Statistical Theory and Related Fields

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

Quantile treatment effect estimation with dimension reduction

Ying Zhang, Lei Wang, Menggang Yu & Jun Shao

To cite this article: Ying Zhang, Lei Wang, Menggang Yu & Jun Shao (2020) Quantile treatment effect estimation with dimension reduction, Statistical Theory and Related Fields, 4:2, 202-213, DOI: 10.1080/24754269.2019.1696645

To link to this article: https://doi.org/10.1080/24754269.2019.1696645

Published online: 28 Nov 2019.

STATISTICAL THEORY AND RELATED FIELDS
2020, VOL. 4, NO. 2, 202–213
https://doi.org/10.1080/24754269.2019.1696645

Taylor & Francis
Taylor & Francis Group

Quantile treatment effect estimation with dimension reduction

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

[a]Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA; [b]School of Statistics and Data Science & LPMC, Nankai University, Tianjin, People's Republic of China; [c]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA

ABSTRACT

Quantile treatment effects can be important causal estimands in evaluation of biomedical treatments or interventions for health outcomes such as medical cost and utilisation. We consider their estimation in observational studies with many possible covariates under the assumption that treatment and potential outcomes are independent conditional on all covariates. To obtain valid and efficient treatment effect estimators, we replace the set of all covariates by lower dimensional sets for estimation of the quantiles of potential outcomes. These lower dimensional sets are obtained using sufficient dimension reduction tools and are outcome specific. We justify our choice from efficiency point of view. We prove the asymptotic normality of our estimators and our theory is complemented by some simulation results and an application to data from the University of Wisconsin Health Accountable Care Organization.

ARTICLE HISTORY
Received 19 August 2019
Revised 23 September 2019
Accepted 20 November 2019

KEYWORDS
Causality; efficiency bound; propensity score; quantile treatment effect; sufficient dimension reduction

1. Introduction

Causal evaluation of treatment or intervention is commonly done by estimating average treatment effect (ATE). However for health outcomes such as medical cost and utilisation, quantile treatment effect (QTE) may be more relevant and informative (Abadie, Angrist, & Imbens, 2002; Cattaneo, 2010; Chernozhukov & Hansen, 2005; Doksum, 1974; Firpo, 2007; Frölich & Melly, 2010, 2013; Lehman, 1975). As outcomes tend to be highly skewed to the right, ATE may not be a proper representative parameter for location. Furthermore, it is often important to learn about distributional impacts beyond ATE, such as the effects on upper (or lower) quantiles of an outcome, which may be of direct interests to policy makers and other stakeholders of a programme.

Our study of QTE is motivated by the following investigation at the University of Wisconsin (UW) Health System. As of January 1, 2013, the UW Health System became an Accountable Care Organization (ACO), which is a network of doctors, clinics and other health care providers that share financial and medical responsibility for providing coordinated care to patients in hopes of limiting unnecessary spending. One strategy pursued by nearly all ACOs is to manage the care to 'high-need, high-cost' patients: those with multiple or complex conditions, often combined with behavioural health problems or socioeconomic challenges. In particular, we are asked to evaluate a particular intervention used by the UW Health System. If the intervention can reduce the upper quantiles of health care utilisation quantified by medical cost, then the next step is to significantly enhance the nurse team so that intervention can be extended to a wider population. In essence, we need to estimate QTEs particularly at an upper level.

To define QTE, we begin with some notation. Let T be a binary treatment indicator, X be a p-dimensional vector of pretreatment covariates, and Y_0 and Y_1 be the potential outcomes under treatments $T = 0$ and $T = 1$, respectively. Since only one treatment is applied, either Y_1 or Y_0 is observed, but not both, i.e. what we observe is $Z = TY_1 + (1 - T)Y_0$. With a fixed $\tau \in (0, 1)$, the $100\tau\%$ QTE is defined as $\theta = q_{1,\tau} - q_{0,\tau}$, where $q_{k,\tau}$ is the τth quantile of Y_k, $k = 0, 1$; e.g. $\tau = 0.5, 0.25$, and 0.75 give the difference of medians, lower quartiles, and upper quartiles, respectively. We focus on the estimation of θ based on a random sample $\{Z_i, X_i, T_i : i = 1, \ldots, n\}$ of n observations from (Z, X, T).

Because we only observe Z, θ is often not identifiable without any condition. Throughout we assume the following assumption that is believed to be reasonable in many applications (Rosenbaum & Rubin, 1983): $T \perp (Y_0, Y_1) \mid X$, i.e. T and the vector of potential outcomes (Y_0, Y_1) are independent conditional on X, which is similar to the ignorable missingness assumption when we treat T as a missingness indicator and unobserved Y_0 or Y_1 as a missing value. Under this assumption, two types of consistent estimators of QTE θ in causal inference or closely related context in missing data have been proposed in the literature. One type is derived through

CONTACT Lei Wang lwangstat@nankai.edu.cn

regression on $(T = k, X)$ for $k = 0, 1$ (Cattaneo, 2010; Chen, Wan, & Zhou, 2015; Cheng & Chu, 1996; Zhou, Wan, & Wang, 2008), and the other type is based on inverse propensity weighting with propensity score $P(T = 1 \mid X)$ (Firpo, 2007). A review is given by Cattaneo, Drukker, and Holland (2013). Since parametric methods rely on correct model specifications, nonparametric estimation of the regression functions or propensity is often preferred and therefore considered in what follows.

In our ACO data, however, the dimension p of X is high and nonparametric estimation of regression or propensity function using for example the kernel method is asymptotically inefficient when Y_k is related with only a lower dimensional function of X. Unnecessarily using a high dimensional X may also affect kernel estimation numerically. Our main task is studying covariate dimension reduction to facilitate stable and efficient estimation of QTE.

If inverse propensity weighting is applied, it seems that covariate dimension reduction is to find a linear function S_T of X with the smallest dimension such that $T \perp X \mid S_T$. Unfortunately, Hahn (1998) indicated that in the estimation of ATE, using such an S_T provides no improvement in estimation efficiency over using the entire X. Because the outcome (Y_1, Y_0) is involved in the estimation of ATE, Hahn (2004) suggested finding a linear function S_{Y_0,Y_1} of X with the smallest possible dimension such that $(Y_0, Y_1) \perp X \mid S_{Y_0,Y_1}$, which also implies $T \perp (Y_0, Y_1) \mid S_{Y_0,Y_1}$. The resulting ATE estimator is asymptotically more efficient than the estimator using the entire X unless $S_{Y_0,Y_1} = X$. De Luna, Waernbaum, and Richardson (2011) further considered an $S_{\min}$ which removes the components in S_{Y_0,Y_1} that are unrelated to T. This $S_{\min}$ is the smallest dimensional $S \subseteq S_{Y_0,Y_1}$ that satisfies $T \perp S_{Y_0,Y_1} \mid S$, which also implies $T \perp (Y_0, Y_1) \mid S_{\min}$. However, it is proved in the Appendix that the asymptotic variance using $S_{\min}$ is larger than that of S_{Y_0,Y_1} unless $S_{\min} = S_{Y_0,Y_1}$; see also Brookhart et al. (2006), Shortreed and Ertefaie (2017).

Note that the estimation of $\theta = q_{1,\tau} - q_{0,\tau}$ can be done by estimating $q_{0,\tau}$ and $q_{1,\tau}$ separately and then taking the difference. If a linear function S_{Y_k} of X satisfies $Y_k \perp X \mid S_{Y_k}$ and has the smallest dimension, then S_{Y_k} has a dimension no larger than that of S_{Y_0,Y_1} defined in Hahn (2004), $k = 0, 1$. Hence, our approach alleviates the curse of dimensionality more and it produces asymptotically more efficient estimator of θ.

In applications, S_{Y_0} and S_{Y_1} have to be estimated using observed data. We adopt the existing nonparametric sufficient dimension reduction methods (Cook & Weisberg, 1991; Li, 1991; Xia, Tong, Li, & Zhu, 2002) to construct estimators $\hat{S}_{Y_k}$ of S_{Y_k}. We establish the asymptotic normality for our estimator of θ based on $\hat{S}_{Y_0}$ and $\hat{S}_{Y_1}$, and compare its efficiency with an asymptotic efficiency bound. We also compare the performances of various estimators in simulation studies and apply our method to the medical cost data from the UW Health System.

2. Methods

Without dimension reduction, three types of nonparametric estimators for θ have been proposed in the literature. The inverse propensity weighting (IPW) method (Firpo, 2007) is a weighed version of the procedure in Koenker and Bassett (1978) for the quantile estimation problem.

The regression (REG) method (Cattaneo, 2010; Chen et al., 2015) estimates the function $m_k(x,t) = E\{\rho(Y_k, t) \mid X = x\} = E\{\rho(Z, t) \mid T = k, X = x\}$ by $\hat{m}_k(x,t)$ using a nonparametric method and data under $T = k$ for $k = 0, 1$ separately, where $\rho(s,t) = (s - t)(\tau - \mathbf{1}\{s \le t\})$ is the check function (Koenker & Bassett, 1978) and $\mathbf{1}\{\cdot\}$ is the indicator function. Finally, Cattaneo et al. (2013) and Chen et al. (2015) combined IPW and REG to obtain the so-called augmented inverse propensity weighting (AIPW) estimator.

For each k, let $S_{Y_k} = B_k^\top X$ with $Y_k \perp X \mid S_{Y_k}$, where $B_k^\top$ denotes the transpose of a $p \times d_k$ deterministic matrix with the smallest possible d_k, $k = 0, 1$. As a consequence of Theorem 2.1 stated below, estimators using S_{Y_k} as covariate sets are asymptotically more efficient than those using X as covariate set when $d_k < p$ (if $d_k = p$, then $S_{Y_k} = X$). In the estimation of ATE, Hahn (2004) recommended to replace X by S_{Y_0,Y_1}, but the dimension of S_{Y_0,Y_1} is no smaller than that of S_{Y_k}, which leads to efficiency loss as a consequence of Theorem 2.1.

In applications, S_{Y_k} has to be estimated by $\hat{S}_{Y_k} = \hat{B}_k^\top X$, and we adopt a nonparametric sufficient dimension reduction method to construct $\hat{B}_k$ (Cook & Weisberg, 1991; Li, 1991; Ma and Zhu, 2012; Xia et al., 2002). Since the distribution of $Y_k \mid X$ is the same as $Z \mid X, T = k$, we separately estimate S_{Y_k} using the observed data (Z_i, X_i) in group $T = k$. To estimate the dimensions of S_{Y_0} and S_{Y_1}, we adopt consistent criteria such as BIC-type criteria introduced by Zhu, Zhu, and Feng (2010) and bootstrap based criteria.

Let $\hat{S}_{Y_k,i} = \hat{B}_k^\top X_i$, $i = 1, \ldots, n$, $k = 0, 1$. In our IPW method, we estimate the propensity $\pi_k(s_k) = P(T = k \mid S_{Y_k} = s_k)$ by $\hat{\pi}_k(s_k)$ using a nonparametric method for $k = 0, 1$ separately. The IPW estimator of θ is $\hat{\theta}_{\text{IPW}} = \hat{q}_{1,\tau}^{\text{IPW}} - \hat{q}_{0,\tau}^{\text{IPW}}$, where

$$\hat{q}_{k,\tau}^{\text{IPW}} = \text{argmin}_t \sum_{i=1}^{n} \frac{T_i^{(k)} \rho(Z_i, t)}{\hat{\pi}_k(\hat{S}_{Y_k,i})}, \quad k = 0, 1, \qquad (1)$$

and $T_i^{(1)} = T_i$, $T_i^{(0)} = 1 - T_i$.

The REG method estimates $m_k(s_k, t) = E\{\rho(Y_k, t) \mid S_{Y_k} = s_k\}$ by $\hat{m}_k(s_k, t)$ using a nonparametric method for $k = 0, 1$ separately, and estimates θ by $\hat{\theta}_{\text{REG}} =$

$\hat{q}^{\mathrm{REG}}_{1,\tau} - \hat{q}^{\mathrm{REG}}_{0,\tau}$, where

$$\hat{q}^{\mathrm{REG}}_{k,\tau} = \mathrm{argmin}_t \sum_{i=1}^{n} \hat{m}_k(\hat{S}_{Y_k,i}, t), \quad k = 0, 1. \tag{2}$$

We can combine IPW and REG to obtain our AIPW estimator, $\hat{\theta}_{\mathrm{AIPW}} = \hat{q}^{\mathrm{AIPW}}_{1,\tau} - \hat{q}^{\mathrm{AIPW}}_{0,\tau}$, where

$$\hat{q}^{\mathrm{AIPW}}_{k,\tau} = \mathrm{argmin}_t \sum_{i=1}^{n} \left[\frac{T_i^{(k)} \rho(Z_i, t)}{\hat{\pi}_k(\hat{S}_{Y_k,i})} - \frac{T_i^{(k)} - \hat{\pi}_k(\hat{S}_{Y_k,i})}{\hat{\pi}_k(\hat{S}_{Y_k,i})} \hat{m}_k(\hat{S}_{Y_k,i}, t) \right], \quad k = 0, 1. \tag{3}$$

To estimate $m_k(s_k, t)$ and $\pi_k(s_k)$ in (1)–(3), we use the nonparametric kernel estimators (Silverman, 1986):

$$\hat{m}_k(s_k, t) = \frac{\sum_{i=1}^n T_i^{(k)} \rho(Z_i, t) \mathcal{K}_{H_k}(\hat{S}_{Y_k,i} - s_k)}{\sum_{i=1}^n T_i^{(k)} \mathcal{K}_{H_k}(\hat{S}_{Y_k,i} - s_k)},$$

$$\hat{\pi}_k(s_k) = \frac{\sum_{i=1}^n T_i^{(k)} \mathcal{K}_{H_k}(\hat{S}_{Y_k,i} - s_k)}{\sum_{i=1}^n \mathcal{K}_{H_k}(\hat{S}_{Y_k,i} - s_k)}, \quad k = 0, 1,$$

where $\mathcal{K}_{H_k}(s_k) = \det(H_k^{-1}) \mathcal{K}_k(H_k^{-1} s_k)$, $\mathcal{K}_k(\cdot)$ is a d_k-dimensional kernel function, d_k is the dimension of $\hat{S}_{Y_k}$, and H_k is the bandwidth matrix. When $\hat{S}_{Y_k}$ is standardised, we consider $H_k = h_{kn} I_{d_k}$ with scalar bandwidth h_{kn} and identity matrix I_{d_k} (Hardle, Muller, Sperlich, & Werwatz, 2004). As in Hu, Follmann, and Wang (2014), the nonparametric kernel estimators are computed using the rth order Gaussian product kernel with standardised covariates. The bandwidth we used here is $h_{kn} = Cn^{-2/(2r_k + d_k)}$, where r_k is the order of $\mathcal{K}_k$, $k = 0, 1$. To determine the constant C we adopt the J-fold cross validation, i.e. we select C that minimises $\sum_{j=1}^J (\hat{\theta} - \hat{\theta}_{-j})^2$, where J is the total number of folds and $\hat{\theta}_{-j}$ is the estimator of θ with all data but not those in the jth fold, $j = 1, \ldots, J$. We use $J = 10$ in our simulations in Section 3.

The following theorem establishes the asymptotic normality of estimators in (1)–(3) and assesses the efficiency of estimators.

Theorem 2.1: *Assume the conditions stated in the Appendix. Let $\hat{\theta}(S_0, S_1)$ be one of $\hat{\theta}_{\mathrm{IPW}}$, $\hat{\theta}_{\mathrm{REG}}$, and $\hat{\theta}_{\mathrm{AIPW}}$ in (1)–(3) with $\hat{S}_{Y_k}$ replaced by $S_k = B_k^\top X$ satisfying $Y_k \perp X \mid S_k$, $k = 0, 1$, and let $\hat{\theta}(\hat{S}_0, \hat{S}_1)$ be the same estimator with S_k replaced by its estimator $\hat{S}_k = \hat{B}_k^\top X$, where $\sqrt{n}\mathrm{vec}(\hat{B}_k - B_k) = n^{-1/2} \sum_{i=1}^n \psi_k(X_i, Z_i, T_i) + o_p(1)$ for some functions ψ_k with $E(\psi_k(X, Z, T)) = 0$, $k = 0, 1$, $\mathrm{vec}(M)$ is a column vector whose components are elements of a matrix M, and $o_p(1)$ denotes a quantity converging to 0 in probability. Then we have the following conclusions.*

(i) $\sqrt{n}\{\hat{\theta}(S_0, S_1) - \theta\}$ is asymptotically normal with mean 0 and variance

$$V^*_{S_0,S_1} = \mathrm{var}\left\{E(g_1(Y_1) \mid S_1) - E(g_0(Y_0) \mid S_0)\right\} + \sum_{k=0,1} E\left\{ \frac{\mathrm{var}(g_k(Y_k) \mid S_k)}{P(T = k \mid S_k)} \right\}, \tag{4}$$

where $g_k(Y_k) = -(\mathbf{1}\{Y_k \le q_{k,\tau}\} - \tau)/f_k(q_{k,\tau})$ and f_k is the p.d.f. of Y_k, $k = 0, 1$.

(ii) $\sqrt{n}\{\hat{\theta}(\hat{S}_0, \hat{S}_1) - \theta\}$ is asymptotically normal with mean 0 and variance

$$V_{S_0,S_1} = V^*_{S_0,S_1} + \mathrm{var}\left\{ \sum_{k=0,1} c_k^\top \psi_k(X, Z, T) \right\} + 2\mathrm{cov}\left\{ \sum_{k=0,1} c_k^\top \psi_k(X, Z, T), S(X, Z, T) \right\}, \tag{5}$$

where

$$c_k = -\mathrm{vec}\left(E\left[\frac{\mathrm{cov}(X, T \mid S_k)}{\pi_k(S_k)} \left\{ \frac{\partial E(g_k(Y_k)|S_k)}{\partial S_k} \right\}^\top \right] \right),$$

and

$$S(X, Z, T) = \sum_{k=0,1} (-1)^{(k-1)} \left[\frac{T^{(k)}}{\pi_k(S_k)} \{g_k(Y_k) - E(g_k(Y_k) \mid S_k)\} + E(g_k(Y_k) \mid S_k) \right].$$

Theorem 2.1(i) justifies our choice of $S_k = S_{Y_k}$. $V^*_{S_0,S_1}$ in (A1) is in fact the semiparametric efficiency bound of estimating θ following the ideas in Bickel, Klaassen, Ritov, and Wellner (1993), Hahn (1998) and Firpo (2007). Details can be found in Lemma A.1 in the Appendix. However, in practice, the IPW estimator may not have enough estimation efficiency, as it does not fully extract the information contained in the auxiliary variables. While, the REG and AIPW estimators use all observed covariates to improve estimation efficiency.

By (A1) and Jensen's inequality, among all linear functions (S_0, S_1) satisfying $Y_k \perp X \mid S_k$, $k = 0, 1$, $V^*_{S_0,S_1}$ is minimised when S_k has the smallest possible dimension, i.e. $S_k = S_{Y_k}$, $k = 0, 1$. In particular, this applies to $S_0 = S_1 = S_{Y_0,Y_1}$ proposed in Hahn (2004), since the dimension of S_{Y_0,Y_1} is no smaller than that of S_{Y_k}.

The sum of last two terms on the right hand side of (5) quantifies the price we may pay for estimating B_k by $\hat{B}_k$. There is an efficiency loss due to estimating S_{Y_k} by $\hat{S}_{Y_k}$ when this sum is positive, while it is possible that this sum is negative so that we have an efficiency gain. If we further include the covariates related to T, i.e. consider $S_{Y_k,T}$ being the

smallest possible dimensional S_k satisfying $T \perp X \mid S_k$ and $Y_k \perp X \mid S_k$, $k = 0, 1$, then $\text{cov}(X, T \mid S_k) = 0$ and $c_k = 0$, hence, $\hat{\theta}(\hat{S}_{Y_0,T}, \hat{S}_{Y_1,T})$ and $\hat{\theta}(S_{Y_0,T}, S_{Y_1,T})$ are asymptotically equivalent. However, it is generally not a good idea to use $(\hat{S}_{Y_0,T}, \hat{S}_{Y_1,T})$, because each $S_{Y_k,T}$ has a dimension no smaller than that of S_{Y_k} and therefore both $\hat{\theta}(\hat{S}_{Y_0,T}, \hat{S}_{Y_1,T})$ and $\hat{\theta}(S_{Y_0,T}, S_{Y_1,T})$ is less efficient than $\hat{\theta}(S_{Y_0}, S_{Y_1})$ according to Theorem 2.1. Although $\hat{\theta}(\hat{S}_{Y_0}, \hat{S}_{Y_1})$ may be less efficient than $\hat{\theta}(S_{Y_0}, S_{Y_1})$ due to the estimation of S_{Y_k}, it may still be more efficient than $\hat{\theta}(\hat{S}_{Y_0,T}, \hat{S}_{Y_1,T})$. Some simulation results are given in Section 3.

In Theorem 2.1, the condition $\sqrt{n}\text{vec}(\hat{B}_k - B_k) = n^{-1/2}\sum_{i=1}^{n} \psi_k(X_i, Z_i, T_i) + o_p(1)$ with $E\psi_k(X, Z, T) = 0$ is satisfied for $\hat{B}_k$ obtained using some sufficient dimension reduction methods (Hsing & Carroll, 1992; Zhu & Ng, 1995).

3. Simulation

We investigate the finite-sample performance of three estimators, $\hat{\theta}_{\text{IPW}}$, $\hat{\theta}_{\text{REG}}$, and $\hat{\theta}_{\text{AIPW}}$, with four choices of linear functions (S_0, S_1), (1) $S_k = S_{Y_k}$, $k = 0, 1$, (2) $S_0 = S_1 = S_{Y_0,Y_1}$, (3) $S_k = S_{Y_k,T}$, $k = 0, 1$, and (4) $S_0 = S_1 = S_T$. For each choice of (S_0, S_1), we consider estimators using the true (S_0, S_1) as well as $(\hat{S}_0, \hat{S}_1)$ by sufficient dimension reduction. Thus, we consider a total of $3 \times 4 \times 2 = 24$ cases. In each case, we consider the estimation of the QTEs with $\tau = 25\%$, 50%, and 75%, under two different sample sizes $n = 200$ and $n = 1000$.

In the first simulation, $X = (X_1, \ldots, X_7)^\top$ with independent $N(0, 1)$ components, $P(T = 1 \mid X) = \exp(2X_4)\{1 + \exp(2X_4)\}^{-1}$, $Y_0 = 3X_1 + 6X_2 + 3X_3 + \epsilon_0$, and $Y_1 = 10 + 3X_1 + 6X_2 + 3X_3 + 3X_4 + \epsilon_1$, where ϵ_k's are independent $N(0, 1)$ and are independent of X. The outcome models are linear in X, the treatment model is logistic, and the log-conditional treatment odds is linear in X. Under this model, S_{Y_0}, S_{Y_1}, and S_T are all one-dimensional, while $S_{Y_0,Y_1} = S_{Y_1,T} = S_{Y_0,T}$ is two-dimensional.

In the second simulation, $X = (X_1, \ldots, X_7)^\top$ with independent $N(0, 1)$ components, $P(T = 1 \mid X) = \exp(-2X_5 + 0.7X_6^2 - 0.5X_7^2)\{1 + \exp(-2X_5 + 0.7X_6^2 - 0.5X_7^2)\}^{-1}$, $Y_0 = 3(X_1 + X_2 + 2X_3 + 2X_4) + 1.5X_6^2 + \epsilon_0$, and $Y_1 = 12 + 3(X_1 + X_2 + 2X_3 + X_4 + X_5) + 1.5X_7^2 + \epsilon_1$, where ϵ_k's are independent $N(0, 1)$ and are independent of X. The outcome models are nonlinear in X, the treatment model is logistic, and the log-conditional treatment odds is nonlinear in X. Under this setting, each S_{Y_k} is two-dimensional, S_T is three-dimensional, while S_{Y_0,Y_1}, $S_{Y_1,T}$, and $S_{Y_0,T}$ are four-dimensional and not the same.

Based on 1000 simulation runs, we calculate the simulated relative bias (RB) and standard deviation (SD) in each scenario. The results for simulations are given in Tables 1–2, respectively. The following conclusions can be obtained from the simulation results in Tables 1–2.

(1) When the true (S_0, S_1) is used, (S_{Y_0}, S_{Y_1}) leads to the best performance overall, followed by S_{Y_0,Y_1}, $(S_{Y_0,T}, S_{Y_1,T})$, and S_T, in agreement with our

Table 1. Relative bias and standard deviation for simulation 1 with true or estimated S_0 and S_1.

		IPW				REG				AIPW			
		Estimated		True		Estimated		True		Estimated		True	
	S_0, S_1	RB	SD	RB	SD	RB	SD	RB	SD	RB	SD	RB	SD
$n = 200$													
$\theta = 9.6$	S_T, S_T	0.11	1.33	0.09	1.77	0.11	1.38	0.08	1.84	0.09	1.46	0.05	2.04
$\tau = 0.25$	$S_{Y_0,T}, S_{Y_1,T}$	0.09	1.18	0.10	1.08	0.06	1.19	0.08	1.03	0.04	1.26	0.06	1.04
	S_{Y_0,Y_1}, S_{Y_0,Y_1}	0.09	1.09	0.10	1.08	0.09	1.09	0.08	1.03	0.07	1.07	0.06	1.04
	S_{Y_0}, S_{Y_1}	0.06	0.96	0.05	0.87	0.05	0.91	0.04	0.85	0.02	0.84	0.01	0.76
$\theta = 10.0$	S_T, S_T	0.12	1.14	0.08	1.51	0.09	1.19	0.06	1.57	0.08	1.25	0.04	1.72
$\tau = 0.5$	$S_{Y_0,T}, S_{Y_1,T}$	0.10	0.97	0.10	0.92	0.07	1.01	0.07	0.95	0.05	1.00	0.05	0.89
	S_{Y_0,Y_1}, S_{Y_0,Y_1}	0.11	0.94	0.10	0.92	0.08	1.01	0.07	0.95	0.06	0.91	0.05	0.89
	S_{Y_0}, S_{Y_1}	0.06	0.75	0.04	0.68	0.04	0.82	0.03	0.75	0.01	0.71	0.01	0.63
$\theta = 10.4$	S_T, S_T	0.10	1.29	0.07	1.63	0.07	1.32	0.05	1.70	0.06	1.35	0.03	1.82
$\tau = 0.75$	$S_{Y_0,T}, S_{Y_1,T}$	0.08	1.12	0.09	1.08	0.06	1.10	0.06	1.01	0.03	1.05	0.04	0.97
	S_{Y_0,Y_1}, S_{Y_0,Y_1}	0.09	1.10	0.09	1.08	0.06	1.06	0.06	1.01	0.05	0.98	0.04	0.97
	S_{Y_0}, S_{Y_1}	0.05	0.83	0.03	0.75	0.03	0.86	0.02	0.81	0.01	0.74	0.01	0.68
$n = 1000$													
$\theta = 9.6$	S_T, S_T	0.07	0.73	0.06	0.94	0.04	0.83	0.04	1.03	0.02	0.95	0.02	1.17
$\tau = 0.25$	$S_{Y_0,T}, S_{Y_1,T}$	0.08	0.50	0.09	0.50	0.06	0.52	0.06	0.51	0.04	0.52	0.03	0.49
	S_{Y_0,Y_1}, S_{Y_0,Y_1}	0.09	0.49	0.09	0.50	0.05	0.52	0.06	0.51	0.03	0.49	0.03	0.49
	S_{Y_0}, S_{Y_1}	0.02	0.36	0.02	0.35	0.01	0.37	0.01	0.36	0.00	0.33	0.00	0.32
$\theta = 10.0$	S_T, S_T	0.05	0.60	0.04	0.77	0.03	0.66	0.02	0.84	0.01	0.74	0.01	0.92
$\tau = 0.5$	$S_{Y_0,T}, S_{Y_1,T}$	0.07	0.42	0.07	0.42	0.04	0.44	0.04	0.44	0.02	0.42	0.02	0.40
	S_{Y_0,Y_1}, S_{Y_0,Y_1}	0.07	0.42	0.07	0.42	0.03	0.44	0.04	0.44	0.01	0.41	0.02	0.40
	S_{Y_0}, S_{Y_1}	0.02	0.30	0.02	0.29	0.01	0.31	0.01	0.30	0.00	0.29	0.00	0.27
$\theta = 10.4$	S_T, S_T	0.04	0.67	0.04	0.80	0.02	0.73	0.02	0.84	0.01	0.80	0.01	0.93
$\tau = 0.75$	$S_{Y_0,T}, S_{Y_1,T}$	0.06	0.45	0.06	0.44	0.03	0.46	0.03	0.45	0.02	0.44	0.02	0.42
	S_{Y_0,Y_1}, S_{Y_0,Y_1}	0.06	0.44	0.06	0.44	0.03	0.45	0.03	0.45	0.01	0.42	0.02	0.42
	S_{Y_0}, S_{Y_1}	0.01	0.31	0.01	0.30	0.01	0.32	0.01	0.32	0.00	0.30	0.00	0.29

RB: relative bias; SD: standard deviation

Table 2. Relative bias and standard deviation for simulation 2 with true or estimated S_0 and S_1.

		IPW				REG				AIPW			
		Estimated		True		Estimated		True		Estimated		True	
	S_0, S_1	RB	SD	RB	SD	RB	SD	RB	SD	RB	SD	RB	SD
$n = 200$													
$\theta = 12.7$	S_T, S_T	−0.07	1.61	−0.06	1.74	−0.06	1.71	−0.04	1.89	−0.05	1.75	−0.03	1.90
$\tau = 0.25$	$S_{Y_0,T}, S_{Y_1,T}$	−0.06	1.31	−0.07	1.38	−0.05	1.29	−0.06	1.34	−0.03	1.25	−0.05	1.24
	S_{Y_0,Y_1}, S_{Y_0,Y_1}	−0.05	1.25	−0.07	1.36	−0.03	1.22	−0.05	1.31	−0.02	1.15	−0.04	1.19
	S_{Y_0}, S_{Y_1}	−0.02	1.04	−0.04	1.10	−0.01	1.09	−0.02	1.12	0.01	0.97	−0.01	0.96
$\theta = 12.0$	S_T, S_T	−0.08	1.47	−0.08	1.64	−0.07	1.54	−0.06	1.79	−0.06	1.58	−0.05	1.81
$\tau = 0.5$	$S_{Y_0,T}, S_{Y_1,T}$	−0.05	1.16	−0.08	1.27	−0.04	1.16	−0.06	1.22	−0.02	1.16	−0.05	1.15
	S_{Y_0,Y_1}, S_{Y_0,Y_1}	−0.05	1.14	−0.08	1.23	−0.03	1.14	−0.05	1.21	−0.02	1.04	−0.04	1.09
	S_{Y_0}, S_{Y_1}	−0.02	0.95	−0.04	1.03	−0.01	1.00	−0.02	1.01	0.01	0.90	−0.01	0.88
$\theta = 11.3$	S_T, S_T	−0.09	1.63	−0.09	1.81	−0.08	1.68	−0.07	1.94	−0.07	1.75	−0.05	2.03
$\tau = 0.75$	$S_{Y_0,T}, S_{Y_1,T}$	−0.06	1.29	−0.09	1.41	−0.04	1.28	−0.06	1.33	−0.03	1.27	−0.06	1.29
	S_{Y_0,Y_1}, S_{Y_0,Y_1}	−0.06	1.26	−0.08	1.37	−0.04	1.25	−0.06	1.32	−0.02	1.20	−0.05	1.27
	S_{Y_0}, S_{Y_1}	−0.02	1.16	−0.04	1.23	−0.01	1.10	−0.02	1.12	0.00	1.06	−0.01	1.07
$n = 1000$													
$\theta = 12.7$	S_T, S_T	−0.05	0.70	−0.05	0.84	−0.03	0.76	−0.03	0.92	−0.02	0.79	−0.02	0.93
$\tau = 0.25$	$S_{Y_0,T}, S_{Y_1,T}$	−0.03	0.55	−0.06	0.62	−0.01	0.57	−0.04	0.62	0.00	0.53	−0.03	0.55
	S_{Y_0,Y_1}, S_{Y_0,Y_1}	−0.02	0.50	−0.06	0.61	0.00	0.52	−0.03	0.59	0.01	0.47	−0.02	0.50
	S_{Y_0}, S_{Y_1}	0.00	0.43	−0.02	0.47	0.01	0.45	−0.01	0.48	0.02	0.42	0.00	0.42
$\theta = 12.0$	S_T, S_T	−0.07	0.60	−0.07	0.76	−0.05	0.66	−0.04	0.84	−0.03	0.68	−0.03	0.88
$\tau = 0.5$	$S_{Y_0,T}, S_{Y_1,T}$	−0.03	0.50	−0.07	0.54	−0.02	0.50	−0.04	0.55	0.00	0.48	−0.03	0.48
	S_{Y_0,Y_1}, S_{Y_0,Y_1}	−0.02	0.42	−0.07	0.51	−0.01	0.44	−0.03	0.51	0.01	0.41	−0.02	0.45
	S_{Y_0}, S_{Y_1}	0.00	0.38	−0.02	0.40	0.01	0.40	−0.01	0.42	0.01	0.37	0.00	0.38
$\theta = 11.3$	S_T, S_T	−0.08	0.72	−0.08	0.87	−0.06	0.76	−0.06	0.96	−0.04	0.81	−0.04	1.00
$\tau = 0.75$	$S_{Y_0,T}, S_{Y_1,T}$	−0.04	0.60	−0.08	0.64	−0.02	0.59	−0.05	0.61	0.00	0.58	−0.04	0.60
	S_{Y_0,Y_1}, S_{Y_0,Y_1}	−0.03	0.53	−0.08	0.62	−0.01	0.53	−0.04	0.58	0.01	0.50	−0.03	0.55
	S_{Y_0}, S_{Y_1}	0.00	0.45	−0.03	0.50	0.01	0.46	−0.01	0.48	0.01	0.43	0.00	0.44

RB: relative bias; SD: standard deviation

asymptotic results discussed in Section 2 and proved in the Appendix.

(2) When estimator $(\hat{S}_{Y_0}, \hat{S}_{Y_1})$ is used, the resulting estimators of θ are in general less efficient than those based on the true (S_{Y_0}, S_{Y_1}), but they are still better than the estimators based on other choices of (S_0, S_1) regardless of whether the true or estimated (S_0, S_1) is used.

(3) The performances of estimators using the true $(S_{Y_0,T}, S_{Y_1,T})$ and $(\hat{S}_{Y_0,T}, \hat{S}_{Y_1,T})$ are quite similar when $n = 1000$, in agreement with the asymptotic results in Theorem 2.1 and our discussion after Theorem 2.1. They are worse than those using $(\hat{S}_{Y_0}, \hat{S}_{Y_1})$.

(4) Consistent with the asymptotic theory, the performance of estimators using S_T is the worst, and the efficiency loss is substantial in most cases. Note that using estimated S_T is actually better than using the true S_T.

(5) Regarding the three different estimation methods, $\hat{\theta}_{\text{REG}}$ and $\hat{\theta}_{\text{AIPW}}$ have very comparable performances and are recommended in practice.

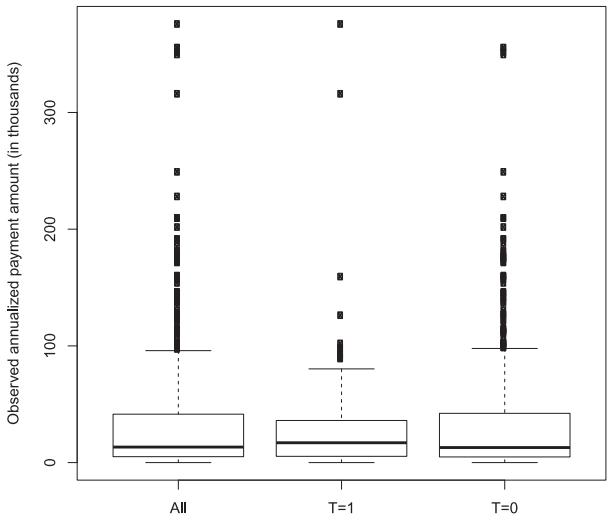

Figure 1. Boxplots of observed annualised payment amount (in thousands) for overall, CCM group $T = 1$, and non-CCM group $T = 0$.

4. Real data analysis

As we mentioned in the introduction, the University of Wisconsin Health System became an Accountable Care Organization (ACO) and implemented a Complex Care Management (CCM) programme since January 1, 2013. In particular, a team of nurses take responsibility for coordinating and implementing complex patients' care plan. The CCM is very intensive in time and resources and therefore it is important to evaluate its specific value.

We demonstrate the proposed estimation methods in a data set resulted from the University of Wisconsin Health ACO study where the primary outcome Z is the annualised payment amount in thousands. The data set consists of 894 patients with 186 in the CCM group $(T = 1)$ and 708 not in the CCM group $(T = 0)$.

Table 3. Estimates and standard errors (SE) for the University of Wisconsin Health ACO data.

		IPW		REG		AIPW	
	S_0, S_1	Estimate	SE	Estimate	SE	Estimate	SE
25% QTE	S_T	0.16	1.00	0.52	1.10	0.39	1.15
	S_{Y_0,Y_1}	0.47	0.91	0.57	0.84	0.47	0.89
	S_{Y_0},S_{Y_1}	0.47	0.92	−0.10	0.81	0.47	0.88
50% QTE	S_T	2.82	2.67	1.59	2.76	3.31	2.64
	S_{Y_0,Y_1}	3.21	2.52	2.34	2.63	3.27	2.46
	S_{Y_0}, S_{Y_1}	3.27	2.41	3.57	2.42	3.27	2.37
75% QTE	S_T	−6.09	4.74	−5.31	4.65	−6.33	4.62
	S_{Y_0,Y_1}	−8.51	3.92	−5.14	3.90	−8.38	3.82
	S_{Y_0}, S_{Y_1}	−8.38	3.83	−8.65	3.72	−8.14	3.76
ATE	S_T	−2.36	3.72	−2.54	3.83	−2.11	3.53
	S_{Y_0,Y_1}, S_{Y_0,Y_1}	−5.68	3.09	−3.41	2.93	−6.03	3.05
	S_{Y_0}, S_{Y_1}	−5.42	2.93	−5.89	2.78	−5.24	2.87

SE: standard error by bootstrapping

Two issues with this dataset actually motivated our study. First, the distribution of annualised payment is right-skewed as shown by the box plots in Figure 1 for all patients and two groups. The overall median, mean, 75% quantile, and maximum of observed payment are about 13, 31, 41, and 376 thousand dollars, respectively. This suggests the need for estimating quantile treatment effects. Second, the dataset consists of three discrete and ninety-four continuous covariates including medicare status, baseline payments, as well as other baseline characteristics of patients. Thus, dimension reduction is needed in nonparametric kernel estimation.

For sufficient dimension reduction, we adopt the semiparametric directional regression method proposed by Ma and Zhu (2012). After sufficient dimension reduction, S_T has 7 dimensions, S_{Y_0} has 5 dimensions, S_{Y_1} has 8 dimensions, and S_{Y_0,Y_1} has 13 dimensions.

Results for three choices of (S_0, S_1) considered in simulation are shown in Table 3 for estimating ATE and QTE with $\tau = 25\%, 50\%$, and 75%. Standard errors (SE) for all estimates are calculated using the bootstrap with 200 replications.

From Table 3, the 25% and 50% QTEs are not significant by all methods. When S_{Y_k} or S_{Y_0,Y_1} is used, the 75% QTE is significantly less than 0, and in terms of SE, the estimate using S_{Y_k}, $k = 0, 1$, is more efficient than the estimate using S_{Y_0,Y_1}. However, the estimate of 75% QTE using S_T is inefficient due to the large variation of using S_T so that the result is insignificant.

Since 75% QTE is significantly negative, the result indicates that receiving CCM intervention effectively helps reducing medical payment for the high-cost patients. But CCM intervention is not so useful for the low-cost or median-cost patients, as 25% and 50% QTEs are not significant. These results may be useful for decision making in ACO.

For comparison, we also include estimates of ATE and SE. The results in the last block of Table 3 show that ATE is not significant by all methods. It is interesting to see that estimates of ATE are all negative whereas estimates of 50% QTE are all positive although they are not significant, which may be caused by fact that the distribution of annualised payment is right-skewed.

The example shows the usefulness of assessing QTEs with different percentages. If we only estimate ATE, no useful conclusion can be made in this example. Even if we check 50% QTE instead of ATE because of the existing skewness, we still cannot obtain any useful conclusion.

Acknowledgments

We are grateful to the editor, the associate editor, and two anonymous referees for their insightful comments and suggestions, which have led to significant improvements.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

Our research was supported by the National Natural Science Foundation of China (11871287, 11831008), the Natural Science Foundation of Tianjin (18JCYBJC41100), the Fundamental Research Funds for the Central Universities, the Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin, the Chinese 111 Project (B14019), the U.S. National Science Foundation (DMS-1612873 and DMS-1914411). This research was partially supported through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1409-21219).

Notes on contributors

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

Dr Lei Wang holds a Ph.D. 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.

Dr Menggang Yu holds a Ph.D. 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 Jun Shao holds a Ph.D. 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.

References

Abadie, A., Angrist, J., & Imbens, G. W. (2002). Instrumental variables estimates of the effect of subsidized training on the quantiles of trainee earnings. *Econometrica*, *70*, 91–117.

Bickel, P. J., Klaassen, C. J., Ritov, Y., & Wellner, J. (1993). *Efficient and adaptive inference in semiparametric models*. Baltimore: Johns Hopkins University Press.

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

Cattaneo, M. D. (2010). Efficient semiparametric estimation of multi-valued treatment effects under ignorability. *Journal of Econometrics*, *155*, 138–154.

Cattaneo, M. D., Drukker, D. M., & Holland, A. D. (2013). Estimation of multivalued treatment effects under conditional independence. *The Stata Journal*, *13*, 407–450.

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.

Cheng, P. E., & Chu, C. (1996). Kernel estiamation of distribution functions and quantiles with missing data. *Statistica Sinica*, *6*, 63–78.

Chernozhukov, V., & Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica*, *73*, 245–261.

Cook, R. D., & Weisberg, S. (1991). Discussion of 'Sliced inverse regression for dimension reduction'. *Journal of the American Statistical Association*, *86*, 328–332.

De Luna, X., Waernbaum, I., & Richardson, T. S. (2011). Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*, *98*, 861–875.

Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics*, *2*, 267–277.

Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica*, *75*, 259–276.

Frölich, M., & Melly, B. (2010). Estimation of quantile treatment effects with Stata. *The Stata Journal*, *10*, 423–457.

Frölich, M., & Melly, B. (2013). Unconditional quantile treatment effects under endogeneity. *Journal of Business and Economic Statistics*, *31*, 346–357.

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.

Hardle, W., Muller, M., Sperlich, S., & Werwatz, A. (2004). *Nonparametric and semiparametric models*. Heidelberg: Springer-Verlag.

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.

Koenker, R., & Bassett, G. (1978). Regression quantiles. *Econometrica*, *46*, 33–50.

Lehman, E. L. (1975). *Nonparametrics: Statistical methods based on ranks*. San Francisco: Holden-Day.

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

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

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

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. London: Chapman and Hall.

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.

Zhou, Y., Wan, A. T. K., & Wang, X. (2008). Estimating equations inference with missing data. *Journal of the American Statistical Association*, *103*, 1187–1199.

Zhu, L. X., & Ng, K. W. (1995). Asymptotics of sliced inverse regression. *Statistica Sinica*, *5*, 727–736.

Zhu, L. P., Zhu, L. X., & Feng, Z. H. (2010). Dimension reduction in regressions through cumulative slicing estimation. *Journal of the American Statistical Association*, *105*, 1455–1466.

Appendices

Appendix 1. Semiparametric efficiency bound of estimating θ with S_k

Throughout the Appendix, the S, S_k, S_{Y_k}, S_{Y_0,Y_1}, S_T, $S_{\min}$ are linear functions of X, i.e. $S = B^\top X$ with B being a $p \times d$ matrix, $S_k = B_k^\top X$ with B_k being a $p \times d_k$ matrix, etc.

Lemma A.1: *Assume $T \perp (Y_0, Y_1) \mid X$ and $Y_k \perp X \mid S_k$, and the distribution of Y_k has density f_k with $f_k(q_{k,\tau}) > 0$, $k = 0, 1$. A lower bound for the asymptotic variance of any asymptotically normal estimator of $\theta = q_{1,\tau} - q_{0,\tau}$ is given by*

$$V^*_{S_0,S_1} = \operatorname{var}\left\{E(g_1(Y_1)|S_1) - E(g_0(Y_0)|S_0)\right\} + \sum_{k=0,1} E\left\{\frac{\operatorname{var}(g_k(Y_k) \mid S_k)}{P(T=k \mid S_k)}\right\}, \tag{A1}$$

where $g_k(Y_k) = -(\mathbf{1}\{Y_k \le q_{k,\tau}\} - \tau)/f_k(q_{k,\tau})$, $k = 0, 1$. If $Y_k \perp X \mid S_k$, $Y_k \perp X \mid S_k'$, and $\mathcal{L}(S_k) \subseteq \mathcal{L}(S_k')$, $k = 0, 1$, then $V^*_{S_0,S_1} \le V^*_{S_0',S_1'}$, where $\mathcal{L}(S)$ denotes the linear space generated by columns of B for $S = B^\top X$.

Proof of Lemma A.1: Our derivation of the efficiency bound mimics the proof in Firpo (2007) which is a direct application of the semiparametric efficiency theory from Bickel et al. (1993). Following the proof of Firpo (2007), one may easily see that knowing $T \mid X = T \mid S_T$ won't change the semiparametric efficiency bound, which is similar with the ATE case in Hahn (1998). In our proof for Lemma A.1, one only needs to carefully keep S_1 and S_0 separate in the derivation. The construction of the efficient influence function is more involved algebraically. We only provide a sketch of the proof for the case $S_k = S_{Y_k}$ here. The density of (Y_0, Y_1, T, X) at (y_0, y_1, k, x) is

$$q(y_0, y_1, k, x) = g(y_0, y_1 \mid x)\pi(x)^k\{1 - \pi(x)\}^{1-k} f(x),$$

where $g(y_0, y_1 \mid x)$ denotes the conditional distribution of (Y_0, Y_1) given X, $f(x)$ denotes the marginal distribution of

X and $\pi(x) = P(T = 1 \mid X = x)$. The density of (Z, T, X) at (z, k, x) is then equal to

$$q(z,k,x) = \{g_1(z \mid x)\pi(x)\}^k \{g_0(z \mid x)(1-\pi(x))\}^{1-k} f(x)$$
$$= \{h_1(z \mid S_{y_1})\pi(x)\}^k \{h_0(z \mid S_{y_0})(1-\pi(x))\}^{1-k} f(x),$$

where $g_1(\cdot \mid x) = \int g(y_0, \cdot \mid x)\, \mathrm{d}y_0$, $g_0(\cdot \mid x) = \int g(\cdot, y_1 \mid x)\, \mathrm{d}y_1$. The second equality holds because by the definition of S_{Y_1}, S_{Y_0}, there exist functions h_1 and h_0 that $g_1(\cdot \mid x) = h_1(\cdot \mid S_{y_1})$ and $g_0(\cdot \mid x) = h_0(\cdot \mid S_{y_0})$. For a regular parametric submodel $q(z, k, x)$ with parameter w,

$$q_\omega(z,k,x) = \{h_1(z \mid S_{y_1}, \omega)\pi(x,\omega)\}^k \times \{h_0(z \mid S_{y_0}, \omega)(1-\pi(x,\omega))\}^{1-k} f(x,\omega).$$

The score function of this parametric submodel is

$$s(z,k,x \mid \omega) = k s_1(z \mid S_{y_1}, \omega) + (1-k) s_0(z \mid S_{y_0}, \omega) + \frac{\{k - \pi(x,\omega)\}\frac{\partial}{\partial\omega}\pi(x,\omega)}{\pi(x,\omega)\{1-\pi(x,\omega)\}} + d(x,\omega),$$

where $d(x,\omega) = f(x,\omega)^{-1}(\partial f(x,\omega)/\partial\omega)$, $s_1(z \mid S_{y_1}, \omega) = h_1(z \mid S_{y_1}, \omega)^{-1}(\partial h_1(z \mid S_{y_1}, \omega)/\partial\omega)$, $s_0(z \mid S_{y_0}, \omega) = h_0(z \mid S_{y_0}, \omega)^{-1}(\partial h_0(z \mid S_{y_0}, \omega)/\partial\omega)$. Therefore, the tangent space is equal to

$$\mathcal{T} = \Big\{ k s_1(z \mid S_{y_1}) + (1-k) s_0(z \mid S_{y_0}) + a(x)(k - \pi(x)) + d(x) : \text{where } (s_0, s_1, d, a) \text{ satisfies } \int s_j(z \mid S_{y_j}) h_j(z \mid S_{y_j})\, \mathrm{d}y = 0, \times \int d(x) f(x)\, \mathrm{d}x = 0 \text{ and } a(x) \text{is unrestricted} \Big\}.$$

For the parametric submodel with parameter ω under consideration, $q_{k,\tau}(\omega)$, the τ-th quantile for $Y_k, k = 0, 1$, satisfies $0 = E_\omega(\mathbf{1}\{Y_k \le q_{k,\tau}(\omega)\} - \tau) = \int\int (\mathbf{1}\{z \le q_{k,\tau}(\omega)\} - \tau) h_k(z \mid S_{Y_k}, \omega)\, \mathrm{d}z f(x,\omega)\, \mathrm{d}x$. Let $\theta(\omega) = q_{1,\tau}(\omega) - q_{0,\tau}(\omega)$, and remember $g_k(Y_k) = -(\mathbf{1}\{Y_k \le q_{k,\tau}\} - \tau)/f_k(q_{k,\tau})$, $k = 0, 1$. By an application of Leibnitz's rule,

$$\frac{\partial\theta(\omega)}{\partial\omega} = \int\int g_1(z) s_1(z \mid S_{y_1}, \omega) h_1(z \mid S_{y_1}, \omega) f(x,\omega)\, \mathrm{d}z\, \mathrm{d}x + \int E_\omega\big(g_1(z) - g_0(z) \mid X = x\big)\, \mathrm{d}(x,\omega) f(x,\omega)\, \mathrm{d}x - \int\int g_0(z) s_0(z \mid S_{y_0}, \omega_0) h_0(z \mid S_{y_0}) f(x,\omega)\, \mathrm{d}z\, \mathrm{d}x.$$

Let

$$F(Z,T,X) = \frac{T\{g_1(Z) - E(g_1(Z) \mid S_{Y_1})\}}{P(T=1 \mid S_{Y_1})} - \frac{(1-T)\{g_0(Z) - E(g_0(Z) \mid S_{Y_0})\}}{1 - P(T=1 \mid S_{Y_0})} + E(g_1(Z) - g_0(Z) \mid X),$$

and the true parameter ω is $\omega = \omega_0$, i.e. $\theta = \theta(\omega_0)$, then we have

$$E\{F(Z,T,X)s(Z,T,X \mid \omega_0)\} = E\left[\frac{T\{g_1(Z) - E(g_1(Z) \mid S_{Y_1})\}}{P(T=1 \mid S_{Y_1})} s(Z,T,X \mid \omega_0)\right] - E\left[\frac{(1-T)\{g_0(Z) - E(g_0(Z) \mid S_{Y_0})\}}{1 - P(T=1 \mid S_{Y_0})} s(Z,T,X \mid \omega_0)\right] + E\big[E(g_1(Z) - g_0(Z) \mid X) s(Z,T,X \mid \omega_0)\big]. \quad \text{(A2)}$$

For the three terms in (A2), after some algebra, we have, respectively,

$$E\left[\frac{T\{g_1(Z) - E(g_1(Z) \mid S_{Y_1})\}}{P(T=1 \mid S_{Y_1})} s(Z,T,X \mid \omega_0)\right]$$
$$= E[\{g_1(Y_1) - E(g_1(Y_1) \mid S_{Y_1})\} s_1(Y_1 \mid S_{Y_1}, \omega_0)],$$
$$\times E\left[\frac{(1-T)\{g_0(Z) - E(g_0(Z) \mid S_{Y_0})\}}{1 - P(T=1 \mid S_{Y_0})} s(Z,T,X \mid \omega_0)\right]$$
$$= E[\{g_0(Z) - E(g_0(Z) \mid S_{Y_0})\} s_0(Y_0 \mid S_{Y_0}, \omega_0)],$$
$$\times E[\{E(g_1(Z) - g_0(Z) \mid X)\} s(Z,T,X \mid \omega_0)]$$
$$= E\{E(g_1(Y_1) - g_0(Y_0) \mid X) d(X, \omega_0)\}.$$

Therefore, $E\{F(Z,T,X)s(Z,T,X \mid \omega_0)\} = \partial\theta(\omega_0)/\partial\omega$. The efficiency bound is the expected square of the projection of F on $\mathcal{T}$. Because $F \in \mathcal{T}$, the projection of F on $\mathcal{T}$ is itself. The conclusion follows. ■

For the second part of Lemma A.1, suppose S_0, S_1 satisfy $\mathcal{L}(S_0) \supseteq \mathcal{L}(S_{Y_0})$, $\mathcal{L}(S_1) \supseteq \mathcal{L}(S_{Y_1})$. Since

$$V^*_{S_{Y_1}, S_{Y_0}} = \mathrm{Var}\{E(g_1(Y_1) \mid X) - E(g_0(Y_0) \mid X)\} + E\left\{\frac{\mathrm{Var}(g_1(Y_1) \mid X)}{P(T=1 \mid S_{Y_1})}\right\} + E\left\{\frac{\mathrm{Var}(g_0(Y_0) \mid X)}{P(T=0 \mid S_{Y_0})}\right\},$$
$$V^*_{S_1, S_0} = \mathrm{Var}\{E(g_1(Y_1) \mid X) - E(g_0(Y_0) \mid X)\} + E\left\{\frac{\mathrm{Var}(g_1(Y_1) \mid X)}{P(T=1 \mid S_1)}\right\} + E\left\{\frac{\mathrm{Var}(g_0(Y_0) \mid X)}{P(T=0 \mid S_0)}\right\}.$$

We only need to prove

$$E\left\{\frac{\mathrm{Var}(g_1(Y_1) \mid X)}{P(T=1 \mid S_{Y_1})}\right\} \le E\left\{\frac{\mathrm{Var}(g_1(Y_1) \mid X)}{P(T=1 \mid S_1)}\right\}.$$

By Jensen's inequity, we have

$$\frac{1}{E\{E(T \mid S_1) \mid S_{Y_1}\}} \le E\left\{\frac{1}{E(T \mid S_1)} \,\middle|\, S_{Y_1}\right\}.$$

Thus the conclusion follows from the inequality below.

$$E\left[\frac{\mathrm{Var}(g_1(Y_1) \mid X)}{P(T=1 \mid S_{Y_1})}\right] = E\left[\frac{\mathrm{Var}(g_1(Y_1) \mid X)}{E\{E(T \mid S_1) \mid S_{Y_1}\}}\right] \le E\left[\mathrm{Var}(g_1(Y_1) \mid X) E\left\{\frac{1}{E(T \mid S_1)} \,\middle|\, S_{Y_1}\right\}\right] = E\left[E\left\{\frac{\mathrm{Var}(g_1(Y_1) \mid X)}{E(T \mid S_1)} \,\middle|\, S_{Y_1}\right\}\right] = E\left\{\frac{\mathrm{Var}(g_1(Y_1) \mid X)}{P(T=1 \mid S_1)}\right\}.$$

Appendix 2. Conditions for Theorem 2.1

(C1) Y_k is a continuous random variable and for any fixed $\tau \in (0,1)$ there exists a unique $q_{k,\tau}$ that $P(Y_k \le q_{k,\tau}) = \tau$ for $k = 0, 1$.
(C2) $\pi_k(S_k)$ is bounded away from 0 and 1.
(C3) S_k has compact support for $k = 0, 1$.
(C4) The function $\pi_k(S_k)$, the density function $f(S_k)$ and $E(\mathbf{1}\{Y_k \le q_{k,\tau}\} \mid S_k)$ all have bounded partial derivatives with respect to S_k up to r_k order, $f(S_k)\pi_k(S_k)$ is bounded away from 0.
(C5) The kernel $\mathcal{K}_k$ is bounded up to second order derivative.

(C6) The smoothing bandwidth h_{kn} satisfies $nh_{kn}^2 \to \infty$, $nh_{kn}^{d_k} \to \infty$ and $\sqrt{n}h_{kn}^{r_k} \to 0$ as $n \to \infty$. Here r_k is the order of the kernel $\mathcal{K}_k$.

Appendix 3. Proof of Theorem 2.1

Proof of Theorem 2.1: (i) In the case that $S_k = X$, Firpo (2007) proved the asymptotics of $\hat{\theta}_{\text{IPW}}$ using kernel method, Chen et al. (2015) proved the asymptotics of $\hat{\theta}_{\text{REG}}$ using kernel method. Following the proofs in their papers and substituting X by S_k, $\sqrt{n}(\hat{\theta}(S_0, S_1) - \theta)$ is asymptotically equivalent to

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{T_i g_1(Z_i)}{\pi_1(S_{1i})} - \frac{E\left(g_1(Y_1)|S_{1i}\right)\{T_i - \pi_1(S_{1i})\}}{\pi_1(S_{1i})}\right] - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\frac{(1-T_i)g_0(Z_i)}{\pi_0(S_{0i})} - \frac{E\left(g_0(Y_0)|S_{0i}\right)\{(1-T_i) - \pi_0(S_{0i})\}}{\pi_{0i}(S_{0i})}\right] + o_p(1), \quad \text{(A3)}$$

where S_{ki} is the ith observation of S_k, $\pi_k(S_{ki}) = P(T = k \mid S_k = S_{ki})$ for $k = 0, 1$. By direct but tedious calculation, the covariance of the two summation terms in (A3) is $-E(g_0(Y_0))E(g_1(Y_1)) + E\{E(g_0(Y_0) \mid S_0)E(g_1(Y_1) \mid S_1)\}$. Their corresponding variances are

$$\begin{aligned}
&\text{Var}\left[\frac{Tg_1(Z)}{\pi_1(S_1)} - \frac{E(g_1(Y_1) \mid S_1)}{\pi_1(S_1)}\{T - \pi_1(S_1)\}\right] \\
&\quad = \text{Var}\{E\left(g_1(Y_1) \mid S_1\right)\} + E\left\{\frac{\text{Var}(g_1(Y_1) \mid S_1)}{\pi_1(S_1)}\right\}, \\
&\text{Var}\left[\frac{(1-T)g_0(Y_0)}{\pi_0(S_0)} - \frac{E(g_0(Y_0)|S_0)}{\pi_0(S_0)}\{(1-T) - \pi_0(S_0)\}\right] \\
&\quad = \text{Var}\{E\left(g_0(Y_0) \mid S_0\right)\} + E\left\{\frac{\text{Var}(g_0(Y_0) \mid S_0)}{\pi_0(S_0)}\right\}.
\end{aligned}$$

Thus the asymptotic variance of $\hat{\theta}(S_0, S_1)$ is

$$\text{Var}\left\{E\left(g_1(Y_1) \mid S_1\right) - E\left(g_0(Y_0) \mid S_0\right)\right\} + E\left\{\frac{\text{Var}(g_1(Y_1) \mid S_1)}{\pi_1(S_1)}\right\} + E\left\{\frac{\text{Var}(g_0(Y_0) \mid S_0)}{\pi_0(S_0)}\right\}.$$

(ii) Here we only list the proof for regression type estimator $\hat{\theta}_{\text{REG}}$ with $d_0 = d_1 = 1$. We only derive the difference of $\hat{q}_{1,\tau}$ between using true B_k and estimated B_k for regression estimator, the proof for the $\hat{q}_{0,\tau}$ is similar. For simplicity, we denote S_1, B_1, h_{1n}, $\mathcal{K}_1$, $g_1(\cdot)$, $\pi_1(\cdot)$ as S, B, h, $\mathcal{K}$, $g(\cdot)$, $\pi(\cdot)$ respectively and define $\mathcal{K}_h(\cdot) = h^{-1}\mathcal{K}(\cdot/h)$ in the following proof. Let $\Delta_{ij} = \mathcal{K}_h(\hat{B}^\top X_j - \hat{B}^\top X_i) - \mathcal{K}_h(B^\top X_j - B^\top X_i)$, it can be verified that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{j=1}^{n} T_j g(Z_j)\mathcal{K}_h(\hat{B}^\top X_j - \hat{B}^\top X_i)}{\sum_{j=1}^{n} T_j \mathcal{K}_h(\hat{B}^\top X_j - \hat{B}^\top X_i)} - \frac{\sum_{j=1}^{n} T_j g(Z_j)\mathcal{K}_h(B^\top X_j - B^\top X_i)}{\sum_{j=1}^{n} T_j \mathcal{K}_h(B^\top X_j - B^\top X_i)}\right\} \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{j=1}^{n} T_j g(Z_j)\mathcal{K}_h(S_j - S_i) + \sum_{j=1}^{n} T_j g(Z_j)\Delta_{ij}}{\sum_{j=1}^{n} T_j \mathcal{K}_h(S_j - S_i) + \sum_{j=1}^{n} T_j \Delta_{ij}} - \frac{\sum_{j=1}^{n} T_j g(Z_j)\mathcal{K}_h(S_j - S_i)}{\sum_{j=1}^{n} T_j \mathcal{K}_h(S_j - S_i)}\right\} \\
&\equiv A_1 + A_2 + A_3,
\end{aligned}$$

where

$$\begin{aligned}
A_1 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{T_j g(Z_j)\Delta_{ij}}{\pi(S_i)f(S_i)} - \frac{T_j E(g(Z_i) \mid S_i)\Delta_{ij}}{\pi(S_i)f(S_i)}\right\}, \\
A_2 &= -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{T_j g(Z_j)\Delta_{ij}}{\pi(S_i)f(S_i)} - \frac{T_j g(Z_j)\Delta_{ij}}{\frac{1}{n}\sum_{l=1}^{n} T_l \mathcal{K}_h(S_l - S_i) + \frac{1}{n}\sum_{l=1}^{n} T_l \Delta_{il}}\right\}, \\
A_3 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} T_j \Delta_{ij}\left\{\frac{E(g(Z_i) \mid S_i)}{\pi(S_i)f(S_i)} - \frac{E(g(Z_i) \mid S_i)}{\frac{1}{n}\sum_{l=1}^{n} T_l \mathcal{K}_h(S_l - S_i) + \frac{1}{n}\sum_{l=1}^{n} T_l \Delta_{il}} + \frac{E(g(Z_i) \mid S_i) - \frac{\sum_{l=1}^{n} T_l g(Z_l)\mathcal{K}_h(S_l - S_i)}{\sum_{l=1}^{n} T_l \mathcal{K}_h(S_l - S_i)}}{\frac{1}{n}\sum_{l=1}^{n} T_l \mathcal{K}_h(S_l - S_i) + \frac{1}{n}\sum_{l=1}^{n} T_l \Delta_{il}}\right\}.
\end{aligned}$$

Since $\Delta_{ij} = \mathcal{K}_h(\hat{B}^\top X_j - \hat{B}^\top X_i) - \mathcal{K}_h(B^\top X_j - B^\top X_i)$, using a Taylor expansion around $B^\top X_j - B^\top X_i$ for Δ_{ij} and plugging in A_1, we have

$$\begin{aligned}
A_1 &= \frac{(\hat{B} - B)^\top}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left\{\frac{T_j\{g(Z_j) - E(g(Z_i) \mid S_i)\}}{\pi(S_i)f(S_i)} \times \frac{1}{h}\left[\mathcal{K}'\left(\frac{B^\top X_j - B^\top X_i}{h}\right)\frac{X_j - X_i}{h}\right]\right\} + o_p(n^{-1/2}) \\
&\equiv \frac{(\hat{B} - B)^\top}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} Q_{ij} + o_p(n^{-1/2}).
\end{aligned}$$

Denote $A_{11} = \sum_{i=1}^{n}\sum_{j=1}^{n} Q_{ij}/n^2$ and $\breve{A}_{11} = \sum_{i=1}^{n}\sum_{j=1}^{n} E(Q_{ij} \mid X_i, g(Z_i), T_i)/n^2$. Note

$$\begin{aligned}
&E\left\{\frac{1}{h}T_j\mathcal{K}'\left(\frac{S_j - S_i}{h}\right)\left(\frac{X_j - X_i}{h}\right)\middle| (X_i, Z_i, T_i) = (x_i, z_i, t_i)\right\} \\
&= E\left[E\left\{\frac{1}{h}T_j\mathcal{K}'\left(\frac{S_j - s_i}{h}\right)\frac{X_j - x_i}{h}\middle| S_j\right\}\right] \\
&= E\left[\frac{1}{h^2}\mathcal{K}'\left(\frac{S_j - s_i}{h}\right)E\{T_j(X_j - x_i) \mid S_j\}\right] \\
&= -\frac{\partial[E\{T(X - x_i) \mid S = t\}f(t)]}{\partial t}\bigg|_{t=s_i} + o_p(1) \\
&= -E(TX \mid S = s_i)f'(s_i) - x_i\pi(s_i)f'(s_i) - x_i\pi'(s_i)f(s_i) + \frac{\partial\{E(TX \mid S = t)\}}{\partial t}\bigg|_{t=s_i} f(s_i) + o_p(1),
\end{aligned}$$

and

$$\begin{aligned}
&E\left\{\frac{1}{h}T_j g(Z_j)\mathcal{K}'\left(\frac{S_j - s_i}{h}\right)\left(\frac{X_j - x_i}{h}\right)\middle| (X_i, Z_i, T_i) = (x_i, z_i, t_i)\right\} \\
&= E\left[E\left\{\frac{1}{h}T_j g(Z_j)\mathcal{K}'\left(\frac{S_j - s_i}{h}\right)\frac{X_j - x_i}{h}\middle| S_j\right\}\right] \\
&= E\left[\frac{1}{h^2}\mathcal{K}'\left(\frac{S_j - s_i}{h}\right)E\{T_j g(Z_j)(X_j - x_i) \mid S_j\}\right] \\
&= -\frac{\partial\{E\left(Tg(Z)(X - x_i) \mid S = t\right)f(t)\}}{\partial t}\bigg|_{t=s_i} + o(1)
\end{aligned}$$

$$
\begin{aligned}
&= -E(Tg(Z)X \mid S{=}s_i)f'(s_i) - x_i\pi(s_i)E(g(Z) \mid S = s_i)f'(s_i) \\
&\quad + \left.\frac{\partial E(Tg(Z)X \mid S = t)}{\partial t}\right|_{t=s_i} f(s_i) \\
&\quad - x_i\pi'(s_i)E(g(Z) \mid S = s_i)f(s_i) - x_i\pi(s_i) \\
&\quad \times \left\{\frac{\partial E(g(Z) \mid S = s_i)}{\partial S}\right\} f(s_i) + o_p(1).
\end{aligned}
$$

Therefore,

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

where

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

It can be seen that the first term in c_1 will equal to 0 if $Y_1 \perp X \mid S$, while the second term in c_1 will equal to 0 if $T \perp X \mid S$. Thus when both $Y_1 \perp X \mid S$ and $T \perp X \mid S$ hold, we will have $c_1 = 0$. Let $A_{11j} = (1/n)\sum_{i=1}^{n} Q_{ij}$, $\breve{A}_{11j} = (1/n)\sum_{i=1}^{n} E(Q_{ij} \mid X_i, g(Z_i), T_i))$, we have

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

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

$$
\sqrt{n}A_1 = c_1^{\top}\{\sqrt{n}(\hat{B} - B)\} + o_p(1).
$$

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

$$
\begin{aligned}
A_2 &\equiv -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \times \left\{\frac{T_j g(Z_j)\Delta_{ij}}{\pi(S_i)f(S_i)}\right. \\
&\quad \left. - \frac{T_j g(Z_j)\Delta_{ij}}{\frac{1}{n}\sum_{l=1}^{n} T_l\mathcal{K}_h(S_l - S_i) + \frac{1}{n}\sum_{l=1}^{n} T_l\Delta_{il}}\right\} \\
&= -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \\
&\quad \times \left[\frac{T_j g(Z_j)}{h}\mathcal{K}'\left(\frac{B^{\top}X_j - B^{\top}X_i}{h}\right)(\hat{B} - B)^{\top}\right. \\
&\quad \times \left(\frac{X_j - X_i}{h}\right)\left\{\frac{1}{\pi(S_i)f(S_i)}\right. \\
&\quad \left.\left. - \frac{1}{\frac{1}{n}\sum_{l=1}^{n} T_l\mathcal{K}_h(S_l - S_i) + \frac{1}{n}\sum_{l=1}^{n} T_l\Delta_{il}}\right\}\right] + o_p(n^{-1/2}).
\end{aligned}
$$

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

$$
\begin{aligned}
\breve{A}_2 &= -\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} \\
&\quad \times \left[\frac{T_j g(Z_j)}{h}\mathcal{K}'\left(\frac{B^{\top}X_j - B^{\top}X_i}{h}\right)(\hat{B} - B)^{\top}\left(\frac{X_j - X_i}{h}\right)\right. \\
&\quad \times E\left\{\frac{1}{\pi(S_i)f(S_i)}\right. \\
&\quad \left.\left. - \frac{1}{\frac{1}{n}\sum_{l=1}^{n} T_l\mathcal{K}_h(S_l - S_i) + \frac{1}{n}\sum_{l=1}^{n} T_l\Delta_{il}}\,\middle|\, X_i, g(Z_i), T_i\right\}\right].
\end{aligned}
$$

Since

$$
\begin{aligned}
&E\left\{\frac{1}{n}\sum_{l=1}^{n} T_l\mathcal{K}_h(S_l - S_i) \mid S_i\right\} \\
&\quad = \pi(S_i)f(S_i) + o_p(1), \\
&E\left\{\frac{1}{n}\sum_{l=1}^{n} T_l g(Z_l)\mathcal{K}_h(S_l - S_i) \mid S_i\right\} \\
&\quad = \pi(S_i)E(g(Z_i) \mid S_i)f(S_i) + o_p(1),
\end{aligned}
$$

using a similar decomposition method as A_1, we can also show $\sqrt{n}A_2 \xrightarrow{p} 0$ and $\sqrt{n}A_3 \xrightarrow{p} 0$. Thus we proved that

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\hat{E}\left[g(Y_{1i}) \mid \hat{S}_i\right] - \frac{1}{n}\sum_{i=1}^{n}\hat{E}\left[g(Y_{1i}) \mid S_i\right] \\
&\quad = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{\sum_{j=1}^{n} T_j g(Z_j)\mathcal{K}_h(\hat{S}_j - \hat{S}_i)}{\sum_{j=1}^{n} T_j\mathcal{K}_h(\hat{S}_j - \hat{S}_i)}\right. \\
&\quad\quad \left. - \frac{\sum_{j=1}^{n} T_j g(Z_j)\mathcal{K}_h(S_j - S_i)}{\sum_{j=1}^{n} T_j\mathcal{K}_h(S_j - S_i)}\right\} \\
&\quad = c_1^{\top}(\hat{B} - B) + o_p(1/\sqrt{n}).
\end{aligned}
$$

Note that the REG estimator for $q_{1,\tau}$ based on estimated S is:

$$
\hat{q}_{1,\tau} = \mathrm{argmin}\frac{1}{n}\sum_{i=1}^{n}\hat{E}\left\{(Y_{1i} - t)(\tau - \mathbf{1}\{Y_{1i} \leq t\}) \mid \hat{S}_i\right\}
$$

$$= \operatorname{argmin} \sum_{i=1}^{n} \Big(\hat{E}\Big[(\mathbf{1}\{Y_{1i} \le q_{1,\tau}\} - \tau)(t - q_{1,\tau}) \,|\, \hat{S}_i \Big] + \hat{E}\Big[(Y_{1i} - t)(\mathbf{1}\{Y_{1i} \le q_{1,\tau}\} - \mathbf{1}\{Y_{1i} \le t\}) \,|\, \hat{S}_i \Big] \Big).$$

Let $u = \sqrt{n}(t - q_{1,\tau})$, $\hat{u} = \sqrt{n}(\hat{q}_{1,\tau} - q_{1,\tau})$, the optimisation will change to

$$\hat{u} = \operatorname{argmin} \left\{ \sum_{i=1}^{n} \frac{u}{\sqrt{n}} \hat{E}\Big[(\mathbf{1}\{Y_{1i} \le q_{1,\tau}\} - \tau) \,|\, \hat{S}_i \Big] + \sum_{i=1}^{n} \hat{E}\big[(Y_{1i} - (q_{1,\tau} + u/\sqrt{n}))(\mathbf{1}\{Y_{1i} \le q_{1,\tau}\} - \mathbf{1}\{Y_{1i} \le q_{1,\tau} + u/\sqrt{n}\}) \,|\, \hat{S}_i \big] \right\}$$

Similar with the proof in Firpo (2007), one may check that the second term equals to $n((f_1(q_{1,\tau})/2)u^2 + o_p(1))$. Hence we have

$$\begin{aligned}
\hat{u} &= \sqrt{n}(\hat{q}_{1,\tau} - q_{1,\tau}) \\
&= \sqrt{n} \left\{ -\frac{1}{n f_1(q_{1,\tau})} \sum_{i=1}^{n} \hat{E}\Big[(\mathbf{1}\{Y_{1i} \le q_{1,\tau}\} - \tau) \,|\, \hat{S}_i \Big] \right\} \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{E}\Big[g(Y_{1i}) \,|\, \hat{S}_i \Big] \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \hat{E}\big[g(Y_{1i}) \,|\, S_i \big] + c_1^{\top} \sqrt{n}\{(\hat{B} - B)\} + o_p(1) \\
&= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left[\frac{T_i g_1(Z_i)}{\pi_1(S_{1i})} - \frac{E\left(g_1(Y_1) \,|\, S_{1i}\right)\{T_i - \pi_1(S_{1i})\}}{\pi_1(S_{1i})} \right] \\
&\quad + c_1^{\top} \sqrt{n}\{(\hat{B} - B)\} + o_p(1).
\end{aligned}$$

The last equation follows from the proof in Chen et al. (2015), which is the linearisation for the REG estimator using true S. Repeat all above procedure for $q_{0,\tau}$, and plug in the linearisation for $(\hat{B} - B)$, we could get the linearisation for $\hat{\theta}(S_0, S_1) - \theta$, hence Theorem 2.1 is proved. ■

Appendix 4. Asymptotic variance comparisons between using $S_{\min}$ and using S_{Y_0,Y_1}

We first prove that the asymptotic variance using S_T will be larger than using X. Following the proof below, one may also easily prove using $S_{\min}$ in De Luna et al. (2011) is larger than $V^*_{S_{Y_0,Y_1}}$ unless $S_{\min} = S_{Y_0,Y_1}$, by replacing original covariate set X with S_{Y_0,Y_1} and replacing S_T with $S_{\min}$. Adapting the proof for Theorem 2.1, we can find that for any S that satisfies $T \perp (Y_0, Y_1) \,|\, S$, the asymptotic variance for using (S, S) in $\hat{\theta}$ is

$$\operatorname{Var}\{E(g(Y_1) \,|\, S) - E(g(Y_0) \,|\, S)\} + E\left\{ \frac{\operatorname{Var}(g_1(Y_1) \,|\, S)}{\pi_1(S)} \right\} + E\left\{ \frac{\operatorname{Var}(g_0(Y_0) \,|\, S)}{\pi_0(S)} \right\}.$$

Since S_T satisfies $T \perp X \,|\, S_T$, we also have $T \perp (Y_0, Y_1) \,|\, X$ thus $T \perp (Y_0, Y_1) \,|\, S_T$. Therefore the asymptotic variance for $\hat{\theta}$ using S_T is:

$$V_{S_T} = \operatorname{Var}\{E(g_1(Y_1) \,|\, S_T) - E(g_0(Y_0) \,|\, S_T)\} + E\left\{ \frac{\operatorname{Var}(g_1(Y_1) \,|\, S_T)}{\pi_1(S_T)} \right\} + E\left\{ \frac{\operatorname{Var}(g_0(Y_0) \,|\, S_T)}{\pi_0(S_T)} \right\}.$$

The asymptotic variance for $\hat{\theta}$ using X is:

$$V_X = \operatorname{Var}\{E(g_1(Y_1) \,|\, X) - E(g_0(Y_0) \,|\, X)\} + E\left\{ \frac{\operatorname{Var}(g_1(Y_1) \,|\, X)}{\pi_1(X)} \right\} + E\left\{ \frac{\operatorname{Var}(g_0(Y_0) \,|\, X)}{\pi_0(X)} \right\}.$$

Therefore

$$\begin{aligned}
V_X - V_{S_T} &= E\big[\{\pi_1^{-1}(X) - 1\}\operatorname{Var}(g_1(Y_1) \,|\, X)\big] \\
&\quad - E\big[\{\pi_1^{-1}(S_T) - 1\}\operatorname{Var}(g_1(Y_1) \,|\, S_T)\big] \qquad \text{(A4)} \\
&\quad + E\big[\{\pi_0^{-1}(X) - 1\}\operatorname{Var}(g_0(Y_0) \,|\, X)\big] \\
&\quad - E\big[\{\pi_0^{-1}(S_T) - 1\}\operatorname{Var}(g_0(Y_0) \,|\, S_T)\big] \qquad \text{(A5)} \\
&\quad + 2E\left\{E(g_1(Y_1) \,|\, S_T)E(g_0(Y_0) \,|\, S_T)\right\} \\
&\quad - 2E\left\{E(g_1(Y_1) \,|\, X)E(g_0(Y_0) \,|\, X)\right\}. \qquad \text{(A6)}
\end{aligned}$$

Let

$$a_1(S_T) = \sqrt{\frac{1}{\pi_1(X)} - 1} = \sqrt{\frac{1}{\pi_1(S_T)} - 1}, \quad a_0(S_T) = \sqrt{\frac{1}{\pi_0(X)} - 1} = \sqrt{\frac{1}{\pi_0(S_T)} - 1}.$$

The expression (A4) equals

$$\begin{aligned}
&E\left\{\operatorname{var}(a_1 g_1(Y_1) \,|\, X) - \operatorname{var}(a_1 g_1(Y_1) \,|\, S_T)\right\} \\
&\quad = -\operatorname{Var}\left\{E(a_1 g_1(Y_1) \,|\, X)\right\} + \operatorname{Var}\left\{E(a_1 g_1(Y_1) \,|\, S_T)\right\} \\
&\quad = -E\left[\operatorname{Var}\left\{E(a_1 g_1(Y_1) \,|\, X) \,|\, S_T\right\}\right].
\end{aligned}$$

Similarly, the expression (A5) equals

$$-E[\operatorname{Var}\{E(a_0 g_0(Y_0) \,|\, X) \,|\, S_T\}],$$

and the expression (A6) equals

$$-2E[\operatorname{cov}\{E(g_0(Y_0) \,|\, X), E(g_1(Y_1) \,|\, X) \,|\, S_T\}].$$

Since $a_0 a_1 = 1$, therefore

$$V_X - V_{S_T} = -E\left(\operatorname{Var}\left[\{a_1 E(g_1(Y_1) \,|\, X) + a_0 E(g_0(Y_0) \,|\, X)\} \,|\, S_T\right]\right) \le 0,$$

which completes the proof.

See Figure 1 for difference choices of S_k and Figure A1 for the comparisons of efficiency of estimator $\hat{\theta}$ based on different S_k, $k = 0, 1$.

Appendix 5. Asymptotic variance comparisons between using S_{Y_0,Y_1} and using $S_{Y_k,T}$

In this section, we prove that the asymptotic variance of $\hat{\theta}(S_{Y_0,Y_1}, S_{Y_0,Y_1})$ is smaller than asymptotic variance of $\hat{\theta}(S_{Y_0,T}, S_{Y_0,T})$, followed by those of $\hat{\theta}(S_T, S_T)$. From the proof of Theorem 2.1, for all S_k satisfying $T \perp Y_k \,|\, S_k$, the asymptotic variance of $\hat{\theta}(S_1, S_0)$ is

$$\operatorname{Var}\left\{E\left(g_1(Y_1) \,|\, S_1\right) - E\left(g_0(Y_0) \,|\, S_0\right)\right\} + E\left\{ \frac{\operatorname{Var}(g_1(Y_1) \,|\, S_1)}{\pi_1(S_1)} \right\} + E\left\{ \frac{\operatorname{Var}(g_0(Y_0) \,|\, S_0)}{\pi_0(S_0)} \right\}.$$

For $\hat{\theta}(S_{Y_0,Y_1}, S_{Y_0,Y_1})$ and $\hat{\theta}(S_{Y_0,T}, S_{Y_1,T})$, the asymptotic variances $V_{S_{Y_0,Y_1}}$ and $V^*_{S_{Y_0,T},S_{Y_1,T}}$ are

$$V^*_{S_{Y_0,Y_1}} = \operatorname{Var}\left\{E\left(g_1(Y_1) \,|\, S_{Y_0,Y_1}\right) - E\left(g_0(Y_0) \,|\, S_{Y_0,Y_1}\right)\right\}$$

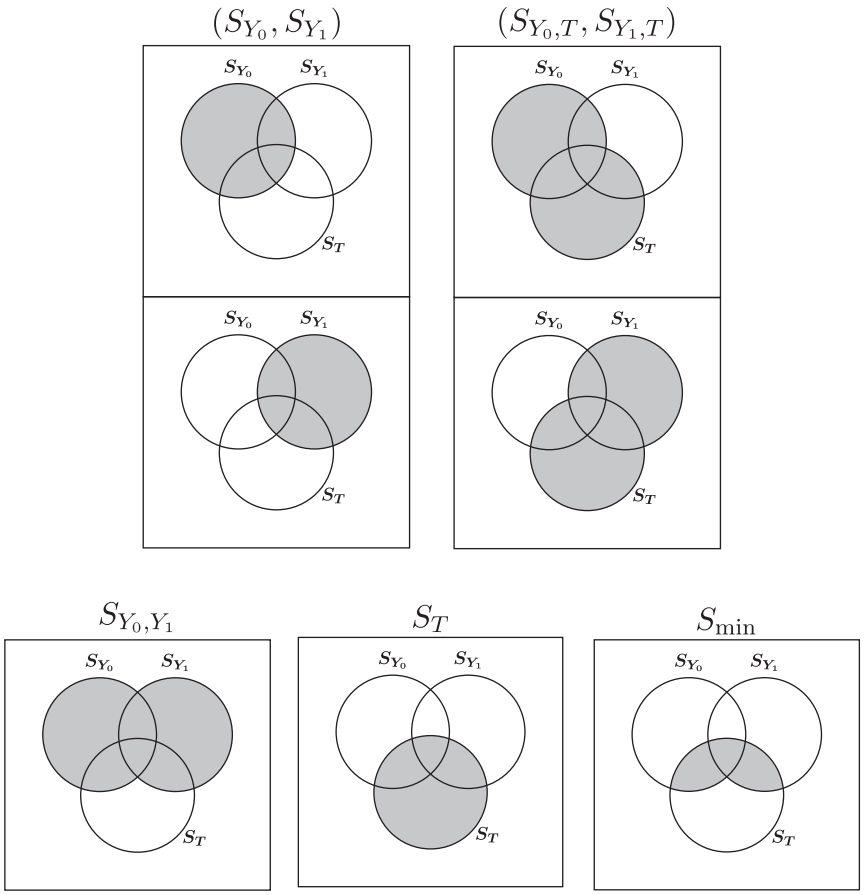

Figure A1. Five choices of S_k in the space of all linear combinations of X. For (S_{Y_0}, S_{Y_1}) and $(S_{Y_0,T}, S_{Y_1,T})$, the first row are S_0 for estimating Y_0 characteristics, the second row are S_1 for estimating Y_1 characteristics.

$$
\begin{aligned}
&+ E\left\{\frac{\mathrm{Var}(g_1(Y_1)\mid S_{Y_0,Y_1})}{\pi_1(S_{Y_0,Y_1})}\right\} + E\left\{\frac{\mathrm{Var}(g_0(Y_0)\mid S_{Y_0,Y_1})}{\pi_0(S_{Y_0,Y_1})}\right\} \\
&= \mathrm{Var}\left\{E\left(g_1(Y_1)\mid S_{Y_1}\right) - E\left(g_0(Y_0)\mid S_{Y_0}\right)\right\} \\
&\quad + E\left\{\frac{\mathrm{Var}(g_1(Y_1)\mid S_{Y_1})}{\pi_1(S_{Y_0,Y_1})}\right\} + E\left\{\frac{\mathrm{Var}(g_0(Y_0)\mid S_{Y_0})}{\pi_0(S_{Y_0,Y_1})}\right\}
\end{aligned}
$$

$$
\begin{aligned}
&V^*_{S_{Y_0,T},S_{Y_1,T}} \\
&= \mathrm{Var}\left\{E\left(g_1(Y_1)\mid S_{Y_1,T}\right) - E\left(g_0(Y_0)\mid S_{Y_0,T}\right)\right\} \\
&\quad + E\left\{\frac{\mathrm{Var}(g_1(Y_1)\mid S_{Y_1,T})}{\pi_1(S_{Y_1,T})}\right\} + E\left\{\frac{\mathrm{Var}(g_0(Y_0)\mid S_{Y_0,T})}{\pi_0(S_{Y_0,T})}\right\} \\
&= \mathrm{Var}\left\{E\left(g_1(Y_1)\mid S_{Y_1}\right) - E\left(g_0(Y_0)\mid S_{Y_0}\right)\right\} \\
&\quad + E\left\{\frac{\mathrm{Var}(g_1(Y_1)\mid S_{Y_1})}{\pi_1(S_{Y_1,T})}\right\} + E\left\{\frac{\mathrm{Var}(g_0(Y_0)\mid S_{Y_0})}{\pi_0(S_{Y_0,T})}\right\} \\
&= \mathrm{Var}\left\{E\left(g_1(Y_1)\mid S_{Y_1}\right) - E\left(g_0(Y_0)\mid S_{Y_0}\right)\right\} \\
&\quad + E\left\{\frac{\mathrm{Var}(g_1(Y_1)\mid S_{Y_1})}{\pi_1(S_T)}\right\} + E\left\{\frac{\mathrm{Var}(g_0(Y_0)\mid S_{Y_0})}{\pi_0(S_T)}\right\}
\end{aligned}
$$

Thus we only need to prove

$$
E\left\{\frac{\mathrm{Var}(g_1(Y_1)\mid S_{Y_1})}{\pi_1(S_T)}\right\} \geq E\left\{\frac{\mathrm{Var}(g_1(Y_1)\mid S_{Y_1})}{\pi_1(S_{Y_0,Y_1})}\right\}
$$

By Jensen's inequity,

$$
\begin{aligned}
E\left\{\frac{\mathrm{Var}(g_1(Y_1)\mid S_{Y_1})}{\pi_1(S_T)}\right\} &= E\left\{\frac{\mathrm{Var}(g_1(Y_1)\mid S_{Y_1})}{\pi_1(X)}\right\} \\
&= E\left\{E\left[\frac{\mathrm{Var}(g_1(Y_1)\mid S_{Y_1})}{\pi_1(X)}\,\middle|\, S_{Y_0,Y_1}\right]\right\}
\end{aligned}
$$

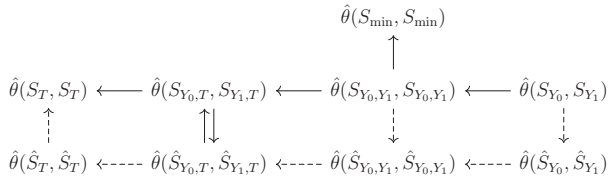

Figure A2. Relative efficiencies of estimators. Solid arrow from A to B means that A is more asymptotically efficient than B. Dashed arrow from A to B means that empirically A is more efficient than B.

$$
\begin{aligned}
&= E\left\{\mathrm{Var}(g_1(Y_1)\mid S_{Y_1})E\left[\frac{1}{\pi_1(X)}\,\middle|\, S_{Y_0,Y_1}\right]\right\} \\
&\geq E\left\{\mathrm{Var}(g_1(Y_1)\mid S_{Y_1})\frac{1}{E\left[\pi_1(X)\mid S_{Y_0,Y_1}\right]}\right\} \\
&= E\left\{\frac{\mathrm{Var}(g_1(Y_1)\mid S_{Y_1})}{\pi_1(S_{Y_0,Y_1})}\right\}
\end{aligned}
$$

Hence $\hat\theta(S_{Y_0,Y_1}, S_{Y_0,Y_1}) \to \hat\theta(S_{Y_0,T}, S_{Y_1,T})$. Note that from Lemma A.1 we have $V^*_{S_{Y_0,T},S_{Y_1,T}} \leq V^*_{X,X}$, i.e. $\hat\theta(S_{Y_0,T}, S_{Y_1,T}) \to \hat\theta(X,X)$. Then the other result follows from $\hat\theta(X,X) \to \hat\theta(S_T, S_T)$, which is proved in the Appendix 4.