2016) but propose a simpler yet equally efficient estimation procedure. Our proposed method does not require explicit nonparametric estimation and hence does not suffer from the curse of dimensionality. The proposed estimator is motivated by the key insight that the model parameter satisfies a parametric conditional moment restriction, of which the semiparametric efficiency bound is identical to the bound derived in Morikawa and Kim (2016). The conditional moment restriction is then turned into an expanding set of unconditional moment restrictions and the parameter of

3

interest is estimated by applying the widely available and easy to compute GMM estimation (see Hansen (1982)). Under some sufficient conditions, we establish that the proposed estimator is consistent and asymptotically normally distributed even if the set of unconditional moment restrictions does not expand, thereby freeing us from the curse of dimensionality and the bandwidth selection problem; when the set does expand, the proposed estimator attains the semiparametric efficiency bound. This is in contrast with the MK2 method of Morikawa and Kim (2016), which is inconsistent if the bandwidth does not go to zero at a certain rate.

The paper is organized as follows. Section 2 describes the estimation, Section 3 derives the large sample properties of the estimator, Section 4 provides a consistent asymptotic variance estimator, Section 5 suggests two data driven approaches to determine the number of unconditional moment restrictions, Section 6 reports on a small scale simulation study, followed by some concluding remarks in Section 7. All technical proofs are relegated to the Appendix.

## 2 Basic Framework and Estimation

We begin by setting up the basic framework. Denote $\boldsymbol{Z} = (\boldsymbol{X}^\top, Y)^\top$. The following assumption shall be maintained throughout the paper:

**Assumption 2.1**. (i) Parameterization of data missing mechanism: $\mathbb{P}(T = 1 | Y, \boldsymbol{X}) = \pi(Y, \boldsymbol{X}; \gamma_0) = \pi(\boldsymbol{Z}; \gamma_0)$ holds for some known function $\pi(.;.)$, where $\gamma_0 \in \mathbb{R}^p$ for some known $p \in \mathbb{N}$ is the true (unknown) value; (ii) exclusion restriction: there exists some nonresponse instrument variables $\boldsymbol{X}_1$ in $\boldsymbol{X} = (\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top)^\top$ so that $\boldsymbol{X}_2$ is independent of $T$ given both $\boldsymbol{X}_1$ and $Y$; and (iii) the parameter of interest is $\theta_0 = \mathbb{E}[U(\boldsymbol{Z})]$ for some known function $U(\cdot)$.

Under Assumption 2.1 and by applying the law of iterated expectations, we obtain the following conditional moment restrictions:

$$\mathbb{E}\left[1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} \bigg| \boldsymbol{X}\right] = 0, \tag{1}$$

$$\mathbb{E}\left[\theta_0 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} U(\boldsymbol{Z})\right] = 0, \tag{2}$$

which will form the basis for the proposed estimation. We notice that the parameters of interest in (1)-(2) are finite dimensional (and there is no explicit infinite dimensional nuisance parameter) and can be easily estimated with GMM estimation. We also notice that it is a special case of the model studied in Ai and Chen

4

(2012). By applying their result (Remark 2.1, pp. 446), we obtain the semiparametric efficiency bound for model (1)-(2), which is identical to the bound derived in Morikawa and Kim (2016), thereby suggesting a simple and efficient estimation.

The (nuisance) parameter $\gamma_0$ is identified by (1) and the parameter of interest $\theta_0$ is identified by (2). The following condition shall also be maintained throughout the paper:

**Assumption 2.2.** The parameter space $\Gamma$ is a compact subset of $\mathbb{R}^p$. The true value $\gamma_0$ lies in the interior of $\Gamma$ and is the only solution to (1). The parameter space $\Theta$ is a compact subset of $\mathbb{R}$ and the true value $\theta_0$ lies in the interior of $\Theta$.

To estimate model (1)-(2), we first turn it into a set of unconditional moment restrictions. We work with a set of known basis functions: for each integer $K \in \mathbb{N}$ with $K \geq p$, let

$$u_K(\boldsymbol{X}) = (u_{1K}(\boldsymbol{X}), \ldots, u_{KK}(\boldsymbol{X}))^\top.$$

Discussion on the choice of $u_K(\boldsymbol{X})$ and its properties can be found in Section 8.2 in Appendix. Model (1)-(2) implies the unconditional moment restrictions:

$$\mathbb{E}\left[\left(1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)}\right) u_K(\boldsymbol{X})\right] = 0, \tag{3}$$

$$\mathbb{E}\left[\theta_0 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} U(\boldsymbol{Z})\right] = 0. \tag{4}$$

To avoid redundant moment restrictions, we require $\mathbb{E}\left[u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top\right]$ to be nonsingular for every $K$. The following somewhat stronger identification condition shall be maintained throughout the paper:

**Assumption 2.2'.** The parameter space $\Gamma$ is a compact subset of $\mathbb{R}^p$. The true value $\gamma_0$ lies in the interior of $\Gamma$ and is the only solution to (3). The parameter space $\Theta$ is a compact subset of $\mathbb{R}$ and the true value $\theta_0$ lies in the interior of $\Theta$.

We can estimate the parameter of interest by the GMM method. Let $\{T_i, \boldsymbol{Z}_i\}_{i=1}^N$ denote an $i.i.d.$ sample drawn from the joint distribution of $(T, \boldsymbol{Z})$. Denote

$$
\begin{aligned}
\boldsymbol{G}_K(\gamma, \theta) : &= \left(\sum_{i=1}^N \left[1 - \frac{T_i}{\pi(\boldsymbol{Z}_i; \gamma)}\right] u_K(\boldsymbol{X}_i)^\top, \sum_{i=1}^N \left[\theta - \frac{T_i}{\pi(\boldsymbol{Z}_i; \gamma)} U(\boldsymbol{Z}_i)\right]\right)^\top \\
&= \sum_{i=1}^N g_K(T_i, \boldsymbol{Z}_i; \gamma, \theta),
\end{aligned}
$$

5

where $g_K(T, \boldsymbol{Z}; \gamma, \theta) := \left(\left[1 - \frac{T}{\pi(\boldsymbol{Z};\gamma)}\right] u_K(\boldsymbol{X})^\top, \theta - \frac{T}{\pi(\boldsymbol{Z};\gamma)} U(\boldsymbol{Z})\right)^\top$. The GMM estimator of $\gamma_0$ and $\theta_0$ is defined as

$$(\check{\gamma}, \check{\theta}) = \arg \min_{\gamma \in \Gamma, \theta \in \Theta} \boldsymbol{G}_K(\gamma, \theta)^T \cdot \mathbf{W} \cdot \boldsymbol{G}_K(\gamma, \theta)$$

where $\mathbf{W}$ is a $(K+1) \times (K+1)$ symmetric weighting matrix. For every fixed $K \geq p$, Hansen (1982) shows that, under some regularity conditions, the estimator

$$(\check{\gamma} - \gamma_0, \check{\theta} - \theta_0) = O_p(N^{-1/2}) \tag{5}$$

is asymptotically normally distributed, but generally not the best unless the best weighting matrix is used. The best weighting matrix is the inverse of

$$\boldsymbol{D}_{(K+1) \times (K+1)} := \mathbb{E}\left[g_K(T, \boldsymbol{Z}; \gamma_0, \theta_0) g_K(T, \boldsymbol{Z}; \gamma_0, \theta_0)^\top\right].$$

The best estimator (within the class defined by the specific unconditional moments) is defined as

$$(\overline{\gamma}, \overline{\theta}) = \arg \min_{\gamma \in \Gamma, \theta \in \Theta} \boldsymbol{G}_K(\gamma, \theta)^T \cdot \boldsymbol{D}^{-1}_{(K+1) \times (K+1)} \cdot \boldsymbol{G}_K(\gamma, \theta).$$

Suppose that the propensity score function is differentiable with respect to $\gamma$. Denote

$$\boldsymbol{B}_{(K+1) \times (p+1)} = \nabla_{\gamma,\theta} \mathbb{E}\left[\frac{1}{N} \boldsymbol{G}_K(\gamma_0, \theta_0)\right] = \begin{pmatrix} \mathbb{E}\left[u_K(\boldsymbol{X}) \frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)^\top}{\pi(\boldsymbol{Z};\gamma_0)}\right], & \boldsymbol{0}_{K \times 1} \\ \mathbb{E}\left[U(\boldsymbol{Z}) \frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)^\top}{\pi(\boldsymbol{Z};\gamma_0)}\right], & 1 \end{pmatrix}$$

and

$$\boldsymbol{V}_K = \left\{\left(\boldsymbol{B}_{(K+1) \times (p+1)}\right)^\top \boldsymbol{D}^{-1}_{(K+1) \times (K+1)} \left(\boldsymbol{B}_{(K+1) \times (p+1)}\right)\right\}^{-1}.$$

Hansen (1982) shows that, for every fixed $K \geq p$,

$$\boldsymbol{V}_K^{-1/2} \begin{pmatrix} \sqrt{N}(\overline{\gamma} - \gamma_0) \\ \sqrt{N}(\overline{\theta} - \theta_0) \end{pmatrix} \to N\left(0, I_{(p+1) \times (p+1)}\right) \text{ in distribution.} \tag{6}$$

Since the best weighting matrix depends on the unknown parameter value, the best estimator $(\overline{\gamma}, \overline{\theta})$ is infeasible. Hansen (1982) suggests the following two-step procedure:

Step I. Compute the initial $\sqrt{N}$-consistent estimator

$$\widehat{\boldsymbol{W}}_0 := \begin{pmatrix} \frac{1}{N}\sum_{i=1}^N u_K(\mathbf{X}_i)u_K(\mathbf{X}_i)^\top & \mathbf{0}_{K\times 1} \\ \mathbf{0}_{K\times 1}^\top & 1 \end{pmatrix},$$

$$(\check{\gamma},\check{\theta}) = \arg\min_{(\gamma,\theta)\in\Gamma\times\Theta} \boldsymbol{G}_K(\gamma,\theta)^T \cdot \widehat{\boldsymbol{W}}_0^{-1} \cdot \boldsymbol{G}_K(\gamma,\theta).$$

Step II. Compute the best weighting matrix and the best estimator

$$\hat{\boldsymbol{D}}_{(K+1)\times(K+1)} := \frac{1}{N}\sum_{i=1}^N g_K(T_i,\boldsymbol{Z}_i;\check{\gamma},\check{\theta})g_K(T_i,\boldsymbol{Z}_i;\check{\gamma},\check{\theta})^\top,$$

$$(\widehat{\gamma},\widehat{\theta}) = \arg\min_{\gamma\in\Gamma,\theta\in\Theta} \boldsymbol{G}_K(\gamma,\theta)^T \cdot \widehat{\boldsymbol{D}}_{(K+1)\times(K+1)}^{-1} \cdot \boldsymbol{G}_K(\gamma,\theta).$$

Hansen (1982) establishes that, for every fixed $K \geq p$,

$$\boldsymbol{V}_K^{-1/2}\begin{pmatrix} \sqrt{N}(\widehat{\gamma}-\gamma_0) \\ \sqrt{N}(\widehat{\theta}-\theta_0) \end{pmatrix} \to N\left(0, I_{(p+1)\times(p+1)}\right) \text{ in distribution.} \qquad (7)$$

Moreover, denote

$$\widehat{\boldsymbol{B}}_{(K+1)\times(p+1)} := \begin{pmatrix} N^{-1}\sum_{i=1}^N u_K(\boldsymbol{X}_i)\frac{\nabla_\gamma\pi(\boldsymbol{Z}_i;\widehat{\gamma})^\top}{\pi(\boldsymbol{Z}_i;\widehat{\gamma})}, & \mathbf{0}_{K\times 1} \\ N^{-1}\sum_{i=1}^N U(\boldsymbol{Z}_i)\frac{\nabla_\gamma\pi(\boldsymbol{Z}_i;\widehat{\gamma})^\top}{\pi(\boldsymbol{Z}_i;\widehat{\gamma})}, & 1 \end{pmatrix}$$

and

$$\widehat{\boldsymbol{V}}_K := \left\{ \left(\widehat{\boldsymbol{B}}_{(K+1)\times(p+1)}\right)^\top \widehat{\boldsymbol{D}}_{(K+1)\times(K+1)}^{-1} \left(\widehat{\boldsymbol{B}}_{(K+1)\times(p+1)}\right) \right\}^{-1}.$$

Hansen (1982) proves that, for every fixed $K \geq p$,

$$\widehat{\boldsymbol{V}}_K \to \boldsymbol{V}_K \text{ in probability.} \qquad (8)$$

The best estimator (within the class defined by the specific unconditional moments) is generally not semiparametrically efficient. To obtain the efficient estimator, we shall allow $K$ to increase with the sample size at the rate $o(N^{1/3})$ so that $\{u_K(\boldsymbol{X})\}$ span the space of measureable functions (see also Geman and Hwang (1982) and Newey (1997)). In the next two sections, we shall establish that results in (5)-(8) still hold with expanding $K = o(N^{1/3})$.

The advantage of our proposed estimator over the existing estimators is evident. Our estimation problem is a parametric one, requiring no modeling of or nonparametric estimation of $f_{Y|\boldsymbol{X},T}(y|x,1)$. In contrast, the estimators proposed in Riddles et al. (2016) and Morikawa and Kim (2016) could be inconsistent if $f_{Y|\boldsymbol{X},T}(y|x,1)$ is incorrectly specified or suffers from the curve of dimensionality and bandwith selection problem of the nonparametric estimation of $f_{Y|\boldsymbol{X},T}(y|x,1)$.

# 3   Asymptotic Theory

In this section, we show that results in (5)- (7) still hold with expanding $K$, all technical proof can be found in the supplemental material Ai et al. (2018). First, we establish the convergence rate of the first step estimator $(\check{\gamma}, \check{\theta})$.

**Theorem 1.** *Under Assumptions 2.1-2.2 and Assumptions 1, 2, 4, 5, 7, and 8 listed in Appendix, with $K = o(N^{1/3})$, the first step estimator satisfies*

$$(\check{\gamma} - \gamma_0, \check{\theta} - \theta_0) = O_p\left(N^{-1/2}\right).$$

Next, we establish the large sample properties of the infeasible best estimator $(\overline{\gamma}, \overline{\theta})$ without imposing the smoothness Assumptions 3 and 6 listed in Appendix.

**Theorem 2.** *Under Assumptions 2.1-2.2 and Assumptions 1, 2, 4, 5, 7, and 8 listed in Appendix, with $K = o(N^{1/3})$, the infeasible best estimator satisfies*

$$\boldsymbol{V}_K^{-1/2}\begin{pmatrix}\sqrt{N}(\overline{\gamma} - \gamma_0) \\ \sqrt{N}(\overline{\theta} - \theta_0)\end{pmatrix} \to N\left(0, I_{(p+1)\times(p+1)}\right) \text{ in distribution.}$$

If in addition the smoothness Assumptions 3 and 6 are satisfied, the next result shows that $\boldsymbol{V}_K \to \boldsymbol{V}_{eff}$ in probability, where $\boldsymbol{V}_{eff}$ is the semiparametric efficiency bound of $(\gamma_0, \theta_0)$ derived in Morikawa and Kim (2016), see Lemma 1 in Section 8.3 of Appendix.

**Theorem 3.** *Under Assumption 2.1-2.2 and Assumption 1-8 listed in Appendix, with $K = o(N^{1/3})$, we obtain*

$$\boldsymbol{V}_K \to \boldsymbol{V}_{eff} \text{ in probability.}$$

By Theorem 1-3, the infeasible best estimator attains the semiparametric efficiency bound. The next result establishes the equivalence between the best estimator $(\widehat{\gamma}, \widehat{\theta})$ and the infeasible best estimator $(\overline{\gamma}, \overline{\theta})$, implying that the best estimator also attains the semiparametric efficiency bound.

**Theorem 4.** *Under Assumption 2.1-2.2 and Assumption 1-8 listed in Appendix, with $K = o(N^{1/3})$, we obtain*

$$\begin{pmatrix}\sqrt{N}(\overline{\gamma} - \widehat{\gamma}) \\ \sqrt{N}(\overline{\theta} - \widehat{\theta})\end{pmatrix} = o_p(1).$$

# 4 Variance Estimation

In order to conduct statistical inference, we need a consistent covariance estimator. Notice that (5) implies that $\widehat{\boldsymbol{V}}_K$ is a consistent estimator of $\boldsymbol{V}_K$ for every fixed $K \geq p$. We now show that this result still holds true with expanding $K$, thereby providing a consistent covariance estimator.

**Theorem 5.** *Under Assumption 2.1-2.2 and Assumption 1-8 listed in Appendix, with $K = o(N^{1/3})$, we obtain*

$$\widehat{\boldsymbol{V}}_K \to \boldsymbol{V}_K \text{ in probability.}$$

We notice that our covariance estimator is much simpler and more natural than the one suggested in Morikawa and Kim (2016), which requires nonprametric estimation of $f_{Y|\boldsymbol{X},T}(y|x,1)$ and tends to have poor performance in finite samples. Our covariance estimator is the GMM covariance estimator and is easily computed by existing statistical packages.

# 5 Selection of $K$

The large sample properties of the proposed estimator established in the previous sections allow for a wide range of values for $K$, and theoretically the sensitivity of the estimator to the choice of $K$ is not so pronounced, it affects higher order terms in a way that does not affect consistency and asymptotic normality. Nevertheless, there may be some higher order effect of the choice of $K$ on perfomance. In this section, we present two data-driven approaches to select $K$. Both approaches strike a balance between bias and variance.

**Covariate balancing approach**. The first approach attempts to balance the distribution of the covariates between the whole population and the non-missing population through weighting. Notice that

$$\mathbb{E}\left[\frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} I(X_j \leq x_j)\right] = \mathbb{E}[I(X_j \leq x_j)] , \ j \in \{1, ..., r\} ,$$

where $X_j$ is the $j^{th}$ component of $\boldsymbol{X}$ and $I(X_j \leq x_j)$ is the indicator function. Obviously the propensity score function $\pi(\boldsymbol{Z}; \gamma_0)$ plays the role of balancing. Notice that the estimator $\hat{\gamma}$ depends on $K$. For a given $K$, we compute

$$\hat{F}^j_{N,K}(x_j) := \frac{1}{N} \sum_{i=1}^{N} \frac{T_i}{\pi(\boldsymbol{X}_i; \hat{\gamma})} I(X_{ij} \leq x_j), \ j \in \{1, \ldots, r\}.$$

We compute the empirical distributions of the covariates

$$\tilde{F}_N^j(x_j) := \frac{1}{N} \sum_{i=1}^N I(X_{ij} \leq x_j), \ j \in \{1, \ldots, r\}.$$

We choose the lowest $K$ so that the difference between $\{\hat{F}_{N,K}^j\}_{j=1}^r$ and $\{\hat{F}_N^j\}_{j=1}^r$ is small. Denote the upper bound of $K$ by $\bar{K}$ (e.g. $\bar{K} = 7$ in our simulation studies). We choose $K \in \{1, ..., \bar{K}\}$ to minimize the aggregate Kolmogorov-Smirnov distance between $\{\hat{F}_{N,K}^j\}_{j=1}^r$ and $\{\hat{F}_N^j\}_{j=1}^r$:

$$\hat{K} = \arg \min_{K \in \{1, \ldots, \bar{K}\}} D_N(K) = \sum_{j=1}^r \sup_{x_j \in \mathbb{R}} \left| \tilde{F}_N^j(x_j) - \hat{F}_{N,K}^j(x_j) \right|.$$

**Higher order MSE approach**. The second approach chooses $K$ to minimize the mean-squared error of the estimator. Donald et al. (2009) derives the higher-order asymptotic mean-square error (MSE) of a linear combination $\mathbf{t}^\top \hat{\gamma}$ for some fixed $\mathbf{t} \in \mathbb{R}^p$.

Let $\check{\gamma}$ be some preliminary estimator. Define:

$$\widehat{\Pi}(K; \boldsymbol{t}) = \sum_{i=1}^N \hat{\xi}_{ii} \rho(T_i, \boldsymbol{X}_i, Y_i; \check{\gamma}) \cdot (\mathbf{t}^\top \hat{\Omega}_{p \times p}^{-1} \tilde{\eta}_i),$$

$$\hat{\Phi}(K; \boldsymbol{t}) = \sum_{i=1}^N \hat{\xi}_{ii} \left\{ \mathbf{t}^\top \hat{\Omega}_{p \times p}^{-1} \left[ \widehat{\boldsymbol{D}}_i^* \rho(T_i, \boldsymbol{X}_i, Y_i; \check{\gamma})^2 - \nabla_\gamma \rho(T_i, \boldsymbol{X}_i, Y_i; \check{\gamma}) \right] \right\}^2$$
$$- \mathbf{t}^\top \hat{\Omega}_{p \times p}^{-1} (\hat{\Gamma}_{K \times p})^\top \hat{\Upsilon}_{K \times K}^{-1} \hat{\Gamma}_{K \times p} \hat{\Omega}_{p \times p}^{-1} \boldsymbol{t}.$$

where $\rho(T_i, \boldsymbol{X}_i, Y_i; \check{\gamma})$, $\hat{\Omega}_{p \times p}$, $\tilde{\eta}_i$, $\hat{\xi}_{ii}$, $\widehat{\boldsymbol{D}}_i^*$, $\hat{\Gamma}_{K \times p}$, and $\hat{\Upsilon}_{K \times K}$ are defined in Section 8.2 of Appendix. Notice that $\widehat{\Pi}(K; \boldsymbol{t})^2/N$ is an estimate of the squared bias term derived in Newey and Smith (2004) and $\hat{\Phi}(K; \boldsymbol{t})$ is the asymptotic variance.

The second approach chooses $K$ to minimize the following higher-order MSEs of $\hat{\gamma}_j$, $j = 1, \ldots, p$:

$$S_{GMM}(K) = \sum_{j=1}^p \left\{ \frac{1}{N} \widehat{\Pi}(K; e_j)^2 + \hat{\Phi}(K; e_j) \right\}, \tag{9}$$

where $e_j$ is the $j^{th}$ column of the $p$-dimensional identity matrix. In practice, we set the upper bound $\bar{K}$ and then choose $K \in \{1, 2, \ldots, \bar{K}\}$ to minimize the criteria (9).

10

Table 1: Simulation results under Scenorio I

| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | | $n = 200$ | | | | | |
| Bias | 0.028 | -0.125 | 0.039 | 0.055 | 0.120 | 0.106 | -0.997 | 0.167 | 0.301 |
| Stdev | 0.254 | 0.413 | 0.129 | 0.229 | 0.272 | 0.118 | 0.197 | 0.266 | 0.101 |
| MSE | 0.065 | 0.186 | 0.018 | 0.055 | 0.088 | 0.025 | 1.033 | 0.099 | 0.101 |
| CP | — | — | 0.908 | — | — | 0.908 | — | — | 0.22 |
| | | | | $n = 500$ | | | | | |
| Bias | 0.011 | -0.067 | 0.016 | 0.048 | 0.058 | 0.063 | -0.966 | 0.220 | 0.299 |
| Stdev | 0.161 | 0.282 | 0.090 | 0.151 | 0.193 | 0.077 | 0.126 | 0.160 | 0.063 |
| MSE | 0.026 | 0.084 | 0.008 | 0.025 | 0.040 | 0.010 | 0.949 | 0.074 | 0.093 |
| CP | — | — | 0.928 | — | — | 0.892 | — | — | 0.034 |
| | | | | $n = 1000$ | | | | | |
| Bias | 0.005 | -0.040 | 0.008 | 0.034 | 0.023 | 0.040 | -0.962 | 0.235 | 0.298 |
| Stdev | 0.103 | 0.187 | 0.065 | 0.102 | 0.132 | 0.055 | 0.078 | 0.099 | 0.045 |
| MSE | 0.010 | 0.036 | 0.004 | 0.011 | 0.018 | 0.004 | 0.932 | 0.065 | 0.091 |
| CP | — | — | 0.934 | — | — | 0.906 | — | — | 0.012 |

Stdev: standard deviation ; MSE: mean squared error; CP: coverage probability. The bandwith used in computing the nonparametric kernel estimators $(\hat{\alpha}_{MK}, \hat{\beta}_{MK}, \hat{\theta}_{MK})$ is $h = 0.15$.

Table 2: Simulation results under Scenorio II

|  | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
|---|---|---|---|---|---|---|---|---|---|
| | | | | | $n = 200$ | | | | |
| Bias | -0.208 | 0.096 | 0.084 | -0.552 | 0.588 | 0.173 | -2.053 | 1.215 | 0.530 |
| Stdev | 0.646 | 0.555 | 0.201 | 0.372 | 0.245 | 0.125 | 0.809 | 0.148 | 0.205 |
| MSE | 0.462 | 0.318 | 0.047 | 0.443 | 0.406 | 0.045 | 4.873 | 1.498 | 0.323 |
| CP | — | — | 0.95 | — | — | 0.784 | — | — | 0.138 |
| | | | | | $n = 500$ | | | | |
| Bias | -0.081 | 0.040 | 0.044 | -0.313 | 0.392 | 0.122 | -1.924 | 1.203 | 0.583 |
| Stdev | 0.406 | 0.363 | 0.131 | 0.261 | 0.186 | 0.085 | 0.175 | 0.064 | 0.132 |
| MSE | 0.171 | 0.134 | 0.019 | 0.166 | 0.188 | 0.022 | 3.732 | 1.451 | 0.357 |
| CP | — | — | 0.932 | — | — | 0.764 | — | — | 0.06 |
| | | | | | $n = 1000$ | | | | |
| Bias | -0.036 | 0.019 | 0.019 | -0.198 | 0.268 | 0.085 | -1.900 | 1.201 | 0.590 |
| Stdev | 0.260 | 0.225 | 0.086 | 0.203 | 0.164 | 0.061 | 0.086 | 0.044 | 0.078 |
| MSE | 0.069 | 0.051 | 0.007 | 0.080 | 0.098 | 0.011 | 3.618 | 1.445 | 0.354 |
| CP | — | — | 0.932 | — | — | 0.768 | — | — | 0.018 |

Stdev: standard deviation ; MSE: mean squared error; CP: coverage probability. The bandwith used in computing the nonparametric kernel estimators $(\hat{\alpha}_{MK}, \hat{\beta}_{MK}, \hat{\theta}_{MK})$ is $h = 0.05$.

Table 3: Simulation results under Scenorio III

| | $n = 200$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | 0.155 | -0.171 | 0.003 | 0.047 | 0.015 | 0.071 | -2.794 | 0.954 | -1.146 |
| Stdev | 0.584 | 0.585 | 0.155 | 0.376 | 0.190 | 0.131 | 1.395 | 0.396 | 0.263 |
| MSE | 0.365 | 0.372 | 0.024 | 0.144 | 0.036 | 0.022 | 9.758 | 1.069 | 1.384 |
| CP | — | — | 0.934 | — | — | 0.884 | — | — | 0.032 |

| | $n = 500$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | 0.034 | -0.036 | 0.000 | 0.012 | 0.012 | 0.034 | 0.782 | 0.355 | 0.123 |
| Stdev | 0.305 | 0.224 | 0.103 | 0.250 | 0.128 | 0.085 | 0.433 | 0.113 | 0.101 |
| MSE | 0.094 | 0.051 | 0.010 | 0.062 | 0.016 | 0.008 | 0.799 | 0.139 | 0.025 |
| CP | — | — | 0.902 | — | — | 0.894 | — | — | 0.698 |

| | $n = 1000$ | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | 0.009 | -0.010 | 0.002 | 0.002 | 0.009 | 0.017 | 0.728 | 0.372 | 0.126 |
| Stdev | 0.215 | 0.157 | 0.069 | 0.167 | 0.083 | 0.056 | 0.302 | 0.078 | 0.067 |
| MSE | 0.046 | 0.024 | 0.004 | 0.028 | 0.007 | 0.003 | 0.621 | 0.144 | 0.020 |
| CP | — | — | 0.932 | — | — | 0.934 | — | — | 0.454 |

Stdev: standard deviation ; MSE: mean squared error; CP: coverage probability. The bandwith used in computing the nonparametric kernel estimators $(\hat{\alpha}_{MK}, \hat{\beta}_{MK}, \hat{\theta}_{MK})$ is $h = 0.1$.

Table 4: Simulation results under Scenorio IV

| | | | | | $n = 200$ | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | 0.097 | -0.114 | 0.005 | -0.018 | 0.027 | 0.043 | -1.002 | 1.003 | 0.136 |
| Stdev | 1.140 | 0.721 | 0.118 | 0.308 | 0.185 | 0.103 | 0.081 | 0.139 | 0.348 |
| MSE | 1.310 | 0.533 | 0.014 | 0.095 | 0.035 | 0.013 | 1.011 | 1.026 | 0.139 |
| CP | — | — | 0.914 | — | — | 0.92 | — | — | 0.998 |
| | | | | | $n = 500$ | | | | |
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | -0.001 | -0.026 | 0.003 | -0.042 | 0.041 | 0.022 | -1.003 | 1.000 | 0.146 |
| Stdev | 0.203 | 0.139 | 0.071 | 0.172 | 0.100 | 0.067 | 0.048 | 0.088 | 0.199 |
| MSE | 0.041 | 0.020 | 0.005 | 0.031 | 0.011 | 0.005 | 1.010 | 1.009 | 0.061 |
| CP | — | — | 0.944 | — | — | 0.946 | — | — | 1.000 |
| | | | | | $n = 1000$ | | | | |
| | $\hat{\alpha}$ | $\hat{\beta}$ | $\hat{\theta}$ | $\hat{\alpha}_{MK}$ | $\hat{\beta}_{MK}$ | $\hat{\theta}_{MK}$ | $\tilde{\alpha}_{MAR}$ | $\tilde{\beta}_{MAR}$ | $\tilde{\theta}_{MAR}$ |
| Bias | 0.010 | -0.034 | -0.001 | -0.027 | 0.024 | 0.011 | -1.000 | 0.997 | 0.134 |
| Stdev | 0.262 | 0.264 | 0.052 | 0.122 | 0.070 | 0.048 | 0.035 | 0.065 | 0.148 |
| MSE | 0.068 | 0.070 | 0.002 | 0.015 | 0.005 | 0.002 | 1.003 | 1.000 | 0.039 |
| CP | — | — | 0.936 | — | — | 0.932 | — | — | 1.000 |

Stdev: standard deviation ; MSE: mean squared error; CP: coverage probability. The bandwith used in computing the nonparametric kernel estimators $(\hat{\alpha}_{MK}, \hat{\beta}_{MK}, \hat{\theta}_{MK})$ is $h = 0.2$.

14

# 6  Simulations

After establishing the large sample properties of the proposed estimator, we now evaluate its finite sample performance through a small scale simulation study. We consider four scenarios. In all scenarios, the parameter of interest is $\theta_0 = \mathbb{E}[Y]$ and the sample size is set respectively at $N = 200, 500$ and $1000$.

- **Scenario I**: $X$ is generated from the normal distribution $N(0,1)$, and the outcome $Y$ is generated from the normal distribution with mean $X + 1$ and unit variance, i.e. $Y \sim N(X + 1, 1)$. The relationship between the outcome variable and the covariate is linear, and the distribution of outcome is normal. The missing mechanism is modeled by

$$\mathbb{P}(T = 1 | Y, X) = [1 + \exp(\alpha_0 + \beta_0 Y)]^{-1} \ ,$$

  with the true value $(\alpha_0, \beta_0) = (0, -1.2)$. The true value of the parameter of interest is $\theta_0 = \mathbb{E}[Y] = 1$.

- **Scenario II**: $X$ is generated from the normal distribution $N(0,1)$, and the outcome $Y$ is generated from the normal distribution with mean $X^2 + 1$ and unit variance, i.e. $Y \sim N(X^2 + 1, 1)$. Thus the relationship between the outcome variable and the covariate is nonlinear, and the distribution of outcome is non-normal. The missing mechanism is modeled as

$$\mathbb{P}(T = 1 | Y, X) = [1 + \exp(\alpha_0 + \beta_0 Y)]^{-1}$$

  with the true value $(\alpha_0, \beta_0) = (1.25, -1.2)$. The true value of the parameter of interest is $\theta_0 = \mathbb{E}[Y] = 2$.

- **Scenario III**. The design follows Qin et al. (2002). We generate the outcome from

$$Y = 0.1X^2 + ZX^{1/2}/5 \ ,$$

  where $Z$ and $X$ are independent, $Z$ is standard normal random variable, and $X$ follows the $\chi^2_{(6)}/2$ distribution. The missing mechanism is modeled as

$$\mathbb{P}(T = 1 | Y, X) = [1 + \exp(\alpha_0 + \beta_0 Y)]^{-1}$$

  with the true value $(\alpha_0, \beta_0) = (3, -1)$. The true value of the target parameter is $\theta_0 = \mathbb{E}[Y] = 1.2$.

- **Scenario IV**. The design is similar to that in Kang and Schafer (2007). $\boldsymbol{Z} = (Z_1, Z_2)$ is generated from the standard bivariate normal distribution, and $Y$ is generated from the normal distribution with mean $2 + Z_1$ and unit variance. The missing mechanism is modeled as

$$\mathbb{P}(T = 1 | Y, X_1, X_2) = [1 + \exp(\alpha_0 Z_1 + \beta_0 Y)]^{-1}$$

with $(\alpha_0, \beta_0) = (1, -1)$. The true value of the parameter of interest is $\theta_0 = \mathbb{E}[Y] = 2$. Instead of directly observing covariates $\boldsymbol{Z}$, we observe a non-linear transformation of $\boldsymbol{Z}$: $X_1 = \exp(Z_1/2)$ and $X_2 = Z_2/(1 + \exp(Z_1))$.

In all scenarios, we generate $J = 500$ random samples, and for each sample, we compute the following three estimators:

1. Naive estimator. We compute the missing at random estimator $(\tilde{\alpha}_{MAR}, \tilde{\beta}_{MAR}, \tilde{\theta}_{MAR})$ as

$$\tilde{\theta}_{MAR} = \frac{1}{N} \sum_{i=1}^{N} \frac{T_i}{\pi(\boldsymbol{X}_i; \tilde{\alpha}_{MAR}, \tilde{\beta}_{MAR})} Y_i \ ,$$

where $\pi(\boldsymbol{X}_i; \tilde{\alpha}_{MAR}, \tilde{\beta}_{MAR})$ is an estimated response model. In Scenarios I, II & III, $\pi(\boldsymbol{X}_i; \tilde{\alpha}_{MAR}, \tilde{\beta}_{MAR}) = \left[1 + \exp(\tilde{\alpha}_{MAR} + \tilde{\beta}_{MAR} X_i)\right]^{-1}$ and in Scenario IV $\pi(\boldsymbol{X}_i; \tilde{\alpha}_{MAR}, \tilde{\beta}_{MAR}) = \left[1 + \exp(\tilde{\alpha}_{MAR} Z_{1i} + \tilde{\beta}_{MAR} X_{2i})\right]^{-1}$, where $(\tilde{\alpha}_{MAR}, \tilde{\beta}_{MAR})$ are estimated by GMM.

2. MK2 estimator. We compute $(\hat{\alpha}_{MK}, \hat{\beta}_{MK}, \hat{\theta}_{MK})$ using the approach of Morikawa and Kim (2016), i.e. $(\hat{\alpha}_{MK}, \hat{\beta}_{MK}, \hat{\theta}_{MK})$ is the solution of

$$\sum_{i=1}^{N} \left( \hat{\boldsymbol{S}}_1(T_i, \boldsymbol{Z}_i; \alpha, \beta)^\top, \hat{S}_2(T_i, \boldsymbol{Z}_i; \alpha, \beta, \theta) \right)^\top = 0 \ ,$$

where

$$\hat{\boldsymbol{S}}_1(T, \boldsymbol{Z}; \alpha, \beta) = -\left(1 - \frac{T}{\pi(\boldsymbol{Z}; \alpha, \beta)}\right) \mathbb{E}^\star \left[ \frac{\nabla_\gamma \pi(\boldsymbol{Z}; \alpha, \beta)}{1 - \pi(\boldsymbol{Z}; \alpha, \beta)} \middle| \boldsymbol{X} \right] \ ,$$

$$\hat{S}_2(T, \boldsymbol{Z}; \alpha, \beta, \theta) = -\frac{T}{\pi(\boldsymbol{Z}; \alpha, \beta)} U(\boldsymbol{Z}) + \theta - \left(1 - \frac{T}{\pi(\boldsymbol{Z}; \alpha, \beta)}\right) \mathbb{E}^\star [U(\boldsymbol{Z}) | \boldsymbol{X}] \ ,$$

and for any function $g(\boldsymbol{Z})$ the quantity $\mathbb{E}^\star[g(\boldsymbol{Z}) | \boldsymbol{X}]$ is defined by

$$\mathbb{E}^\star[g(\boldsymbol{Z}) | \boldsymbol{X} = \boldsymbol{x}] := \frac{\sum_{j=1}^{N} T_j K_h(\boldsymbol{x} - \boldsymbol{X}_j) T_j \pi(\boldsymbol{Z}_j; \alpha, \beta)^{-1} O(\boldsymbol{x}, Y_j; \alpha, \beta) g(\boldsymbol{x}, Y_j)}{\sum_{j=1}^{N} K_h(\boldsymbol{x} - \boldsymbol{X}_j) T_j \pi(\boldsymbol{Z}_j; \alpha, \beta)^{-1} O(\boldsymbol{x}, Y_j; \alpha, \beta)} \ ;$$

$$O(\boldsymbol{z}; \alpha, \beta) = \frac{1 - \pi(\boldsymbol{z}; \alpha, \beta)}{\pi(\boldsymbol{z}; \alpha, \beta)} \ ,$$

16

$K_h(\boldsymbol{x} - \boldsymbol{w}) = K\left((\boldsymbol{x} - \boldsymbol{w}/h)\right)$, $K(\cdot)$ is Gaussian kernel function and $h$ is the bandwidth.

3. Our GMM estimator. We compute $(\hat{\alpha}, \hat{\beta}, \hat{\theta})$ using the proposed approach and the covariate balancing approach to select $K$, with $\bar{K} = 7$ in Scenarios I, II, III, and with $\bar{K} = 10$ in Scenario IV. Here $\bar{K}$ is the maximal number of candidate moments to be considered.

The simulation results (the bias, the standard deviation (Stdev), the mean squared error (MSE), and the coverage probability (CP) (for significance level $\alpha = 0.05$) of the point estimates) for all scenarios are reported in Tables 1, 2, 3, and 4 respectively. The histogram of selected $K'$s (based on 500 Monte Carlo samples) in all scenarios is reported in Figure 1. Glancing at these tables, we find:

1. In all scenarios, the naive estimator using the missing at random assumption has a large bias, because this assumption does not hold.

2. In all scenarios, our proposed estimator of $\mathbb{E}[Y]$ out-performs the MK estimator.

3. In all scenarios, our proposed variance estimator has coverage probability close to 95%, even the sample size is small. The MK's variance estimator performs well in Scenario IV, but badly in other scenarios: in Scenarios I, the coverage probability using MK's approach converges to 90% rather than 95%; in Scenarios II, the CP values are far from 95% in Scenario 2 no matter the sample size is small or large; in Scenarios III, the MK's variance estimaotr is consistent only when the sample size is large.

4. When the sample size is small the optimal $K$ tends to be 2. When the sample size is large, the optimal $K$ tends to be 3. The growing rate of $K$ is extremely slow comparing to that of the sample size $n$, which is consistent with our theoretical Assumption 8.

These results clearly show that the proposed approach has better finite sample performance.

Figure 1: Histogram of $K$
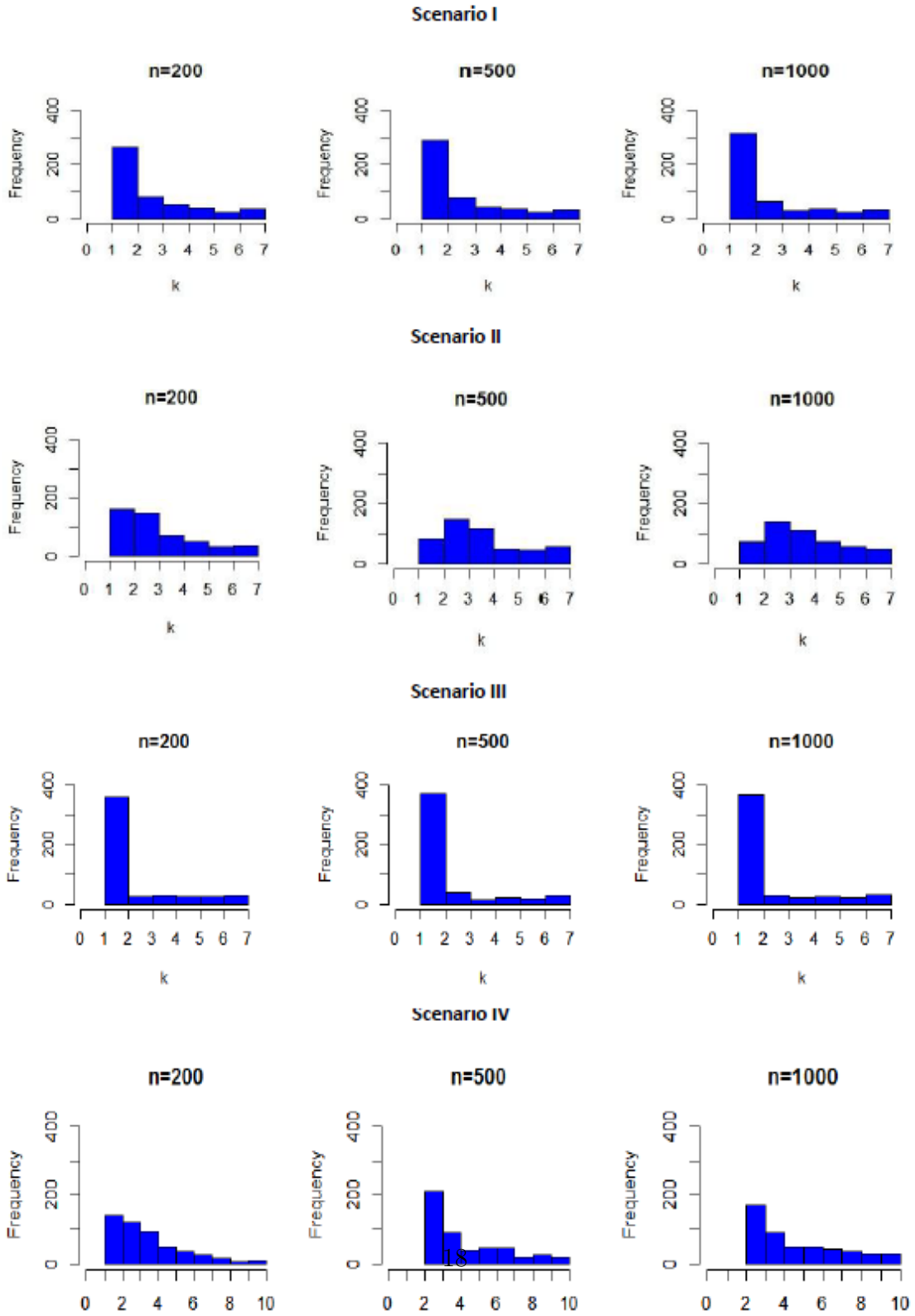


The Monte Carlo sample size used to plot the histogram of $K$ is $J = 500$.

# 7 Discussion

The data missing not at random problem is common in applications. Morikawa and Kim (2016) studies the efficient estimation of a class of missing not at random problems. But their approach requires nonparametric estimation of the conditional density function and thus suffers from the curse of dimensionality and smoothing parameter selection problem. In this paper, we study the same class of missing not at random problems but present a much simpler and more natural efficient estimator. Our approach is based on a parametric moment restriction model that does not require nonparametric estimation and hence does not suffer from the curse of dimensionality problem nor the bandwidth selection problem. Indeed the simulation results confirm that the proposed approach out-performs their approach in finite samples. The GMM approach is also easy to adapt to stratified sampling and other sampling schemes common in survey data.

Both approaches require correct parameterization of the propensity score function. If the propensity score function is misspecified, then both approaches yield inconsistent estimates. There is some attempt in the literature to mitigate this problem. For instance, Zhao and Shao (2015) introduce a partial linear index to model missing mechanism. The proposed approach can be extended in this direction. Such extension shall be pursued in a future study.

# 8 Appendix

## 8.1 Assumptions

We first introduce the smoothness classes of functions used in the nonparametric estimation; see e.g. Stone (1982, 1994), Robinson (1988), Newey (1997), Horowitz (2012) and Chen (2007). Suppose that $\mathcal{X}$ is the Cartesian product of $r$-compact intervals. Let $0 < \delta \leq 1$. A fucntion $f$ on $\mathcal{X}$ is said to satisfy a Hölder condition with exponent $\delta$ if there is a positive constant $L$ usch that $\|f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2)\| \leq L\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|^{\delta}$ for all $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$. Given a $r$-tuple $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_r)$ of nonnegative integer, denote $[\boldsymbol{\alpha}] = \alpha_1 + \cdots + \alpha_r$ and let $D^{\boldsymbol{\alpha}}$ denote the differential operator defined by $D^{\boldsymbol{\alpha}} = \frac{\partial^{[\boldsymbol{\alpha}]}}{\partial x_1^{\alpha_1}...\partial x_r^{\alpha_r}}$, where $\boldsymbol{x} = (x_1, ..., x_r)$.

**Definition 1.** Let $s$ be a nonnegative integer and $s := s_0 + \delta$. The function $f$ on $\mathcal{X}$ is said to be *s-smooth* if it is $s$ times continuously differentiable on $\mathcal{X}$ and $D^{\boldsymbol{\alpha}} f$ satisfies a Hölder condition with exponent $\delta$ for all $\boldsymbol{\alpha}$ with $[\boldsymbol{\alpha}] = s_0$.

The following notations are needed for presenting the efficiency bounds:

$$O(\boldsymbol{Z}) := \frac{1 - \pi(\boldsymbol{Z}; \gamma_0)}{\pi(\boldsymbol{Z}; \gamma_0)}, \quad \boldsymbol{S}_0(\boldsymbol{Z}) := -\frac{\nabla_\gamma \pi(\boldsymbol{Z}; \gamma_0)}{1 - \pi(\boldsymbol{Z}; \gamma_0)} , \tag{10}$$

$$m(\boldsymbol{X}) := \frac{\mathbb{E}[O(\boldsymbol{Z})\boldsymbol{S}_0(\boldsymbol{Z})|\boldsymbol{X}]}{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}, \quad R(\boldsymbol{X}) := \frac{\mathbb{E}[O(\boldsymbol{Z})U(\boldsymbol{Z})|\boldsymbol{X}]}{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]} , \tag{11}$$

$$\boldsymbol{S}_1(T, \boldsymbol{Z}; \gamma_0) := \left(1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)}\right) m(\boldsymbol{X}) , \tag{12}$$

$$S_2(T, \boldsymbol{Z}; \gamma_0, \theta_0) := -\frac{T}{\pi(\boldsymbol{Z}; \gamma_0)}U(\boldsymbol{Z}) + \theta_0 - \left(1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)}\right) R(\boldsymbol{X}) . \tag{13}$$

The following assumptions are required in this paper:

**Assumption 1.** *There exists a nonresponse instrumental variable* $\mathbf{X_2}$, *i.e.,* $\mathbf{X} = (\mathbf{X}_1^\intercal, \mathbf{X}_2^\intercal)^\intercal$, *such that* $\boldsymbol{X}_2$ *is independent of* $T$ *given* $\boldsymbol{X}_1$ *and* $Y$; *furthermore,* $\boldsymbol{X}_2$ *is correlated with* $Y$.

**Assumption 2.** *The support of* $\mathbf{X}$, *which is denoted by* $\mathcal{X}$, *is a Cartesian product of* $r$-*compact intervals, and we denote* $\boldsymbol{X} = (X_1, ..., X_r)^\top$.

**Assumption 3.** *The functions* $\mathbb{E}[O(\boldsymbol{Z})S_0(\boldsymbol{Z})|\mathbf{X} = \boldsymbol{x}]$, $\mathbb{E}[O(\boldsymbol{Z})U(\boldsymbol{Z})|\mathbf{X} = \boldsymbol{x}]$ *and* $\mathbb{E}[O(\boldsymbol{Z})|\mathbf{X} = \boldsymbol{x}]$ *are* $s$-*smooth in* $\boldsymbol{x}$, *where* $s > 0$.

**Assumption 4.** *There exists two finite positive constants* $\underline{a}$ *and* $\bar{a}$ *such that the smallest (resp. largest) eigenvalue of* $\mathbb{E}[u_K(\boldsymbol{X})u_K^\top(\boldsymbol{X})]$ *is bounded away from* $\underline{a}$ *(resp.* $\bar{a}$*) uniformly in* $K$, *i.e.,*

$$0 < \underline{a} \le \lambda_{\min}(\mathbb{E}[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^\top]) \le \lambda_{\max}(\mathbb{E}[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^\top]) \le \bar{a} < \infty .$$

**Remark**: Asssumption 4 implies that following results:

1.

$$\mathbb{E}[\|u_K(\boldsymbol{X})\|^2] = \mathrm{tr}\left(\mathbb{E}\left[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^\top\right]\right) = O(K) ; \tag{14}$$

2. the matrices $\bar{a} \cdot I_{K \times K} - \mathbb{E}[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^\top]$ and $\mathbb{E}[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^\top] - \underline{a} \cdot I_{K \times K}$ are positive definite, and

$$\underline{a} \le \inf_{k \in \{1, ..., K\}} \mathbb{E}[u_{kK}(\boldsymbol{X})^2] \le \sup_{k \in \{1, ..., K\}} \mathbb{E}[u_{kK}(\boldsymbol{X})^2] \le \bar{a} . \tag{15}$$

**Assumption 5.** *The full data* $\{(T_i, \boldsymbol{X}_i, Y_i)\}_{i=1}^N$ *are independently and identically distributed.*

**Assumption 6.** $\boldsymbol{S}_{eff}(T, \boldsymbol{Z}; \gamma, \theta) := (\boldsymbol{S}_1^{\mathsf{T}}(T, \boldsymbol{Z}; \gamma), S_2(T, \boldsymbol{Z}; \gamma, \theta))^{\mathsf{T}}$ *is continuously differentiable at each* $(\gamma, \theta) \in \Gamma \times \Theta$ *with probability one, and* $\mathbb{E}\left[\partial \boldsymbol{S}_{eff}(\gamma, \theta)/\partial(\gamma^{\top}, \theta)\right]$ *is nonsingular at* $(\gamma_0, \theta_0)$.

**Assumption 7.** *The response probability* $\pi$ *satisfies the following conditions:*

1. *there exists two positive constants* $\bar{c}$ *and* $\underline{c}$ *such that* $0 < \underline{c} \leq \pi(\boldsymbol{x}, y; \gamma) \leq \bar{c} < 1$ *for all* $\gamma \in \Gamma$ *and* $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathbb{R}$;

2. *the propensity score* $\pi(\boldsymbol{x}, y; \gamma)$ *is twice continuously differentiable in* $\gamma \in \Gamma$, *and the derivatives are uniformly bounded.*

3. *for any* $\gamma \in \Gamma$, *the conditional functions* $\mathbb{E}\left[1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma)} \mid \boldsymbol{X} = \boldsymbol{x}\right]$ *and* $\mathbb{E}\left[\frac{\nabla_{\gamma} \pi(\boldsymbol{Z}; \gamma)}{\pi(\boldsymbol{Z}; \gamma)} \mid \boldsymbol{X} = \boldsymbol{x}\right]$ *are s-smooth in* $\boldsymbol{x}$, *where* $s > 0$.

**Assumption 8.** *Suppose* $K \to \infty$ *and* $K^3/N \to 0$.

The Assumption 1 is used for the identification of the model, which was discussed in Wang et al. (2014). Assumptions 2 and 3 are required for uniform boundedness of approximations. Assumption 4 is a standard assumption used in nonparametric sieve approximation, see also Newey (1997). Assumption 5 is a standard condition for statistical sampling. Assumptions 6-7 are required for the convergence of our estimator as well as the boundness of the asymptotic variance. Assumption 8 is the same as Assumption 2 in Newey and Powell (2003), it is required for controlling the stochastic order of the residual terms, which is desirable in practice because $K$ grows very slowly with $N$ so a relatively small number of moment conditions is sufficient for the method proposed to perform well.

## 8.2 Discussion on $u_K$

To construct the GMM estimator, we need to specify the matching function $u_K(\boldsymbol{X})$. Although the approximation theory is derived for general sequences of approximating functions, the most common class of functions are power series. Suppose the dimension of covariate $\boldsymbol{X}$ is $r \in \mathbb{N}$, namely $\boldsymbol{X} = (X_1, ..., X_r)^{\top}$. Let $\lambda = (\lambda_1, \ldots \lambda_r)^{\top}$ be an $r$-dimensional vector of nonnegative integers (multi-indices), with norm $|\lambda| = \sum_{j=1}^{r} \lambda_j$. Let $(\lambda(k))_{k=1}^{\infty}$ be a sequence that includes all distinct multi-indicesand satisfies $|\lambda(k)| \leq |\lambda(k+1)|$, and let $\boldsymbol{X}^{\lambda} = \prod_{j=1}^{r} X_j^{\lambda_j}$. For a sequence $\lambda(k)$ we consider the series $u_{kK}(\boldsymbol{X}) = \boldsymbol{X}^{\lambda(k)}$, $k \in \{1, ..., K\}$. Newey (1997) showed the following property for the power series: there exists a universal constant $C > 0$ such that

$$\zeta(K) := \sup_{\boldsymbol{x} \in \mathcal{X}} \|u_K(\boldsymbol{x})\| \leq CK , \tag{16}$$

where $\| \cdot \|$ denotes the usual matrix norm $\|A\| = \sqrt{\mathrm{tr}(A^\top A)}$.

Another important issue is choosing the number of the matching function $K$ in finite sample experiment. Donald et al. (2009) proposed a strategy for an appropriate choice of $K$ by minimizing the higher order MSE defined in (9), and the following notations are needed to describe this criteria:

$$\rho(T_i, \boldsymbol{X}_i, Y_i; \check{\gamma}) = 1 - \frac{T_i}{\pi(\boldsymbol{X}_i, Y_i; \check{\gamma})}, \quad \hat{\Upsilon}_{K \times K} = \frac{1}{N} \sum_{i=1}^{N} \rho(T_i, \boldsymbol{X}_i, Y_i; \check{\gamma})^2 u_K(\boldsymbol{X}_i)^{\otimes 2},$$

$$\widehat{\Gamma}_{K \times p} = \frac{1}{N} \sum_{i=1}^{N} u_K(\boldsymbol{X}_i) \nabla_\gamma \rho(T_i, \boldsymbol{X}_i, Y_i; \check{\gamma})^\top, \quad \hat{\Omega}_{p \times p} = (\hat{\Gamma}_{K \times p})^\top \hat{\Upsilon}_{K \times K}^{-1} \hat{\Gamma}_{K \times p},$$

$$\tilde{\mathbf{d}}_i = (\hat{\Gamma}_{K \times p})^\top \left( \frac{1}{N} \sum_{j=1}^{N} u_K(\boldsymbol{X}_j)^{\otimes 2} \right)^{-1} u_K(\boldsymbol{X}_i), \quad \tilde{\eta}_i = \nabla_\gamma \rho(T_i, \boldsymbol{X}_i, Y_i; \check{\gamma}) - \tilde{\mathbf{d}}_i,$$

$$\hat{\xi}_{ij} = \frac{1}{N} u_K(\boldsymbol{X}_i)^\top \hat{\Upsilon}_{K \times K}^{-1} u_K(\boldsymbol{X}_j), \quad \widehat{\mathbf{D}}_i^* = (\hat{\Gamma}_{K \times p})^\top \hat{\Upsilon}_{K \times K}^{-1} u_K(\boldsymbol{X}_i) .$$

## 8.3 Semiparametric Efficiency Bounds

The following lemma is Theorem 1 in Morikawa and Kim (2016).

**Lemma 1** (Morikawa and Kim (2016))**.** *The efficient variance bounds of $(\gamma_0, \theta_0)$ is $\boldsymbol{V}_{eff} := \mathbb{E}\left[\boldsymbol{S}_{eff}(T, \boldsymbol{Z}; \gamma_0, \theta_0)^{\otimes 2}\right]^{-1}$, where $\boldsymbol{S}_{eff} = (\boldsymbol{S}_1^\top, S_2)^\top$ and $\boldsymbol{S}_1, S_2$ are defined in (12) and (13) respectively.*

Let $\boldsymbol{V}_{\gamma_0}$ (resp. $V_{\theta_0}$) be the efficient variance bound of $\gamma_0$ (resp. $\theta_0$). After some simple computation, we can find out

$$\boldsymbol{V}_{\gamma_0} = \mathbb{E}\left[ \frac{1 - \pi(\boldsymbol{Z}; \gamma_0)}{\pi(\boldsymbol{Z}; \gamma_0)} m(\boldsymbol{X})^{\otimes 2} \right]^{-1} \tag{17}$$

and

$$V_{\theta_0} = Var\left( S_2(T, \boldsymbol{Z}; \gamma_0, \theta_0) - \kappa^\top \boldsymbol{S}_1(T, \boldsymbol{Z}; \gamma_0) \right) . \tag{18}$$

where

$$\kappa^\top := \mathbb{E}\left[ \frac{\nabla_\gamma \pi(\boldsymbol{Z}; \gamma_0)^\top}{\pi(\boldsymbol{Z}; \gamma_0)} \{R(\boldsymbol{Z}) - U(\boldsymbol{X})\} \right] \cdot \mathbb{E}\left[ \frac{m(\boldsymbol{X})}{\pi(\boldsymbol{Z}; \gamma_0)} \nabla_\gamma \pi(\boldsymbol{Z}; \gamma_0)^\top \right]^{-1} . \tag{19}$$

# References

Ai, C. and Chen, X. (2012). The semiparametric efficiency bound for models of sequential moment restrictions containing unknown functions, *Journal of Econometrics* **170**(2): 442–457.

Ai, C., Linton, O. and Zhang, Z. (2018). Supplemental material for "a simple and efficient estimation method for models with nonignorable missing data", *Technical report* .

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

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models, *Handbook of econometrics* **6**: 5549–5632.

Chen, X., Hong, H. and Tarozzi, A. (2008). Semiparametric efficiency in gmm models with auxiliary data, *Ann. Statist.* **36**(2): 808–843.

Donald, S. G., Imbens, G. W. and Newey, W. K. (2009). Choosing instrumental variables in conditional moment restriction models, *Journal of Econometrics* **152**(1): 28–36.

Geman, S. and Hwang, C.-R. (1982). Nonparametric maximum likelihood estimation by the method of sieves, *The Annals of Statistics* pp. 401–414.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators, *Econometrica: Journal of the Econometric Society* pp. 1029–1054.

Heitjan, D. F. and Basu, S. (1996). Distinguishing missing at random and missing completely at random, *The American Statistician* **50**(3): 207–213.

Horowitz, J. L. (2012). *Semiparametric methods in econometrics*, Vol. 131, Springer Science & Business Media.

Kang, J. and Schafer, J. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data, *Statistical science* **22**(4): 523–539.

Kim, J. K. and Yu, C. L. (2011). A semiparametric estimation of mean functionals with nonignorable missing data, *Journal of the American Statistical Association* **106**(493): 157–165.

Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values, *Journal of the American Statistical Association* **83**(404): 1198–1202.

Little, R. J. and Rubin, D. B. (1989). The analysis of social science data with missing values, *Sociological Methods & Research* **18**(2-3): 292–326.

Little, R. J. and Rubin, D. B. (2014). *Statistical analysis with missing data*, John Wiley & Sons.

Morikawa, K. and Kim, J. K. (2016). Semiparametric adaptive estimation with nonignorable nonresponse data, *arXiv preprint arXiv:1612.09207* .

Newey, W. K. (1997). Convergence rates and asymptotic normality for series estimators, *Journal of econometrics* **79**(1): 147–168.

Newey, W. K. and Powell, J. L. (2003). Instrumental variable estimation of nonparametric models, *Econometrica* **71**(5): 1565–1578.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of gmm and generalized empirical likelihood estimators, *Econometrica* **72**(1): 219–255.

Qin, J., Leung, D. and Shao, J. (2002). Estimation with survey data under nonignorable nonresponse or informative sampling, *Journal of the American Statistical Association* **97**(457): 193–200.

Qin, J. and Zhang, B. (2007). Empirical-likelihood-based inference in missing response problems and its application in observational studies, *J. R. Statist. Soc. B (Statistical Methodology)* **69**(1): 101–122.

Riddles, M. K., Kim, J. K. and Im, J. (2016). A propensity-score-adjustment method for nonignorable nonresponse, *Journal of Survey Statistics and Methodology* **4**(2): 215–245.

Robins, J. M., Ritov, Y. et al. (1997). Toward a curse of dimensionality appropriate(coda) asymptotic theory for semi-parametric models, *Statistics in medicine* **16**(3): 285–319.

Robins, J. M. and Rotnitzky, A. (1995). Semiparametric efficiency in multivariate regression models with missing data, *Journal of the American Statistical Association* **90**(429): 122–129.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data, *Journal of the american statistical association* **90**(429): 106–121.

Robinson, P. M. (1988). Root-n-consistent semiparametric regression, *Econometrica: Journal of the Econometric Society* pp. 931–954.

Rotnitzky, A., Lei, Q., Sued, M. and Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models, *Biometrika* **99**(2): 439–456.

Rubin, D. B. (1976a). Comparing regressions when some predictor values are missing, *Technometrics* **18**(2): 201–205.

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

Shao, J. and Wang, L. (2016). Semiparametric inverse propensity weighting for nonignorable missing data, *Biometrika* **103**(1): 175–187.

Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression, *The annals of statistics* pp. 1040–1053.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation, *The Annals of Statistics* pp. 118–171.

Sverchkov, M. (2008). A new approach to estimation of response probabilities when missing data are not missing at random, *Proceedings of the Survey Research Methods Section*, pp. 867–874.

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

Tang, G., Little, R. J. and Raghunathan, T. E. (2003). Analysis of multivariate missing data with nonignorable nonresponse, *Biometrika* **90**(4): 747–764.

Wang, S., Shao, J. and Kim, J. K. (2014). An instrumental variable approach for identification and estimation with nonignorable nonresponse, *Statistica Sinica* pp. 1097–1116.

Zhao, J. and Shao, J. (2015). Semiparametric pseudo-likelihoods in generalized linear models with nonignorable missing data, *Journal of the American Statistical Association* **110**(512): 1577–1590.

# Supplemental Material for

## "*A Simple and Efficient Estimation Method for Models with Nonignorable Missing Data* "

Chunrong Ai[*]– University of Florida
Oliver Linton[†]– University of Cambridge
Zheng Zhang[‡]– Renmin University of China

[*]Department of Economics, University of Florida. E-mail: `tsinghua@ufl.edu`
[†]Faculty of Economics, University of Cambridge. E-mail: `obl20@cam.ac.uk`
[‡]Institute of Statistics and Big Data, Renmin University of China. E-mail: `zhengzhang@ruc.edu.cn`

1

# Contents

# 1   Assumptions

The following notations are needed for our proof:

$$O(\mathbf{Z}) := \frac{1 - \pi(\mathbf{Z}; \gamma_0)}{\pi(\mathbf{Z}; \gamma_0)}, \quad \mathbf{S}_0(\mathbf{Z}) := -\frac{\nabla_\gamma \pi(\mathbf{Z}; \gamma_0)}{1 - \pi(\mathbf{Z}; \gamma_0)} , \tag{1}$$

$$m(\mathbf{X}) := \frac{\mathbb{E}[O(\mathbf{Z})\mathbf{S}_0(\mathbf{Z})|\mathbf{X}]}{\mathbb{E}[O(\mathbf{Z})|\mathbf{X}]}, \quad R(\mathbf{X}) := \frac{\mathbb{E}[O(\mathbf{Z})U(\mathbf{Z})|\mathbf{X}]}{\mathbb{E}[O(\mathbf{Z})|\mathbf{X}]} , \tag{2}$$

$$\mathbf{S}_1(T, \mathbf{Z}; \gamma_0) := \left(1 - \frac{T}{\pi(\mathbf{Z}; \gamma_0)}\right) m(\mathbf{X}) , \tag{3}$$

$$S_2(T, \mathbf{Z}; \gamma_0, \theta_0) := -\frac{T}{\pi(\mathbf{Z}; \gamma_0)} U(\mathbf{Z}) + \theta_0 - \left(1 - \frac{T}{\pi(\mathbf{Z}; \gamma_0)}\right) R(\mathbf{X}) , \tag{4}$$

$$\kappa^{\mathsf{T}} := \mathbb{E}\left[\frac{\nabla_\gamma \pi(\mathbf{Z}; \gamma_0)^{\mathsf{T}}}{\pi(\mathbf{Z}; \gamma_0)} \{R(\mathbf{Z}) - U(\mathbf{X})\}\right] \cdot \mathbb{E}\left[\frac{m(\mathbf{X})}{\pi(\mathbf{Z}; \gamma_0)} \nabla_\gamma \pi(\mathbf{Z}; \gamma_0)^{\mathsf{T}}\right]^{-1} , \tag{5}$$

$$\mathbf{V}_{\gamma_0} = \mathbb{E}\left[\frac{1 - \pi(\mathbf{Z}; \gamma_0)}{\pi(\mathbf{Z}; \gamma_0)} m(\mathbf{X})^{\otimes 2}\right]^{-1} \text{ (the efficient variance bound of } \gamma_0) , \tag{6}$$

$$V_{\theta_0} = Var\left(S_2(T, \mathbf{Z}; \gamma_0, \theta_0) - \kappa^{\mathsf{T}} \mathbf{S}_1(T, \mathbf{Z}; \gamma_0)\right) \text{ (the efficient variance bound of } \theta_0) . \tag{7}$$

The following assumptions are required in this paper:

**Assumption 1**  *There exists a nonresponse instrumental variable $\mathbf{X_2}$, i.e., $\mathbf{X} = (\mathbf{X}_1^{\mathsf{T}}, \mathbf{X}_2^{\mathsf{T}})^{\mathsf{T}}$, such that $\mathbf{X}_2$ is independent of $T$ given $\mathbf{X}_1$ and $Y$; furthermore, $\mathbf{X}_2$ is correlated with $Y$.*

**Assumption 2**  *The support of $\mathbf{X}$, which is denoted by $\mathcal{X}$, is a Cartesian product of $r$-compact intervals, and we denote $\mathbf{X} = (X_1, ..., X_r)^{\top}$.*

**Assumption 3**  *The functions $\mathbb{E}[O(\mathbf{Z})S_0(\mathbf{Z})|\mathbf{X} = \mathbf{x}]$, $\mathbb{E}[O(\mathbf{Z})U(\mathbf{Z})|\mathbf{X} = \mathbf{x}]$ and $\mathbb{E}[O(\mathbf{Z})|\mathbf{X} = \mathbf{x}]$ are $s$-smooth in $\mathbf{x}$, where $s > 0$.*

**Assumption 4**  *There exists two finite positive constants $\underline{a}$ and $\overline{a}$ such that the smallest (resp. largest) eigenvalue of $\mathbb{E}[u_K(\mathbf{X})u_K^{\top}(\mathbf{X})]$ is bounded away from $\underline{a}$ (resp. $\overline{a}$) uniformly in $K$, i.e.,*

$$0 < \underline{a} \leq \lambda_{\min}(\mathbb{E}[u_K(\mathbf{X})u_K(\mathbf{X})^{\top}]) \leq \lambda_{\max}(\mathbb{E}[u_K(\mathbf{X})u_K(\mathbf{X})^{\top}]) \leq \overline{a} < \infty .$$

**Remark**: Asssumption 4 implies that following results:

1.

$$\mathbb{E}[\|u_K(\mathbf{X})\|^2] = \operatorname{tr}\left(\mathbb{E}\left[u_K(\mathbf{X})u_K(\mathbf{X})^{\top}\right]\right) = O(K) ; \tag{8}$$

3

2. the matrices $\bar{a} \cdot I_{K \times K} - \mathbb{E}[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^\top]$ and $\mathbb{E}[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^\top] - \underline{a} \cdot I_{K \times K}$ are positive definite, and

$$\underline{a} \leq \inf_{k \in \{1,\dots,K\}} \mathbb{E}[u_{kK}(\boldsymbol{X})^2] \leq \sup_{k \in \{1,\dots,K\}} \mathbb{E}[u_{kK}(\boldsymbol{X})^2] \leq \bar{a} . \tag{9}$$

**Assumption 5** *The full data $\{(T_i, \boldsymbol{X}_i, Y_i)\}_{i=1}^N$ are independently and identically distributed.*

**Assumption 6** $\boldsymbol{S}_{eff}(T, \boldsymbol{Z}; \gamma, \theta) := (\boldsymbol{S}_1^\top(T, \boldsymbol{Z}; \gamma), S_2(T, \boldsymbol{Z}; \gamma, \theta))^\top$ *is continuously differentiable at each $(\gamma, \theta) \in \Gamma \times \Theta$ with probability one, and $\mathbb{E}\left[\partial \boldsymbol{S}_{eff}(\gamma, \theta)/\partial(\gamma^\top, \theta)\right]$ is nonsingular at $(\gamma_0, \theta_0)$.*

**Assumption 7** *The response probability $\pi$ satisfies the following conditions:*

1. *there exists two positive constants $\bar{c}$ and $\underline{c}$ such that $0 < \underline{c} \leq \pi(\boldsymbol{x}, y; \gamma) \leq \bar{c} < 1$ for all $\gamma \in \Gamma$ and $(\boldsymbol{x}, y) \in \mathcal{X} \times \mathbb{R}$;*

2. *the propensity score $\pi(\boldsymbol{x}, y; \gamma)$ is twice continuously differentiable in $\gamma \in \Gamma$, and the derivatives are uniformly bounded.*

3. *for any $\gamma \in \Gamma$, the conditional functions $\mathbb{E}\left[1 - \frac{T}{\pi(\boldsymbol{Z};\gamma)}\Big|\boldsymbol{X} = \boldsymbol{x}\right]$ and $\mathbb{E}\left[\frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma)}{\pi(\boldsymbol{Z};\gamma)}\Big|\boldsymbol{X} = \boldsymbol{x}\right]$ are s-smooth in $\boldsymbol{x}$, where $s > 0$.*

**Assumption 8** *Suppose $K \to \infty$ and $K^3/N \to 0$.*

# 2 Some useful results

We present some results which will be used in the proof of Theorems 1 and 2.

## 2.1 Matrix inversion formula

- (General Formula) Let $\mathbf{A}$, $\mathbf{C}$, and $\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B}$ be non-singular square matrices; then

$$(\mathbf{A} + \mathbf{B}\mathbf{C}\mathbf{D})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{C}^{-1} + \mathbf{D}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{D}\mathbf{A}^{-1} . \tag{10}$$

- (Matrix Inversion in Block form) Let a $(m+1) \times (m+1)$ matrix $\boldsymbol{M}$ be partitioned into a block form

$$\boldsymbol{M} = \begin{bmatrix} \boldsymbol{A} , & \boldsymbol{b} \\ \boldsymbol{b}^\top , & d \end{bmatrix}$$

4

where $\boldsymbol{A}$ is a $m \times m$ matrix, $\boldsymbol{b}$ is a $m$ dimensional column vector, $d$ is a constant. Then

$$\boldsymbol{M}^{-1} = \begin{bmatrix} \boldsymbol{A}^{-1} + \frac{1}{k}\boldsymbol{A}^{-1}\boldsymbol{b}\boldsymbol{b}^{\top}\boldsymbol{A}^{-1}, & -\frac{1}{k}\boldsymbol{A}^{-1}\boldsymbol{b} \\ -\frac{1}{k}\boldsymbol{b}^{\top}\boldsymbol{A}^{-1}, & \frac{1}{k} \end{bmatrix} , \tag{11}$$

where $k = d - \boldsymbol{b}^{\top}\boldsymbol{A}^{-1}\boldsymbol{b}$.

## 2.2 Discussion on $u_K$

To construct the GMM estimator, we need to specify the matching function $u_K(\boldsymbol{X})$. Although the approximation theory is derived for general sequences of approximating functions, the most common class of functions are power series. Suppose the dimension of covariate $\boldsymbol{X}$ is $r \in \mathbb{N}$, namely $\boldsymbol{X} = (X_1, ..., X_r)^{\top}$. Let $\lambda = (\lambda_1, \ldots \lambda_r)^{\top}$ be an $r$-dimensional vector of nonnegative integers (multi-indices), with norm $|\lambda| = \sum_{j=1}^{r} \lambda_j$. Let $(\lambda(k))_{k=1}^{\infty}$ be a sequence that includes all distinct multi-indicesand satisfies $|\lambda(k)| \le |\lambda(k+1)|$, and let $\boldsymbol{X}^{\lambda} = \prod_{j=1}^{r} X_j^{\lambda_j}$. For a sequence $\lambda(k)$ we consider the series $u_{kK}(\boldsymbol{X}) = \boldsymbol{X}^{\lambda(k)}$, $k \in \{1, ..., K\}$. Newey (1997) showed the following property for the power series: there exists a universal constant $C > 0$ such that

$$\zeta(K) := \sup_{\boldsymbol{x} \in \mathcal{X}} \|u_K(\boldsymbol{x})\| \le CK , \tag{12}$$

where $\| \cdot \|$ denotes the usual matrix norm $\|A\| = \sqrt{\text{tr}(A^{\top}A)}$.

## 2.3 Convergence rate of $L^2$ approximation

Suppose $f : \mathbb{R}^r \to \mathbb{R}$ is the function we want to approximate. Let $f_K(\mathbf{X})$ be the $L^2$-projection of $f(\boldsymbol{X})$ on the space linearly spanned by $u_K(\boldsymbol{X})$, i.e.

$$f_K(\mathbf{X}) = \beta_K^{\top} u_K(\boldsymbol{X}) \tag{13}$$

where

$$\beta_K := \mathbb{E}\left[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^{\top}\right]^{-1} \mathbb{E}\left[u_K(\boldsymbol{X})f(\boldsymbol{X})\right] .$$

In this section, we establish the $L^2$-convergence rate of $f_K(\boldsymbol{X})$ to $f(\boldsymbol{X})$, which will be used for proving the theorems of our paper.

**Lemma 2.1** *Under Assumpitons 2 and 4, suppose the function $f : \mathbb{R}^r \to \mathbb{R}$ is s-smooth and $f_K$ is defined by (13), then we have*

$$\mathbb{E}\left[|f(\boldsymbol{X}) - f_K(\boldsymbol{X})|^2\right] = O\left(K^{-\frac{2s}{r}}\right) .$$

**Proof.** Since $f(\boldsymbol{x})$ is $s$-smooth and the support $\mathcal{X}$ is compact by Assumpiton 2, from Section 2.3.1 of Chen (2007), we know that there exists $\beta^* \in \mathbb{R}^K$ such that

$$\sup_{\boldsymbol{x} \in \mathcal{X}} \left| f(\boldsymbol{x}) - (\beta^*)^\top u_K(\boldsymbol{x}) \right| = O(K^{-\frac{s}{r}}) \ .$$

We first claim that

$$\|\beta_K - \beta^*\| = O(K^{-\frac{s}{r}}) \ . \tag{14}$$

With the claim (14), Cauchy-Schwarz inequality, and Assumption 4, we can obtain that

$$\begin{aligned}
&\mathbb{E}\left[ |f(\boldsymbol{X}) - f_K(\boldsymbol{X})|^2 \right] \\
=& \int_{\mathcal{X}} \left\{ (\beta_K - \beta^*)^\top u_K(\boldsymbol{x}) + \left[ (\beta^*)^\top u_K(\boldsymbol{x}) - f(\boldsymbol{x}) \right] \right\}^2 dF_X(\boldsymbol{x}) \\
\leq& 2(\beta_K - \beta^*)^\top \int_{\mathcal{X}} u_K(\boldsymbol{x}) u_K(\boldsymbol{x})^\top dF_X(\boldsymbol{x})(\beta_K - \beta^*) + 2 \int_{\mathcal{X}} \left[ (\beta^*)^\top u_K(\boldsymbol{x}) - f(\boldsymbol{x}) \right]^2 dF_X(\boldsymbol{x}) \\
\leq& 2\|\beta_K - \beta^*\|^2 \cdot \lambda_{\max}\left( \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right] \right) + 2 \sup_{\boldsymbol{x} \in \mathcal{X}} \left\| f(\boldsymbol{x}) - (\beta^*)^\top u_K(\boldsymbol{x}) \right\|^2 \\
=& 2\|\beta_K - \beta^*\|^2 \cdot O(1) + O(K^{-\frac{2s}{r}}) = O(K^{-\frac{2s}{r}}) \ .
\end{aligned}$$

We now prove the claim (14). Note that

$$\begin{aligned}
\beta_K - \beta^* =& \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right]^{-1} \mathbb{E}\left[ u_K(\boldsymbol{X}) f(\boldsymbol{X})^\top \right] - \beta^* \\
=& \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right]^{-1} \mathbb{E}\left[ u_K(\boldsymbol{X}) f(\boldsymbol{X})^\top \right] - \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right]^{-1} \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \beta^* \right] \\
=& \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right]^{-1} \mathbb{E}\left[ u_K(\boldsymbol{X}) \left\{ f(\boldsymbol{X}) - (\beta^*)^\top u_K(\boldsymbol{X}) \right\} \right] \ .
\end{aligned}$$

Then

$$\begin{aligned}
&\|\beta_K - \beta^*\|^2 \\
=& \operatorname{tr}\left( \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right]^{-1} \mathbb{E}\left[ u_K(\boldsymbol{X}) \left\{ f(\boldsymbol{X}) - (\beta^*)^\top u_K(\boldsymbol{X}) \right\} \right]^{\otimes 2} \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right]^{-1} \right) \\
\leq& \lambda_{\max}\left( \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right]^{-1} \right) \cdot \operatorname{tr}\left( \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right]^{-\frac{1}{2}} \mathbb{E}\left[ u_K(\boldsymbol{X}) \left\{ f(\boldsymbol{X}) - (\beta^*)^\top u(\boldsymbol{X}) \right\} \right]^{\otimes 2} \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right]^{-\frac{1}{2}} \right) \\
=& \lambda_{\max}\left( \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right]^{-1} \right) \cdot \mathbb{E}\left[ \left\{ f(\boldsymbol{X}) - (\beta^*)^\top u_K(\boldsymbol{X}) \right\} u_K(\boldsymbol{X})^\top \right] \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right]^{-1} \mathbb{E}\left[ u_K(\boldsymbol{X}) \left\{ f(\boldsymbol{X}) - (\beta^*)^\top u(\boldsymbol{X}) \right\} \right] \\
=& \lambda_{\max}\left( \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right]^{-1} \right) \cdot \mathbb{E}\left[ \left| \mathbb{E}\left[ \left\{ f(\boldsymbol{X}) - (\beta^*)^\top u_K(\boldsymbol{X}) \right\} u_K(\boldsymbol{X})^\top \right] \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right]^{-1} u_K(\boldsymbol{X}) \right|^2 \right] \\
\leq& \lambda_{\max}\left( \mathbb{E}\left[ u_K(\boldsymbol{X}) u_K(\boldsymbol{X})^\top \right]^{-1} \right) \mathbb{E}\left[ \left| f(\boldsymbol{X}) - (\beta^*)^\top u_K(\boldsymbol{X}) \right|^2 \right]
\end{aligned}$$

$$\leq \lambda_{\max}\left(\mathbb{E}\left[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^\top\right]^{-1}\right)\sup_{\boldsymbol{x}\in\mathcal{X}}\left|f(\boldsymbol{x})-(\beta^*)^\top u_K(\boldsymbol{x})\right|^2 = O(K^{-\frac{2s}{r}}),$$

where the first inequality follow from the fact that $\mathrm{tr}(AB) \leq \lambda_{\max}(B)\mathrm{tr}(A)$ for any symmetric matrix $B$ and positive semidefinite matrix $A$; the second inequality follows from the fact that $\mathbb{E}\left[\{f(\boldsymbol{X})-(\beta^*)^\top u_K(\boldsymbol{X})\}u_K(\boldsymbol{X})^\top\right]\mathbb{E}\left[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^\top\right]^{-1}u_K(\boldsymbol{X})$ is the $L^2$-projection of $f(\boldsymbol{X})-(\beta^*)^\top u_K(\boldsymbol{X})$ on the space spanned by $u_K(\boldsymbol{X})$, which implies

$$\left\|\mathbb{E}\left[\{f(\boldsymbol{X})-(\beta^*)^\top u_K(\boldsymbol{X})\}u_K(\boldsymbol{X})^\top\right]\mathbb{E}\left[u_K(\boldsymbol{X})u_K(\boldsymbol{X})^\top\right]^{-1}u_K(\boldsymbol{X})\right\|_{L^2} \leq \left\|f(\boldsymbol{X})-(\beta^*)^\top u_K(\boldsymbol{X})\right\|_{L^2}.$$

This complete the proof of the lemma.

∎

# 3 Proof of Theorem 1

Define the objective function:

$$\hat{Q}_N(\gamma,\theta) := \left\{\frac{1}{N}\sum_{i=1}^N g_K(T_i,\boldsymbol{Z}_i;\gamma,\theta)\right\}^\top \widehat{\boldsymbol{W}}_0^{-1}\left\{\frac{1}{N}\sum_{i=1}^N g_K(T_i,\boldsymbol{Z}_i;\gamma,\theta)\right\},$$

and

$$Q_0(\gamma,\theta) := \mathbb{E}\left[\left(\mathbb{E}\left[1-\frac{T}{\pi(\boldsymbol{Z};\gamma)}\Big|\mathbf{X}\right]\right)^2\right] + \left\{\mathbb{E}\left[\theta-\frac{T}{\pi(\boldsymbol{Z};\gamma)}U(\boldsymbol{Z})\right]\right\}^2.$$

By definition, $(\check{\gamma},\check{\theta})$ (resp. $(\gamma_0,\theta_0)$) is a unique minimizer of $\hat{Q}_N(\gamma,\theta)$ (resp. $Q_0(\gamma,\theta)$).

**We first show that** $\|(\check{\gamma},\check{\theta})-(\gamma_0,\theta_0)\|\xrightarrow{p}0$. Since $\Gamma\times\Theta$ is compact by definition, in order to prove the consistency, it is sufficient to show the following conditions hold:

(a) $\hat{Q}_N(\gamma,\theta)$ is continuous in $(\gamma,\theta)$;

(b) $Q_0(\gamma,\theta)$ is continuous in $(\gamma,\theta)$;

(c) $\sup_{(\gamma,\theta)\in\Gamma\times\Theta}|\hat{Q}_N(\gamma,\theta)-Q_0(\gamma,\theta)|\xrightarrow{p}0$.

Condition (a) is implied by Assumption 7.2. Condition (b) follows from Assumption 7, and Dominated Convergence Theorem. Therefore, we are left to show that Condition (c) holds.

Note that

$$
\hat{Q}_N(\gamma, \theta)
$$

$$
= \left\{ \frac{1}{N} \sum_{i=1}^{N} \left( 1 - \frac{T_i}{\pi(\boldsymbol{Z}_i; \gamma)} \right) u_K(\boldsymbol{X}_i) \right\}^{\top} \left[ \frac{1}{N} \sum_{i=1}^{N} u_K(\boldsymbol{X}_i) u_K(\boldsymbol{X}_i)^{\top} \right]^{-1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left( 1 - \frac{T_i}{\pi(\boldsymbol{Z}_i; \gamma)} \right) u_K(\boldsymbol{X}_i) \right\}
$$

$$
+ \left\{ \frac{1}{N} \sum_{i=1}^{N} \left( \theta_0 - \frac{T_i}{\pi(\boldsymbol{Z}_i; \gamma)} U(\boldsymbol{Z}_i) \right) \right\}^2
$$

$$
= \frac{1}{N} \sum_{i=1}^{N} \left( \widehat{\mathbb{E}} \left[ 1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma)} \middle| \boldsymbol{X} = \boldsymbol{X}_i \right] \right)^2 + \left\{ \frac{1}{N} \sum_{i=1}^{N} \left( \theta_0 - \frac{T_i}{\pi(\boldsymbol{Z}_i; \gamma)} U(\boldsymbol{Z}_i) \right) \right\}^2 ,
$$

where

$$
\widehat{\mathbb{E}} \left[ 1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma)} \middle| \boldsymbol{X} = \boldsymbol{x} \right] := \left\{ \frac{1}{N} \sum_{l=1}^{N} \left( 1 - \frac{T_l}{\pi(\boldsymbol{Z}_l; \gamma)} \right) u_K(\boldsymbol{X}_l) \right\}^{\top} \left[ \frac{1}{N} \sum_{l=1}^{N} u_K(\boldsymbol{X}_l) u_K(\boldsymbol{X}_l)^{\top} \right]^{-1} u_K(\boldsymbol{x}) .
$$

Therefore, to show $\sup_{(\gamma, \theta) \in \Gamma \times \Theta} |\hat{Q}_N(\gamma, \theta) - Q_0(\gamma, \theta)| \xrightarrow{p} 0$, it is sufficient to show that

$$
\sup_{\gamma \in \Gamma} \left| \frac{1}{N} \sum_{i=1}^{N} \left( \widehat{\mathbb{E}} \left[ 1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma)} \middle| \boldsymbol{X} = \boldsymbol{X}_i \right] \right)^2 - \mathbb{E} \left[ \left( \mathbb{E} \left[ 1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma)} \middle| \boldsymbol{X} \right] \right)^2 \right] \right| \xrightarrow{p} 0 , \qquad (15)
$$

$$
\sup_{(\gamma, \theta) \in \Gamma \times \Theta} \left| \left\{ \frac{1}{N} \sum_{i=1}^{N} \left( \theta - \frac{T_i}{\pi(\boldsymbol{Z}_i; \gamma)} U(\boldsymbol{Z}_i) \right) \right\}^2 - \left\{ \mathbb{E} \left[ \theta - \frac{T}{\pi(\boldsymbol{Z}; \gamma)} U(\boldsymbol{Z}) \right] \right\}^2 \right| \xrightarrow{p} 0 . \qquad (16)
$$

Since $\widehat{\mathbb{E}} \left[ 1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma)} \middle| \boldsymbol{X} \right]$ is the least square regression of $1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma)}$ on the basis $u_K(\boldsymbol{X})$, and Assumptions 2.2, 7, 8 hold, then the claim (15) follows from Corollary 4.2 of Newey (1991) .

Next we consider to prove (16). From Corrollary 2.2 of Newey (1991), to prove (16) it suffices to show the following conditions hold:

1. (Compactness) $\Gamma \times \Theta$ is compact;

2. (Pointwise Convergence) for each fixed $(\gamma, \theta) \in \Gamma \times \Theta$,

$$
\left\{ \frac{1}{N} \sum_{i=1}^{N} \left( \theta - \frac{T_i}{\pi(\boldsymbol{Z}_i; \gamma)} U(\boldsymbol{Z}_i) \right) \right\}^2 - \left\{ \mathbb{E} \left[ \theta - \frac{T}{\pi(\boldsymbol{Z}; \gamma)} U(\boldsymbol{Z}) \right] \right\}^2 \xrightarrow{p} 0 .
$$

3. For all $(\gamma, \theta) \in \Gamma \times \Theta$, the derivative of $\frac{1}{N} \sum_{i=1}^{N} \left( \theta - \frac{T_i}{\pi(\boldsymbol{Z}_i; \gamma)} U(\boldsymbol{Z}_i) \right)$ w.r.t. $(\gamma, \theta)$ is of $O_p(1)$.

The above three conditions can be easily verified based on our proposed assumptions. Condition 1 simply follows from our Assumption 2.2. Condition 2 follows from the weak law of large number theorem (WLLN). Condition 3 follows from Assumption 7 and Markov's inequality. Therefore, we can obtain our desired result (16). Combining (15) and (16), we complete the verification of Condition (c). Finally, we can obtain the consistency result

$$\|(\check{\gamma}, \check{\theta}) - (\gamma_0, \theta_0)\| \xrightarrow{p} 0 . \tag{17}$$

**Next, we establish the rate of convergence** $\|(\check{\gamma}, \check{\theta}) - (\gamma_0, \theta_0)\| = O_p(N^{-1/2})$. Using the first order condition of optimization in Step I, we obtain that

$$\mathbf{G}_K(\check{\gamma}, \check{\theta})^T \cdot \widehat{\boldsymbol{W}}_0^{-1} \cdot \nabla_{\gamma, \theta} \mathbf{G}_K(\check{\gamma}, \check{\theta}) = \mathbf{0} . \tag{18}$$

An application of Taylor's expansion yields

$$\begin{aligned}
\mathbf{0}_{1 \times (p+1)} =& \frac{1}{\sqrt{N}} \boldsymbol{G}_K^\intercal(\gamma_0, \theta_0) \cdot \widehat{\boldsymbol{W}}_0^{-1} \cdot \left[ \frac{1}{N} \nabla_{\gamma, \theta} \boldsymbol{G}_K(\gamma_0, \theta_0) \right] \\
&+ \left( \sqrt{N}(\check{\gamma} - \gamma_0)^\intercal, \sqrt{N}(\check{\theta} - \theta_0) \right) \cdot \left[ \frac{1}{N} \nabla_{\gamma, \theta} \boldsymbol{G}_K(\tilde{\gamma}, \tilde{\theta}) \right]^\intercal \cdot \widehat{\boldsymbol{W}}_0^{-1} \cdot \left[ \frac{1}{N} \nabla_{\gamma, \theta} \boldsymbol{G}_K(\tilde{\gamma}, \tilde{\theta}) \right] \\
&+ \left[ \frac{1}{N} \mathbf{G}_K(\tilde{\gamma}, \tilde{\theta}) \right]^\intercal \cdot \widehat{\boldsymbol{W}}_0^{-1} \cdot \left[ \frac{1}{N} \nabla_{\gamma, \theta}^2 \boldsymbol{G}_K(\tilde{\gamma}, \tilde{\theta}) \right] \left( \sqrt{N}(\check{\gamma} - \gamma_0)^\intercal, \sqrt{N} \cdot (\check{\theta} - \theta_0) \right)^\intercal , \quad (19)
\end{aligned}$$

where $(\tilde{\gamma}, \tilde{\theta})$ lies on the line joining $(\check{\gamma}, \check{\theta})$ and $(\gamma_0, \theta_0)$. Obviously, in order to show $\|(\check{\gamma}, \check{\theta}) - (\gamma_0, \theta_0)\| = O_p(N^{-1/2})$, it suffices to establish the following resutls:

$$\frac{1}{\sqrt{N}} \boldsymbol{G}_K^\intercal(\gamma_0, \theta_0) \cdot \widehat{\boldsymbol{W}}_0^{-1} \cdot \left[ \frac{1}{N} \nabla_{\gamma, \theta} \boldsymbol{G}_K(\gamma_0, \theta_0) \right] = O_p(1) , \tag{20}$$

$$\left\{ \left[ \frac{1}{N} \nabla_{\gamma, \theta} \boldsymbol{G}_K(\tilde{\gamma}, \tilde{\theta}) \right]^\intercal \cdot \widehat{\boldsymbol{W}}_0^{-1} \cdot \left[ \frac{1}{N} \nabla_{\gamma, \theta} \boldsymbol{G}_K(\tilde{\gamma}, \tilde{\theta}) \right] \right\}^{-1} = O_p(1) , \tag{21}$$

$$\left[ \frac{1}{N} \mathbf{G}_K(\tilde{\gamma}, \tilde{\theta}) \right]^\intercal \cdot \widehat{\boldsymbol{W}}_0^{-1} \cdot \left[ \frac{1}{N} \nabla_{\gamma, \theta}^2 \boldsymbol{G}_K(\tilde{\gamma}, \tilde{\theta}) \right] = o_p(1) . \tag{22}$$

We first prove (20). Using Chebyshev's inequality, Inequality (12), and Assumption 8, we can

9

obtain that

$$\left\| \frac{1}{N} \boldsymbol{G}_K(\gamma_0, \theta_0) \right\| = O_p\left( \sqrt{\frac{K}{N}} \right) = o_p(1) , \tag{23}$$

$$\left\| \widehat{\boldsymbol{W}}_0 - \mathbb{E}[\widehat{\boldsymbol{W}}_0] \right\| = O_p\left( \sqrt{\frac{K^3}{N}} \right) = o_p(1) , \tag{24}$$

$$\left\| \frac{1}{N} \nabla_{\gamma,\theta} \boldsymbol{G}_K(\gamma_0, \theta_0) - \boldsymbol{B}_{(K+1)\times(p+1)} \right\| = O_p\left( \sqrt{\frac{K}{N}} \right) = o_p(1) , \tag{25}$$

$$\left\| \frac{1}{N} \nabla^2_{\gamma,\theta} \boldsymbol{G}_K(\gamma_0, \theta_0) - \mathbb{E}\left[ \frac{1}{N} \nabla^2_{\gamma,\theta} \boldsymbol{G}_K(\gamma_0, \theta_0) \right] \right\| = O_p\left( \sqrt{\frac{K}{N}} \right) = o_p(1) . \tag{26}$$

Using (23) and (24), we can deduce that

$$\frac{1}{\sqrt{N}} \boldsymbol{G}_K^{\intercal}(\gamma_0, \theta_0) \cdot \widehat{\boldsymbol{W}}_0^{-1} \cdot \left[ \frac{1}{N} \nabla_{\gamma,\theta} \boldsymbol{G}_K(\gamma_0, \theta_0) \right] = \frac{1}{\sqrt{N}} \boldsymbol{G}_K^{\intercal}(\gamma_0, \theta_0) \cdot \mathbb{E}[\widehat{\boldsymbol{W}}_0]^{-1} \cdot \boldsymbol{B}_{(K+1)\times(p+1)} + o_p(1) .$$

Computing the variance of $N^{-1/2} \boldsymbol{G}_K^{\intercal}(\gamma_0, \theta_0) \cdot \mathbb{E}[\widehat{\boldsymbol{W}}_0]^{-1} \cdot \boldsymbol{B}_{(K+1)\times(p+1)}$ yields:

$$\left\| Var\left( \frac{1}{\sqrt{N}} \boldsymbol{G}_K^{\intercal}(\gamma_0, \theta_0) \cdot \mathbb{E}[\widehat{\boldsymbol{W}}_0]^{-1} \cdot \boldsymbol{B}_{(K+1)\times(p+1)} \right) \right\| \tag{27}$$

$$= \left\| \boldsymbol{B}_{(K+1)\times(p+1)}^{\intercal} \mathbb{E}[\widehat{\boldsymbol{W}}_0]^{-1} \mathbb{E}\left[ g_K(T, \boldsymbol{Z}; \gamma_0, \theta_0)^{\otimes 2} \right] \mathbb{E}[\widehat{\boldsymbol{W}}_0]^{-1} \boldsymbol{B}_{(K+1)\times(p+1)} \right\|$$

$$= \mathbb{E}\left[ \left\| \boldsymbol{B}_{(K+1)\times(p+1)}^{\intercal} \mathbb{E}[\widehat{\boldsymbol{W}}_0]^{-1} g_K(T, \boldsymbol{Z}; \gamma_0, \theta_0) \right\|^2 \right]$$

$$= \mathbb{E}\left[ \left\| \mathbb{E}\left[ \frac{\nabla_\gamma \pi(\boldsymbol{Z}; \gamma_0)}{\pi(\boldsymbol{Z}; \gamma_0)} u_K(\boldsymbol{X})^{\intercal} \right] \mathbb{E}\left[ u_K(\boldsymbol{X})^{\otimes 2} \right]^{-1} u_K(\boldsymbol{X}) \left\{ 1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} \right\} \right. \right.$$

$$\left. \left. + \mathbb{E}\left[ \frac{\nabla_\gamma \pi(\boldsymbol{Z}; \gamma)}{\pi(\boldsymbol{Z}; \gamma)} \right] \left\{ \theta_0 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} U(\boldsymbol{Z}) \right\} \right\|^2 \right]$$

$$+ \mathbb{E}\left[ \left| \theta_0 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} U(\boldsymbol{Z}) \right|^2 \right] .$$

Note that $\mathbb{E}\left[ \frac{\nabla_\gamma \pi(\boldsymbol{Z}; \gamma_0)}{\pi(\boldsymbol{Z}; \gamma_0)} u_K(\boldsymbol{X})^{\intercal} \right] \mathbb{E}\left[ u_K(\boldsymbol{X})^{\otimes 2} \right]^{-1} u_K(\boldsymbol{X})$ is the $L^2$-projection of $\mathbb{E}\left[ \frac{\nabla_\gamma \pi(\boldsymbol{Z}; \gamma_0)}{\pi(\boldsymbol{Z}; \gamma_0)} \middle| \boldsymbol{X} \right]$ on the space spanned by $u_K(\boldsymbol{X})$, by Lemma 2 in Appendix 2, we can have

$$\left\| Var\left( \frac{1}{\sqrt{N}} \boldsymbol{G}_K^{\intercal}(\gamma_0, \theta_0) \cdot \mathbb{E}[\widehat{\boldsymbol{W}}_0]^{-1} \cdot \boldsymbol{B}_{(K+1)\times(p+1)} \right) \right\|$$

$$= \mathbb{E}\left[ \left\| \mathbb{E}\left[ \frac{\nabla_\gamma \pi(\boldsymbol{Z}; \gamma_0)}{\pi(\boldsymbol{Z}; \gamma_0)} \middle| \boldsymbol{X} \right] \left\{ 1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} \right\} + \mathbb{E}\left[ \frac{\nabla_\gamma \pi(\boldsymbol{Z}; \gamma)}{\pi(\boldsymbol{Z}; \gamma)} \right] \left\{ \theta_0 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} U(\boldsymbol{Z}) \right\} \right\|^2 \right]$$

$$+ \mathbb{E}\left[\left\|\theta_0 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)}U(\boldsymbol{Z})\right\|^2\right] = O(1) \ ,$$

which implies our claim (20) by Chebyshev's inequality.

We next consider to prove (21). Note that

$$\left[\frac{1}{N}\nabla_{\gamma,\theta}\boldsymbol{G}_K(\gamma,\theta)\right]^{\top} \cdot \widehat{\boldsymbol{W}}_0^{-1} \cdot \left[\frac{1}{N}\nabla_{\gamma,\theta}\boldsymbol{G}_K(\gamma,\theta)\right] = \begin{pmatrix} \widehat{\boldsymbol{M}}_{p\times p}^{(1)}(\gamma), & \widehat{\boldsymbol{M}}_{p\times 1}^{(2)}(\gamma) \\ \left(\widehat{\boldsymbol{M}}_{p\times 1}^{(2)}(\gamma)\right)^{\top}, & 1 \end{pmatrix} \ ,$$

where

$$\widehat{\boldsymbol{M}}_{p\times p}^{(1)}(\gamma) := \frac{1}{N}\sum_{i=1}^{N}\widehat{\mathbb{E}}\left[\left.\frac{T_i}{\pi(\boldsymbol{Z}_i;\gamma)^2}\nabla_{\gamma}\pi(\boldsymbol{Z}_i;\gamma)\right|\boldsymbol{X}_i\right]^{\otimes 2} + \left\{\frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\pi(\boldsymbol{Z}_i;\gamma)^2}U(\boldsymbol{Z}_i)\nabla_{\gamma}\pi(\boldsymbol{Z}_i;\gamma)\right\}^{\otimes 2} \ ,$$

$$\widehat{\boldsymbol{M}}_{p\times 1}^{(2)}(\gamma) := \frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\pi(\boldsymbol{Z}_i;\gamma)^2}U(\boldsymbol{Z}_i)\nabla_{\gamma}\pi(\boldsymbol{Z}_i;\gamma) \ ,$$

and

$$\widehat{\mathbb{E}}\left[\left.\frac{T_i}{\pi(\boldsymbol{Z}_i;\gamma)^2}\nabla_{\gamma}\pi(\boldsymbol{Z}_i;\gamma)\right|\boldsymbol{X}\right]$$
$$:= \left\{\frac{1}{N}\sum_{i=1}^{N}\frac{T_i}{\pi(\boldsymbol{Z}_i;\gamma)^2}\nabla_{\gamma}\pi(\boldsymbol{Z}_i;\gamma)u_K(\boldsymbol{X}_i)\right\}\left[\frac{1}{N}\sum_{i=1}^{N}u_K(\boldsymbol{X}_i)u_K(\boldsymbol{X}_i)^{\top}\right]^{-1}u_K(\boldsymbol{X}) \ .$$

Similar to (15) and (16), we can obtain that

$$\sup_{\gamma\in\Gamma}\left|\widehat{\boldsymbol{M}}_{p\times p}^{(1)}(\gamma) - \boldsymbol{M}_{p\times p}^{(1)}(\gamma)\right| = o_p(1) \ , \tag{28}$$

$$\sup_{\gamma\in\Gamma}\left|\widehat{\boldsymbol{M}}_{p\times 1}^{(2)}(\gamma) - \boldsymbol{M}_{p\times 1}^{(2)}(\gamma)\right| = o_p(1) \ . \tag{29}$$

where

$$\boldsymbol{M}_{p\times p}^{(1)}(\gamma) := \mathbb{E}\left[\mathbb{E}\left[\left.\frac{T}{\pi(\boldsymbol{Z};\gamma)^2}\nabla_{\gamma}\pi(\boldsymbol{Z};\gamma)\right|\boldsymbol{X}\right]^{\otimes 2}\right] + \mathbb{E}\left[\frac{T}{\pi(\boldsymbol{Z};\gamma)^2}U(\boldsymbol{Z})\nabla_{\gamma}\pi(\boldsymbol{Z};\gamma)\right]^{\otimes 2} \ ,$$

$$\boldsymbol{M}_{p\times 1}^{(2)}(\gamma) := \mathbb{E}\left[\frac{T}{\pi(\boldsymbol{Z};\gamma)^2}U(\boldsymbol{Z})\nabla_{\gamma}\pi(\boldsymbol{Z};\gamma)\right] \ .$$

Combining (28), (29), and the fact that $\|\check\gamma - \gamma_0\| \xrightarrow{p} 0$, we can obtain that

$$
\left[\frac{1}{N}\nabla_{\gamma,\theta}\boldsymbol{G}_K(\tilde\gamma, \tilde\theta)\right]^{\mathsf{T}} \cdot \widehat{\boldsymbol{W}}_0^{-1} \cdot \left[\frac{1}{N}\nabla_{\gamma,\theta}\boldsymbol{G}_K(\tilde\gamma, \tilde\theta)\right] \xrightarrow{p} \begin{pmatrix} \boldsymbol{M}_{p\times p}^{(1)}(\gamma_0), & \boldsymbol{M}_{p\times 1}^{(2)}(\gamma_0) \\ \left(\boldsymbol{M}_{p\times 1}^{(2)}(\gamma_0)\right)^{\mathsf{T}}, & 1 \end{pmatrix} , \tag{30}
$$

where the limiting matrix is finite and non-degenerate. Therefore, we can justify the claim (21).

The proof of (22) is analogous to that of (21). Finally, by combining (20), (21), (22), and (19), we can conclude our desired result $\|(\tilde\gamma, \tilde\theta) - (\gamma_0, \theta_0)\| = O_p(N^{-1/2})$.

# 4   Proof of Theorem 2

The consistency result $\|(\bar\gamma, \bar\theta) - (\gamma_0, \theta_0)\| \xrightarrow{p} 0$ holds by using a similar argument for showing $\|(\tilde\gamma, \check\theta) - (\gamma_0, \theta_0)\| \xrightarrow{p} 0$ in Theorem 1.

We next show asymptotic normality for the infeasible estimator $(\bar\gamma, \bar\theta)$. Using the first order condition of optimization, we can obtain that

$$
\mathbf{G}_K(\bar\gamma, \bar\theta)^T \cdot \boldsymbol{D}_{(K+1)\times(K+1)}^{-1} \cdot \nabla_{\gamma,\theta}\mathbf{G}_K(\bar\gamma, \bar\theta) = \mathbf{0} ,
$$

then an application of Taylor's theorem yields:

$$
\begin{aligned}
\mathbf{0}_{1\times(p+1)} = &\frac{1}{\sqrt{N}}\boldsymbol{G}_K^{\mathsf{T}}(\gamma_0, \theta_0) \cdot \boldsymbol{D}_{(K+1)\times(K+1)}^{-1} \cdot \left[\frac{1}{N}\nabla_{\gamma,\theta}\boldsymbol{G}_K(\gamma_0, \theta_0)\right] \\
&+ \left(\sqrt{N}(\bar\gamma - \gamma_0)^{\mathsf{T}}, \sqrt{N}(\bar\theta - \theta_0)\right) \cdot \left[\frac{1}{N}\nabla_{\gamma,\theta}\boldsymbol{G}_K(\gamma^*, \theta^*)\right]^{\mathsf{T}} \cdot \boldsymbol{D}_{(K+1)\times(K+1)}^{-1} \cdot \left[\frac{1}{N}\nabla_{\gamma,\theta}\boldsymbol{G}_K(\gamma^*, \theta^*)\right] \\
&+ \left[\frac{1}{N}\mathbf{G}_K(\gamma^*, \theta^*)\right]^{\mathsf{T}} \cdot \boldsymbol{D}_{(K+1)\times(K+1)}^{-1} \cdot \left[\frac{1}{N}\nabla_{\gamma,\theta}^2\boldsymbol{G}_K(\gamma^*, \theta^*)\right] \left(\sqrt{N}(\bar\gamma - \gamma_0)^{\mathsf{T}}, \sqrt{N}\cdot(\bar\theta - \theta_0)\right)^{\mathsf{T}} .
\end{aligned} \tag{31}
$$

where $(\gamma^*, \theta^*)$ lies on the line joining $(\bar\gamma, \bar\theta)$ and $(\gamma_0, \theta_0)$. Note that the expression (31) has the same structure as (19), except for that the weighting matrix used in (31) is $\boldsymbol{D}_{(K+1)\times(K+1)}^{-1}$ while the weighting matrix used in (19) is $\widehat{\boldsymbol{W}}_0^{-1}$. Using a similar argument to show (30) and (22), we can obtain the following results:

$$
\left[\frac{1}{N}\nabla_{\gamma,\theta}\boldsymbol{G}_K(\gamma^*, \theta^*)\right]^{\mathsf{T}} \cdot \boldsymbol{D}_{(K+1)\times(K+1)}^{-1} \cdot \left[\frac{1}{N}\nabla_{\gamma,\theta}\boldsymbol{G}_K(\gamma^*, \theta^*)\right]
$$

$$=(\boldsymbol{B}_{(K+1)\times(p+1)})^{\mathsf{T}} \cdot \boldsymbol{D}^{-1}_{(K+1)\times(K+1)} \cdot \boldsymbol{B}_{(K+1)\times(p+1)} + o_p(1) \tag{32}$$

and

$$\left[\frac{1}{N}\mathbf{G}_K(\gamma^*, \theta^*)\right]^{\mathsf{T}} \cdot \boldsymbol{D}^{-1}_{(K+1)\times(K+1)} \cdot \left[\frac{1}{N}\nabla^2_{\gamma,\theta}\mathbf{G}_K(\gamma^*, \theta^*)\right] = o_p(1) . \tag{33}$$

Combining (25), (31), (32), and (33) together, we can deduce that

$$\begin{pmatrix} \sqrt{N}(\bar{\gamma} - \gamma_0) \\ \sqrt{N}(\bar{\theta} - \theta_0) \end{pmatrix} = - \left\{ (\boldsymbol{B}_{(K+1)\times(p+1)})^{\mathsf{T}} \cdot \boldsymbol{D}^{-1}_{(K+1)\times(K+1)} \cdot \boldsymbol{B}_{(K+1)\times(p+1)} \right\}^{-1}$$

$$\cdot (\boldsymbol{B}_{(K+1)\times(p+1)})^{\mathsf{T}} \boldsymbol{D}^{-1}_{(K+1)\times(K+1)} \cdot \frac{1}{\sqrt{N}}\boldsymbol{G}_K(\gamma_0, \theta_0) + o_p(1) .$$

Then

$$Cov\left(\begin{pmatrix} \sqrt{N}(\bar{\gamma} - \gamma_0) \\ \sqrt{N}(\bar{\theta} - \theta_0) \end{pmatrix}\right) = \left((\boldsymbol{B}_{(K+1)\times(p+1)})^{\mathsf{T}} \cdot \boldsymbol{D}^{-1}_{(K+1)\times(K+1)} \cdot \boldsymbol{B}_{(K+1)\times(p+1)}\right)^{-1} + o(1) = \boldsymbol{V}_K + o(1) .$$

Therefore, we can obtain that

$$\sqrt{N}\boldsymbol{V}_K^{-1/2}\begin{pmatrix} \bar{\gamma} - \gamma_0 \\ \bar{\theta} - \theta_0 \end{pmatrix} = - \left\{ \boldsymbol{B}^{\mathsf{T}}_{(K+1)\times(p+1)} \cdot \boldsymbol{D}^{-1}_{(K+1)\times(K+1)} \cdot \boldsymbol{B}_{(K+1)\times(p+1)} \right\}^{-1/2} \tag{34}$$

$$\cdot \boldsymbol{B}^{\mathsf{T}}_{(K+1)\times(p+1)} \boldsymbol{D}^{-1}_{(K+1)\times(K+1)} \cdot \frac{1}{\sqrt{N}}\boldsymbol{G}_K(\gamma_0, \theta_0) + o_p(1) .$$

We next show that the normalized estimator $\sqrt{N}\boldsymbol{V}_K^{-1/2}\begin{pmatrix} \bar{\gamma} - \gamma_0 \\ \bar{\theta} - \theta_0 \end{pmatrix}$ behaves asymptotically normal. The key is to verify the Lindeberg type conditions in Eicker (1966). Note that $\frac{1}{\sqrt{N}}\boldsymbol{G}_K(\gamma_0, \theta_0)$ can be written as the following matrix form:

$$\frac{1}{\sqrt{N}}\boldsymbol{G}_K(\gamma_0, \theta_0) = \begin{pmatrix} \frac{1}{\sqrt{N}}\sum_{i=1}^N \left[1 - \frac{T_i}{\pi(\boldsymbol{Z}_i;\gamma_0)}\right] u_K(\boldsymbol{X}_i) \\ \frac{1}{\sqrt{N}}\sum_{i=1}^N \left[\theta_0 - \frac{T_i}{\pi(\boldsymbol{Z}_i;\gamma_0)}U(\boldsymbol{Z}_i)\right] \end{pmatrix} = \boldsymbol{A}_{(K+1)\times N(K+1)} \cdot \mathcal{E}_{N(K+1)\times 1} ,$$

where

$$\boldsymbol{A}_{(K+1)\times N(K+1)} := \begin{pmatrix} \frac{1}{\sqrt{N}}\mathbf{1}_{1\times N}, & \mathbf{0}_{1\times N}, & \mathbf{0}_{1\times N}, & \cdots\mathbf{0}_{1\times N}, & \mathbf{0}_{1\times N} \\ \mathbf{0}_{1\times N}, & \frac{1}{\sqrt{N}}\mathbf{1}_{1\times N}, & \mathbf{0}_{1\times N}, & \cdots\mathbf{0}_{1\times N}, & \mathbf{0}_{1\times N} \\ \vdots & \vdots & \vdots & \vdots & \\ \mathbf{0}_{1\times N}, & \mathbf{0}_{1\times N} & \mathbf{0}_{1\times N}, & \frac{1}{\sqrt{N}}\mathbf{1}_{1\times N}, & \mathbf{0}_{1\times N} \\ \mathbf{0}_{1\times N}, & \mathbf{0}_{1\times N}, & \mathbf{0}_{1\times N}, & \cdots\mathbf{0}_{1\times N} & \frac{1}{\sqrt{N}}\mathbf{1}_{1\times N} \end{pmatrix} \text{ is a} (K+1)\times N(K+1) \text{matrix },$$

$$\mathcal{E}_{N(K+1)\times 1} := (\boldsymbol{v}_{1\times N}(1),\ldots,\boldsymbol{v}_{1\times N}(K),\boldsymbol{w}_{1\times N})^{\top},$$

$$\boldsymbol{v}_{1\times N}(k) := \left(\left[1-\frac{T_1}{\pi(\boldsymbol{Z}_1;\gamma_0)}\right]u_{kK}(\boldsymbol{X}_1),\ldots,\left[1-\frac{T_N}{\pi(\boldsymbol{Z}_N;\gamma_0)}\right]u_{kK}(\boldsymbol{X}_N)\right),\ k\in\{1,\ldots,K\},$$

$$\boldsymbol{w}_{1\times N} := \left(\theta_0-\frac{T_1}{\pi(\boldsymbol{Z}_1;\gamma_0)}U(\boldsymbol{Z}_1),\cdots,\theta_0-\frac{T_N}{\pi(\boldsymbol{Z}_N;\gamma_0)}U(\boldsymbol{Z}_N)\right)$$

and $\mathbf{1}_{1\times N}$ (resp. $\mathbf{0}_{1\times N}$) is a $N$-dimensional row vector whose elements are all of 1's (resp. 0's). From Eicker (1966), the following Lindeberg type conditions are sufficient to ensure $\sqrt{N}\boldsymbol{V}_K^{-1/2}(\bar{\gamma}-\gamma_0, \bar{\theta}-\theta_0)^{\top} \xrightarrow{d} N(0,I_{(p+1)\times(p+1)})$, namely,

1. $\displaystyle\max_{i\in\{1,\ldots,N(K+1)\}} \boldsymbol{a}_i^{\top}\left(\boldsymbol{A}_{(K+1)\times N(K+1)}\boldsymbol{A}_{(K+1)\times N(K+1)}^{\top}\right)^{-1}\boldsymbol{a}_i \to 0$, where $\boldsymbol{a}_i$ is the $i^{th}$ column of $\boldsymbol{A}_{(K+1)\times N(K+1)}$;

2. both $\displaystyle\sup_{k\in\{1,\ldots,K\}}\mathbb{E}\left[\left(1-\frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right)^2 u_{kK}(\boldsymbol{X})^2 I\left(\left|\left(1-\frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right)u_{kK}(\boldsymbol{X})\right|\geq s\right)\right] \to 0$ and

$$\mathbb{E}\left[\left(\theta_0-\frac{T}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})\right)^2 I\left(\left|\theta_0-\frac{T}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})\right|\geq s\right)\right] \to 0$$

as $s\to\infty$;

3. both $\displaystyle\inf_{k\in\{1,\ldots,K\}}\mathbb{E}\left[\left(1-\frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right)^2 u_{kK}(\boldsymbol{X})^2\right] > 0$ and $\mathbb{E}\left[\left(\theta_0-\frac{T}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})\right)^2\right] > 0$ .

Conditions 1 is natually satisfied by the definition of $\boldsymbol{A}_{(K+1)\times N(K+1)}$. Condition 2 holds because

$$\mathbb{E}\left[\left(1-\frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right)^2 u_{kK}(\boldsymbol{X})^2 I\left(\left|\left(1-\frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right)u_{kK}(\boldsymbol{X})\right|\geq s\right)\right]$$

$$\leq \left(1+\frac{1}{\underline{c}}\right)^2 \cdot \mathbb{E}\left[u_{kK}(\boldsymbol{X})^2 I\left(|u_{kK}(\boldsymbol{X})|\geq s\left(1+\frac{1}{\underline{c}}\right)^{-1}\right)\right] \xrightarrow{s\to\infty} 0 \,,$$

where $\underline{c} > 0$ is lower bound of the propensity socre $\pi(\boldsymbol{z};\gamma_0)$ (Assumption 7), and the last convergence holds from the fact $\displaystyle\sup_{k\in\{1,..,K\}}\mathbb{E}[u_{kK}^2(\boldsymbol{X})] \leq \bar{a}$ (see (9)) and Dominated Convergence Theorem;

the second part in Condition 2 also follows from Assumption 7 and Dominated Convergence Theorem. Condition 3 holds because

$$\mathbb{E}\left[\left(1 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right)^2 u_K(\boldsymbol{X})^{\otimes 2}\right] = \mathbb{E}\left[\frac{1 - \pi(\boldsymbol{Z})}{\pi(\boldsymbol{Z})} u_K(\boldsymbol{X})^{\otimes 2}\right] \geq \frac{1 - \bar{c}}{\bar{c}}\mathbb{E}\left[u_K(\boldsymbol{X})^{\otimes 2}\right] \geq \frac{(1 - \bar{c})a}{\bar{c}} \cdot I_{K \times K},$$

which implies that $\inf\limits_{k \in \{1,\ldots,K\}} \mathbb{E}\left[\left(1 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right)^2 u_{kK}(\boldsymbol{X})^2\right] > 0$; the second part in Condition 3 is obvious. Therefore, Conditions 1, 2, and 3 are all satisfied, and we can conclude our desired result that $\sqrt{N}\boldsymbol{V}_K^{-1/2}\left(\bar{\gamma} - \gamma_0, \bar{\theta} - \theta_0\right)^\top \xrightarrow{d} N\left(0, I_{(p+1)\times(p+1)}\right)$.

# 5  Proof of Theorem 3

From Theorem 2 we know have that $\sqrt{N}\boldsymbol{V}_K^{-1/2}\left(\bar{\gamma} - \gamma_0, \bar{\theta} - \theta_0\right)^\top \xrightarrow{d} N\left(0, I_{(p+1)\times(p+1)}\right)$, therefore to prove Theorem 3, it suffices to show $\boldsymbol{V}_K$ converges to the efficient variance bound of $(\gamma_0, \theta_0)$.

Since $\boldsymbol{V}_K = \left\{\boldsymbol{B}_{(K+1)\times(p+1)}^\top \cdot \boldsymbol{D}_{(K+1)\times(K+1)}^{-1} \cdot \boldsymbol{B}_{(K+1)\times(p+1)}\right\}^{-1}$, we first find the explicit expression of $\boldsymbol{V}_K$. Using the inverse matrix formula (11), we can have:

$$\mathbf{D}_{(K+1)\times(K+1)}^{-1} = \begin{pmatrix} \boldsymbol{A}_{K\times K}^{-1} + \frac{1}{c}\boldsymbol{A}_{K\times K}^{-1}\boldsymbol{b}_K\boldsymbol{b}_K^\top\boldsymbol{A}_{K\times K}^{-1}, & -\frac{1}{c}\boldsymbol{A}_{K\times K}^{-1}\boldsymbol{b}_K \\ -\frac{1}{c}\mathbf{b}_K^T\boldsymbol{A}_{K\times K}^{-1}, & \frac{1}{c} \end{pmatrix}, \tag{35}$$

where

$$\boldsymbol{A}_{K\times K} := \mathbb{E}\left[\frac{1 - \pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^{\otimes 2}\right], \quad \boldsymbol{b}_K := \mathbb{E}\left[\frac{1 - \pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})u_K(\boldsymbol{X})\right], \tag{36}$$

$$c := \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})\right)^2\right] - \boldsymbol{b}_K^T\boldsymbol{A}_{K\times K}^{-1}\boldsymbol{b}_K. \tag{37}$$

Then we have

$$\boldsymbol{B}_{(K+1)\times(p+1)}^\top \cdot \boldsymbol{D}_{(K+1)\times(K+1)}^{-1} \cdot \boldsymbol{B}_{(K+1)\times(p+1)} = \begin{pmatrix} \tilde{\boldsymbol{A}}_{p\times p}, & \tilde{\boldsymbol{b}}_p \\ \tilde{\boldsymbol{b}}_p^\top & \frac{1}{c} \end{pmatrix},$$

where

$$\tilde{\boldsymbol{A}}_{p\times p} := \mathbb{E}\left[\frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^T\right]\cdot \boldsymbol{A}_{K\times K}^{-1}\cdot \mathbb{E}\left[u_K(\boldsymbol{X})\frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)^T}{\pi(\boldsymbol{Z};\gamma_0)}\right] + c\cdot \tilde{\boldsymbol{b}}_p\tilde{\boldsymbol{b}}_p^T , \qquad (38)$$

$$\tilde{\boldsymbol{b}}_p := -\frac{1}{c}\mathbb{E}\left[\frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^T\right]\cdot \boldsymbol{A}_{K\times K}^{-1}\cdot \boldsymbol{b}_K + \frac{1}{c}\mathbb{E}\left[\frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})\right] . \qquad (39)$$

Using the matrix inversion formula (11) again, we can obtain that

$$\boldsymbol{V}_K = \left(\boldsymbol{B}_{(K+1)\times(p+1)}^{\mathsf{T}}\cdot \boldsymbol{D}_{(K+1)\times(K+1)}^{-1}\cdot \boldsymbol{B}_{(K+1)\times(p+1)}\right)^{-1}$$

$$= \begin{pmatrix} \tilde{\boldsymbol{A}}_{p\times p}^{-1} + \frac{1}{\tilde{c}}\tilde{\boldsymbol{A}}_{p\times p}^{-1}\tilde{\boldsymbol{b}}_p\tilde{\boldsymbol{b}}_p^T\tilde{\boldsymbol{A}}_{p\times p}^{-1}, & -\frac{1}{\tilde{c}}\tilde{\boldsymbol{A}}_{p\times p}^{-1}\tilde{\boldsymbol{b}}_p \\ -\frac{1}{\tilde{c}}\tilde{\boldsymbol{b}}_p^T\tilde{\boldsymbol{A}}_{p\times p}^{-1} & \frac{1}{\tilde{c}} \end{pmatrix} ,$$

where

$$\tilde{c} := \frac{1}{c} - \tilde{\boldsymbol{b}}_p^T\tilde{\boldsymbol{A}}_{p\times p}^{-1}\tilde{\boldsymbol{b}}_p .$$

Then we can obtain that

$$Var(\sqrt{N}(\bar{\gamma} - \gamma_0)) = \tilde{\boldsymbol{A}}_{p\times p}^{-1} + \frac{1}{\tilde{c}}\tilde{\boldsymbol{A}}_{p\times p}^{-1}\tilde{\boldsymbol{b}}_p\tilde{\boldsymbol{b}}_p^T\tilde{\boldsymbol{A}}_{p\times p}^{-1} + o(1), \qquad (40)$$

$$Var(\sqrt{N}(\bar{\theta} - \theta_0)) = \frac{1}{\tilde{c}} + o(1) . \qquad (41)$$

In order to establish our Theorem 3, it suffices to show that (40) and (41) converge to $V_{\gamma_0}$ and $V_{\theta_0}$ respectively. The proof is constituted of two parts:

**Proof of Part (I)**

**We first show that (40) converges to $V_{\gamma_0}$ as $N$ goes to infinity**. Note that

$$Var(\sqrt{N}(\bar{\gamma} - \gamma_0)) = \tilde{\boldsymbol{A}}_{p\times p}^{-1} + \frac{1}{\tilde{c}}\tilde{\boldsymbol{A}}_{p\times p}^{-1}\tilde{\boldsymbol{b}}_p\tilde{\boldsymbol{b}}_p^T\tilde{\boldsymbol{A}}_{p\times p}^{-1} + o(1)$$

$$= \left[\tilde{\boldsymbol{A}}_{p\times p} - c\cdot \tilde{\boldsymbol{b}}_p\tilde{\boldsymbol{b}}_p^T\right]^{-1} + o(1)$$

$$= \left\{\mathbb{E}\left[\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^T\right]\cdot \boldsymbol{A}_{K\times K}^{-1}\cdot \mathbb{E}\left[u_K(\boldsymbol{X})\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)^T}{\pi(\boldsymbol{Z};\gamma_0)}\right]\right\}^{-1} + o(1) \quad \text{(using (38))} ,$$

$$(42)$$

16

where the second equality can be straightforwardly verified as follows:

$$\left( \tilde{A}_{p\times p}^{-1} + \frac{1}{\tilde{c}} \tilde{A}_{p\times p}^{-1} \tilde{b}_p \tilde{b}_p^T \tilde{A}_{p\times p}^{-1} \right) \cdot \left( \tilde{A}_{p\times p} - c \cdot \tilde{b}_p \tilde{b}_p^T \right)$$

$$=I_{p\times p} - c \cdot \tilde{A}_{p\times p}^{-1} \tilde{b}_p \tilde{b}_p^T + \frac{1}{\tilde{c}} \tilde{A}_{p\times p}^{-1} \tilde{b}_p \tilde{b}_p^T - \frac{c}{\tilde{c}} \tilde{A}_{p\times p}^{-1} \tilde{b}_p \tilde{b}_p^T \tilde{A}_{p\times p}^{-1} \tilde{b}_p \tilde{b}_p^T$$

$$=I_{p\times p} - \frac{1}{\tilde{c}} \cdot \left[ \tilde{c}c \cdot \tilde{A}_{p\times p}^{-1} \tilde{b}_p \tilde{b}_p^T + c \cdot \tilde{A}_{p\times p}^{-1} \tilde{b}_p \tilde{b}_p^T \cdot \tilde{b}_p^T \tilde{A}_{p\times p}^{-1} \tilde{b}_p \right] + \frac{1}{\tilde{c}} \tilde{A}_{p\times p}^{-1} \tilde{b}_p \tilde{b}_p^T$$

$$=I_{p\times p} - \frac{1}{\tilde{c}} \cdot \left[ \tilde{A}_{p\times p}^{-1} \tilde{b}_p \tilde{b}_p^T \right] + \frac{1}{\tilde{c}} \tilde{A}_{p\times p}^{-1} \tilde{b}_p \tilde{b}_p^T \qquad \left( \text{since } c\tilde{c} = 1 - c \cdot \tilde{b}_p^T \tilde{A}_{p\times p}^{-1} \tilde{b}_p \right)$$

$$=I_{p\times p} \ .$$

Therefore,

$$Var(\sqrt{N}(\bar{\gamma} - \gamma_0)) = \left\{ \mathbb{E}\left[ \frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)} u_K(\boldsymbol{X})^T \right] \cdot \boldsymbol{A}_{K\times K}^{-1} \cdot \mathbb{E}\left[ u_K(\boldsymbol{X}) \frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)^T}{\pi(\boldsymbol{Z};\gamma_0)} \right] \right\}^{-1}$$

$$=\mathbb{E}\left[ f_K(\boldsymbol{Z}) f_K(\boldsymbol{Z})^\mathsf{T} \right]^{-1} + o(1) \ , \tag{43}$$

where

$$f_K(\boldsymbol{Z}) := - \mathbb{E}\left[ \frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)} u_K(\boldsymbol{X})^T \right] \cdot \mathbb{E}\left[ \frac{1 - \pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)} u_K(\boldsymbol{X})^{\otimes 2} \right]^{-1} u_K(\boldsymbol{X}) \cdot \left( 1 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)} \right) \ .$$

Recalling the definitions of $\boldsymbol{S}_0(\boldsymbol{Z};\gamma_0)$ and $O(\boldsymbol{Z})$ in (1), we can obtain that

$$f_K(\boldsymbol{Z}) = \mathbb{E}\left[ \boldsymbol{S}_0(\boldsymbol{Z};\gamma_0) O(\boldsymbol{Z}) u_K(\boldsymbol{X})^T \right] \cdot \mathbb{E}\left[ O(\boldsymbol{Z}) u_K(\boldsymbol{X})^{\otimes 2} \right]^{-1} u_K(\boldsymbol{X}) \left( 1 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)} \right)$$

$$= \left\{ \mathbb{E}\left[ \frac{\mathbb{E}[\boldsymbol{S}_0(\boldsymbol{Z};\gamma_0) O(\boldsymbol{Z})|\boldsymbol{X}]}{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}} \sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]} u_K(\boldsymbol{X})^T \right] \cdot \mathbb{E}\left[ \left( \sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]} u_K(\boldsymbol{X}) \right)^{\otimes 2} \right]^{-1} \right.$$

$$\left. \cdot \sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]} u_K(\boldsymbol{X}) \right\} \frac{1}{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}} \left( 1 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)} \right)$$

$$= h_K(\boldsymbol{X}) \frac{1}{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}} \left( 1 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)} \right) \ ,$$

where

$$h_K(\boldsymbol{X}) := \mathbb{E}\left[ \frac{\mathbb{E}[\boldsymbol{S}_0(\boldsymbol{Z};\gamma_0) O(\boldsymbol{Z})|\boldsymbol{X}]}{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}} \sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]} u_K(\boldsymbol{X})^T \right] \cdot \mathbb{E}\left[ \left( \sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]} u_K(\boldsymbol{X}) \right)^{\otimes 2} \right]^{-1} \cdot \sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]} u_K(\boldsymbol{X}) \ .$$

Note that $h_K(\boldsymbol{X})$ is the least square projection of $\frac{\mathbb{E}[\boldsymbol{S}_0(\boldsymbol{Z};\gamma_0) O(\boldsymbol{Z})|\boldsymbol{X}]}{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}}$ on the space linearly spanned

by $\{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}u_K(\boldsymbol{X})\}$, by Assumption 3 and Lemma 2 in Appendix 2, we can have

$$\mathbb{E}\left[\left\|h_K(\boldsymbol{X}) - \frac{\mathbb{E}[\boldsymbol{S}_0(\boldsymbol{Z};\gamma_0)O(\boldsymbol{Z})|\boldsymbol{X}]}{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}}\right\|^2\right] = O\left(K^{-\frac{2s}{r}}\right) = o(1) \ . \tag{44}$$

Therefore, we can have that

$$\begin{aligned}
\mathbb{E}\left[f_K(\boldsymbol{Z})f_K(\boldsymbol{Z})^\top\right] =& \mathbb{E}\left[h_K(\boldsymbol{X})h_K(\boldsymbol{X})^\top \frac{1}{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}\left(1 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right)^2\right] \\
=& \mathbb{E}\left[h_K(\boldsymbol{X})h_K(\boldsymbol{X})^\top\right] \\
\rightarrow& \mathbb{E}\left[\frac{\mathbb{E}[\boldsymbol{S}_0(\boldsymbol{Z};\gamma_0)O(\boldsymbol{Z})|\boldsymbol{X}] \cdot \mathbb{E}[\boldsymbol{S}_0(\boldsymbol{Z};\gamma_0)O(\boldsymbol{Z})|\boldsymbol{X}]^\top}{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}\right] \\
=& \mathbb{E}\left[m(\boldsymbol{X})m(\boldsymbol{X})^\top\left(1 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right)^2\right] \ ,
\end{aligned}$$

and then in light of (43) we can obtain that

$$Var(\sqrt{N}(\bar{\gamma} - \gamma_0)) \rightarrow \mathbb{E}\left[\left(1 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right)^2 \cdot m(\boldsymbol{X})^{\otimes 2}\right]^{-1} \ .$$

Therefore, $\sqrt{N}(\bar{\gamma} - \gamma_0)$ attains the efficiency variance bound (6).

**Proof of Part (II)**

**Next, we show that** (41) **converges to** $V_{\theta_0}$ **as** $N$ **goes to infinity.** Applying the matrix inverse formula (10) to (41), we have

$$\begin{aligned}
Var(\sqrt{N}(\bar{\theta} - \theta_0)) =& \left[\frac{1}{c} - \tilde{\boldsymbol{b}}_p^T \tilde{\boldsymbol{A}}_{p\times p}^{-1} \tilde{\boldsymbol{b}}_p\right]^{-1} + o(1) \\
=& c - c \cdot \tilde{\boldsymbol{b}}_p^T \cdot \left(-\tilde{\boldsymbol{A}}_{p\times p} + c \cdot \tilde{\boldsymbol{b}}_p\tilde{\boldsymbol{b}}_p^T\right)^{-1} \tilde{\boldsymbol{b}}_p \cdot c + o(1) \qquad \text{(using the matrix inverse formula (10))} \\
=& c + \left(\mathbb{E}\left[\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^T\right] \cdot \boldsymbol{A}_{K\times K}^{-1} \cdot \boldsymbol{b}_K - \mathbb{E}\left[\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})\right]\right)^T + o(1) \\
& \cdot \left\{\mathbb{E}\left[\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^T\right] \cdot \boldsymbol{A}_{K\times K}^{-1} \cdot \mathbb{E}\left[u_K(\boldsymbol{X})\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)^T}{\pi(\boldsymbol{Z};\gamma_0)}\right]\right\}^{-1} + o(1) \\
& \cdot \left(\mathbb{E}\left[\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^T\right] \cdot \boldsymbol{A}_{K\times K}^{-1} \cdot \boldsymbol{b}_K - \mathbb{E}\left[\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})\right]\right) + o(1) \ , \tag{45}
\end{aligned}$$

18

where the last equality follows from the definitions of $\tilde{A}_{p \times p}$ and $\tilde{b}_p$ in (38) and (39), namely,

$$\tilde{A}_{p \times p} - c \cdot \tilde{b}_p \tilde{b}_p^T = \mathbb{E}\left[\frac{\nabla \pi(Z; \gamma_0)}{\pi(Z; \gamma_0)} u_K(X)^T\right] \cdot A_{K \times K}^{-1} \cdot \mathbb{E}\left[u_K(X) \frac{\nabla \pi(Z; \gamma_0)^T}{\pi(Z; \gamma_0)}\right] ,$$

$$\tilde{b}_p \cdot c = -\mathbb{E}\left[\frac{\nabla \pi(Z; \gamma_0)}{\pi(Z; \gamma_0)} u_K(X)^T\right] \cdot A_{K \times K}^{-1} \cdot b_K + \mathbb{E}\left[\frac{\nabla \pi(Z; \gamma_0)}{\pi(Z; \gamma_0)} U(Z)\right] .$$

In the following we show that

$$c = \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(Z)} U(Z)\right)^2\right] - \mathbb{E}\left[\left\{\left(1 - \frac{T}{\pi(Z; \gamma_0)}\right) R(X)\right\}^2\right] + o(1) ; \tag{46}$$

$$\mathbb{E}\left[\frac{\nabla \pi(Z; \gamma_0)}{\pi(Z; \gamma_0)} u_K(X)^T\right] \cdot A_{K \times K}^{-1} \cdot b_K - \mathbb{E}\left[\frac{\nabla \pi(Z; \gamma_0)}{\pi(Z; \gamma_0)} U(Z)\right] = \mathbb{E}\left[\frac{\nabla \pi(Z; \gamma_0)}{\pi(Z; \gamma_0)} (R(X) - U(Z))\right] + o(1) ; \tag{47}$$

$$\left\{\mathbb{E}\left[\frac{\nabla \pi(Z; \gamma_0)}{\pi(Z; \gamma_0)} u_K(X)^T\right] \cdot A_{K \times K}^{-1} \cdot \mathbb{E}\left[u_K(X) \frac{\nabla \pi(Z; \gamma_0)^T}{\pi(Z; \gamma_0)}\right]\right\}^{-1} = V_{\gamma_0} + o(1) . \tag{48}$$

where $c$ is defined in (37).

**For the term** (46): Note that

$$
\begin{aligned}
c =& \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(Z; \gamma_0)} U(Z)\right)^2\right] \\
& - \mathbb{E}\left[O(Z)U(Z)u_K(X)^T\right] \cdot \mathbb{E}\left[O(Z)u_K(X)^{\otimes 2}\right]^{-1} \cdot \mathbb{E}\left[O(Z)U(Z)u_K(X)\right] \quad \text{(by definition (37))} \\
=& \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(Z; \gamma_0)} U(Z)\right)^2\right] \\
& - \mathbb{E}\left[\left\{\mathbb{E}\left[O(Z)U(Z)u_K(X)^T\right] \cdot \mathbb{E}\left[O(Z)u_K(X)^{\otimes 2}\right]^{-1} \cdot u_K(X)\left(1 - \frac{T}{\pi(Z; \gamma_0)}\right)\right\}^2\right] \\
=& \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(Z; \gamma_0)} U(Z)\right)^2\right] \\
& - \mathbb{E}\left[\left\{\mathbb{E}\left[\mathbb{E}\left[O(Z)U(Z)|X\right] \cdot u_K(X)^T\right] \cdot \mathbb{E}\left[\mathbb{E}[O(Z)|X] \cdot u_K(X)^{\otimes 2}\right]^{-1} \cdot u_K(X)\left(1 - \frac{T}{\pi(Z; \gamma_0)}\right)\right\}^2\right] \\
=& \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(Z; \gamma_0)} U(Z)\right)^2\right] \\
& - \mathbb{E}\left[\left\{\mathbb{E}\left[\frac{\mathbb{E}\left[O(Z)U(Z)|X\right]}{\sqrt{\mathbb{E}[O(Z)|X]}} \cdot \sqrt{\mathbb{E}[O(Z)|X]} u_K(X)^T\right] \cdot \mathbb{E}\left[\mathbb{E}[O(Z)|X] \cdot u_K(X)^{\otimes 2}\right]^{-1}\right.\right.
\end{aligned}
$$

19

$$\times \sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}u_K(\boldsymbol{X}) \cdot \frac{1}{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}}\left(1 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right)\bigg\}^2\bigg] .$$

Considering the last term in above expression, we note that

$$\mathbb{E}\left[\frac{\mathbb{E}\left[O(\boldsymbol{Z})U(\boldsymbol{Z})|\boldsymbol{X}\right]}{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}} \cdot \sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}u_K(\boldsymbol{X})^T\right] \cdot \mathbb{E}\left[\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}] \cdot u_K(\boldsymbol{X})^{\otimes 2}\right]^{-1} \cdot \sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}u_K(\boldsymbol{X})$$

is the $L^2$-projection of $\frac{\mathbb{E}[O(\boldsymbol{Z})U(\boldsymbol{Z})|\boldsymbol{X}]}{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}}$ on the space linearly spanned by $\{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}u_K(\boldsymbol{X})\}$, by Assumption 3 and Lemma 2 in Appendix 2, we can have

$$\mathbb{E}\left[\left\|\mathbb{E}\left[\frac{\mathbb{E}\left[O(\boldsymbol{Z})U(\boldsymbol{Z})|\boldsymbol{X}\right]}{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}} \cdot \sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}u_K(\boldsymbol{X})^T\right] \cdot \mathbb{E}\left[\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}] \cdot u_K(\boldsymbol{X})^{\otimes 2}\right]^{-1}\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}u_K(\boldsymbol{X})\right.\right.$$
$$\left.\left.- \frac{\mathbb{E}\left[O(\boldsymbol{Z})U(\boldsymbol{Z})|\boldsymbol{X}\right]}{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}}\right\|^2\right] = O(K^{-\frac{2s}{r}}) = o(1) . \tag{49}$$

Then we can have that

$$c \to \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})\right)^2\right] - \mathbb{E}\left[\left\{\frac{\mathbb{E}\left[O(\boldsymbol{Z})U(\boldsymbol{Z})|\boldsymbol{X}\right]}{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}} \cdot \frac{1}{\sqrt{\mathbb{E}[O(\boldsymbol{Z})|\boldsymbol{X}]}}\left(1 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right)\right\}^2\right]$$
$$= \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})\right)^2\right] - \mathbb{E}\left[\left\{\left(1 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right)R(\boldsymbol{X})\right\}^2\right] ,$$

which is (46).

**For the term** (47). Next, we consider the term

$$\mathbb{E}\left[\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^T\right] \cdot \boldsymbol{A}_{K\times K}^{-1} \cdot \boldsymbol{b}_K - \mathbb{E}\left[\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})\right] .$$

Using (49), we can have

$$\mathbb{E}\left[\frac{\nabla_\gamma\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^\top\right] \cdot \boldsymbol{A}_{K\times K}^{-1} \cdot \boldsymbol{b}_K$$
$$= \mathbb{E}\left[O(\boldsymbol{Z})U(\boldsymbol{Z})u_K(\boldsymbol{X})^\top\right] \cdot \mathbb{E}\left[O(\boldsymbol{Z})u_K(\boldsymbol{X})^{\otimes 2}\right]^{-1} \cdot \mathbb{E}\left[u_K(\boldsymbol{X})\frac{\nabla_\gamma\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}\right]$$
$$= \mathbb{E}\left[R(\boldsymbol{X})\frac{\nabla_\gamma\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}\right] + o(1) \qquad \text{(using (49))} .$$

Therefore, we can obtain that

$$\mathbb{E}\left[\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^T\right]\cdot\boldsymbol{A}_{K\times K}^{-1}\cdot\boldsymbol{b}_K - \mathbb{E}\left[\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})\right] \to \mathbb{E}\left[\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}(R(\boldsymbol{X})-U(\boldsymbol{Z}))\right],$$

which justifies our claim (47).

**For the term** (48). From Theorem 1 in Morikawa and Kim (2016), the efficient influence function of $\gamma_0$ is

$$\boldsymbol{\psi}_{eff}(T,\boldsymbol{Z};\gamma_0) = \mathbb{E}\left[\boldsymbol{S}_1(T,\boldsymbol{Z};\gamma_0)^{\otimes 2}\right]^{-1}\cdot\boldsymbol{S}_1(T,\boldsymbol{Z};\gamma_0) = -\mathbb{E}\left[\frac{\partial}{\partial\gamma}\boldsymbol{S}_1(T,\boldsymbol{Z};\gamma_0)\right]^{-1}\cdot\boldsymbol{S}_1(T,\boldsymbol{Z};\gamma_0),$$

thus the efficient variance bound of $\gamma_0$ is

$$\begin{aligned}
\boldsymbol{V}_{\gamma_0} =& Var(\boldsymbol{\psi}_{eff}(T,\boldsymbol{Z};\gamma_0)) \\
=& \mathbb{E}\left[\frac{1-\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}m(\boldsymbol{X})^{\otimes 2}\right]^{-1} \quad (50) \\
=& \mathbb{E}\left[\frac{m(\boldsymbol{X})}{\pi(\boldsymbol{Z};\gamma_0)}\nabla_\gamma\pi(\boldsymbol{Z};\gamma_0)^{\intercal}\right]^{-1}\mathbb{E}\left[\frac{1-\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}m(\boldsymbol{X})^{\otimes 2}\right]\left\{\mathbb{E}\left[\frac{m(\boldsymbol{X})}{\pi(\boldsymbol{Z};\gamma_0)}\nabla_\gamma\pi(\boldsymbol{Z};\gamma_0)^{\intercal}\right]^{-1}\right\}^{\intercal}. \quad (51)
\end{aligned}$$

Using (42), the fact that $Var(\sqrt{N}(\tilde{\gamma}-\gamma_0)) \to \boldsymbol{V}_{\gamma_0}$, and (51), we can obtain that

$$\begin{aligned}
\text{L.H.S. of (48)} =& \left\{\mathbb{E}\left[\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^T\right]\cdot\boldsymbol{A}_{K\times K}^{-1}\cdot\mathbb{E}\left[u_K(\boldsymbol{X})\frac{\nabla\pi(\boldsymbol{Z};\gamma_0)^T}{\pi(\boldsymbol{Z};\gamma_0)}\right]\right\}^{-1} \\
=& Var(\sqrt{N}(\tilde{\gamma}-\gamma_0)) \\
\to& \boldsymbol{V}_{\gamma_0} = \mathbb{E}\left[\frac{m(\boldsymbol{X})}{\pi(\boldsymbol{Z};\gamma_0)}\nabla_\gamma\pi(\boldsymbol{Z};\gamma_0)^{\intercal}\right]^{-1}\mathbb{E}\left[\boldsymbol{S}_1(T,\boldsymbol{Z};\gamma_0)^{\otimes 2}\right]\left(\mathbb{E}\left[\frac{m(\boldsymbol{X})}{\pi(\boldsymbol{Z};\gamma_0)}\nabla_\gamma\pi(\boldsymbol{Z};\gamma_0)^{\intercal}\right]^{-1}\right)^{\intercal},
\end{aligned}$$

which justifies our claim (48).

Combining (45), (46), (47), (48) and the definition of $\boldsymbol{\kappa}$ in (5), we can obtain that

$$\begin{aligned}
& Var(\sqrt{N}(\bar{\theta}-\theta_0)) \\
\to& \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})\right)^2\right] - \mathbb{E}\left[\left\{\left(1 - \frac{T}{\pi(\boldsymbol{Z};\gamma_0)}\right)R(\boldsymbol{X})\right\}^2\right] + \boldsymbol{\kappa}^{\intercal}\mathbb{E}[\boldsymbol{S}_1(T,\boldsymbol{Z};\gamma_0)^{\otimes 2}]\boldsymbol{\kappa}.
\end{aligned}$$

(52)

21

To complete the proof, we remain to show

$$(52) = Var(S_2(T, \boldsymbol{Z}; \gamma_0, \theta_0) - \boldsymbol{\kappa}^T \boldsymbol{S}_1(T, \boldsymbol{Z}; \gamma_0)) ,$$

which can be easily verified in the rest of the proof. Note that

$$
\begin{aligned}
&Var\left(S_2(T, \boldsymbol{Z}; \gamma_0, \theta_0) - \boldsymbol{\kappa}^T \boldsymbol{S}_1(T, \boldsymbol{Z}; \gamma_0)\right) \\
&= \mathbb{E}\left[S_2(T, \boldsymbol{Z}; \gamma_0, \theta_0)^2\right] - 2\boldsymbol{\kappa}^T \mathbb{E}\left[\boldsymbol{S}_1(T, \boldsymbol{Z}; \gamma_0) S_2(T, \boldsymbol{Z}; \gamma_0, \theta_0)\right] + \boldsymbol{\kappa}^T \mathbb{E}\left[\boldsymbol{S}_1(T, \boldsymbol{Z}; \gamma_0)^{\otimes 2}\right] \boldsymbol{\kappa} .
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\mathbb{E}\left[S_2(T, \boldsymbol{Z}; \gamma_0, \theta_0)^2\right] \\
&= \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} U(\boldsymbol{Z})\right)^2\right] - 2 \cdot \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} U(\boldsymbol{Z})\right)\left(1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)}\right) R(\boldsymbol{X})\right] \\
&\quad + \mathbb{E}\left[\left(1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)}\right)^2 R(\boldsymbol{X})^2\right] \\
&= \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} U(\boldsymbol{Z})\right)^2\right] + 2 \cdot \mathbb{E}\left[\frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} U(\boldsymbol{Z})\left(1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)}\right) R(\boldsymbol{X})\right] \\
&\quad + \mathbb{E}\left[\frac{1 - \pi(\boldsymbol{Z}; \gamma_0)}{\pi(\boldsymbol{Z}; \gamma_0)} R(\boldsymbol{X})^2\right] \\
&= \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} U(\boldsymbol{Z})\right)^2\right] - 2 \cdot \mathbb{E}\left[U(\boldsymbol{Z})\left(\frac{1 - \pi(\boldsymbol{Z}; \gamma_0)}{\pi(\boldsymbol{Z}; \gamma_0)}\right) R(\boldsymbol{X})\right] + \mathbb{E}\left[\frac{1 - \pi(\boldsymbol{Z}; \gamma_0)}{\pi(\boldsymbol{Z}; \gamma_0)} R(\boldsymbol{X})^2\right] \\
&= \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} U(\boldsymbol{Z})\right)^2\right] - 2 \cdot \mathbb{E}\left[\left(\frac{1 - \pi(\boldsymbol{Z}; \gamma_0)}{\pi(\boldsymbol{Z}; \gamma_0)}\right) R(\boldsymbol{X})^2\right] + \mathbb{E}\left[\frac{1 - \pi(\boldsymbol{Z}; \gamma_0)}{\pi(\boldsymbol{Z}; \gamma_0)} R(\boldsymbol{X})^2\right] \\
&= \mathbb{E}\left[\left(\theta_0 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} U(\boldsymbol{Z})\right)^2\right] - \mathbb{E}\left[\left(\frac{1 - \pi(\boldsymbol{Z}; \gamma_0)}{\pi(\boldsymbol{Z}; \gamma_0)}\right) R(\boldsymbol{X})^2\right]
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}\left[\boldsymbol{S}_1(T, \boldsymbol{Z}; \gamma_0) S_2(T, \boldsymbol{Z}; \gamma_0, \theta_0)\right] \\
&= -\mathbb{E}\left[\left(1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)}\right) m(\boldsymbol{X}) \cdot \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)} U(\boldsymbol{Z})\right] - \mathbb{E}\left[\left(1 - \frac{T}{\pi(\boldsymbol{Z}; \gamma_0)}\right)^2 m(\boldsymbol{X}) R(\boldsymbol{X})\right] \\
&= \mathbb{E}\left[\frac{1 - \pi(\boldsymbol{Z}; \gamma_0)}{\pi(\boldsymbol{Z}; \gamma_0)} m(\boldsymbol{X}) \cdot U(\boldsymbol{Z})\right] - \mathbb{E}\left[\frac{1 - \pi(\boldsymbol{Z}; \gamma_0)}{\pi(\boldsymbol{Z}; \gamma_0)} m(\boldsymbol{X}) \cdot R(\boldsymbol{X})\right]
\end{aligned}
$$

$$
\begin{aligned}
=&\mathbb{E}\left[\frac{1-\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}m(\boldsymbol{X})\cdot U(\boldsymbol{Z})\right] - \mathbb{E}\left[\mathbb{E}\left[\frac{1-\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}\bigg|\boldsymbol{X}\right]m(\boldsymbol{X})\cdot R(\boldsymbol{X})\right]\\
=&\mathbb{E}\left[\frac{1-\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}m(\boldsymbol{X})\cdot U(\boldsymbol{Z})\right] - \mathbb{E}\left[m(\boldsymbol{X})\cdot\mathbb{E}[O(\boldsymbol{Z})U(\boldsymbol{Z})|\boldsymbol{X}]\right]\\
=&\mathbb{E}\left[\frac{1-\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}m(\boldsymbol{X})\cdot U(\boldsymbol{Z})\right] - \mathbb{E}\left[\frac{1-\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}m(\boldsymbol{X})\cdot U(\boldsymbol{Z})\right]\\
=&\boldsymbol{0}\ .
\end{aligned}
$$

Thus, we can conclude $(52) = Var(S_2(T,\boldsymbol{Z};\gamma_0,\theta_0) - \boldsymbol{\kappa}^T\boldsymbol{S}_1(T,\boldsymbol{Z};\gamma_0))$. The proof is completed.

# 6 Proof of Theorem 4

From the equation (31), in order to prove $\|\sqrt{N}(\hat{\gamma}-\bar{\gamma},\hat{\theta}-\bar{\theta})\| \xrightarrow{p} 0$, it is sufficient to show

$$
\widehat{\boldsymbol{D}}_{(K+1)\times(K+1)} \xrightarrow{p} \boldsymbol{D}_{(K+1)\times(K+1)}\ . \tag{53}
$$

Note that

$$
\hat{\boldsymbol{D}}_{(K+1)\times(K+1)} := \begin{pmatrix} N^{-1}\sum_{i=1}^{N}\frac{1-\pi(\boldsymbol{Z}_i;\check{\gamma})}{\pi(\boldsymbol{Z}_i;\check{\gamma})}u_K(\boldsymbol{X}_i)^{\otimes 2}, & N^{-1}\sum_{i=1}^{N}\frac{1-\pi(\boldsymbol{Z}_i;\check{\gamma})}{\pi(\boldsymbol{Z}_i;\check{\gamma})}u_K(\boldsymbol{X}_i)U(\boldsymbol{Z}_i) \\ N^{-1}\sum_{i=1}^{N}\frac{1-\pi(\boldsymbol{Z}_i;\check{\gamma})}{\pi(\boldsymbol{Z}_i;\check{\gamma})}u_K(\boldsymbol{X}_i)^{\mathsf{T}}U(\boldsymbol{Z}_i), & N^{-1}\sum_{i=1}^{N}\left(\check{\theta}-\frac{T}{\pi(\boldsymbol{Z}_i;\check{\gamma})}U(\boldsymbol{Z}_i)\right)^2 \end{pmatrix}
$$

and

$$
\boldsymbol{D}_{(K+1)\times(K+1)} := \begin{pmatrix} \mathbb{E}\left[\frac{1-\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^{\otimes 2}\right], & \mathbb{E}\left[\frac{1-\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})U(\boldsymbol{Z})\right] \\ \mathbb{E}\left[\frac{1-\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^{\mathsf{T}}U(\boldsymbol{Z})\right], & \mathbb{E}\left[\left(\theta_0-\frac{T}{\pi(\boldsymbol{Z};\gamma_0)}U(\boldsymbol{Z})\right)^2\right] \end{pmatrix}
$$

For simplicity, we show that the upper left block of $\widehat{\boldsymbol{D}}_{(K+1)\times(K+1)}$ converges in probability to that of $\boldsymbol{D}_{(K+1)\times(K+1)}$, namely

$$
\left\|\frac{1}{N}\sum_{i=1}^{N}\frac{1-\pi(\boldsymbol{Z}_i;\check{\gamma})}{\pi(\boldsymbol{Z}_i;\check{\gamma})}u_K(\boldsymbol{X}_i)^{\otimes 2} - \mathbb{E}\left[\frac{1-\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^{\otimes 2}\right]\right\| \xrightarrow{p} 0\ , \tag{54}
$$

and similar argument can be applied to the other three blocks are also of convergence.

Using Taylor's expansion, we can obtain that

$$\frac{1}{N}\sum_{i=1}^{N}\frac{1-\pi(\boldsymbol{Z}_i;\check{\gamma})}{\pi(\boldsymbol{Z}_i;\check{\gamma})}u_K(\boldsymbol{X}_i)^{\otimes 2}-\mathbb{E}\left[\frac{1-\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^{\otimes 2}\right]$$

$$=\frac{1}{N}\sum_{i=1}^{N}\frac{1-\pi(\boldsymbol{Z}_i;\gamma_0)}{\pi(\boldsymbol{Z}_i;\gamma_0)}u_K(\boldsymbol{X}_i)^{\otimes 2}-\mathbb{E}\left[\frac{1-\pi(\boldsymbol{Z};\gamma_0)}{\pi(\boldsymbol{Z};\gamma_0)}u_K(\boldsymbol{X})^{\otimes 2}\right] \tag{55}$$

$$-(\check{\gamma}-\gamma_0)^{\top}\cdot\frac{1}{N}\sum_{i=1}^{N}\frac{\nabla_{\gamma}\pi(\boldsymbol{Z}_i;\gamma^*)}{\pi(\boldsymbol{Z}_i;\gamma^*)^2}u_K(\boldsymbol{X}_i)^{\otimes 2}\ , \tag{56}$$

where $\gamma^*$ lies on the line joining $\gamma_0$ and $\check{\gamma}$. By computing the second moments of (55), and using Chebyshev's inequality and Assumption 8, we can claim that the term (55) is of $o_p(1)$.

Consider the term (56). From Assumption 7, we know that the funciton $\nabla_{\gamma}\pi(\boldsymbol{Z};\gamma)$ is uniformly bounded and the propensity score $\pi(\boldsymbol{Z};\gamma)$ are uniformly bounded away from zero, thus we can find a finte constant $C > 0$ such that

$$|(56)|\leq\|\check{\gamma}-\gamma_0\|\cdot C\cdot\left\|\frac{1}{N}\sum_{i=1}^{N}u_K(\boldsymbol{X}_i)^{\otimes 2}\right\|\ .$$

Using Chebyshev's inequality, Inequality (12), and Assumption 8, we can deduce that

$$\left\|\frac{1}{N}\sum_{i=1}^{N}u_K(\boldsymbol{X}_i)^{\otimes 2}-\mathbb{E}\left[u_K(\boldsymbol{X})^{\otimes 2}\right]\right\|=O_p\left(\sqrt{\frac{K^3}{N}}\right)=o_p(1)\ .$$

We also note that $\|\mathbb{E}\left[u_K(\boldsymbol{X})^{\otimes 2}\right]\|\leq\lambda_{\max}\left(\mathbb{E}\left[u_K(\boldsymbol{X})^{\otimes 2}\right]\right)\cdot\|I_{K\times K}\|=O(\sqrt{K})$. Therefore, in light of Theorem 1 and Assumption 8 we can deduce that

$$|(56)|\leq O_p(N^{-1/2})\cdot C\cdot O(\sqrt{K})=o_p(1)\ .$$

Since the terms (55) and (56) are all of $o_p(1)$, we can justify (54).

# 7 Proof of Theorem 5

Note that

$$\widehat{\boldsymbol{V}}_K=\left\{\widehat{\boldsymbol{B}}_{(K+1)\times(p+1)}^{\top}\widehat{\boldsymbol{D}}_{(K+1)\times(K+1)}^{-1}\widehat{\boldsymbol{B}}_{(K+1)\times(p+1)}\right\}^{-1}\ ,$$

$$\boldsymbol{V}_K=\left\{\boldsymbol{B}_{(K+1)\times(p+1)}^{\top}\boldsymbol{D}_{(K+1)\times(K+1)}^{-1}\boldsymbol{B}_{(K+1)\times(p+1)}\right\}^{-1}\ ,$$

and $\boldsymbol{V}_K \xrightarrow{p} \boldsymbol{V}_{eff}$. From (53), we know that $\widehat{\boldsymbol{D}}_{(K+1)\times(K+1)} \xrightarrow{p} \boldsymbol{D}_{(K+1)\times(K+1)}$. Therefore, to prove the consistency result $\widehat{\boldsymbol{V}}_K \xrightarrow{p} \boldsymbol{V}_{eff}$, it suffices to show $\widehat{\boldsymbol{B}}_{(K+1)\times(p+1)} \xrightarrow{p} \boldsymbol{B}_{(K+1)\times(p+1)}$.

We show that the upper left block of $\widehat{\boldsymbol{B}}_{(K+1)\times(p+1)}$ converges in probability to that of $\boldsymbol{B}_{(K+1)\times(p+1)}$, namely,

$$\frac{1}{N}\sum_{i=1}^{N} u_K(\boldsymbol{X}_i)\frac{\nabla_\gamma \pi(\boldsymbol{Z}_i;\hat{\gamma})^\top}{\pi(\boldsymbol{Z}_i;\hat{\gamma})^2} \xrightarrow{P} \mathbb{E}\left[u_K(\boldsymbol{X})\frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)^\top}{\pi(\boldsymbol{Z};\gamma_0)}\right] , \tag{57}$$

and similar arguments can be applied to show that the other three blocks are also of convergence. Using Taylor's expansion, we can have

$$\frac{1}{N}\sum_{i=1}^{N} u_K(\boldsymbol{X}_i)\frac{\nabla_\gamma \pi(\boldsymbol{Z}_i;\hat{\gamma})^\top}{\pi(\boldsymbol{Z}_i;\hat{\gamma})}$$

$$= \frac{1}{N}\sum_{i=1}^{N} u_K(\boldsymbol{X}_i)\frac{\nabla_\gamma \pi(\boldsymbol{Z}_i;\gamma_0)^\top}{\pi(\boldsymbol{Z}_i;\gamma_0)} \tag{58}$$

$$- \frac{1}{N}\sum_{i=1}^{N} u_K(\boldsymbol{X}_i)\frac{\nabla_\gamma \pi(\boldsymbol{Z}_i;\gamma^{**})^\top}{\pi(\boldsymbol{Z}_i;\gamma^{**})^2}\nabla_\gamma \pi(\boldsymbol{Z}_i;\gamma^{**})^\top(\hat{\gamma}-\gamma_0) \tag{59}$$

$$+ \frac{1}{N}\sum_{i=1}^{N} u_K(\boldsymbol{X}_i)\frac{1}{\pi(\boldsymbol{Z}_i;\gamma^{**})}\cdot\nabla_\gamma^2 \pi(\boldsymbol{Z}_i;\gamma^{**})(\hat{\gamma}-\gamma_0)\cdot\,, \tag{60}$$

where $\gamma^{**}$ lies on the line joining $\hat{\gamma}$ and $\gamma_0$. Consider the term (58), we have

$$\mathbb{E}\left[\left\|\frac{1}{N}\sum_{i=1}^{N} u_K(\boldsymbol{X}_i)\frac{\nabla_\gamma \pi(\boldsymbol{Z}_i;\gamma_0)^\top}{\pi(\boldsymbol{Z}_i;\gamma_0)} - \mathbb{E}\left[u_K(\boldsymbol{X})\frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)^\top}{\pi(\boldsymbol{Z};\gamma_0)}\right]\right\|^2\right]$$

$$\leq \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\left\|u_K(\boldsymbol{X}_i)\frac{\nabla_\gamma \pi(\boldsymbol{Z}_i;\gamma_0)^\top}{\pi(\boldsymbol{Z}_i;\gamma_0)}\right\|^2\right]$$

$$\leq \frac{1}{N}\mathbb{E}\left[\frac{\|\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)\|^2}{\pi(\boldsymbol{Z};\gamma_0)^2}\cdot\|u_K(\boldsymbol{X})\|^2\right] \leq O(1)\cdot\frac{1}{N}\mathbb{E}\left[\|u_K(\boldsymbol{X})\|^2\right] = O(K/N) = o(1) \,,$$

where the second equality holds because $\frac{\|\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)\|^2}{\pi(\boldsymbol{Z};\gamma_0)}$ is uniformly bounded, while the last equality holds because of Assumption 8. Then in light of Markov's inequality, we can have

$$\left\|\frac{1}{N}\sum_{i=1}^{N} u_K(\boldsymbol{X}_i)\frac{\nabla_\gamma \pi(\boldsymbol{Z}_i;\gamma_0)^\top}{\pi(\boldsymbol{Z}_i;\gamma_0)} - \mathbb{E}\left[u_K(\boldsymbol{X})\frac{\nabla_\gamma \pi(\boldsymbol{Z};\gamma_0)^\top}{\pi(\boldsymbol{Z};\gamma_0)}\right]\right\| = o_p(1) \,.$$

For the terms (59) and (60), by using a similar argument of showing (56) $= o_p(1)$, we can obtain that both (59) and (60) are of $o_p(1)$. Therefore, we can justify the validity of (57). Finally, we can claim our consistency result $\widehat{\boldsymbol{V}}_K \xrightarrow{p} \boldsymbol{V}_{eff}$.

# References

CHEN, X. (2007): "Large sample sieve estimation of semi-nonparametric models," *Handbook of econometrics*, 6, 5549–5632.

EICKER, F. (1966): "A multivariate central limit theorem for random linear vector forms," *The Annals of Mathematical Statistics*, pp. 1825–1828.

MORIKAWA, K., AND J. K. KIM (2016): "Semiparametric Adaptive Estimation With Nonignorable Nonresponse Data," *arXiv preprint arXiv:1612.09207*.

NEWEY, W. K. (1991): "Uniform convergence in probability and stochastic equicontinuity," *Econometrica: Journal of the Econometric Society*, pp. 1161–1167.

——— (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of econometrics*, 79(1), 147–168.