The role of propensity score structure in asymptotic efficiency of estimated conditional quantile treatment effect

Niwen Zhou1 | Xu Guo2 | Lixing Zhu1,3

1 Center for Statistics and Data Science, Beijing Normal University, Zhuhai, China
2 School of Statistics, Beijing Normal University, Beijing, China
3 Department of Mathematics, Hong Kong Baptist University, Hongkong, China

Correspondence
Lixing Zhu, Department of Mathematics, Hong Kong Baptist University, Hongkong, China.
Email: lzhu@hkbu.edu.hk

Funding information
Beijing Natural Science Foundation, Grant/Award Number: 1212004; Hong Kong Baptist University, Grant/Award Number: HKBU12302720; National Natural Science Foundation of China, Grant/Award Numbers: 11771048, 11871263, 12071038

Abstract
When a strict subset of covariates is given, we propose conditional quantile treatment effect (CQTE) to offer, compared with the unconditional quantile treatment effect (QTE) and conditional average treatment effect (CATE), a more complete and informative view of the heterogeneity of treatment effects via the quantile sheet that is a function of the given covariates and quantile levels. Even though either one or both QTE and CATE are not significant, CQTE could still show some impact of the treatment on the upper and lower tails of subpopulations’ (defined by the covariates subset) distribution. To the best of our knowledge, this is the first to consider such a low-dimensional conditional quantile treatment effect in the literature. We focus on deriving the asymptotic normality of propensity score-based estimators under parametric, nonparametric, and semiparametric structure. We make a systematic study on the estimation efficiency to check the importance of propensity score structure and the essential differences from the unconditional counterparts. The derived unique properties can answer: what is the general ranking of these estimators? how does the affiliation of the given covariates to the set of covariates of the propensity score affect the efficiency? how does the convergence rate of the estimated propensity score affect the efficiency? and why
1 | INTRODUCTION

Treatment effect is a vital issue in diverse research and applied fields. In the literature, most of the existing studies focus on average treatment effects (ATE) defined by the population mean of potential outcomes as well as quantile treatment effects (QTE) by the population quantile. QTE can capture the heterogeneity of treatment effect. For example, reducing class sizes may have a positive effect on the academic performance of excellent students but opposite on that of the weaker, or vice versa (Koenker, 2017). Doksum (1974) and D’Abrera and Lehmann (1975) defined the \( \tau \)th QTE as the difference between quantiles of the two marginal potential outcome distributions. As Firpo (2007) commented, the \( \tau \)th quantile in this definition is not exactly equal to the \( \tau \)th quantile of the distribution of the potential outcomes difference unless the rank preservation assumption is satisfied. Yet it is still reasonable and informative in studying QTE and some references include Koenker (2005), Firpo (2007), and Zhang (2018). Firpo (2007) proposed the QTE and QTT (quantile treatment effect on the treated) estimator by minimizing the expectation of proper weighting check functions, where the weight is based on the inverse of propensity score. Furthermore, the conditional QTE (CQTE) can provide a quantile sheet, that is a function of quantile levels and given covariates, to examine one more type of heterogeneity: QTE in specific subpopulations decided by the covariates, which can reflect their influence at different quantiles. Such a heterogeneity is also informative, see, for example, Wager and Athey (2018) and Luo et al. (2019).

In this article, we propose CQTE in a general situation in which the conditioning continuous covariates \( X_1 \) form a strict subset of covariates \( X (X_1 \subsetneq X) \). This is a generalization of CQTE proposed by Imbens and Wooldridge (2009) who referred to it as the difference of quantiles for two potential outcome distributions conditional on the whole \( X \) to guarantee the unconfoundedness assumption. As mentioned in Imbens and Wooldridge (2009), although QTE was proposed in the 1970s by Doksum (1974) and D’Abrera and Lehmann (1975), it has received much more attention only in recent decades and let alone CQTE. Note that we can get ATE by averaging CATE(X) over \( X \), that is, \( ATE = E[CATE(X)] \), but cannot simply obtain QTE from CQTE(X). It is probably a reason why there are few results on conditional quantile treatment effects in the literature. With the development of personalized medicine and customized marketing recommendations, there have been increasing interests in estimating heterogeneous treatment effects. Thus the generalized CQTE(X_1) could become a powerful tool to explore the heterogeneity of treatment effects. By using conditional quantile treatment effects (CQTE(X_1), hereafter), we can know how the quantile treatment effects change with \( X_1 \), which can help not only a detailed program evaluation, but also the investigation on the importance of \( X_1 \). Note that the unconfoundedness assumption
on $X_1$ may not hold. Based on this indirect unconfoundedness assumption the technical skills for theoretical development have to be more sophisticated than those for $QTE$. Also, the quantile sheet of $CQTE(X_1)$ (denoted as $\Delta_\tau(X_1)$), as a function of both $\tau$ and $X_1$, is more informative than conditional average treatment effect ($CATE(X_1)$) proposed by Abrevaya et al. (2015) as shown in Figure 1.

It is well known that in unconditional cases, with estimating propensity score, the estimation efficiency for $ATE$ and $QTE$ can be enhanced. See Hirano et al. (2003) for instance. This type of estimator is referred as the inverse probability weighting-based (IPW, hereafter) estimator. Based on a nonparametrically estimated propensity score, Firpo (2007) proposed an IPW-type estimator for $QTE$ that can achieve the semiparametric efficiency bound. Based on the potential outcome model, we in this article will construct the pointwise $CQTE(X_1)$ estimator via minimizing a properly weighted sum of check functions with the estimated propensity score. Therefore, we will first estimate $CQTE$ and then investigate asymptotic behaviors of the estimated $CQTE$ when the propensity score is under parametric, nonparametric, and semiparametric dimension reduction structure. See the relevant references such as Yao et al. (2010), Abrevaya et al. (2015), and Guo et al. (2018).

As under the regularity conditions designed in this article, all estimators are asymptotically unbiased, we then discuss asymptotic efficiencies by asymptotic variances. According to the research for $ATE$ and $QTE$, we consider the efficiency bound and efficient estimation construction for the $CQTE$ function $\Delta_\tau(X_1)$ defined in the next section. As pointed out by Kennedy et al. (2017), if we only assume mild smoothness conditions on $\Delta_\tau(X_1)$, there is no existing theory in the literature to derive the efficiency bound and an efficient estimator for $\Delta_\tau(X_1)$ in the sense $ATE$ or $QTE$ shares. Furthermore, because $\Delta_\tau(X_1)$ is not pathwise differentiable, any estimator cannot achieve $\sqrt{n}$-consistent. However, we can have some information on the estimation efficiency in the following sense. For brevity, write $\hat{CQTE}_O$, $\hat{CQTE}_P$, $\hat{CQTE}_S$, and $\hat{CQTE}_N$ as the estimators with true, parametric, semiparametric, and nonparametric estimated propensity score, respectively.

**FIGURE 1** The quantile sheet of $CQTE(X_1)$ for Model 1 in simulation with $\tau = \{0.05, 0.25, 0.5, 0.75, 0.95\}$
Let $A \preceq B$ mean that the asymptotic variance of estimator $A$ is not greater than that of estimator $B$ and $A \cong B$ stand for that $A$ has the same asymptotic variance function as $B$. We have the following when propensity score is properly estimated for every estimator to derive asymptotic normal distribution.

1. In general, the asymptotic efficiency of the four estimators has the ranking:
\[ \hat{CQTE}_N \preceq \hat{CQTE}_S \preceq \hat{CQTE}_P \cong \hat{CQTE}_O. \]

2. When the estimated propensity score has $X_1$ as a strict subset of its true arguments,
\[ \hat{CQTE}_N \preceq \hat{CQTE}_S \preceq \hat{CQTE}_P \cong \hat{CQTE}_O. \]

When $X_1$ is not a strict subset of its true arguments,
\[ \hat{CQTE}_N \cong \hat{CQTE}_S \cong \hat{CQTE}_P \cong \hat{CQTE}_O. \]

Subsection 2.5.2 presents some more detail.

3. If the propensity score function is very smooth and the kernel functions and tuning parameters in the nonparametric estimation are selected delicately,
\[ \hat{CQTE}_N \cong \hat{CQTE}_S \cong \hat{CQTE}_P \cong \hat{CQTE}_O. \]

Subsection 2.5.1 presents the results. It is worthwhile to point out that the research in this scenario basically serves as a theoretical exploration and provides an insight into the nature of $CQTE$. For practical use, we may not consider such ways to estimate $\hat{CQTE}_N$ and $\hat{CQTE}_S$ to lose their estimation efficiency.

4. We recommend $\hat{CQTE}_S$ for practical use as it can greatly alleviate the curse of dimensionality which is a very serious problem for $\hat{CQTE}_N$, and is robust against model misspecification, particularly, of parametric propensity score structure.

These newly found phenomena show the unique properties of $CQTE$ and demonstrate the essential differences from their unconditional counterparts. As is well known, estimating propensity scores can always enhance, with smaller asymptotic variances, the estimation efficiencies of their unconditional counterparts. Furthermore, the nonparametrically estimated propensity score can make a better efficiency than the parametric/semiparametric one. Relevant references include Hirano et al. (2003), Guo et al. (2018), and Liu et al. (2018).

It should be mentioned that some parts of this research are extensions, but not trivial, of existing works. $\hat{CQTE}_N$ is an extension of the procedure of Firpo (2007) from $QTE$ to $CQTE$. Since $CQTE$ is a function of the given covariates $X_1$, it makes the asymptotic analysis essentially different from that of $QTE$. $\hat{CQTE}_P$ and $\hat{CQTE}_N$ also extend, with more information, the approach of Abrevaya et al. (2015) from $CATE$. But the unsmoothness of the quantile loss function causes the asymptotic analysis more complex than that for $CATE$. The new $\hat{CQTE}_S$ has a very important feature of dimension reduction nature in estimating propensity score. This feature can simultaneously alleviate the risk of misspecification and the curse of dimensionality.

The rest of the article is organized as follows. In Section 2, we introduce the estimation procedures for $CQTE(X_1)$ and investigate their asymptotic properties. Subsection 2.5 presents more
detailed results, for the asymptotic normality of the estimators, about the role of convergence rate of estimated propensity score and the role of the affiliation of \( X_1 \) to the set of the arguments of propensity score. Section 3 contains some numerical studies to examine the performances of the three estimators. In Section 4, we apply our methods to analyze a real data set for illustration and find some phenomena that \( CATE \) does not share. Conclusions and discussions are given in Section 5. Due to the space limitation, all the technical proofs and some additional simulation results are relegated to the supplementary material.

## 2 | ESTIMATION PROCEDURES AND ASYMPTOTIC PROPERTIES

### 2.1 | Definition and preparation

Let \( D \) be the indicator variable of treatment and \( Y \) the outcome. \( D_i = 0, 1 \) respectively means the \( i \)th individual does not receive or receives treatment. Denote the corresponding potential outcome as \( Y_i(0) \) or \( Y_i(1) \) and write the observed outcome as \( Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \). Let \( X \) be a \( k \)-dimensional vector of covariates with \( k \geq 2 \) and \( X_1 \in R^l \) be a subvector of \( X \in R^k \) with \( 1 \leq l < k \). Write \( p(X) \) as the propensity score \( E(D|X) \). Further assume that \( (X_i, Y_i(1), Y_i(0), D_i), i = 1, \ldots, n \), are independent identically distributed (i.i.d.) random vectors. Let \( \tau \) be a real value in \((0, 1)\) and the \( CQTE \) function \( \Delta_{\tau}(x_{10}) = q_{1,\tau}(x_{10}) - q_{0,\tau}(x_{10}) \) with

\[
q_{i,\tau}(x_{10}) = \inf_a E[\rho_{i}(Y(j) - a)|X_1 = x_{10}], \quad j = 0, 1.
\]

Here \( \rho_{i}(u) = u(\tau - \mathbb{I}(u < 0)) \) is the check function, \( \mathbb{I}(\cdot) \) is an indicator function, and \( x_{10} \in \Omega \) with \( \Omega \) containing all the interior points of the support of \( X_1 \). Denote the conditional distribution of \( Y(1)|X_1 \) and of \( Y(0)|X_1 \) as \( F_{1}(Y(1)|X_1) \) and \( F_{0}(Y(0)|X_1) \), respectively.

To introduce the estimation procedures and theoretical results smoothly, the following assumptions are required.

- **Assumption 1 (strong ignorability):**
  - (i) Unconfoundedness: \((Y(0), Y(1)) \perp D|X).
  - (ii) Common support: For some very small \( c > 0 \), \( c < p(X) < 1 - c \).
- **Assumption 2 (conditional quantile function):** For any \( \tau \in (0, 1), j = 0, 1 \) and \( x_{10} \in \Omega \),
  - (i) \( q_{j,\tau}(x_{10}) \) is the unique \( \tau \)th conditional quantile of \( Y(j)|X_1 = x_{10} \).
  - (ii) \( q_{j,\tau}(X_1), j = 0, 1 \) is \( s_1 \geq 2 \) times continuously differentiable for \( X_1 \).
  - (iii) The conditional distribution function \( F_j(y|X_1) \) and density function \( f_j(y|X_1) \) are bounded and uniformly continuous in \( y_j \) for \( X_1 \).

Assumption 1 is commonly used, see, for example, Rosenbaum and Rubin (1983). Part(ii) of Assumption 1 implies that the observed vector \( X \) can fully control for any endogeneity in treatment choice and part(ii) of Assumption 1 means that there is overlap between the supports of the conditional distributions of \( X \) given \( D = 0 \) and \( D = 1 \), respectively. Part(i) of Assumption 2 guarantees the identifiability of \( q_{j,\tau}(x_{10}), j = 0, 1 \). It follows that \( q_{1,\tau}(x_{10}) = F_1^{-1}(\tau|X_1 = x_{10}) \) and \( q_{0,\tau}(x_{10}) = F_0^{-1}(\tau|X_1 = x_{10}), \forall \tau \in (0, 1), x_{10} \in \Omega \). Furthermore, part(ii) of Assumption 2 is required to ensure the function smoothness which will be used, particularly for nonparametric-based estimations.
Under the unconfoundedness assumption, $q_{i,r}(x_{10})$ can also be rewritten as

$$q_{0,r}(x_{10}) = \inf_{a_0} E \left[ \frac{1 - D}{1 - p(X)} \rho_r(Y - a_0) | X_1 = x_{10} \right],$$

$$q_{1,r}(x_{10}) = \inf_{a_1} E \left[ \frac{D}{p(X)} \rho_r(Y - a_1) | X_1 = x_{10} \right],$$

where $p(x) = P(D = 1|X = x)$, $j = 0, 1$.

Note that estimating $\Delta_r(x_{10}) = q_{1,r}(x_{10}) - q_{0,r}(x_{10})$ does not involve estimating the conditional distributions $F_1(Y(1)|X_1)$ and $F_0(Y(0)|X_1)$ that can be nonparametric. Thus, we can estimate $\Delta_r(x_{10})$ in a simpler manner. First, as a benchmark, we consider the oracle case with the given $p(X)$ and denote the corresponding oracle $CQTE$ estimator ($\hat{CQTE}_O(x_{10})$) as $\hat{\Delta}_r^O(x_{10})$. Note that for any value $X_{1i}$ that is close to $x_{10}$, Taylor expansion yields that $q_{1,r}(X_{1i}) \approx q_{1,r}(x_{10}) + q'_{1,r}(x_{10})(X_{1i} - x_{10})$. Thus, we can use the minimizer of the following loss function to define an estimator of $q_{1,r}(x_{10})$:

$$\left( \hat{q}_{1,r}^O(x_{10}), \hat{q}'_{1,r}^O(x_{10}) \right) = \arg \min_{a,b} \sum_{i=1}^{n} \frac{D_i}{p(X_i)} \rho_r(Y_i - a - b(X_{1i} - x_{10})) K \left( \frac{X_{1i} - x_{10}}{h} \right),$$

where $K(\cdot)$ is the kernel function and $h$ is the bandwidth. Similarly, we can define an estimator of $q_{0,r}(x_{10})$ under the same paradigm:

$$\left( \hat{q}_{0,r}^O(x_{10}), \hat{q}'_{0,r}^O(x_{10}) \right) = \arg \min_{a,b} \sum_{i=1}^{n} \frac{1 - D_i}{1 - p(X_i)} \rho_r(Y_i - a - b(X_{1i} - x_{10})) K \left( \frac{X_{1i} - x_{10}}{h} \right).$$

Hence, obtain the oracle $CQTE$ estimator as

$$\hat{CQTE}_O(x_{10}) = \hat{q}_{1,r}^O(x_{10}) - \hat{q}_{0,r}^O(x_{10}).$$

For unknown propensity score $p(X)$, there are several ways to get its estimator. If $p(X) = \pi(X, \beta)$ is known up to some unknown parameters $\beta$ such as the popular logistic model or probit model, we then need to estimate $\beta$. If we do not have such prior information on its structure, a nonparametric estimation is required such as the Nadaraya–Watson (N–W) estimation. Furthermore, when it has a semiparametric structure: $p(X) = p(\alpha^T X)$, where both the function $p(\cdot)$ and the $k \times q$ orthonormal matrix $\alpha$ are unknown with $q \leq k$. From the definition of $p(X)$, we can see that the information about $D$ from $X$ can be completely captured by the projected variables $\alpha^T X$. Thus, we can use the following conditional independence to present the above semiparametric structure:

$$\text{Constraint 1 : } D \perp X | \alpha^T X.$$  \hspace{1cm} (4)

It follows that $(Y(0), Y(1)) \perp D | \alpha^T X$. Note that (4) still holds if we replace $\alpha$ with any $\alpha C$, where $C \in \mathbb{R}^{q \times q}$ is any nonsingular matrix. In general, the matrix $\alpha$ can only be identifiable up to a rotation matrix $C$. Thus, under this dimension reduction framework, Li (2018) pointed out that the identifiable parameter in (4) is $\alpha C$ or in the other words, the space spanned by the columns of $\alpha$. In the literature, various methods have been proposed to estimate this space including sliced inverse regression (SIR; Li, 1991), and minimum average variance estimation (MAVE; Xia et al., 2002). As for determining the structural dimension $q$, several eigendecomposition-based
methodologies have been proposed in the literature, such as the sequential test method (Li, 1991) and the BIC-type method (Zhu et al., 2006). For ease of exposition, we assume the dimension \( q \) of \( \alpha \) is given. This semiparametric dimension reduction structure can not only alleviate the curse of dimensionality, but also maintain the model interpretation and flexibility simultaneously to greatly avoid model misspecification.

The three estimators of \( p(X) \) in parametric, nonparametric, and semiparametric scenarios are respectively as

\[
\hat{p}(X_i) = \pi(X_i, \hat{\beta}), \quad \hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} (D_i \log \pi(X_i, \beta) + (1 - D_i)(1 - \log \pi(X_i, \beta)));
\]

\[
\hat{p}(X_i) = \frac{1}{nh_{0,j}^q} \sum_{j \neq i} D_j L \left( \frac{X_j - X_i}{h_0} \right) / \frac{1}{nh_{0,j}^q} \sum_{j \neq i} L \left( \frac{X_j - X_i}{h_0} \right);
\]

\[
\hat{p}(X_i) = \hat{p}(\hat{\alpha}^\top X_i) = \frac{1}{nh_{2,j}^q} \sum_{j \neq i} D_j H \left( \frac{\hat{\alpha}^\top X_j - \hat{\alpha}^\top X_i}{h_2} \right) / \frac{1}{nh_{2,j}^q} \sum_{j \neq i} H \left( \frac{\hat{\alpha}^\top X_j - \hat{\alpha}^\top X_i}{h_2} \right),
\]

(5)

where \( L(\cdot) \) and \( H(\cdot) \) are two kernel functions, \( h_0 \) and \( h_2 \) are bandwidths, and \( \hat{\alpha} \) is an estimator derived by a sufficient dimension reduction method. To be specific, taking MAVE as an example of dimension reduction method, we can get the estimator \( \hat{\alpha} \) by solving the minimization problem

\[
\min_{a,a,b} \sum_{i=1}^{n} \{D_i - a_j - b_j \hat{\alpha}^\top (X_i - X_j)\}^2 \omega_{ij}.
\]

Here \( \omega_{ij} = H(\alpha^\top (X_i - X_j)/h_2) / \sum_{i=1}^{n} H(\alpha^\top (X_i - X_j)/h_2), a = (a_1, \ldots, a_n), b = (b_1, \ldots, b_n) \). Also, the estimation procedure can be implemented in practice by using the R package MAVE (Hang & Xia, 2019).

After having the estimation of \( p(\cdot) \), we now proceed to the step of estimating \( \Delta_r(x_{10}) \). Use \( \hat{p}(X) \) to replace \( p(X) \) in (2) and (3) and get:

\[
(\hat{q}_{1,r}(x_{10}), \hat{q}_{1,r}'(x_{10})) = \arg\min_{a,b} \sum_{i=1}^{n} \frac{D_i}{\hat{p}(X_i)} \rho_r(Y_i - a - b(X_{1i} - x_{10})) K_i,
\]

\[
(\hat{q}_{0,r}(x_{10}), \hat{q}_{0,r}'(x_{10})) = \arg\min_{a,b} \sum_{i=1}^{n} \frac{1 - D_i}{1 - \hat{p}(X_i)} \rho_r(Y_i - a - b(X_{1i} - x_{10})) K_i.
\]

(6)

Note that we use the local constant (Nadaraya–Watson [N–W] method) and local linear smoother to estimate \( p(\cdot) \) and the function \( q_{0,r}(x_{10}) \), respectively. This is mainly because of the following considerations. We note that the asymptotic bias of \( q_{0,r}(x_{10}) \) has no relationship with \( \hat{p}(X) \) as long as its rate of convergence can be fast sufficiently. We then use a simpler estimation for \( \hat{p}(x) \) for ease of exposition, and the local linear smoother for \( q_{0,r}(x_{10}) \) such that the asymptotic analysis can be carefully worked out.

For convenience, denote the estimator \( \hat{\Delta}_r(x_{10}) \) incorporated with the parametric estimator \( \hat{p}(X) \) as \( \hat{CQTE}_P(x_{10}) \), and with the other two nonparametric and semiparametric estimators \( \hat{p}(X) \) separately as \( \hat{CQTE}_N(x_{10}) \) and \( \hat{CQTE}_S(x_{10}) \). As the asymptotic results vary with the different estimators \( \hat{p}(X) \) we present them in the separate subsections. For the sake of comparison, all the CQTE estimators are based on the same bandwidth, \( h \) and kernel function, \( K(\cdot) \).
2.2 Asymptotic properties of $\hat{CQTE}_{p}$

Before stating the asymptotic results, define some important quantities:

1. $f_{j}(y_{j}|x_{1})$ to be the value of the conditional density function of $(Y(j)|X_{1})$, $j = 0, 1$ at the point $(Y(j) = y(j), X_{1} = x_{1})$;
2. $\hat{q}_{j}(x_{1}) = q_{j}(x_{10}) + q_{j}(x_{10})(X_{1} - x_{10})$, $m_{j}(X) = \frac{E[\{Y(j)\leq q_{j}(x_{10})\} - I]}{f_{j}(q_{j}(x_{10})(X_{1} - x_{10}))}$, $j = 0, 1$;
3. $\psi(p(X_{i}), Z_{i}) = \frac{D_{i}}{p(X_{i})} \frac{\eta_{j}(Y_{i}) - \frac{1 - D_{i}}{1 - p(X_{i})} \eta_{0}(Y_{i})}{\frac{E[(\{Y(i)\leq q_{j}(x_{10})\}) - I]}{f_{j}(q_{j}(x_{10})(X_{1} - x_{10}))}}$, $\sigma_{j}^{2}(x_{10}) = E(\psi^{2}(p(X_{i}), Z)|X_{1} = x_{10}) = \frac{E((\{Y(i)\leq q_{j}(x_{10})\}) - I)^{2}|X_{1} = x_{10})}{(1 - p(X_{i}))^{2}f_{j}(q_{j}(x_{10})(X_{1} - x_{10}))}$ with $Z_{i} = (X_{i}, D_{i}, Y_{i})$, and $\eta_{j}(Y_{i}) = \frac{1}{f_{j}(q_{j}(x_{10}))}$;
4. $\mu_{s}(K) = \int u_{1}^{p_{1}} \cdots u_{l}^{p_{l}}K(u)du$ for integers $p_{1}, \ldots, p_{l}$ such that $\sum_{i=1}^{l} p_{i} = s_{1}$, $\|K\|_{2}^{2} = \int K^{2}(u)du$, and $K_{l} = K(\frac{X_{11} - x_{10}}{h})$.

Give the following assumptions. Recall the definition of high-order kernel in the literature. We say a function $g : R^{r} \rightarrow R$ is a kernel of order $s$ if it integrates to one over $R^{r}$, and $\int u^{p_{1}} \cdots u^{p_{r}}g(u)du = 0$ for all nonnegative integers $p_{1}, \ldots, p_{r}$ such that $1 \leq \sum_{i} p_{i} < s$, and it is nonzero when $\sum_{i} p_{i} = s$.

- Assumption 3 (on distribution):
  (i) The support of the $k$-dimensional covariate vector $X, \chi$, is a Cartesian product of compact intervals. The density functions of $X$ and $(X_{1}, \chi^{T}X)$ are bounded away from zero and infinity and $s_{1} \geq 2$ times continuously differentiable.
  (ii) The density function of $X_{1}, f(X_{1})$, and the conditional density $f_{j}(Y(j)|X_{1})$, are bounded away from zero and infinity and continuously differentiable.
- Assumption 4 (on kernel function):
  $K(u)$ is a kernel of order $s_{1}$, is symmetric around zero, and is $s^{s}$ times continuously differentiable.
- Assumption 5: $h \rightarrow 0$, $nh^{l} \rightarrow \infty$, $nh^{2s_{1}+l+2} \rightarrow 0$.
- Assumption 6 (parametric propensity score estimator): The estimator $\hat{\beta}$ of the propensity score model $\pi(X, \beta)$, $\beta \in \Theta \subset R^{d}$, $d < \infty$, satisfies $\sup_{X \in X} |\pi(X, \hat{\beta}) - \pi(X, \beta_{0})| = O_{p}(n^{-1/2})$, where $\beta_{0} \in \Theta$ such that $p(X) = \pi(X, \beta_{0})$ for all $X \in \chi$.

Assumption 3 is commonly used for nonparametric estimation in the literature. Assumption 4 is for high-order kernel. When $l = 1$ and $s_{1} = 2$, Gaussian kernel satisfies this assumption. Furthermore, the value of $s^{s}$ depends on the estimation procedure to ensure the function smoothness which will be used in studying the asymptotic behaviors of the estimators. To be more specific, $s^{s} \geq 2$ in the case of $\hat{CQTE}_{p}$, while $s^{s} \geq s$ and $s^{s} \geq s_{2}$ in the case of $\hat{CQTE}_{N}$ and $\hat{CQTE}_{S}$, respectively. Assumption 5 is a condition on the bandwidth selection. Obviously, if we assume $nh^{2s_{1}+l} \rightarrow 0$, the $CQTE$ estimators can be asymptotically unbiased. However, to better analyze the bandwidth selection rule, we only assume $nh^{2s_{1}+l+2} \rightarrow 0$. Assumption 6 is a typical result if we estimate $p(X) = \pi(X, \beta_{0})$ by a parametric model like a logit model or a probit model based on a linear index via the maximum likelihood method. Based on the above discussion, the asymptotic results of $\hat{CQTE}_{p}$ are stated in the following theorem.
Theorem 1. Suppose that Assumptions 1–6 are satisfied for $s_1 \geq 2$. Then, $\widehat{CQTE}_p(X_{10})$ has the asymptotically linear representation as
\[
\sqrt{nh^l} \left( \widehat{\Delta}_r(x_{10}) - \Delta_r(x_{10}) \right) = -\frac{1}{\sqrt{nh^l}} \frac{1}{f(x_{10})} \sum_{i=1}^{n} \phi_1(p(X_i), Z_i)K_i + o_p(1),
\]
and the asymptotic distribution is
\[
\sqrt{nh^l} \left( \widehat{\Delta}_r(x_{10}) - \Delta_r(x_{10}) - b_1(x_{10}) \right) \xrightarrow{D} N \left( 0, \frac{\|K\|^2_2 \sigma_2^2(x_{10})}{f(x_{10})} \right), \quad \forall x_{10} \in \Omega,
\]
where $\phi_1(p(X), Z) = \psi(p(X), Z)$, $b_1(x_{10}) = O_p(\mu_{s_1}(K)h^{s_1})$, $\sigma_2^2(x_{10}) = \sigma_2^2(x_{10})$.

Furthermore, $\widehat{CQTE}_p$ is asymptotically equivalent to $CQTE_0$ in distribution.

Remark 1. To implement the estimation procedure, we also need to choose the bandwidth, $h$. It is well known that having a balance between asymptotic bias and variance of $\widehat{\Delta}_r(x_{10})$, we can select an optimal bandwidth by minimizing the asymptotic weighted mean integrated squared error (AMISE) (e.g., Fan & Gijbels, 1992). For example, when the order of kernel function $K(\cdot)$ is $s_1 = 2$, the corresponding AMISE is
\[
\int \left( \frac{1}{4} \|\Delta''_r(x_1)\mu_2(K)\|^2 h^4 + \frac{1}{nh^l} \frac{\|K\|^2_2 \sigma_2^2(x_1)}{f(x_1)} \right) w(x_1) dx_1,
\]
and then the corresponding optimal bandwidth is $h^p_{opt} = \left( \frac{\|K\|^2_2 \mu_2(K) f(x_1) w(x_1) dx_1}{\|\Delta''_r(x_1)\|^4 w(x_1) dx_1} \right)^{\frac{1}{2}} \rightarrow n \rightarrow \frac{1}{\sqrt{n}}$, where $w(x_1)$ is a nonnegative, bounded weight function with bounded support, which is contained in the interior of the support of $f(X_i)$.

To further conduct statistical inference, a consistent estimator of the asymptotic variance functions is required. A commonly used method is by plugging appropriate estimators for those unknown parameters into the asymptotic variance formula. For example, a plug-in estimator for $\sigma_2^2(x_{10})$ is
\[
\hat{\sigma}_2^2(x_{10}) = \left[ \frac{1}{nh^l} \sum_{i=1}^{n} \left( \hat{\phi}_1(p(X_i), Z_i) \right)^2 K_i \right] / \hat{f}(x_{10}),
\]
where $\hat{f}(x_{10}) = \sum_{i=1}^{n} K_i / (nh^l)$, and $\hat{\phi}_1(p(X_i), Z_i) = \frac{D \hat{\eta}_r(Y_i)}{\hat{p}(X_i)} - \frac{(1-D) \hat{\eta}_r(Y_i)}{1-\hat{p}(X_i)}$ with $\hat{p}(X_i) = \pi(X_i, \hat{\beta})$ and $\hat{\eta}_r(Y_i) = [\|Y_i \leq \hat{q}_{j, r}(x_{10})\| - \tau] / \hat{f}(\hat{q}_{j, r}(x_{10}))$, $j = 0, 1$.

Under certain regularity conditions, it is expected that the sample analogues are consistent. Hence we omit the technical details in this article. Meanwhile, we also see that this estimation method involves many unknowns and thus may have relatively large estimation error in finite sample scenarios. An alternative is the nonparametric bootstrap approximation (Efron, 1979). Given $X_1 = x_{10} \in \Omega$, a bootstrap-based estimator for the asymptotic variance of $\hat{\Delta}_r(x_{10})$ can be obtained by:
\[
\widehat{\text{Var}}[\Delta^p_r(x_{10})] = \frac{1}{B - 1} \sum_{b=1}^{B} \left[ \hat{\Delta}^p_r(x_{10}) - \tilde{\Delta}^p_r(x_{10}) \right]^2,
\]
where $\hat{\Delta}^p_r(x_{10})$ is obtained from the $b$th bootstrapped sample for $b = 1, \ldots , B$. 
Based on the above two asymptotic variance estimators, we can obtain two pointwise consistent estimators for the standard error of \( \sqrt{nhl} (\hat{\Delta}_r^P(x_{10}) - \Delta_r(x_{10})) \) and then two \((1 - \alpha)100\%\) pointwise confidence intervals for \( \Delta_r(x_{10}) \), that is,

\[
\Delta_r^P(x_{10}) \pm c_{\alpha/2} \left( \frac{\|K\|_\infty^2 \hat{\sigma}_r^2(x_{10})}{nh^l f(x_{10})} \right)^{1/2},
\]

and

\[
\hat{\Delta}_r(x_{10}) \pm c_{\alpha/2} \left( \text{Var}[\hat{\Delta}_r^P(x_{10})] \right)^{1/2},
\]

where \( c_{\alpha/2} \) is the \((1 - \alpha/2)\) quantile of the standard normal distribution.

### 2.3 Asymptotic properties of \( \widehat{CQTE}_N \)

We make some additional assumptions about kernel functions \( L(\cdot) \) and bandwidths \( h \) and \( h_0 \) to back up the theoretical development.

- Assumption 7: \( L(u) \) is a kernel of order \( s \geq k + l \), is symmetric around zero, has finite support \([-M, M]^k\), and its \((s + 1)\)th derivative is continuous. Furthermore, the density function of \( X \), \( f_s(X) \), is \( s \) times continuously differentiable and bounded away from zero and infinity.
- Assumption 8: \( h \to 0 \) and \( \log(n)/(nh_0^{k+s}) \to 0 \).
- Assumption 9: \( h_0^2h^{-2s-l} \to 0 \), \( nh^{-1}h_0^{2s} \to 0 \).

Assumption 7 is also to ensure the smoothness of the density function. Assumptions 8 and 9 are the technical conditions to guarantee the existence of the limiting distribution when we need to prove the asymptotic negligibility of all remainders. These are because of the involvement of two bandwidths.

**Theorem 2.** Suppose that Assumptions 1–5 and 7–9 are satisfied for some \( s^* \geq s \geq k + l \), for each point \( x_{10} \in \Omega \), the asymptotically linear representation of \( \widehat{CQTE}_N(x_{10}) \) is

\[
\sqrt{nhl} \left( \Delta_r^N(x_{10}) - \Delta_r(x_{10}) \right) = -\frac{1}{\sqrt{nhl} f(x_{10})} \sum_{i=1}^{n} \phi_2(p(X_i), Z_i)K_i + o_p(1).
\]

The asymptotic distribution of \( \hat{\Delta}_r^N(x_{10}) \) is

\[
\sqrt{nhl} \left( \hat{\Delta}_r^N(x_{10}) - \Delta_r(x_{10}) - b_2(x_{10}) \right) \xrightarrow{D} N \left( 0, \frac{\|K\|_\infty^2 \hat{\sigma}_r^2(x_{10})}{f(x_{10})} \right),
\]

where \( \phi_2(p(X_i), Z_i) = \psi(p(X_i), Z_i) - n_p(X_i)e_i \), \( n_p(X_i) = \frac{h_0}{p(X_i)} + \frac{m_0(X_i)}{p(X_i)} \), \( b_2(x_{10}) = O_p(\mu_s(K)h^{s_1}) \), \( \epsilon = D - p(X) \).

Rewrite \( \phi_2(p(X_i), Z_i) \) as \( \phi_2(p(X_i), Z_i) = \frac{Dp_0(X_i) - m_{0,r}(X_i)}{p(X_i)} - \frac{(1-D)(m_{0,r}(X_i) - p_0(X_i))}{1-p(X_i)} + m_{1,r}(X_i) - m_{0,r}(X_i) \). We then get the following corollary.
Corollary 1. Under the regularity conditions in the previous theorems,

\begin{align*}
(1) & \quad \sigma^2_N(x_{10}) = E \left\{ (m_{1,r}(X) - m_{0,r}(X))^2 + \frac{\sigma^2(X)}{p(X)} + \frac{\sigma^2(X)}{1-p(X)} \middle| X_1 = x_{10} \right\}, \\
(2) & \quad \sigma^2_\tau(x_{10}) = \sigma^2_N(x_{10}) + E \left\{ \left( \frac{m_{1,r}(X)}{p(X)} + \frac{m_{0,r}(X)}{1-p(X)} \right)^2 \middle| X_1 = x_{10} \right\} \geq \sigma^2_N(x_{10}),
\end{align*}

where \( \sigma^2_\tau(X) = \text{Var} \left( \frac{I(y_j \leq q_{ij}(x_{10})) - \tau}{f_j(q_{ij}(x_{10}))} \right) | X \).

Remark 2. From all the results above, the asymptotic variances are comparable as they have the same constant \(|K|_2\) and the density function \(f(x_{10})\) except for \(\sigma^2_\tau(x_{10})\)’s, \(i = O, P, N\). Corollary 1 implies that for any \(x_{10}, \text{CQTE}_N \leq \text{CQTE}_P \approx \text{CQTE}_O\). As discussed before, we cannot show whether \(\text{CQTE}_N(x_{10})\) is the most efficient CQTE estimator as the standard semiparametric estimation theory derived in this article cannot provide the idea on the optimal bandwidth selection. This phenomenon is very different from the unconditional quantities, for example, \(\text{ATE}\) and \(\text{QTE}\), which can achieve the semiparametric efficient bound when nonparametrically estimated propensity score is used. But looking at all the asymptotic variance functions, \(\text{CQTE}_N(x_{10})\) is the most efficient estimator and thus, we conjecture that \(\text{CQTE}_N(x_{10})\) would achieve an efficient bound in certain sense. This deserves a further study.

Remark 3. From the estimation procedure for \(\hat{\Delta}_N^r(x_{10})\), we can see that delicately choosing the bandwidth \(h_0\) in the estimated propensity score for \(\text{CQTE}_N\) is necessary. Also, for the final estimator the bandwidth \(h\) is in need. Since the bandwidth selection procedure is not easy to balance the magnitudes between these bandwidths, \(\text{CQTE}_N\) can be sensitive to the selected bandwidths. The results in simulation show this phenomenon. Furthermore, we note that the first-order asymptotic theory derived in this article cannot provide the idea on the optimal bandwidth selection. Thus, we use the rule of thumb to guide the selection. Recall that the corresponding bandwidths \(h_0\) and \(h\) should satisfy Assumptions 5 and 7–9. Thus to fulfill the assumptions, take

\[ h = a \cdot n^{1/4(\delta+\delta_0)}, \quad h_0 = a_1 \cdot n^{1/\delta_0}, \quad \text{for } a > 0, \delta > 0, a_1 > 0, \delta_0 > 0. \]  

(10)

Note that \(\delta\) and \(\delta_0\) can be made as small as necessary or desired, thus we set them as zero. Furthermore, the orders of kernel function are set as: \(s_1 = s\) and \(s = k + l + \delta^*\). Here when \(k + l\) is even, \(\delta^* = 1\), otherwise \(\delta^* = 0\).

As for the final estimator \(\hat{\Delta}_N^r(x_{10})\), we can also construct the two types of asymptotic variance estimators similarly as \(\hat{\Delta}_P(x_{10})\). A plug-in estimator of \(\sigma^2_N(x_{10})\) can be constructed as

\[ \hat{\sigma}^2_N(x_{10}) = \left[ \frac{1}{nh^l} \sum_{i=1}^n \{ \hat{\phi}_2(\hat{p}(X_i), Z_i) \}^2 K_i \right] \hat{f}(x_{10}), \]  

(11)

where \(\hat{\phi}_2(\hat{p}(X_i), Z_i) = \frac{D_i(\hat{\eta}_{1r}(Y_i) - \hat{m}_{1,r}(X_i))}{\hat{p}(X_i)} - \frac{(1-D_i(\hat{\eta}_{0r}(Y_i) - \hat{m}_{0,r}(X_i))}{1-\hat{p}(X_i)} + \hat{m}_{1,r}(X_i) - \hat{m}_{0,r}(X_i)\) with \(\hat{p}(X_i)\) being the nonparametric estimator in (5) and \(\hat{m}_{1,r}(X_i)\) also being a consistent nonparametric estimator for \(m_{1,r}(X_i), j = 0, 1\). Meanwhile, a bootstrap-based estimator for the asymptotic variance of \(\hat{\Delta}_r^N(x_{10})\) can also be similarly obtained as:

\[ \text{Var}[\hat{\Delta}_r^N(x_{10})] = \frac{1}{B - 1} \sum_{b=1}^B \left[ \hat{\Delta}_r^N(x_{10}) - \Delta_r^N(x_{10}) \right]^2. \]
We can then construct two types of the $(1 - \alpha)100\%$ pointwise confidence intervals for $\Delta_r(x_{10})$ as (8) and (9) by replacing $\hat{\Delta}_r^P(x_{10})$ with $\hat{\Delta}_r^N(x_{10})$.

### 2.4 Asymptotic properties of $\widehat{\text{CQTE}}_S$

If we postulate that the information about $D$ from $X$ can be completely captured by $q$ linear combinations $a^T X$ of $X$ with $q \ll k$, we can then estimate the propensity score function $p(X) = p(a^T X)$ with $a^T X$ rather than the original $X$ to avoid the curse of dimensionality. To this end, we can use a lower dimensional kernel function $H(u)$, instead of a high-dimensional kernel function $L(X)$ to get the local smooth estimator $\hat{p}(a^T X)$ of $p(X)$.

We first correspondingly rectify the assumptions related to $L(X)$ and bandwidth in Subsection 2.3.

- **Assumption 7**: $H(u)$ is symmetric around zero, has finite support $[-M, M]^q$, and is $s_2 \geq q + l$ times continuously differentiable. The density function of $a^T X$, $f_a(a^T X)$ is $s_2$ times continuously differentiable.

- **Assumption 8**: $h_2 \to 0$ and $\log(n)/(nh_2^{s_2+q}) \to 0$.

- **Assumption 9**: $h_2^{2s_2} h^{-2s_2-l} \to 0$, $nh^{-1} h_2^{2s_2} \to 0$.

- **Assumption 10**: $\hat{\alpha}$ is a root-$n$ consistent estimator of $\alpha$ and $q$, the dimension of $\alpha$, is given with $q \ll k$.

Since the propensity score of $\widehat{\text{CQTE}}_S(x_{10})$ is based on $a^T X$, Assumptions 7’–9’ are adjusted to those in the case of $\widehat{\text{CQTE}}_N(x_{10})$ and play the same role.

Furthermore, we give some notation. Let $\text{card}(A)$ be the cardinality of set $A$. $A \subseteq B$ means $A \cap B = A$, that is, all elements of $A$ are also elements of $B$, while $A \not\subseteq B$ means $A \subseteq B$ but $\text{card}(A) < \text{card}(B)$. $A \sqsubset B$ means $A \cap B = C$ with the cardinality $\text{card}(C) = t$, that is, there only exist $t$ elements of $A$ belonging to $B$. Especially $t = 0$, means $A$ and $B$ are disjoint.

**Theorem 3.** Suppose Assumptions 1–4 and 7’–10’ are satisfied for $s^* \geq s_2 \geq q + l$, the following statements hold for each point $x_{10} \in \Omega$:

1. When $X_1 \sqsubseteq a^T X$ with $0 < r \leq l$ and $s_2(2 - l/r) + l > 0$, the asymptotically linear representation of $\widehat{\text{CQTE}}_S(x_{10})$ is

$$
\sqrt{nh^l} \left( \hat{\Delta}_r^S(x_{10}) - \Delta_r(x_{10}) \right) = -\frac{1}{\sqrt{nh^l} f(x_{10})} \frac{1}{n} \sum_{i=1}^{n} \phi_3(p(a^T X_i), Z_i) K_i + o_p(1),
$$

and the asymptotic distribution of $\hat{\Delta}_r^S(x_{10})$ is

$$
\sqrt{nh^l} \left( \hat{\Delta}_r^S(x_{10}) - \Delta_r(x_{10}) - b_3(x_{10}) \right) \overset{D}{\to} \mathcal{N} \left( 0, \frac{\|K\|^2 \sigma^2_3(x_{10})}{f(x_{10})} \right).
$$

2. When $X_1 \subseteq a^T X$, the asymptotically linear representation of $\widehat{\text{CQTE}}_S(x_{10})$ is

$$
\sqrt{nh^l} \left( \hat{\Delta}_r^S(x_{10}) - \Delta_r(x_{10}) \right) = -\frac{1}{\sqrt{nh^l} f(x_{10})} \frac{1}{n} \sum_{i=1}^{n} \phi_3^*(p(a^T X_i), Z_i) K_i + o_p(1),
$$

where $\phi_3^*$ and $\phi_3$ are defined in (A.12) and (A.10), respectively.
and the asymptotic normality of \( \hat{\Delta}_S^S(x_{10}) \) is

\[
\sqrt{nh^l} \left( \hat{\Delta}_S^S(x_{10}) - \Delta_e(x_{10}) - b_3^*(x_{10}) \right) \xrightarrow{D} N \left( 0, \frac{\|K\| \sigma^2_3(x_{10})}{\hat{f}(x_{10})} \right).
\]

(3) \( \hat{\text{CQTE}}_S(x_{10}) \) has a limiting variance that is smaller or equal to those of \( \hat{\text{CQTE}}_p(x_{10}) \) and the oracle \( \hat{\text{CQTE}}_O(x_{10}) \) as below:

\[
\sigma^2_S(x_{10}) = \sigma^2_O(x_{10}) = \sigma^2_S(x_{10}) = \sigma^2_S^*(x_{10})
\]

\[
+ E \left[ p(a^TX)(1 - p(a^TX)) \left( \frac{m_{1,r}(a^TX)}{p(a^TX)} + \frac{m_{0,r}(a^TX)}{1 - p(a^TX)} \right)^2 |X_1 = x_{10} \right] \geq \sigma^2_S(x_{10}).
\]

where \( \phi_3(p(a^TX), Z_i) = \psi(p(X_i), Z_i), b_3(x_{10}) = O_p(\mu_{e_1}(K)h^3) \). \( \phi_3^*(p(a^TX), Z_i) = \psi(p(a^TX), Z_i) - e_p(a^TX) \), \( e_p(a^TX) = \frac{m_{1,r}(a^TX)}{p(a^TX)} + \frac{m_{0,r}(a^TX)}{1 - p(a^TX)} \), \( b_3^*(x_{10}) = O_p(\mu_{e_1}(K)h^3) \), and \( e = D - p(a^TX) \).

Remark 4. We should also note that when \( a^TX = X_1, E(m_{j,r}(a^TX)|X_1 = x_{10}) = 0, j = 0, 1 \), and the asymptotic variance is

\[
\sigma^2_p(x_{10}) = \sigma^2_S(x_{10}) = \frac{\tau(1 - \tau)}{p(x_{10})f'_1(q_{1,r}(x_{10})|x_{10})} + \frac{\tau(1 - \tau)}{(1 - p(x_{10}))f'_2(q_{0,r}(x_{10})|x_{10})}.
\]

That implies, when \( a^TX = X_1, \hat{\text{CQTE}}_S \) cannot be more efficient than \( \hat{\text{CQTE}}_p \) even when the propensity score is estimated nonparametrically. This is an essential difference from the unconditional counterpart. But when \( X_1 \not\subset a^TX \), the nonparametric structure of the \( \hat{\text{CQTE}}_S(x_{10}) \) estimator does play a positive role in estimation efficiency. In Subsection 2.5, we give some more discussion and more general results to provide a relatively complete picture of estimation efficiency in this field.

Remark 5. To choose the bandwidths \( h_2 \) and \( h_0 \) in \( \hat{\text{CQTE}}_S \), we can just replace the role of \( k \) in \( \hat{\text{CQTE}}_N \) with \( q \), that is,\( h = b \cdot n^{-\frac{1}{2 + \delta}}, h_2 = b_1 \cdot n^{-\frac{1}{4 + 2\delta + b}} \), for \( b > 0, \delta > 0, b_1 > 0, \delta_0 > 0 \).

Next, we also consider to make further statistical inference based on \( \hat{\Delta}_S^S(x_{10}) \). Note that the specification formula of asymptotic variance for \( \hat{\Delta}_S^S(x_{10}) \) in Theorem 3 depends on the value of \( |X_1 \cap a^TX| \). Hence, we can construct the plug-in asymptotic variance estimator based on the value of \( |X_1 \cap a^TX| \) in practice. When \( |X_1 \cap a^TX| \approx l \), we can get the estimator as

\[
\hat{\sigma}^2_S(x_{10}) = \left[ \frac{1}{nh^l} \sum_{i=1}^n \{ \hat{\phi}_3(\hat{p}(X_i), Z_i) \}^2 K_i \right] / \hat{f}(x_{10}),
\]

where \( \hat{\phi}_3(\hat{p}(X_i), Z_i) \) is similar to the one in (7). When \( |X_1 \cap a^TX| \approx l \),

\[
\hat{\sigma}^2_S^{*}(x_{10}) = \left[ \frac{1}{nh^l} \sum_{i=1}^n \{ \hat{\phi}_3^*(\hat{p}(X_i), Z_i) \}^2 K_i \right] / \hat{f}(x_{10}),
\]
where \( \hat{\phi}_3^*(\hat{p}(X_i), Z_i) \) is similar to the one in (11). As for the bootstrap-based estimator for the asymptotic variance of \( \hat{\Delta}_r^S(x_{10}) \), it can still be obtained as

\[
\hat{\text{Var}}[\hat{\Delta}_r^S(x_{10})] = \frac{1}{B - 1} \sum_{b=1}^{B} \left[ \hat{\Delta}_{r,b}^S(x_{10}) - \hat{\Delta}_r^S(x_{10}) \right]^2.
\]

Hence, the further inference procedure can be implemented.

### 2.5 Further studies about the role of propensity score in estimation efficiency

The above results about \( \text{CQTE} \) present two interesting phenomena. In the scenario Theorem 2 presents, \( \hat{\text{CQTE}}_N(x_{10}) \) can be asymptotically more efficient than \( \hat{\text{CQTE}}_P(x_{10}) \) and the oracle one \( \hat{\text{CQTE}}_O \). Yet, in the scenario Theorem 3 shows, \( \hat{\text{CQTE}}_S(x_{10}) \) cannot always be so although \( \hat{\text{CQTE}}_S(x_{10}) \) also uses the nonparametric method to estimate the propensity score.

This motivates us to further investigate the role of the estimated propensity score in the asymptotic behaviors of the \( \text{CQTE} \) estimators. At first glance, it seems that the different asymptotic behaviors are because of different estimation methods for propensity score. Comparing \( \hat{\text{CQTE}}_S(x_{10}) \) with \( \hat{\text{CQTE}}_N(x_{10}) \) and the technical proofs for Theorem 2 with that for Theorem 3 in supplementary material, we note that this is not completely true while two factors play the important role in the estimation efficiency: How fast is the convergence rate of the estimated propensity score and whether \( X_1 \) is a strict subset of the true arguments of the propensity score. We then separately discuss them.

#### 2.5.1 The role of convergence rate of the estimated propensity score

From the technical proofs and the main differences between \( \hat{\text{CQTE}}_P \) and \( \hat{\text{CQTE}}_N / \hat{\text{CQTE}}_S \) we can see that fast rate of convergence can make the first-order expansion of the estimated propensity score such as \( \hat{\text{CQTE}}_P \) vanish while slow rate such as for \( \hat{\text{CQTE}}_N / \hat{\text{CQTE}}_S \) cannot cancel off it and thus enhance the asymptotic efficiency due to a negative correlation with the leading term. Thus, \( \hat{\text{CQTE}}_N \) and \( \hat{\text{CQTE}}_S \) can lose their efficiency superiority if the nonparametric estimates converge faster with higher order smoothness as concluded in the following corollary.

**Corollary 2.** In addition to the conditions in Theorems 2 and 3 respectively with replacing the assumptions on the bandwidths by \( \sqrt{nh}_{0}^1 \left( h_0^s + \sqrt{\log(n)/nh_0^k} \right) = o(1) \) and \( \sqrt{nh}_{2}^1 \left( h_2^s + \sqrt{\log(n)/nh_2^k} \right) = o(1) \), \( \hat{\text{CQTE}}_N(x_{10}) \) and \( \hat{\text{CQTE}}_S(x_{10}) \) have the same asymptotic distribution as \( \hat{\text{CQTE}}_P(x_{10}) \). That is,

\[
\hat{\text{CQTE}}_O(x_{10}) \cong \hat{\text{CQTE}}_P(x_{10}) \cong \hat{\text{CQTE}}_N(x_{10}) \cong \hat{\text{CQTE}}_S(x_{10}).
\]

**Remark 6.** Obviously, the above discussion is mainly for theoretical investigation. In practice, it makes no sense to choose such bandwidths to use. But the discussion is still helpful for us to better understand the estimation mechanisms. These results further reveal the essential differences between the conditional and unconditional structure. As we know, under the unconditional
structure, the estimator requires a standardizing constant $\sqrt{n}$ to derive its limiting distribution. Any estimator of propensity score is at the rate of order $1/\sqrt{n}$ or slower, the impact from the estimated propensity score will play a role in reducing the asymptotic variance. See Hirano et al. (2003) and Liu et al. (2018). By contrast, under the conditional structure, the estimator requires a standardizing constant $\sqrt{nh^l}$ to approximate its limiting distribution. When an estimator of propensity score is at a rate $b_n$ faster than $1/\sqrt{nh^l}$ such as the case where propensity score is parametric and an estimator has the rate of order $1/\sqrt{n}$, the impact from the estimated propensity score will play no role in the asymptotic variance reduction. For CATE, Abrevaya et al. (2015) showed the case with parametric propensity score, but did not include the discussion on nonparametric and semiparametric cases. Thus, for the estimated semiparametric and non-parametric propensity score, delicately choosing bandwidth to obtain a proper rate of convergence becomes vital for the estimation efficiency. It is clear that in the above corollary, the conditions $\sqrt{nh^l} \left( h_2^2 + \sqrt{\log(n)/nh^k_2} \right) = o(1)$ and $\sqrt{nh^l} \left( h_2^2 + \sqrt{\log(n)/nh^k_2} \right) = o(1)$ are much stronger than the assumptions in Theorems 2 and 3, but is still possible to choose so as long as the involved functions are sufficiently smooth and high-order kernels are used. This is because for the non-parametric estimation, the rate of convergence can be as close to $1/\sqrt{n}$ as possible when the function is very smooth. However, utilizing a high-order kernel for regression fit means we would assign negative weights to some range of the data, which is an undesirable side effect. See Li and Racine (2007).

Remark 7. The theorems in this article also add new insights about the superefficiency phenomenon found in missing data and treatment effect area. That is, for unconditional treatment effect, generally inverse of propensity score-based estimators with estimated propensity score is more efficient than the one with true propensity score. As discussed above, estimating propensity score is not necessary to play a role in the asymptotic variance reduction.

2.5.2 The effect of the affiliation of $X_1$ to the set of true arguments of propensity score function

As pointed out before, $CQTE$ is a function of $X_1$, and its affiliation to the set of all arguments of the propensity score plays a role for estimation efficiency. Recall that in the scenario Theorem 3 discusses, the asymptotic distribution of $\hat{CQTE}_S(x_10)$ depends on the relationship between $X_1$ and $a^\top X$. Therefore, under Constraint 2 below, it can be expected that $\hat{CQTE}_N(x_10)$ will have similar properties, namely, the affiliation of $X_1$ to $X$ (or $a^\top X$) should affect the asymptotic distribution of $\hat{CQTE}_N(x_10)$ (or $\hat{CQTE}_S(x_10)$).

Constraint 2 : $D \perp X|\bar{X}$. \hfill (13)

Thus we call $\bar{X}$ the set of true arguments of propensity score, $p(X) = p(\bar{X})$ and $\bar{k} = \text{dim}(\bar{X})$. Obviously, when $\bar{X} = X$, $X_1 \subseteq \bar{X}$, $\hat{CQTE}_N(x_10)$ can be more efficient than $\hat{CQTE}_P(x_10)$ by Theorem 2. Recall that $X_1 \sqsubset^{l-r} \bar{X}$ means that $X_1$ has $l - r$ components of $\bar{X}$. We will see that when $X_1 \sqsubset^{l-r} \bar{X} \not\subseteq X$, the situation becomes different as concluded by the following corollary.

Corollary 3. Suppose that there is a given $\bar{X}$ such that $D \perp X|\bar{X}$ with $X_1 \sqsubset^{l-r} \bar{X} \not\subseteq X$ and $0 < r \leq l$. Then if the propensity score $p(\bar{X})$ is estimated by basing on $\bar{X}$ rather than $X$, under the conditions in Theorem 2 and $s(2 - 1/r) + l > 0$, $\hat{CQTE}_N(x_10)$ has the same asymptotic distribution as $\hat{CQTE}_P(x_10)$. Then $\hat{CQTE}_N \cong \hat{CQTE}_P$. 
Furthermore, we clarify the relation between $\widehat{CQTE}_N(x_{10})$ and $\widehat{CQTE}_S(x_{10})$ when both $X_1 \subsetneq \bar{X}$ and $X_1 \subsetneq a^T X$ hold.

**Corollary 4.** Suppose all the assumptions listed above and the two assumptions (4) and (13) are satisfied, namely $p(X) = \hat{p}(X) = p(a^T X)$, and $X_1 \subsetneq \bar{X}$ and $X_1 \subsetneq a^T X$, we have the following asymptotic variance functions of $CQTE_S$ and $CQTE_N$:

$$
\sigma^2_S(x_{10}) = \sigma^2_N(x_{10}) + E \left[ p(a^T X)(1 - p(a^T X)) \left\{ \frac{\Delta m_{1,r}}{p(a^T X)} + \frac{\Delta m_{0,r}}{1 - p(a^T X)} \right\} \right]^2 | X_1 = x_{10}.
$$

where $\Delta m_{j,r} = m_{j,r}(\hat{X}) - m_{j,r}(a^T X)$.

We are now in the position to summarize all results about the affiliation effect of $X_1$.

1. $\widehat{CQTE}_N(x_{10}) \lhd CQTE_S(x_{10}) \lhd \widehat{CQTE}_P(x_{10}) \lhd \widehat{CQTE}_O(x_{10})$, for $X_1 \subsetneq \bar{X}$ and $X_1 \subsetneq a^T X$;
2. $\widehat{CQTE}_N(x_{10}) \lhd \widehat{CQTE}_S(x_{10}) \lhd \widehat{CQTE}_P(x_{10}) \lhd \widehat{CQTE}_O(x_{10})$, for $X_1 \subsetneq \bar{X}$ and $X_1 \subsetneq a^T X$;
3. $\widehat{CQTE}_N(x_{10}) \lhd \widehat{CQTE}_S(x_{10}) \lhd \widehat{CQTE}_P(x_{10}) \lhd \widehat{CQTE}_O(x_{10})$, for $X_1 \subsetneq a^T X \subsetneq \bar{X}$ and $X_1 \subsetneq a^T X$.

**Remark 8.** It is very interesting that whether $\widehat{CQTE}_N(x_{10})$ and $\widehat{CQTE}_S(x_{10})$ can be asymptotically more efficient also relies on whether the given covariates are a strict subset of the arguments of the propensity score. This important observation is not easy to explain, but might be because of the following. Note that when it does not include the given covariates, then under the conditional structure, the estimated propensity score is, in the large sample sense and in probability, conditionally independent of the conditional treatment effect and then plays little role for the asymptotic property of the estimated treatment effect. It deserves further study to confirm this explanation.

## 3 SIMULATION STUDIES

### 3.1 Preliminary of the simulation

In this section, we aim to compare the finite sample performances of the proposed estimators, taking $\widehat{CQTE}_O$ as the benchmark to examine the aforementioned theoretical results. For ease of exposition, we only consider the case of $X_1 \in R$, that is, $l = 1$.

Consider the following heteroscedasticity models respectively:

**Model 1:** $Y(0) = 0$, and $Y(1) = X_1 + \ldots + X_k + |X_1|\epsilon_1, p_1(X) = \frac{\exp(a_1^T X)}{1 + \exp(a_1^T X)}$.

**Model 2:** $Y(0) = 0$, and $Y(1) = X_1 + X_2 + |X_1|\epsilon_1, p_2(X) = \frac{\exp(a_2^T X)}{1 + \exp(a_2^T X)}$ with $k = 2$.

**Model 3:** $Y(0) = 0$, and $Y(1) = X_1 + X_2 + X_3 + X_4 + |X_1|\epsilon_1, p_3(X) = \frac{\exp(X_1^2 + a_3^T X)}{1 + \exp(X_1^2 + a_3^T X)}$ with $k = 4$.

Here $a_1 = (1, \ldots, 1)^T / \sqrt{k}$ with $k = \dim(X) \in \{2, 3, 4, 5\}$, $a_2 = (0, 1)^T$, $a_3 = (0, 1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})^T$. As for the generation of $\{X_1, \ldots, X_5\}$ and $\epsilon_1$, let $X_1 \sim U(-0.5, 0.5), X_2 = (1 + X_3^2 + \epsilon_2, X_3 = X_1(1 + X_1) + \epsilon_3, X_4 = \exp(-1 - X_1) + \epsilon_3, X_5 = (1 + X_1)^3 + \epsilon_4$, with $\epsilon_j \sim N(0, 0.25^2), j = 1, \ldots, 4$, and $\epsilon_1 \sim N(0, 1)$.

Obviously, under Model 1, when $p(X) = p_1(X), X_1 \subsetneq \bar{X} = X$ but $X_1 \varsubsetneq a_1^T X$, which is designed to examine whether $\widehat{CQTE}_N$ is the most efficient estimator while $\widehat{CQTE}_S$ is asymptotically similar to $\widehat{CQTE}_P$ and $\widehat{CQTE}_O$. As for $p_2(X)$ in Model 2, since $X_1 \varsubsetneq a_1^T X$, it can
be expected that all CQTE are asymptotically similar. \(p_j(X)\) in Model 3 is set to verify that \(\hat{CQTE}_S\) can be more efficient than \(\hat{CQTE}_P\). In this propensity score model, \(D \perp X|\alpha^TX\) with \(\alpha^T = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1/\sqrt{3} & 1/\sqrt{3} & 1/\sqrt{3} \end{pmatrix}\). Here, we use the aforementioned sufficient dimension reduction method, MAVE, to estimate the index \(\alpha^TX\).

In the procedures described here, we give the estimated values of \(\Delta_\tau(x_{10})\) at \(x_{10} \in \{-0.2, 0, 0.2\}\) and \(\tau \in \{0.5, 0.75, 0.99\}\). The sample size \(n = \{500, 1000\}\) in some cases and \(n = \{100, 300, 500, 700, 1000\}\) in other cases. The experiments are repeated 1500 times and the size of bootstrap samples is 200. We choose a Gaussian kernel and then high-order kernels are derived from it throughout this section. Next, the values of the order and smoothness \(s_1, s_2, s\) of the kernels and the bandwidths \(h, h_0, h_2\) are chosen via the rules in (10) and (12). To examine the performances fairly, the parameters \(s, h\) for \(K(u)\) are the same for all four CQTE estimators. Since the conditions for \(\hat{CQTE}_N\) are more restrictive than those for the other estimators, we first select all parameters for \(\hat{CQTE}_N\), that is, the tuning parameters in (10) and (12), where we set \(a = b\). That means we just need to confirm the values of turning parameters \([a, a_1, b_1]\) about the bandwidths. By the rule of thumb, we try two groups of values of \(a, a_1, b_1\) to see which ones could make stable performances of estimators: Group 1 : \([a = 0.5, a_1 = 1.1, b_1 = 1.2]\), Group 2 : \([a = 0.5, a_1 = 0.9, b_1 = 1.1]\).

Furthermore, we should point out that the estimated propensity score is trimmed to lie in the interval \([0.005, 0.995]\). Although we introduce a plug-in method to estimate the asymptotic variances, we should note that its estimation procedure is very complex, since it involves the computation of many unknown functions, such as \(f(Y(j)|X_1)\) and \(m_{j,\tau}(X)\), \(j = 0, 1\). Hence, similarly as Chen et al. (2015), we turn to estimate the asymptotic variance based on the bootstrap method. Therefore, we use the following indices to evaluate the performances of the involved estimators: bias, SD, the bootstrap-based estimated standard deviation (ESD), mean squared error (MSE) of \(\hat{\Delta}_\tau(x_1)\), and the 95% confidence interval coverage probability based on ESD, denoting as CP.

### 3.2 Simulation results

For space saving and better illustration, we in this section only report the simulation results under some specific setting, while the rest are relegated to the supplementary material. By analyzing the simulation results reported in Figures 2–4 and Tables 1 and 2, we summarize the observations as follows.

*The effect of sample size.* To examine the effect of sample size, we focus on Model 1 with different values of \(n\). The dimension of \(X\) is chosen to be both \(k = 2\) and \(k = 4\). Figure 2 presents the values of three evaluation indices: bias, SD, and MSE against \(n\) when \(x_{10} = 0\) and \(\tau = 0.5\). From this figure, we can see that larger sample size leads to smaller MSE, so do bias and SD. Meanwhile, we can see that the improvement under \(k = 4\) is more significant than that under \(k = 2\). These can also be observed when \(x_{10} \in \{-0.2, 0.2\}\) and \(\tau \in \{0.75, 0.99\}\). Meanwhile, from the tables in the supplementary material, ESD is closer to SD in most cases, indicating the bootstrap approximation method performs well. In addition, Table 1 presents the values of empirical coverage probability across different sample sizes when \(\tau = 0.5\). The empirical values of CP are closer to the nominal level 0.95 even though the sample size \(n = 100\). This implies that the normal approximation works as well. When the quantile level is \(\tau = 0.99\), and the sample size is small, the performance of the approximation is slightly weakened by this extremal quantile level, while the negative impact would be alleviated when the sample size is large.
The simulation results of bias, SD, MSE against sample size with $k \in \{2, 4\}, \tau = 0.5, x_{10} = 0$ and the tuning parameters in Group 1. MSE, mean squared error

**FIGURE 2**  The simulation results of bias, SD, MSE against sample size with $k \in \{2, 4\}, \tau = 0.5, x_{10} = 0$ and the tuning parameters in Group 1. MSE, mean squared error

**The effect of dimensionality.** To better illustrate the influence caused by dimension on the evaluation indices, we summarize the simulation results under Model 1 in Figure 3 when $\tau = 0.5$ and $x_{10} = 0$. It is expectable that when the dimension $k$ increases, the values of bias, SD, and MSE obviously increase. But we also found that, even when $k = 5$, the values of these evaluation indices are still small. Furthermore, comparing the case under $k = 2$ setting with that under $k = 4$ or $k = 5$ in terms of SD, we can see that when $k = 2$, $\hat{CQTE}_N$ is almost uniformly more efficient than the other $CQTE$ estimators even when $n = 500$. This is consistent with the asymptotic results in Theorem 2. However, when the dimension increases to $k = 4$ or $k = 5$, $\hat{CQTE}_N$ loses its efficiency superiority in most cases even when $n = 1000$. This would be due to the estimation inaccuracy when the dimension is high. When $\tau = 0.75$, we can get a similar conclusion. When it comes to the extremal quantile level, that is, $\tau = 0.99$, we can see that $\hat{CQTE}_N$ loses its efficiency superiority, which is mainly because of the sparsity of the data in extremal quantile level. We also present, when $\tau = 0.5$, the empirical coverage probability (CP) for different dimensions in Table 2. The empirical coverage probabilities are close to the nominal level 0.95 in most cases, which also indicates the usefulness of normal approximation.
The simulation results of bias, SD, MSE against the dimension of $X$ with $n \in \{500, 1000\}$, $\tau = 0.5$, $x_{10} = 0$ and the tuning parameters in Group 1. MSE, mean squared error

**FIGURE 3** The simulation results of bias, SD, MSE against the dimension of $X$ with $n \in \{500, 1000\}$, $\tau = 0.5$, $x_{10} = 0$ and the tuning parameters in Group 1. MSE, mean squared error

*The effect of the estimation method.* In terms of all the evaluation indices, $\hat{CQTE}_P$ has, in most cases, similar performance to $\hat{CQTE}_O$. This coincides with the asymptotic properties presented in Theorem 1. Furthermore, as discussed before, the performances of $\hat{CQTE}_N$ and $\hat{CQTE}_S$ are related to whether the given $X_1$ is included in the set of arguments of the propensity score. In Figure 4, we summarize the asymptotic relative efficiency of the estimators against $\hat{CQTE}_O$ based on SD under Group 1 values of $\{a, a_1, b_1\}$ and $\tau = 0.5$. Specifically, when $p(X) = p_1(X)$, that is, under Model 1, Figure 4 shows that $\hat{CQTE}_N$ is uniformly the most efficient one. While when $p(X) = p_2(X)$, that is, under Model 2, as $X_1$ is not fully included in the set of the arguments of $p_2(X)$, Figure 4 shows that the estimation efficiency of $\hat{CQTE}_N$ loses. When $p(x) = p_3(x)$, Figure 4 shows that both $\hat{CQTE}_N$ and $\hat{CQTE}_S$ are generally more efficient than $\hat{CQTE}_O$ and $\hat{CQTE}_P$. Furthermore, when $k$ is larger, $\hat{CQTE}_S$ can be, in some cases, more efficient than $\hat{CQTE}_N$. That is, $\hat{CQTE}_N$ is no longer superior to $\hat{CQTE}_S$. This is mainly because of the dimension reduction structure in $\hat{CQTE}_S$ and thus, less estimation inaccuracy. Thus, the performance of $\hat{CQTE}_S$ could be more robust than $\hat{CQTE}_N$ against dimensionality. When $\tau = 0.75$, we can get a similar conclusion. Similarly as the previous discussion, in case with the extremal quantile level $\tau = 0.99$, ...
we cannot have some definitive conclusions to say which one should be better in finite sample scenarios.

In summary, we highlight that, if $X_1$ is included in the set of arguments of $p(X)$, both $\hat{CQE}_N$ and $\hat{CQE}_S$ can be useful as they can be robust against misspecification. Furthermore, owing to the dimension reduction structure, $\hat{CQE}_S$ is worth of recommendation in high-dimensional scenarios. When $X_1$ is not included in the set of arguments of $p(X)$, all the estimators perform similarly in most cases.

4 | A REAL DATA EXAMPLE

In this section, we estimate the $CQTE$ function to investigate the quantile effect of maternal smoking on birth weight over the mother's age. We adopt a data set based on the records between 1988 and 2002 by the North Carolina State Center Health Services, which can be obtained from Robert Lieli’s website http://www.personal.ceu.hu/staff/Robert_Lieli/cate-birthdata.zip. Noted
TABLE 1  The simulation results of empirical coverage probability (CP) against sample size with $k \in \{2, 4\}$, $\tau = 0.5$ and the tuning parameters in Group 1

| $k$ | $n$   | $CQTE_O$ | $CQTE_P$ | $CQTE_N$ | $CQTE_S$ |
|-----|-------|----------|----------|----------|----------|
|     |       | -0.2 | 0 | 0.2 | -0.2 | 0 | 0.2 | -0.2 | 0 | 0.2 | -0.2 | 0 | 0.2 |
| $k = 2$ | $n = 100$ | 0.9560 | 0.9467 | 0.9467 | 0.9513 | 0.9587 | 0.9533 | 0.9500 | 0.9500 | 0.9500 | 0.9507 | 0.9540 | 0.9493 |
|     | $n = 300$ | 0.9533 | 0.9547 | 0.9387 | 0.9580 | 0.9453 | 0.9500 | 0.9573 | 0.9447 | 0.9473 | 0.9580 | 0.9587 | 0.9473 |
|     | $n = 500$ | 0.9373 | 0.9420 | 0.9447 | 0.9440 | 0.9440 | 0.9427 | 0.9473 | 0.9480 | 0.9413 | 0.9487 | 0.9487 | 0.9453 |
|     | $n = 700$ | 0.9567 | 0.9600 | 0.9487 | 0.9533 | 0.9560 | 0.9513 | 0.9580 | 0.9600 | 0.9480 | 0.9620 | 0.9687 | 0.9553 |
|     | $n = 1000$ | 0.9520 | 0.9480 | 0.9487 | 0.9513 | 0.9540 | 0.9487 | 0.9480 | 0.9547 | 0.9520 | 0.9673 | 0.9680 | 0.9560 |
| $k = 4$ | $n = 100$ | 0.9553 | 0.9580 | 0.9547 | 0.9567 | 0.9653 | 0.9547 | 0.9547 | 0.9580 | 0.9560 | 0.9633 |
|     | $n = 300$ | 0.9500 | 0.9500 | 0.9487 | 0.9567 | 0.9387 | 0.9420 | 0.9533 | 0.9413 | 0.9387 | 0.9593 | 0.9533 | 0.9453 |
|     | $n = 500$ | 0.9553 | 0.9480 | 0.9507 | 0.9587 | 0.9593 | 0.9500 | 0.9560 | 0.9480 | 0.9520 | 0.9673 | 0.9680 | 0.9560 |
|     | $n = 700$ | 0.9540 | 0.9613 | 0.9487 | 0.9427 | 0.9447 | 0.9553 | 0.9460 | 0.9560 | 0.9507 | 0.9627 | 0.9700 | 0.9573 |
|     | $n = 1000$ | 0.9553 | 0.9540 | 0.9513 | 0.9513 | 0.9553 | 0.9527 | 0.9500 | 0.9527 | 0.9487 | 0.9660 | 0.9687 | 0.9627 |
Table 2: The simulation results of empirical coverage probability (CP) against the dimension of $X$ with $n \in \{500, 1000\}$, $\tau = 0.5$ and the tuning parameters in Group 1.

| $n$   | $k$   | $CQTE_O$ | $CQTE_P$ | $CQTE_N$ | $CQTE_S$ |
|-------|-------|----------|----------|----------|----------|
|       |       | $-0.2$   | $0$      | $0.2$    | $-0.2$   | $0$      | $0.2$    | $-0.2$   | $0$      | $0.2$    | $-0.2$   | $0$      | $0.2$    |
| $n = 500$ | $k = 2$ | 0.9373   | 0.9420   | 0.9447   | 0.9440   | 0.9440   | 0.9427   | 0.9473   | 0.9480   | 0.9413   | 0.9487   | 0.9487   | 0.9453   |
|       | $k = 3$ | 0.9400   | 0.9467   | 0.9380   | 0.9513   | 0.9480   | 0.9453   | 0.9367   | 0.9467   | 0.9493   | 0.9533   | 0.9547   | 0.9553   |
|       | $k = 4$ | 0.9553   | 0.9480   | 0.9507   | 0.9587   | 0.9593   | 0.9500   | 0.9560   | 0.9480   | 0.9500   | 0.9507   | 0.9547   | 0.9513   |
|       | $k = 5$ | 0.9513   | 0.9387   | 0.9467   | 0.9467   | 0.9393   | 0.9447   | 0.9473   | 0.9420   | 0.9500   | 0.9513   | 0.9540   | 0.9527   | 0.9513   |
| $n = 1000$ | $k = 2$ | 0.9520   | 0.9480   | 0.9487   | 0.9513   | 0.9540   | 0.9487   | 0.9480   | 0.9533   | 0.9460   | 0.9533   | 0.9647   | 0.9513   |
|       | $k = 3$ | 0.9540   | 0.9620   | 0.9687   | 0.9493   | 0.9560   | 0.9593   | 0.9460   | 0.9527   | 0.9613   | 0.9633   | 0.9640   | 0.9593   |
|       | $k = 4$ | 0.9553   | 0.9540   | 0.9513   | 0.9513   | 0.9553   | 0.9527   | 0.9500   | 0.9527   | 0.9487   | 0.9660   | 0.9687   | 0.9627   |
|       | $k = 5$ | 0.9413   | 0.9427   | 0.9540   | 0.9407   | 0.9473   | 0.9493   | 0.9360   | 0.9480   | 0.9487   | 0.9513   | 0.9593   | 0.9580   |
that this data set was also analyzed by Abrevaya et al. (2015), who aimed to estimate the conditional average treatment effect (CATE) of maternal smoking on birth weight by selecting $X_1$ as mother’s age. Focusing on first-time mothers, they observed that the CATE function is mostly negative, and Abrevaya et al. (2015) also noted that CATE is stronger (more negative) for older mothers.

We also choose the mother’s age as $X_1$, and aim to explore more information about the conditional smoking effect besides the average treatment effects provided. Both the low-birthweight (LBW) (weighing less than 2500 g) and high-birthweight (HBW) (weighing more than 4000 g) babies should receive attention in the literature. For example, we may also want to know, when mother is older, whether the smoking effect will be stronger for her LBW babies, or whether there exist different trends of the smoking effect over mother’s age for the LBW and HBW babies. Thus, we estimate the conditional quantile treatment effect (CQTE), under $\tau = 0.1, 0.5, 0.9$, respectively, to investigate how the quantile treatment effect varies with different values of mother’s age and different groups of babies.

Before estimation, we first introduce some details and settings about the data set. We restrict our sample to White and first-time mothers, thus the sample size is $n = 433,558$ while the smoking sample size is 74,386. The outcome $Y$ here is birth weight measured in grams and the treatment indicator variable $D$ is a binary variable. When $D = 1$, it means mother smokes and $D = 0$ otherwise. Furthermore, to ensure the unconfoundedness assumption, we choose a large set of variables as $X$, including mother’s age, education level, the month of the first prenatal visit (=10 if the prenatal care is foregone), the number of prenatal visits, and indicators for baby’s gender, mother’s marital status, whether or not father’s age is missing, gestational diabetes, hypertension, amniocentesis, ultrasound exams, the previous (terminated) pregnancies, and alcohol use.

We estimate the CQTE function $\Delta_\tau(X_1)$ in the interval between ages 15 and 35 years under three different quantiles, that is, $\Delta_{0.1}(X_1), \Delta_{0.5}(X_1)$, and $\Delta_{0.9}(X_1)$, respectively. Since the dimension of $X$ is high, we use a semiparametric model for the propensity score that has a single index structure such that the dimensionality and model misspecification problem can be alleviated. Thus, we first use the sufficient dimension reduction method, SIR, to estimate the index. However, in order to capture the nonlinear information of $p(x)$, the covariates $X^*$ used in estimation consist of all the elements of $X$, the square of mother’s age, and the interaction terms between mother’s age and all other elements of $X$. When it comes to the selection of bandwidth, we set $h_2 = \hat{\sigma}_d n^{-1/3}$ and $h = \hat{\sigma}_1 n^{-1/5}$, where $\hat{\sigma}_d = \sqrt{\text{var}(\hat{\alpha}^\top X^*)}$, $\hat{\alpha}$ is the estimated linear index direction and $\hat{\sigma}_1 = 2\sqrt{\text{var}(\overline{X_1})}$. As for kernel function, we use the Gaussian kernel.

Figure 5 displays the results of the estimated CQTE($X_1$) as a function of mother’s age in the range of 15–35 years. There are several points we want to highlight: (1) The CQTE($X_1$) for the effect of maternal smoking on birth weight is remarkably negative. All three CQTE($X_1$) curves range from about −140 to −300 g, which means if a mother smokes during her pregnancy period, the birth weight of her baby will most likely decrease. This finding is in accordance with the conclusion of Abrevaya et al. (2015). (2) The LBW babies suffer the most from maternal smoking and get thinner across mother’s age. When we focus on the $\Delta_{0.1}(X_1)$ curve, it is at the bottom of all the three curves. Furthermore, we can observe it has a decreasing trend over mother’s age. Thus this suggests that older mothers would more urgently quit smoking to avoid the ultra-low-weight baby occurring. (3) The trend of $\Delta_{0.5}(X_1)$ varies with mother’s age. As for $\Delta_{0.5}(X_1)$, we can also find a decreasing trend from 16 to around 22 years of age, while the curve is rather stable between the age of 23 and
FIGURE 5 Three conditional quantile treatment effects (CQTE) curves over mother's age: $\hat{\Delta}_{0.9}(X_1)$ (dotted line), $\hat{\Delta}_{0.5}(X_1)$ (solid line), and $\hat{\Delta}_{0.1}(X_1)$ (dashed line)

28 years. As the relationship between median and average, it can be expected that $\Delta_{0.5}(X_1)$ is like $CATE(X_1)$.

5 | CONCLUSION

In this article, we investigate estimating conditional quantile treatment effect (CQTE), aimed to capture the conditional treatment effect in a specific subgroup. Four estimators are constructed when propensity score is under true function, parametric, nonparametric, and semiparametric structure: $CQTE_O$, $CQTE_P$, $CQTE_N$, and $CQTE_S$ where $CQTE_O$ mainly serves as the benchmark for the comparison among the other three estimators. The asymptotic properties of the estimators are systematically studied. The new findings show that the estimations under the conditional framework are rather different from their unconditional counterparts. More importantly, under conditional framework, two factors play important roles in the estimation efficiency of nonparametric and semiparametric-based estimators: 1) the convergence rate of the estimated propensity score and 2) the affiliation of the given covariates to the set of arguments of propensity score. These are not the cases for the unconditional counterparts in studying treatment effect. One more issue is about semiparametric efficiency the unconditional counterparts can achieve when nonparametric estimation is used for propensity score. Under conditional framework, CQTEs are functions of the given covariates, it is unclear what could be defined as a semiparametric efficiency. It seems to involve uniformly asymptotic efficiency over a function. Thus, we leave it to further study.

Furthermore, in this field, potential outcome regression and doubly robust estimation are also the basic methodologies, and the relevant studies are worthwhile. However, as this article mainly focuses on a systematic investigation on the asymptotic efficiencies of different propensity score-based estimations, the systematic studies about potential outcome regression and doubly robust estimation will be the topics in further study. Furthermore, it would be interesting to develop a confidence band of CQTE over $x_1$ or quantile index, which is also a topic in further study.
ACKNOWLEDGMENTS
The authors are grateful to the editor, the associate editor, and the anonymous referees for the constructive comments and suggestions that led to significant improvement of an early manuscript. Xu Guo was supported by Beijing Natural Science Foundation (1212004) and National Natural Science Foundation of China grants (NSFC 12071038, 11871263, 11771048); Lixing Zhu was supported by a grant from the University Grants Council of Hong Kong (HKBU 12302720).

ORCID
Lixing Zhu https://orcid.org/0000-0003-1611-4267

REFERENCES
Abrevaya, J., Hsu, Y.-C., & Lieli, R. P. (2015). Estimating conditional average treatment effects. *Journal of Business & Economic Statistics, 33*(4), 485–505.
Chen, X., Wan, A. T. K., & Zhou, Y. (2015). Efficient quantile regression analysis with missing observations. *Journal of the American Statistical Association, 110*(510), 723–741.
D’Abrera, H., & Lehmann, E. (1975). *Nonparametrics: Statistical methods based on ranks*. Holden-Day.
Doksum, K. (1974). Empirical probability plots and statistical inference for nonlinear models in the two-sample case. *The Annals of Statistics, 2*(2), 267–277.
Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics, 7*(1), 1–26.
Fan, J., & Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *The Annals of Statistics, 20*, 2008–2036.
Firpo, S. (2007). Efficient semiparametric estimation of quantile treatment effects. *Econometrica, 75*(1), 259–276.
Guo, X., Fang, Y., Zhu, X., Xu, W., & Zhu, L. (2018). Semiparametric double robust and efficient estimation for mean functionals with response missing at random. *Computational Statistics & Data Analysis, 128*, 325–339.
Hang, W., & Xia Y. (2019). MAVE: Methods for dimension reduction. R package version 1.3.10.
Hirano, K., Imbens, G. W., & Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica, 71*(4), 1161–1189.
Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47*(1), 5–86.
Kennedy, E. H., Ma, Z., McHugh, M. D., & Small, D. S. (2017). Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79*(4), 1229–1245.
Koenker, R. (2005). *Quantile regression*. Cambridge University Press.
Koenker, R. (2017). Quantile regression: 40 years on. *Annual Review of Economics, 9*(1), 155–176.
Li, B. (2018). *Sufficient dimension reduction: Methods and applications with R*. CRC Press.
Li, K. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association, 86*(414), 316–327.
Li, Q., & Racine, J. S. (2007). *Nonparametric econometrics: Theory and practice*. Princeton University Press.
Liu, J., Ma, Y., & Wang, L. (2018). An alternative robust estimator of average treatment effect in causal inference. *Biometrics, 74*(3), 910–923.
Luo, W., Wu, W., & Zhu, Y. (2019). Learning heterogeneity in causal inference using sufficient dimension reduction. *Journal of Causal Inference, 7*(1), 20180015.
Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41–55.
Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association, 113*(523), 1228–1242.
Xia, Y., Tong, H., Li, W., & Zhu, L. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*(3), 363–410.
Yao, L., Sun, Z., & Wang, Q. (2010). Estimation of average treatment effects based on parametric propensity score model. *Journal of Statistical Planning and Inference, 140*(3), 806–816.
Zhang, Y. (2018). Extremal quantile treatment effects. *The Annals of Statistics, 46*(6), 3707–3740.
Zhu, L., Miao, B., & Peng, H. (2006). On sliced inverse regression with high-dimensional covariates. *Journal of the American Statistical Association, 101*(474), 630–643.

**SUPPORTING INFORMATION**
Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Zhou, N., Guo, X., & Zhu, L. (2022). The role of propensity score structure in asymptotic efficiency of estimated conditional quantile treatment effect. *Scandinavian Journal of Statistics, 49*(2), 718–743. https://doi.org/10.1111/sjos.12539