Statistical Theory and Related Fields

ISSN: 2475-4269 (Print) 2475-4277 (Online) Journal homepage: https://www.tandfonline.com/loi/tstf20

Impact of sufficient dimension reduction in nonparametric estimation of causal effect

Ying Zhang, Jun Shao, Menggang Yu & Lei Wang

To cite this article: Ying Zhang, Jun Shao, Menggang Yu & Lei Wang (2018) Impact of sufficient dimension reduction in nonparametric estimation of causal effect, Statistical Theory and Related Fields, 2:1, 89-95, DOI: 10.1080/24754269.2018.1466100

To link to this article: https://doi.org/10.1080/24754269.2018.1466100

Published online: 18 May 2018.

STATISTICAL THEORY AND RELATED FIELDS
2018, VOL. 2, NO. 1, 89–95
https://doi.org/10.1080/24754269.2018.1466100

Taylor & Francis
Taylor & Francis Group

Impact of sufficient dimension reduction in nonparametric estimation of causal effect

Ying Zhang[a], Jun Shao[a,b], Menggang Yu[c] and Lei Wang[d]

[a]Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA; [b]School of Statistics, East China Normal University, Shanghai, People's Republic of China; [c]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA; [d] Institute of Statistics and LPMC, Nankai University, Tianjin, People's Republic of China

ABSTRACT

We consider the estimation of causal treatment effect using nonparametric regression or inverse propensity weighting together with sufficient dimension reduction for searching low-dimensional covariate subsets. A special case of this problem is the estimation of a response effect with data having ignorable missing response values. An issue that is not well addressed in the literature is whether the estimation of the low-dimensional covariate subsets by sufficient dimension reduction has an impact on the asymptotic variance of the resulting causal effect estimator. With some incorrect or inaccurate statements, many researchers believe that the estimation of the low-dimensional covariate subsets by sufficient dimension reduction does not affect the asymptotic variance. We rigorously establish a result showing that this is not true unless the low-dimensional covariate subsets include some covariates superfluous for estimation, and including such covariates loses efficiency. Our theory is supplemented by some simulation results.

ARTICLE HISTORY
Received 10 November 2017
Accepted 14 April 2018

KEYWORDS
Asymptotic variance; causal treatment effect; nonparametric regression or propensity weighting; $n^{1/2}$-consistency

1. Introduction

Consider the estimation of an unknown parameter θ based on a sample of size n from a given population. Many estimators are of the form $\hat{\theta}_n(\hat{\lambda}_n)$, a function of $\hat{\lambda}_n$ that is an estimator of another parameter λ, where both $\hat{\theta}_n$ and $\hat{\lambda}_n$ are functions of the sample (e.g., Gong & Samaniego, 1981; Randles, 1982). Under some conditions both $n^{1/2}\{\hat{\theta}_n(\hat{\lambda}_n) - \theta\}$ and $n^{1/2}\{\hat{\theta}_n(\lambda) - \theta\}$ are asymptotically normal with mean zero as n increases to infinity. A question of both theoretical and practical interest is whether the estimation efficiency is affected by the fact that λ is estimated, i.e., whether $\hat{\theta}_n(\hat{\lambda}_n)$ and $\hat{\theta}_n(\lambda)$ have the same asymptotic variance. Examples with equal asymptotic variance were given in Raghavachari (1965), Adichie (1974), De Wet, and Van Wyk (1979) and Randles (1982). Examples in which $\hat{\theta}_n(\hat{\lambda}_n)$ and $\hat{\theta}_n(\lambda)$ have different asymptotic variances can be found in Gong and Samaniego (1981) and Randles (1982).

In the problem of causal evaluation of treatment (Hahn, 1998, 2004; Hirano, Imbens, & Ridder, 2003; Imbens, Newey, & Ridder, 2006; Rosenbaum & Rubin, 1983; Wang & Chen, 2009; Wang, 2007), the previously described issue is not well addressed in the literature, and some incorrect or inaccurate statements are given with incorrect proofs. The problem can be described as follows. Let T be a binary treatment indicator, X be a p-dimensional vector of pretreatment covariates and Y_k be the potential outcome under treatment $T = k$. We focus on the causal effect $\theta = E(Y_1) - E(Y_0)$, other causal effects such as quantile treatment effects can be similarly considered. Since only one treatment is applied, what we can observe is $Y = TY_1 + (1 - T)Y_0$, not both Y_1 and Y_0. Based on a random sample from the distribution of (Y, X, T), we can estimate θ under the assumption that $T \perp Y_k \,|\, X$ (Rosenbaum & Rubin, 1983), i.e., T and Y_k are independent conditional on X, $k = 0,1$. In the special case where $Y_0 \equiv 0$, this problem reduces to the well-known missing data problem where $T = 0$ indicates a missing Y_1, $\theta = E(Y_1)$, and $T \perp Y_1 \,|\, X$ is simply the missing at random assumption.

Estimators based on nonparametric regression or nonparametric inverse propensity weighting as described in Section 2 require almost no model assumption on (Y, X) but they do not perform well when the covariate dimension p is not very small. Since frequently only a few linear combinations of X are actually related with Y_k, it is attractive to first find a lower dimensional $B_k^{\mathrm{T}}X$ satisfying $Y_k \perp X \,|\, B_k^{\mathrm{T}}X$, where B_k is a $p \times d_k$ constant matrix with a small d_k, $k = 0,1$, and then apply nonparametric regression or inverse propensity weighting with X replaced by $B_k^{\mathrm{T}}X$. If B_0 and B_1 are known, then the resulting estimator of θ is denoted as $\hat{\theta}_n(\lambda)$ with $\lambda = (B_0, B_1)$. However, λ is usually unknown and a sufficient dimension reduction method (e.g., Cook & Weisberg, 1991; Li, 1991; Xia Tong, Li, & Zhu, 2002) is typically applied to estimate it

CONTACT Lei Wang lwangstat@nankai.edu.cn

by $\hat{\lambda}_n = (\hat{B}_0, \hat{B}_1)$. Under some conditions, both $\hat{\theta}_n(\hat{\lambda}_n)$ and $\hat{\theta}_n(\lambda)$ are asymptotically normal with mean zero and hence the relevant question is whether the estimation of λ by $\hat{\lambda}_n$ affects the asymptotic efficiency of estimating θ. There is no precise conclusion in the literature regarding this issue, but some researchers implicitly assume that using $n^{1/2}$-consistent estimators of λ is asymptotically the same as using the true λ. For instance, Hu, Follmann, and Wang (2014) and Deng & Wang (2017) claimed that $\hat{\theta}_n(\hat{\lambda}_n)$ and $\hat{\theta}_n(\lambda)$ have the same asymptotic variance if $\hat{B}_k$ is $n^{1/2}$-consistent, which is an incorrect conclusion in general. In a very recent publication, Luo, Zhu, and Ghosh (2017) made the same wrong conclusion.

We rigorously establish a result showing that under the additional condition $T \perp X \mid B_k^{\mathrm{T}}X$, $k=0,1$, $n^{1/2}\{\hat{\theta}_n(\hat{\lambda}_n) - \hat{\theta}_n(\lambda)\} = o_p(1)$. Although this condition is sufficient but not necessary for the asymptotic equivalence between $\hat{\theta}_n(\hat{\lambda}_n)$ and $\hat{\theta}_n(\lambda)$, we provide an example showing that without $T \perp X \mid B_k^{\mathrm{T}}X$, $k=0,1$, $\hat{\theta}_n(\hat{\lambda}_n)$ and $\hat{\theta}_n(\lambda)$ have different asymptotic variances. Our theory is supplemented by simulation results showing that $\hat{\theta}_n(\hat{\lambda}_n)$ can be substantially less efficient than $\hat{\theta}_n(\lambda)$. However, our simulation results also show that finding a B_k satisfying the additional condition $T \perp X \mid B_k^{\mathrm{T}}X$ may not be a good idea, because, although the resulting estimator is not affected by the estimation of B_k, $B_k^{\mathrm{T}}X$ may include some covariates superfluous for estimation and have an unnecessarily high dimension to lose efficiency.

2. Theory

To study the asymptotic behaviour of $\hat{\theta}_n(\hat{\lambda}_n)$ and $\hat{\theta}_n(\lambda)$, we first described three popular nonparametric estimators $\hat{\theta}_n$. We adopt the notation in Section 1. The regression method (Hu et al., 2014; Imbens et al., 2006) estimates $\theta = E\{E(Y_1 \mid B_1^{\mathrm{T}}X, T=1) - E(Y_0 \mid B_0^{\mathrm{T}}X, T=0)\}$ through estimating the function $m_k(s) = E(Y_k \mid B_k^{\mathrm{T}}X = s, T=k)$ by the usual kernel estimator $\hat{m}_k(s) = \sum_{i=1}^n T_i^{(k)} Y_i \mathcal{K}_{h_k}(B_k^{\mathrm{T}}X_i - s) / \sum_{i=1}^n T_i^{(k)} \mathcal{K}_{h_k}(B_k^{\mathrm{T}}X_i - s)$, $k=0,1$, where $T_i^{(1)} = T_i$, $T_i^{(0)} = 1 - T_i$, $\mathcal{K}_{h_k}(s) = h_k^{-d_k}\mathcal{K}_k(h_k^{-1}s)$, $\mathcal{K}_k$ is a d_k-dimensional kernel function and h_k is the bandwidth. The regression estimator of θ is

$$\hat{\theta}_{\mathrm{REG}}(\lambda) = \frac{1}{n}\sum_{i=1}^n \{\hat{m}_1(B_1^{\mathrm{T}}X_i) - \hat{m}_0(B_0^{\mathrm{T}}X_i)\}.$$

The inverse propensity weighting method (Imai & Ratkovic, 2014; Imbens, 2004; Kang & Schafer, 2007) estimates the probability $\pi_k(s) = \mathrm{pr}(T=k \mid B_k^{\mathrm{T}}X = s)$ by the kernel estimator $\hat{\pi}_k(s) = \sum_{i=1}^n T_i^{(k)} \mathcal{K}_{h_k}(B_k^{\mathrm{T}}X_i - s) / \sum_{i=1}^n \mathcal{K}_{h_k}(B_k^{\mathrm{T}}X_i - s)$, $k=0,1$, and obtains the following estimator of θ by inverse propensity weighting,

$$\hat{\theta}_{\mathrm{IPW}}(\lambda) = \sum_{k=0,1}(-1)^{k-1}\left\{\sum_{i=1}^n \frac{T_i^{(k)}}{\hat{\pi}_k(B_k^{\mathrm{T}}X_i)}\right\}^{-1} \times \sum_{i=1}^n \frac{T_i^{(k)}Y_i}{\hat{\pi}_k(B_k^{\mathrm{T}}X_i)}.$$

However, this estimator often does not have good empirical performance and can be improved by the estimator combining the regression and inverse propensity weighting, the so-called augmented inverse propensity weighting estimator,

$$\hat{\theta}_{\mathrm{AIPW}}(\lambda) = \frac{1}{n}\sum_{i=1}^n \sum_{k=0,1}(-1)^{k-1} \times \left\{\frac{T_i^{(k)}Y_i}{\hat{\pi}_k(B_k^{\mathrm{T}}X_i)} - \frac{T_i^{(k)} - \hat{\pi}_k(B_k^{\mathrm{T}}X_i)}{\hat{\pi}_k(B_k^{\mathrm{T}}X_i)}\hat{m}_k(B_k^{\mathrm{T}}X_i)\right\}.$$

In what follows, we use $\hat{\theta}_n(\lambda)$ to denote one of $\hat{\theta}_{\mathrm{IPW}}(\lambda)$, $\hat{\theta}_{\mathrm{REG}}(\lambda)$ and $\hat{\theta}_{\mathrm{AIPW}}(\lambda)$. Under the conditions $T \perp Y_k \mid X$, $Y_k \perp X \mid B_k^{\mathrm{T}}X$, $k=0,1$, and some regularity conditions, it has been shown that $n^{1/2}\{\hat{\theta}_n(\lambda) - \theta\}$ is asymptotically normal with mean 0 and variance

$$\sigma^2(\lambda) = \mathrm{var}\{E(Y_1 \mid B_1^{\mathrm{T}}X) - E(Y_0 \mid B_0^{\mathrm{T}}X)\} + E\left\{\frac{\mathrm{var}(Y_1 \mid B_1^{\mathrm{T}}X)}{\mathrm{pr}(T=1 \mid B_1^{\mathrm{T}}X)} + \frac{\mathrm{var}(Y_0 \mid B_0^{\mathrm{T}}X)}{\mathrm{pr}(T=0 \mid B_0^{\mathrm{T}}X)}\right\} \quad (1)$$

(e.g., Hu et al., 2014; Luo et al., 2017; Wang & Chen, 2009; Wang, 2007).

Our main result is about the asymptotic behaviour of $\hat{\theta}_n(\hat{\lambda}_n)$ with a $n^{1/2}$-consistent estimator $\hat{\lambda}_n$ of $\lambda = (B_0, B_1)$, which leads to a sufficient condition under which $\hat{\theta}_n(\hat{\lambda}_n)$ and $\hat{\theta}_n(\lambda)$ are asymptotically equivalent. In the following, $\mathrm{vec}(B)$ denotes a column vector whose components are elements of a matrix B and $o_p(1)$ denotes a term converging to 0 in probability. A proof of the following theorem is given in the appendix.

Theorem 2.1: *Assume $T \perp Y_k \mid X$, $Y_k \perp X \mid B_k^{\mathrm{T}}X$, $k = 0, 1$, and the regularity conditions in the appendix.*

(i) *If $\hat{B}_k$ is a $n^{1/2}$-consistent estimator of B_k, $k = 0, 1$, then,*

$$n^{1/2}\{\hat{\theta}_n(\hat{\lambda}_n) - \hat{\theta}_n(\lambda)\} = \sum_{k=0,1} n^{1/2} c_k^{\mathrm{T}} \mathrm{vec}(\hat{B}_k - B_k) + o_p(1), \quad (2)$$

where

$$c_k = -\text{vec}\left[E\left\{\frac{\text{cov}(T, X \mid B_k^{\mathrm{T}}X)}{\pi_k(B_k^{\mathrm{T}}X)}\frac{\partial m_k(s)}{\partial s^{\mathrm{T}}}\bigg|_{s=B_k^{\mathrm{T}}X}\right\}\right]. \tag{3}$$

(ii) *If*

$$n^{1/2}\text{vec}(\hat{B}_k - B_k) = n^{-1/2}\sum_{i=1}^{n}\psi_k(X_i, Y_i, T_i) + o_p(1) \tag{4}$$

for some functions ψ_k *with* $E\{\psi_k(X, Y, T)\} = 0$, $k = 0, 1$, *then,* $n^{1/2}\{\hat{\theta}_n(\hat{\lambda}_n) - \theta\}$ *is asymptotically normal with mean 0 and variance*

$$\sigma^2(\lambda) + \text{var}\left\{\sum_{k=0,1} c_k^{\mathrm{T}}\psi_k(X, Y, T)\right\} + 2\text{cov}\left\{\sum_{k=0,1} c_k^{\mathrm{T}}\psi_k(X, Y, T), S(X, Y, T)\right\}, \tag{5}$$

where $\sigma^2(\lambda)$ *is given by Equation (1) and*

$$S(X, Y, T) = \frac{T\{Y_1 - m_1(B_1^{\mathrm{T}}X)\}}{\pi_1(B_1^{\mathrm{T}}X)} - \frac{(1-T)\{Y_0 - m_0(B_0^{\mathrm{T}}X)\}}{\pi_0(B_0^{\mathrm{T}}X)} + m_1(B_1^{\mathrm{T}}X) - m_0(B_0^{\mathrm{T}}X).$$

Condition (4) is satisfied for some sufficient dimension reduction methods (Hsing & Carroll, 1992; Zhu & Ng, 1995).

Theorem 2.1 shows that the asymptotic difference between $\hat{\theta}_n(\hat{\lambda}_n)$ and $\hat{\theta}_n(\lambda)$ is related to the magnitude of $\hat{B}_k - B_k$, $k = 0,1$, through Equation (2). From the sufficient dimension reduction literature, $\hat{B}_k - B_k$ is at most of the order $n^{-1/2}$. Hence, a sufficient condition under which $\hat{\theta}_n(\hat{\lambda}_n)$ and $\hat{\theta}_n(\lambda)$ are asymptotically equivalent is that both c_0 and c_1 in Equation (3) are equal to 0. By formula (3), the only realistic situation where $c_k = 0$ is when $\text{cov}(T, X \mid B_k^{\mathrm{T}}X) = 0$, which is implied by $T \perp X \mid B_k^{\mathrm{T}}X$.

Hence, if we choose B_k satisfying both $Y_k \perp X \mid B_k^{\mathrm{T}}X$ and $T \perp X \mid B_k^{\mathrm{T}}X$, then $\hat{\theta}_n(\hat{\lambda}_n)$ and $\hat{\theta}_n(\lambda)$ are asymptotically equivalent, provided that Equation (4) holds. However, we may pay a price for doing so, because $B_k^{\mathrm{T}}X$ satisfying the additional requirement $T \perp X \mid B_k^{\mathrm{T}}X$ may include some covariates superfluous for estimation and, thus, have an unnecessarily high dimension and lose efficiency. Let $\lambda = (B_0, B_1)$ with B_k satisfying $Y_k \perp X \mid B_k^{\mathrm{T}}X$, and let $\lambda' = (B_0', B_1')$ with $B'k$ satisfying both $Y_k \perp X \mid B_k'^{\mathrm{T}}X$ and $T \perp X \mid B_k'^{\mathrm{T}}X$. Although $\hat{\theta}_n(\hat{\lambda}_n')$ and $\hat{\theta}_n(\lambda')$ are asymptotically equivalent, their asymptotic variance is $\sigma^2(\lambda')$ given by Equation (1) with B_k replaced by B_k', which is larger than $\sigma^2(\lambda)$ when $\dim(B'_k)$ is larger than $\dim(B_k)$ due to the extra requirement of $T \perp X \mid B_k'^{\mathrm{T}}X$. Furthermore, even when $\hat{\theta}_n(\hat{\lambda}_n)$ is less efficient than $\hat{\theta}_n(\lambda)$ due to the estimation of λ, it may still be more efficient than $\hat{\theta}_n(\lambda')$. The following is an example for illustration.

Example 2.2: Let $X = (X^{(1)}, X^{(2)})^{\mathrm{T}}$, $Y_1 = X^{(1)} + \epsilon$, $Y_0 \equiv 0$, where $X^{(1)}$ and $X^{(2)}$ are independent and uniform on the interval $[0, 1]$, $\epsilon \sim N(0, 1)$ and is independent of X. Let $\text{pr}(T = 1 \mid X) = \exp(-2 + 3X^{(2)})/\{1 + \exp(-2 + 3X^{(2)})\}$. Then $B_1^{\mathrm{T}}X = X^{(1)}$ satisfying $Y_1 \perp X \mid B_1^{\mathrm{T}}X$, but not $T \perp X \mid B_1^{\mathrm{T}}X$. Let $B_1'^{\mathrm{T}}X = X$. Then both $Y_1 \perp X \mid B_1'^{\mathrm{T}}X$ and $T \perp X \mid B_1'^{\mathrm{T}}X$ hold. However, $\dim(B_1^{\mathrm{T}}X) = 1 < 2 = \dim(B_1'^{\mathrm{T}}X)$, and $B_1'^{\mathrm{T}}X$ contains $X^{(2)}$ that is not useful for estimating $\theta = E(Y_1)$. In this case, $\sigma^2(\lambda) = 1/12 + \{\text{pr}(T = 1)\}^{-1} = 2.612$, smaller than $\sigma^2(\lambda') = 1/12 + E\{1 + \exp(2 - 3X^{(2)})\} = 3.424$. The c_1 vector defined by Equation (3) is a two-dimensional vector whose first component is 0 and second component $= -E\{\text{cov}(T, X^{(2)}|X^{(1)})/\text{pr}(T = 1|X_1)\} = -0.136 \neq 0$, so the asymptotic variance of $\hat{\theta}_n(\hat{\lambda}_n)$ given by Equation (5) differs from $\sigma^2(\lambda)$. Calculating the asymptotic variance in Equation (5) requires further information about $\hat{B}_1$.

In next section, we provide some numerical results for the variance in Equation (5).

3. Simulation

To support our theory we investigate the finite-sample performances of $\hat{\theta}_{\text{REG}}$ and $\hat{\theta}_{\text{AIPW}}$ with two choices of B_k discussed in Section 2, i.e., B_k satisfies $Y_k \perp X \mid B_k^{\mathrm{T}}X$ with smallest possible $\dim(B_k^{\mathrm{T}}X)$, and B_k' satisfies $Y_k \perp X \mid B_k'^{\mathrm{T}}X$ and $T \perp X \mid B_k'^{\mathrm{T}}X$ with smallest possible $\dim(B_k'^{\mathrm{T}}X)$. We consider estimators using the true B_k and B_k' as well as estimated B_k and B_k' by applying the sliced inverse regression method (Li, 1991). According to Theorem 2.1, estimators using the true B_k and estimated B_k have different asymptotic variances, whereas estimators using the true B_k' and estimated B_k' are asymptotically equivalent. We try two sample sizes, $n = 200$ and $n = 1000$. As in Hu et al. (2014), the nonparametric kernel estimators $\hat{\pi}_k(\hat{S}_k)$ and $\hat{m}_k(\hat{S}_k)$ are computed using the rth order Gaussian product kernel with standardised covariates. The bandwidth we used here is $h_k = 1.5n^{-2/(2r_k + d_k)}$ (Chen, Wan, & Zhou, 2015; Hu et al., 2014).

We consider the following three simulation models.

(1) $X = (X_1, X_2, X_3)^{\mathrm{T}}$ with independent $N(0, 10^2)$ components, $Y_0 = 10X_1 + \epsilon_0$, $Y_1 = 10 + X_2 + \epsilon_1$, where ϵ_k's are independent $N(0, 1)$ and are independent of X, and $\text{pr}(T = 1 \mid X) = \exp(3X_2)/\{1 + \exp(3X_2)\}$. The outcome models are linear in X

Table 1. Relative bias (RB) and standard deviation (SD) of $\hat{\theta}_n$ based on 1000 simulations.

			$\hat{\theta}_n(\hat{\lambda}_n)$		$\hat{\theta}_n(\lambda)$		$\hat{\theta}_n(\hat{\lambda}'_n)$		$\hat{\theta}_n(\lambda')$	
Method	Model	n	RB	SD	RB	SD	RB	SD	RB	SD
REG	(1)	200	0.06	7.48	0.06	7.23	0.07	7.66	0.06	7.53
		1000	0.03	3.39	0.03	3.22	0.04	3.31	0.03	3.30
	(2)	200	0.03	0.37	0.02	0.29	0.06	0.58	0.05	0.41
		1000	0.01	0.17	0.01	0.12	0.04	0.22	0.03	0.20
	(3)	200	−0.01	0.52	−0.02	0.44	−0.04	0.66	−0.05	0.62
		1000	0.01	0.19	−0.01	0.18	−0.01	0.23	−0.03	0.24
AIPW	(1)	200	0.04	7.50	0.04	7.21	0.05	7.65	0.04	7.52
		1000	0.02	3.40	0.02	3.23	0.03	3.31	0.02	3.29
	(2)	200	0.02	0.34	0.01	0.28	0.05	0.57	0.05	0.41
		1000	0.00	0.15	0.00	0.12	0.03	0.20	0.02	0.17
	(3)	200	0.00	0.48	−0.01	0.43	−0.04	0.61	−0.05	0.60
		1000	0.01	0.18	0.00	0.17	0.00	0.22	−0.03	0.22

Notes: $\lambda = (B_0, B_1)$, $Y_k \perp X \mid B_k^{\mathrm{T}}X, k = 0,1$; $\lambda' = (B'_0, B'_1)$, $Y_k \perp X \mid B_k'^{\mathrm{T}}X$ and $T \perp X \mid B_k'^{\mathrm{T}}X, k = 0,1$. $\hat{\lambda}_n$ and $\hat{\lambda}'_n$: estimates of λ and λ' by sufficient dimension reduction.

and the log-conditional treatment odds is linear in X. Under this model, $\dim(B_0^{\mathrm{T}}X) = \dim(B_1^{\mathrm{T}}X) = 1$, $\dim(B_0'^{\mathrm{T}}X) = 2$ and $\dim(B_1'^{\mathrm{T}}X) = 1$.

(2) $X = (X_1, \ldots, X_7)^{\mathrm{T}}$ with independent $N(0,1)$ components, $Y_0 = 3X_1 + 6X_2 + 3X_3 + \epsilon_0$, $Y_1 = 10 + 3X_1 + 6X_2 + 3X_3 + 3X_4 + \epsilon_1$, where ϵ_k's are independent $N(0,1)$ and are independent of X, and $\mathrm{pr}(T = 1 \mid X) = \exp(2X_4)/\{1 + \exp(2X_4)\}$. The outcome models are linear in X and the log-conditional treatment odds is linear in X. Under this model, $\dim(B_0^{\mathrm{T}}X) = \dim(B_1^{\mathrm{T}}X) = 1$ but $\dim(B_0'^{\mathrm{T}}X) = \dim(B_1'^{\mathrm{T}}X) = 2$.

(3) $X = (X_1, \ldots, X_7)^{\mathrm{T}}$ with independent $N(0,1)$ components, $Y_0 = 3(X_1 + X_2 + 2X_3 + 2X_4) + 1.5X_6^2 + \epsilon_0$, $Y_1 = 12 + 3(X_1 + X_2 + 2X_3 + X_4 + X_5) + 1.5X_7^2 + \epsilon_1$, where ϵ_k's are independent $N(0,1)$ and are independent of X, and $\mathrm{pr}(T = 1 \mid X) = \exp(-2X_5 + 0.7X_6^2 - 0.5X_7^2)/\{1 + \exp(-2X_5 + 0.7X_6^2 - 0.5X_7^2)\}$. The outcome models are nonlinear in X and the log-conditional treatment odds is also nonlinear in X. Under this model, $\dim(B_0^{\mathrm{T}}X) = \dim(B_1^{\mathrm{T}}X) = 2$ but $\dim(B_0'^{\mathrm{T}}X) = \dim(B_1'^{\mathrm{T}}X) = 4$.

Table 1 shows the simulated relative bias and standard deviation in each scenario based on 1000 simulation runs. It can be seen that the simulation results are in agreement with the asymptotic result (Theorem 2.1), especially when $n = 1000$, i.e., the SD of $\hat{\theta}_n(\hat{\lambda}'_n)$ and $\hat{\theta}_n(\lambda')$ are very close while the SD of $\hat{\theta}_n(\hat{\lambda}_n)$ and $\hat{\theta}_n(\lambda)$ may be quite different. Although $\hat{\theta}_n(\hat{\lambda}_n)$ may be worse than $\hat{\theta}_n(\lambda)$, it may be better than $\hat{\theta}_n(\hat{\lambda}'_n)$; hence, it is not a good idea to search for a $\hat{\lambda}'_n$ that does not affect the asymptotic variance. Regarding the two different estimation methods, $\hat{\theta}_{\mathrm{REG}}$ and $\hat{\theta}_{\mathrm{AIPW}}$ have very comparable performances.

Acknowledgments

We are grateful to the Editor, the Associate Editor and two anonymous referees for their insightful comments and suggestions on this article, which have led to significant improvements. The views in this publication are solely the responsibility of the authors and do not necessarily represent the views of the PCORI, its Board of Governors or Methodology Committee.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This research was partially supported through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1409-21219). This research was also supported by the National Natural Science Foundation of China (11501208), Fundamental Research Funds for the Central Universities, National Social Science Foundation (13BTJ009), the Chinese 111 Project grant (B14019) and the U.S. National Science Foundation (DMS-1305474 and DMS-1612873).

Notes on contributors

Ying Zhang is a PhD candidate, Department of Statistics, University of Wisconsin-Madison.

Dr Jun Shao holds a PhD in statistics from the University of Wisconsin-Madison. He is a professor of statistics at the University of Wisconsin-Madison. His research interests include variable selection and inference with high dimensional data, sample surveys, and missing data problems.

Dr Menggang Yu holds a PhD in biostatistics from the University of Michigan. He is now a professor of biostatistics at the University of Wisconsin-Madison. Besides developing statistical methodology related to cancer research and clinical trials, Dr Yu is also very interested in health services research.

Dr Lei Wang holds a PhD in statistics from East China Normal University. He is an assistant professor of statistics at Nankai University. His research interests include empirical likelihood and missing data problems.

References

Adichie, J. N. (1974). Rank score comparison of several regression parameters. *The Annals of Statistics*, *2*, 396–402.

Chen, X., Wan, A. T. K., & Zhou, Y. (2015). Efficient quantile regression analysis with missing observations. *Journal of the American Statistical Association*, *110*, 723–741.

Cook, R. D., & Weisberg, S. (1991). Discussion of 'Sliced inverse regression for dimension reduction'. *Journal of the American Statistical Association*, *86*, 328–332.
Deng, J., & Wang, Q. (2017). Dimension reduction estimation for probability density with data missing at random when covariables are present. *Journal of Statistical Planning and Inference*, *181*, 11–29.
De Wet, T., & Van Wyk, J. W. J. (1979). Efficiency and robustness of Hogg's adaptive trimmed means. *Communications in Statistics: Theory and Methods*, *8*, 117–128.
Gong, G., & Samaniego, F. J. (1981). Pseudo maximum likelihood estimation: Theory and applications. *The Annals of Statistics*, *9*, 861–869.
Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, *66*, 315–331.
Hahn, J. (2004). Functional restriction and efficiency in causal inference. *Review of Economics and Statistics*, *86*, 73–76.
Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, *71*, 1161–1189.
Hsing, T., & Carroll, R. J. (1992). An asymptotic theory for sliced inverse regression. *The Annals of Statistics*, *20*, 1040–1061.
Hu, Z., Follmann, D. A., & Wang, N. (2014). Estimation of mean response via the effective balancing score. *Biometrika*, *101*, 613–624.
Imai, K., & Ratkovic, M. (2014). Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *76*, 243–263.
Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*, 4–29.
Imbens, G. W., Newey, W., & Ridder, G. (2006). *Mean-squared-error calculations for average treatment effects* (Working Paper).
Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, *22*, 523–539.
Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, *86*, 316–327.
Luo, W., Zhu, Y., & Ghosh, D. (2017). On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika*, *104*, 51–65.
Raghavachari, M. (1965). On the efficiency of the normal scores test relative to the *F*-test. *The Annals of Mathematical Statistics*, *36*, 1306–1307.
Randles, R. H. (1982). On the asymptotic normality of statistics with estimated parameters. *The Annals of Statistics*, *10*, 462–474.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41–55.
Wang, D., & Chen, S. X. (2009). Empirical likelihood for estimating equations with missing values. *The Annals of Statistics*, *37*, 490–517.
Wang, Q. (2007). M-estimators based on inverse probability weighted estimating equations with response missing at random. *Communications in Statistics: Theory and Methods*, *36*, 1091–1103.
Xia, Y., Tong, H., Li, W. K., & Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *64*, 363–410.
Zhu, L. X., & Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica*, *5*, 727–736.

Appendix

The following regularity conditions are assumed for Theorem 2.1, where conditions (1)–(4) are the same as conditions C1–C5 in Wang & Chen (2009) with X replaced by $S_k = B_k^{\mathrm{T}}X$, $k=0,1$:

(1) $\pi_k(S_k) = \mathrm{pr}(T = k \mid S_k)$ is bounded away from 0 and 1.
(2) The propensity function $\pi_k(S_k)$, the S_k-density function $f(S_k)$ and $m_k(S_k)$ all have bounded partial derivatives with respect to S_k up to order r_k with $r_k \geq 2$, where r_k is the order of the kernel $\mathcal{K}_{h_k}$.
(3) $E(Y_k^4) < \infty$.
(4) The smoothing bandwidth h_k satisfies $nh_k^{d_k} \to \infty$ and $n^{1/2}h_k^{r_k} \to 0$ as $n \to \infty$.
(5) The kernel $\mathcal{K}_k$ is bounded up to the second-order derivative.
(6) The smoothing bandwidth h_k satisfies $nh_k^2 \to \infty$ as $n \to \infty$.

Proof of Theorem 2.1: For purposes of simplicity, we focus only on the proof for regression type estimator $\hat{\theta}_{\mathrm{REG}}$ with $d_0 = d_1 = 1$ and show the difference of first term in regression estimator between using true B_1 and estimated B_1. Denote $h_1, \mathcal{K}_1, m_1(\cdot), \pi_1(\cdot), B_1$ as $h, \mathcal{K}, m(\cdot), \pi(\cdot), B$, respectively, and define $\mathcal{K}_h(\cdot) = h^{-1}\mathcal{K}(\cdot/h)$ in the following proof. Let $\Delta_{ij} = \mathcal{K}_h(\hat{B}^{\mathrm{T}}X_j - \hat{B}^{\mathrm{T}}X_i) - \mathcal{K}_h(B^{\mathrm{T}}X_j - B^{\mathrm{T}}X_i)$; it can be verified that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\{\hat{m}(B^{\mathrm{T}}X_i) - \hat{m}(\hat{B}^{\mathrm{T}}X_i)\} \\
&\quad= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{j=1}^{n} T_jY_j\mathcal{K}_h(\hat{B}^{\mathrm{T}}X_j - \hat{B}^{\mathrm{T}}X_i)}{\sum_{j=1}^{n} T_j\mathcal{K}_h(\hat{B}^{\mathrm{T}}X_j - \hat{B}^{\mathrm{T}}X_i)}\right. \\
&\qquad\left. - \frac{\sum_{j=1}^{n} T_jY_j\mathcal{K}_h(B^{\mathrm{T}}X_j - B^{\mathrm{T}}X_i)}{\sum_{j=1}^{n} T_j\mathcal{K}_h(B^{\mathrm{T}}X_j - B^{\mathrm{T}}X_i)}\right\} \\
&\quad= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{j=1}^{n} T_jY_j\mathcal{K}_h(S_j - S_i) + \sum_{j=1}^{n} T_jY_j\Delta_{ij}}{\sum_{j=1}^{n} T_j\mathcal{K}_h(S_j - S_i) + \sum_{j=1}^{n} T_j\Delta_{ij}}\right. \\
&\qquad\left. - \frac{\sum_{j=1}^{n} T_jY_j\mathcal{K}_h(S_j - S_i)}{\sum_{j=1}^{n} T_j\mathcal{K}_h(S_j - S_i)}\right\} \\
&\quad= A_1 + A_2 + A_3,
\end{aligned}
$$

where

$$
A_1 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{T_jY_j\Delta_{ij}}{\pi(S_i)f(S_i)} - \frac{T_jm(S_i)\Delta_{ij}}{\pi(S_i)f(S_i)}\right\},
$$

$$
\begin{aligned}
A_2 = -\frac{1}{n^2}&\sum_{i=1}^{n}\sum_{j=1}^{n} \\
&\times\left\{\frac{T_jY_j\Delta_{ij}}{\pi(S_i)f(S_i)} - \frac{T_jY_j\Delta_{ij}}{\begin{array}{r}n^{-1}\sum_{l=1}^{n} T_l\mathcal{K}_h(S_l - S_i) \\ +n^{-1}\sum_{l=1}^{n} T_l\Delta_{il}\end{array}}\right\},
\end{aligned}
$$

$$A_3 = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} T_j\Delta_{ij} \times \left\{ \frac{m(S_i)}{\pi(S_i)f(S_i)} - \frac{m(S_i)}{\begin{array}{c} n^{-1}\sum_{l=1}^{n} T_l\mathcal{K}_h(S_l - S_i) \\ +n^{-1}\sum_{l=1}^{n} T_l\Delta_{il} \end{array}} + \frac{\begin{array}{c} m(S_i) - \sum_{l=1}^{n} T_l Y_l \mathcal{K}_h(S_l - S_i)/ \\ \sum_{l=1}^{n} T_l\mathcal{K}_h(S_l - S_i) \end{array}}{n^{-1}\sum_{l=1}^{n} T_l\mathcal{K}_h(S_l - S_i) + n^{-1}\sum_{l=1}^{n} T_l\Delta_{il}} \right\}.$$

Using a Taylor expansion around $B^{\mathrm{T}}X_j - B^{\mathrm{T}}X_i$ for Δ_{ij} and plugging in A_1, we have

$$\begin{aligned} A_1 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{T_jY_j\Delta_{ij}}{\pi(S_i)f(S_i)} - \frac{T_jm(S_i)\Delta_{ij}}{\pi(S_i)f(S_i)}\right\} \\ &= \frac{(\hat{B}-B)^{\mathrm{T}}}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{T_j\{Y_j - m(S_i)\}}{\pi(S_i)f(S_i)}\frac{1}{h} \times \left[\mathcal{K}'\left(\frac{B^{\mathrm{T}}X_j - B^{\mathrm{T}}X_i}{h}\right)\frac{X_j - X_i}{h}\right]\right\} + o_p(n^{-1/2}) \\ &= \frac{(\hat{B}-B)^{\mathrm{T}}}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Q_{ij} + o_p(n^{-1/2}). \end{aligned}$$

Denote $A_{11} = (1/n^2)\sum_{i=1}^{n}\sum_{j=1}^{n} Q_{ij}$ and $\breve{A}_{11} = (1/n^2)\sum_{i=1}^{n}\sum_{j=1}^{n} E(Q_{ij} \mid X_i, Y_i, T_i)$. Simple calculation entails that

$$\begin{aligned} &E\left\{\frac{1}{h}T_j\mathcal{K}'\left(\frac{S_j - S_i}{h}\right)\left(\frac{X_j - X_i}{h}\right) \,\middle|\, (X_i, Y_i, T_i) = (x_i, y_i, t_i)\right\} \\ &= -E(TX \mid S = s_i)f'(s_i) - x_i\pi(s_i)f'(s_i) - x_i\pi'(s_i)f(s_i) \\ &\quad + \frac{\partial\{E(TX \mid S = t)\}}{\partial t}\bigg|_{t=s_i} f(s_i) + o_p(1), \end{aligned}$$

and

$$\begin{aligned} &E\left\{\frac{1}{h}T_jY_j\mathcal{K}'\left(\frac{S_j - s_i}{h}\right)\left(\frac{X_j - x_i}{h}\right) \,\middle|\, (X_i, Y_i, T_i) = (x_i, y_i, t_i)\right\} \\ &= -E(TYX \mid S = s_i)f'(s_i) - x_i\pi(s_i)m(s_i)f'(s_i) \\ &\quad + \frac{\partial E(TYX \mid S = t)}{\partial t}\bigg|_{t=s_i} f(s_i) \\ &\quad - x_i\pi'(s_i)m(s_i)f(s_i) - x_i\pi(s_i)m'(s_i)f(s_i) + o_p(1). \end{aligned}$$

Therefore,

$$\begin{aligned} \breve{A}_{11} &= \frac{1}{n}\sum_{i=1}^{n}\left\{\mathrm{cov}(TX, Y \mid S = s_i)f'(s_i) + \frac{\partial E(TYX \mid S = t)}{\partial t}\bigg|_{t=s_i} f(s_i) - \frac{\partial E(TX \mid S = t)}{\partial t}\bigg|_{t=s_i} m(s_i)f(s_i) - x_i\pi(s_i)m'(s_i)f(s_i)\right\} + o_p(1) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\left\{\mathrm{cov}(TX, Y \mid S = s_i)f'(s_i) + \frac{\partial \mathrm{cov}(TX, Y \mid S = t)}{\partial t}\bigg|_{t=s_i} f(s_i) + E(TX \mid S = s_i)m'(s_i)f(s_i) - x_i\pi(s_i)m'(s_i)f(s_i)\right\} + o_p(1) \\ &= (c_1)_{p\times 1} + o_p(1), \end{aligned}$$

where

$$\begin{aligned} c_1 &= -E\left\{\frac{\mathrm{cov}(TX, Y \mid S)f'(S) + \partial\,\mathrm{cov}(TX, Y \mid S)/\partial S f(S)}{\pi(S)f(S)}\right\} + E\left[\frac{\{E(TX \mid S) - X\pi(S)\}m'(S)}{\pi(S)}\right] \\ &= E\left[\frac{\partial\{\pi(S)^{-1}\}}{\partial S}\mathrm{cov}(TX, Y \mid S) - \frac{\mathrm{cov}(T, X \mid S)m'(S)}{\pi(S)}\right]. \end{aligned}$$

It can be seen that the first term in c_1 will be equal to 0 if $Y_1 \perp X \mid S$, while the second term in c_1 will be equal to 0 if $T \perp X \mid S$. Thus, it leads to $c_1 = 0$ when both $Y_1 \perp X \mid S$ and $T \perp X \mid S$ hold.

Let $A_{11j} = (1/n)\sum_{i=1}^{n} Q_{ij}$ and $\breve{A}_{11j} = (1/n)\sum_{i=1}^{n} E(Q_{ij} \mid X_i, Y_i, T_i)$. We have

$$\begin{aligned} &E(A_{11} - \breve{A}_{11})^2 \\ &= \frac{1}{n^2}\sum_{j=1}^{n} E(A_{11j} - \breve{A}_{11j})^2 + \frac{2}{n(n-1)}\sum_{j\neq k} E(A_{11j} - \breve{A}_{11j})E(A_{11k} - \breve{A}_{11k}) \\ &= \frac{1}{n}E(A_{11j} - \breve{A}_{11j})^2 = \frac{1}{n}\{E(A_{11j}^2) - E(\breve{A}_{11j}^2)\} \\ &\leq \frac{1}{n}E(A_{11j}^2) = o_p(1). \end{aligned}$$

Thus, $A_{11} = c_1 + o_p(1)$, which leads to

$$n^{1/2}A_1 = c_1^{\mathrm{T}}\{n^{1/2}(\hat{B} - B)\} + o_p(1).$$

For A_2, we also use a Taylor expansion for Δ_{ij}:

$$\begin{aligned} A_2 &= -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{T_jY_j\Delta_{ij}}{\pi(S_i)f(S_i)} - \frac{T_jY_j\Delta_{ij}}{\begin{array}{c} n^{-1}\sum_{l=1}^{n} T_l\mathcal{K}_h(S_l - S_i) \\ +n^{-1}\sum_{l=1}^{n} T_l\Delta_{il} \end{array}}\right\} \\ &= -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{T_jY_j}{h}\mathcal{K}'\left(\frac{B^{\mathrm{T}}X_j - B^{\mathrm{T}}X_i}{h}\right) \times (\hat{B} - B)^{\mathrm{T}}\left(\frac{X_j - X_i}{h}\right) \times \left\{\frac{1}{\pi(S_i)f(S_i)} - \frac{1}{\begin{array}{c} n^{-1}\sum_{l=1}^{n} T_l\mathcal{K}_h(S_l - S_i) \\ +n^{-1}\sum_{l=1}^{n} T_l\Delta_{il} \end{array}}\right\}\right] \\ &\quad + o_p(n^{-1/2}). \end{aligned}$$

We then decompose A_2 by conditioning on indexes i, j, that is, we define

$$\breve{A}_2 = -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\frac{T_jY_j}{h}\mathcal{K}'\left(\frac{B^{\mathrm{T}}X_j - B^{\mathrm{T}}X_i}{h}\right)\right.$$
$$\times (\hat{B}-B)^{\mathrm{T}}\left(\frac{X_j - X_i}{h}\right)\times E\left\{\frac{1}{\pi(S_i)f(S_i)}\right.$$
$$\left.\left. - \frac{1}{\begin{array}{c} n^{-1}\sum_{l=1}^{n} T_l\mathcal{K}_h(S_l - S_i) + \\ n^{-1}\sum_{l=1}^{n} T_l\Delta_{il}\end{array}}\,\middle|\, X_i, Y_i, T_i\right\}\right].$$

Since

$$E\left\{\frac{1}{n}\sum_{l=1}^{n} T_l\mathcal{K}_h(S_l - S_i)\mid S_i\right\} = \pi(S_i)f(S_i) + o_p(1),$$
$$E\left\{\frac{1}{n}\sum_{l=1}^{n} T_lY_l\mathcal{K}_h(S_l - S_i)\mid S_i\right\} = \pi(S_i)m(S_i)f(S_i) + o_p(1),$$

using a similar decomposition method as A_1, we can also show $n^{1/2}A_2 \xrightarrow{p} 0$ and $n^{1/2}A_3 \xrightarrow{p} 0$. Theorem 2.1 is proved. ■