Ensemble and calibration multiply robust estimation for quantile treatment effect

Xiaohong He and Lei Wang

School of Statistics and Data Science & LPMC, Nankai University, Tianjin, People’s Republic of China

ABSTRACT
Quantile treatment effects can be important causal estimands in the evaluation of biomedical treatments or interventions for health outcomes such as birthweight and medical cost. However, the existing estimators require either a propensity score model or a conditional density vector model is correctly specified, which is difficult to verify in practice. In this paper, we allow multiple models for propensity score and conditional density vector, then construct a class of calibration estimators based on multiple imputation and inverse probability weighting approaches via empirical likelihood. The resulting estimators multiply robust in the sense that they are consistent if any one of these models is correctly specified. Moreover, we propose another class of ensemble estimators to reduce computational burden while ensuring multiple robustness. Simulations are performed to evaluate the finite sample performance of the proposed estimators. Two applications to the birthweight of infants born in the United States and AIDS Clinical Trials Group Protocol 175 data are also presented.

1. Introduction
Causal evaluation of treatment or intervention is commonly done by estimating the average treatment effect (ATE). However, in some studies about the low weight in children growth, medical cost and so on, outcomes tend to be highly skewed such that the ATE may not be a proper representative parameter. Our study is motivated by the birthweight data based on June 1997 Detailed Natality Data published by the National Center for Health Statistics. Low birthweight can cause a series of subsequent health problems, long duration of healthcare and other economic problems according to previous studies. We want to investigate the impact of maternal behavior on the birthweight of babies born in the United States under the conditions of various demographic characteristics. In specific, similar to Xie et al. [31], we consider the association between the birthweight and smoking status of the mother. It can be checked that the distribution of birthweights is skewed such that the ATE may not be appropriate. In this case, it is often important to learn about distributional impacts beyond the ATE, such as the effects on upper (or lower) quantiles of an
outcome, while quantile treatment effect (QTE) may be more relevant and informative [2,4,10,11,14–16,26]. For the birthweight data, the QTE is defined as the quantile difference of birthweight distributions between the treatment (mother smokes) and control (mother does not smoke) groups. Another example is based on the AIDS Clinical Trials Group Protocol 175 (ACTG 175 [17]) data. Our purpose is to examine whether the new regimens work or not. Since the outcome distributions are skewed, it is more reasonable to investigate the QTEs between the three new regimens and control arm, respectively.

To define the QTE, we begin with some notations. Let $T$ be a binary treatment indicator, $X$ be a $p$-dimensional pretreatment covariate, $Y_1$ and $Y_0$ be the potential outcomes under treatment $T = 1$ and $T = 0$, respectively. For each unit, either $Y_1$ or $Y_0$ is observed, but not both, i.e. we only observe $Y = TY_1 + (1 - T)Y_0$. Given $\tau \in (0, 1)$, the $\tau$th QTE is defined as

$$ q_\tau = q^1_\tau - q^0_\tau, $$

where $q^t_\tau = \inf\{q : \Pr(Y_t \leq q) \geq \tau \}$ is the $\tau$th quantile of $Y_t$, $t = 0, 1$. The problem of interest is to estimate $q_\tau$ based on $n$ independent and identically distributed samples $\{(X_i, Y_i, T_i), i = 1, \ldots, n\}$ from $(X, Y, T)$. For simplicity, we omit $\tau$ from the rest of this paper, but we should remember that they are $\tau$-specific. In order to identify the QTE, we assume that the treatment assignment is independent of the potential outcomes when conditioning on observed covariates, i.e. $T \perp (Y_1, Y_0) \mid X$, which is usually called ignorability of treatment or unconfoundedness assumption. Under this assumption, the propensity score (PS) is defined as $\pi(X) = \Pr(T = 1 \mid X)$, i.e. the probability of treatment assignment conditional on covariates.

There exist several methods to estimate the QTE, such as the inverse propensity weighting (IPW), instrumental variable approach, matching, imputation, calibration and so on [1,14,18,23,24,28]. One of the most commonly used is the IPW estimator $\hat{q}_{ipw} = \hat{q}^1_{ipw} - \hat{q}^0_{ipw}$, where $\hat{q}^1_{ipw}$ and $\hat{q}^0_{ipw}$ are the solutions to

$$ \frac{1}{n} \sum_{i=1}^{n} T_i \psi(Y_i - \hat{q}^1) = \frac{1}{1 - \pi(X_i, \hat{\beta})}, \quad \frac{1}{n} \sum_{i=1}^{n} (1 - T_i) \psi(Y_i - \hat{q}^0) = \frac{1}{1 - \pi(X_i, \hat{\beta})}, $$

(1)

$\psi(u) = \tau - I(u \leq 0)$ is the check function, $\pi(X, \beta)$ is a parametric model for $\pi(X)$ with unknown parameter $\beta$ and $\hat{\beta}$ is a consistent estimator of $\beta$ [8]. On the other hand, imputation [29] is also a widely adopted approach. For example, Chen and Yu [9] and many others considered the imputation method for quantile regression estimation. Write $f(X) = \{f^1(X), f^0(X)\}$ with $f^t(X) = f(Y_t \mid X), t = 0, 1$. Gaussian or more complex parametric conditional density vector $f(X, \gamma) = \{f^1(X, \gamma^1), f^0(X, \gamma^0)\}$ with unknown parameters $\gamma = \{\gamma^1, \gamma^0\}$ can be assumed for predicting unobservable potential outcomes. We define the multiple imputation (MI) estimator as $\hat{q}_{mi} = \hat{q}^1_{mi} - \hat{q}^0_{mi}$, where $\hat{q}^1_{mi}$ and $\hat{q}^0_{mi}$ are the solutions to

$$ \frac{1}{n} \sum_{i=1}^{n} T_i \psi(Y_i - \hat{q}^1_{mi}) + (1 - T_i) \frac{1}{L} \sum_{l=1}^{L} \psi(\tilde{Y}^l_{i}(\gamma^1_{mi}) - \hat{q}^1_{mi}) = 0, $$

(2)

$$ \frac{1}{n} \sum_{i=1}^{n} (1 - T_i) \psi(Y_i - \hat{q}^0_{mi}) + T_i \frac{1}{L} \sum_{l=1}^{L} \psi(\tilde{Y}^l_{i}(\gamma^0_{mi}) - \hat{q}^0_{mi}) = 0, $$

(3)
respectively, where $\hat{\gamma}^t$ is a consistent estimator of $\gamma^t$, $\{\tilde{Y}_l^t(\hat{\gamma}^t)\}_{l=1}^L$ are independently imputed values given $X_i$ from the estimated conditional distribution $f^t(X_i, \hat{\gamma}^t)$ and $L$ is the number of multiple imputations. To improve the estimation efficiency of the IPW and MI estimators, augmented inverse probability weighting (AIPW [28]) estimator is defined as $\hat{q}_{aipw} = \hat{q}_{aipw}^1 - \hat{q}_{aipw}^0$, where $\hat{q}_{aipw}^1$ and $\hat{q}_{aipw}^0$ are the solutions to

$$
\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{T_i \psi_\tau(Y_i - q^1)}{\pi(X_i, \hat{\beta})} - \frac{T_i}{\pi(X_i, \hat{\beta})} \frac{1}{L} \sum_{l=1}^{L} \psi_\tau(\tilde{Y}_l^t(\hat{\gamma}^1) - q^1) \right] = 0, \tag{4}
$$

$$
\frac{1}{n} \sum_{i=1}^{n} \left[ \frac{(1 - T_i) \psi_\tau(Y_i - q^0)}{1 - \pi(X_i, \hat{\beta})} - \frac{\pi(X_i, \hat{\beta})}{1 - \pi(X_i, \hat{\beta})} \frac{T_i}{L} \sum_{l=1}^{L} \psi_\tau(\tilde{Y}_l^t(\hat{\gamma}^0) - q^0) \right] = 0. \tag{5}
$$

It can be verified that the IPW estimator $\hat{q}_{ipw}$ is consistent only if $\pi(X)$ is correctly modeled and the MI estimator $\hat{q}_{mi}$ is consistent only if $f(X)$ is correctly modeled. However, if the candidate model is wrongly selected, the resulting IPW or MI estimator will be biased. While the consistency of the AIPW estimator is provided as long as either $\pi(X)$ or $f(X)$ is correctly modeled, which is also called the double robustness (DR) property [3]. In conclusion, the existing methods require the specification of either propensity score $\pi(X)$ or conditional density vector $f(X).

### 1.1. Related literature

In practice, however, the data structure may be complicated, which makes it difficult to impose a correct candidate model. Beyond the DR, Han and Wang [22] proposed to maximize an empirical likelihood (EL) function subject to multiple moment calibration equations and introduced the concept of multiple robustness (MR) in missing data problem, which can be viewed as an extension of the DR (see also [5,6,19,20] and so on) Alternatively, Xie et al. [31] proposed multiply robust estimation of the QTE based on the multiple propensity score models and multiple conditional distribution functions (CDFs). However, Xie et al. [31] placed some restrictions on the CDFs and estimated the unknown parameters in the CDFs based on quantile-based equations. Alternatively, different from specifying the conditional distribution functions by Xie et al. [31], Tang and Qin [30] and Han et al. [21] found that it is more easy and convenient to specify the conditional density models, and the unknown parameters can be easily estimated through a complete-case maximum likelihood estimator (MLE). Moreover, when the number of models increases, the computational cost in the EL optimization procedure will increase, the estimation efficiency and stability will decrease. If the number of models is large, the EL may not be properly defined because of the so-called empty set problem.

To address these problems, in this paper, we propose two classes of efficient MR estimators based on the multiple imputation and inverse probability weighting approaches. In specific,

1. we first propose a class of calibration estimators based on multiple propensity score models and multiple conditional density vector models. Motivated by Han et al. [21], the weights are obtained through the empirical likelihood (EL) approach by matching
population moments of model-based fitted values from the treated and untreated subsamples to the full sample. Under the ignorability of treatment assumption, we prove that the proposed estimators have multiple protection on estimation consistency when any one of the candidate models is correctly specified. Furthermore, the asymptotic normality is established under some mild conditions.

(2) Recognizing the computational burden induced by a large number of models in the EL procedure, we extend the ensemble approach in [12] to the QTE estimation and propose another class of MR estimators by refitting working models using the least-square (LS) approach based on the complete subsamples in treated and untreated groups. Through such refitting, each set of fitted working models is reduced to a scalar value.

(3) Simulation results show that the proposed two classes of estimators provide extra protection against model misspecification, while estimation efficiency is not compromised.

This article is organized as follows. The proposed calibration MR estimators and asymptotic theories are studied in Section 2. The ensemble MR estimators are discussed in Section 3. Sections 4 and 5 present the numerical studies and two real data applications, respectively. Some relevant discussions are given in Section 6. The Supplementary Material contains all proofs and additional simulation results.

2. Multiply robust QTE estimation

In general, true functional forms of propensity score \( \pi(X) \) and conditional densities of \( Y_1, Y_0 \) given \( X \) are all unknown. Let

\[
\mathcal{W} = \{ \pi_m(X, \beta_m), m = 1, 2, \ldots, M \},
\]

\[
\mathcal{F} = \{ f_k(X, \gamma_k) = \{ f_k^1(X, \gamma_k^1), f_k^0(X, \gamma_k^0) \}, k = 1, 2, \ldots, K \},
\]

be two collections of candidate models for \( \pi(X) \) and \( f(X) \), respectively, with unknown parameters \( \beta_m \) and \( \gamma_k = \{ \gamma_k^1, \gamma_k^0 \} \). Since covariate \( X_i \) is always observed, a consistent MLE of \( \beta_m \), denoted as \( \hat{\beta}_m \), can be obtained by maximizing the binomial likelihood function:

\[
\hat{\beta}_m = \arg\max_{\beta} \prod_{i=1}^{n} (\pi_m(X_i, \beta_m))^T_i (1 - \pi_m(X_i, \beta_m))^{1-T_i}, \quad m = 1, 2, \ldots, M. \quad (6)
\]

If \( \hat{\beta}_m \to \beta_{ms} \) as \( n \to \infty \), then \( \pi_m(X, \beta_{ms}) = \pi(X) \) only if \( \pi_m(X, \beta_m) \) is just the correct model for \( \pi(X) \). Write \( S_t = \{ i : T_i = t, i = 1, \ldots, n \} \) as the set of observations in the control \((t = 0)\) and treatment \((t = 1)\) groups, respectively, with sample sizes \( n_0 = \sum_{i=1}^{n} (1 - T_i) \) and \( n_1 = \sum_{i=1}^{n} T_i \). Similarly, the MLE of \( \gamma_k^t \), denoted as \( \hat{\gamma}_k^t \), can be obtained, i.e.

\[
\hat{\gamma}_k^t = \arg\max_{\gamma} \prod_{i \in S_t} f_k^t(X_i, \gamma), \quad t = 0, 1. \quad (7)
\]

by fitting the conditional density model based on subsamples \( \{ X_i, Y_i, T_i = t, i = 1, \ldots, n \} \) for \( t = 0, 1 \). If \( \hat{\gamma}_k^t \to \gamma_{kn}^t \) as \( n \to \infty \), then \( f_k^t(X, \gamma_{kn}^t) = f^t(X) \) only if \( f_k^t(X, \gamma_k^t) \) is just the correct model for \( f^t(X) \). Once \( \hat{\gamma}_k^t \) is obtained, as in (2) – (3), we can obtain the MI estimators...
of $q^1$ and $q^0$ based on the $k$th working model $f_k(X, \gamma_k)$, denoted as $\hat{q}_{1k}$ and $\hat{q}_{0k}$, respectively. The MI estimators are consistent only if $f_k(X, \gamma_k)$ is correctly modeled.

**Remark 2.1:** When the dimension of $X$ is not low, the ordinary MLE may not work well. To solve this problem, we assume that only a few of these available covariates are actually related to the parametric propensity and conditional density models, respectively, i.e. the parameters are sparse, and then some regularization penalties, such as Lasso, SCAD and MCP [13], can be applied in the maximum likelihood estimation to efficiently select significant variables and estimate parameters simultaneously.

**Remark 2.2:** If the outcome variable $Y$ is not normally distributed, we can try to fit the data by the commonly used Gamma, Beta, Chi-square distributions or exponential family of distributions. In addition, Box–Cox transformations can be applied to the outcome such that the transformed outcome is normally distributed.

### 2.1. Calibration MR estimator based on empirical likelihood

The proposed calibration MR estimators for the QTE can be expressed as the following steps.

(S1) Calculate the MLEs $\hat{\beta}_m$ and $\hat{\gamma}_t^k$ for $m = 1, \ldots, M$, $k = 1, 2, \ldots, K$, $t = 0, 1$.

(S2) Independently draw $L$ samples $\{\tilde{Y}_l^i(\hat{\gamma}_t^k)\}_{l=1}^L$ from the estimated conditional density distribution $f_k^t(X_i, \hat{\gamma}_t^k)$ and $L$ is the number of multiple imputations. Calculate $\hat{q}_{tk}^l$, $t = 0, 1$ and $k = 1, 2, \ldots, K$, as an imputation estimator of $q^t$ by solving

$$\frac{1}{n} \sum_{i=1}^{n} \left[ T_i \psi_\tau (Y_i - q^1) + (1 - T_i) \frac{1}{L} \sum_{l=1}^{L} \psi_\tau (\tilde{Y}_l^i(\hat{\gamma}_t^k) - q^1) \right] = 0,$$

$$\frac{1}{n} \sum_{i=1}^{n} \left[ (1 - T_i) \psi_\tau (Y_i - q^0) + T_i \frac{1}{L} \sum_{l=1}^{L} \psi_\tau (\tilde{Y}_l^i(\hat{\gamma}_t^k) - q^0) \right] = 0,$$

respectively.

(S3) Obtain the weights $w_i$'s, $i \in S_t$, for each group by maximizing $\prod_{i \in S_t} w_i$ under the following calibration constraints,

$$\sum_{i \in S_t} w_i = 1, \quad \sum_{i \in S_t} w_i \pi_m(X_i, \hat{\beta}_m) = \hat{\theta}_m, \quad \sum_{i \in S_t} w_i \left[ \frac{1}{L} \sum_{l=1}^{L} \psi_\tau (\tilde{Y}_l^i(\hat{\gamma}_t^k) - \hat{q}_{tk}^l) \right] = \hat{\eta}_t^k,$$

where $\hat{\theta}_m = \sum_{i=1}^{n} \pi_m(X_i, \hat{\beta}_m) / n$ and

$$\hat{\eta}_t^k = \sum_{i=1}^{n} \sum_{l=1}^{L} \psi_\tau (\tilde{Y}_l^i(\hat{\gamma}_t^k) - \hat{q}_{tk}^l) / (nL).$$

(S4) The proposed estimator is defined as $\hat{q}_{mi} = \hat{q}_{mi}^1 - \hat{q}_{mi}^0$, where $\hat{q}_{mi}^t$ is the solution to

$$\sum_{i \in S_t} \hat{w}_i \psi_\tau (Y_i - q^t) = 0. \quad (8)$$
In (S3), the first constraint is imposed for regularization, the second and third calibration constraints balance the weighted average of each candidate model evaluated on the observed samples with the corresponding unweighted sample average. This coincides with calibration method in [6] based on the fact that $E[(T/\pi(X) - 1)f(x)] = 0$ with $f(X)$ being $\pi_m(X)$ and $\sum_{l=1}^L \psi_l (\hat{Y}_l^t (\hat{q}_l^t) - q_l^t)/L$, respectively. Since $\hat{w}_i$ is a discrete probability measure, the estimation for $\hat{w}_i$ can be naturally considered as a typical form of an EL problem. In specific, let

$$g_i(X_i, \hat{\beta}, \hat{q}_t, \hat{\gamma}_t^t) = \begin{cases} (\pi_1(X_i, \hat{\beta}_1) - \hat{\theta}_1)(-1)^{t+1} \\ \vdots \\ (\pi_M(X_i, \hat{\beta}_M) - \hat{\theta}_M)(-1)^{t+1} \\ \frac{1}{L} \sum_{l=1}^L \psi_l (\hat{Y}_l^t (\hat{q}_l^t) - \hat{q}_l^t) - \hat{\eta}_l^t \\ \vdots \\ \frac{1}{L} \sum_{l=1}^L \psi_l (\hat{Y}_l^t (\hat{q}_K^t) - \hat{q}_K^t) - \hat{\eta}_K^t \end{cases}, \quad \text{for } i \in S_t,$$

where $\hat{\beta} = (\hat{\beta}_1^T, \ldots, \hat{\beta}_M)^T$, $\hat{\gamma}_t^t = (\hat{\gamma}_1^t)^T, \ldots, (\hat{\gamma}_K^t)^T$ and $\hat{q}_t^t = (\hat{q}_1^t, \ldots, \hat{q}_K^t)^T$. For fixed $(\hat{\beta}, \hat{\gamma}_t^t, \hat{q}_t^t)$, using the Lagrange multiplier method in [27], it can be seen that

$$\hat{w}_i = \frac{1}{n_t(1 + \hat{\rho}_i^T g_i(X_i, \hat{\beta}, \hat{q}_t^t, \hat{\gamma}_t^t))}, \quad \text{for } i \in S_t,$$

where $\hat{\rho}_t = (\hat{\rho}_{t1}, \hat{\rho}_{t2}, \ldots, \hat{\rho}_{tJ})^T$ is a $J = M + K$ dimensional vector satisfying the equations

$$\sum_{i \in S_t} \frac{g_i(X_i, \hat{\beta}, \hat{q}_t^t, \hat{\gamma}_t^t)}{1 + \hat{\rho}_i^T g_i(X_i, \hat{\beta}, \hat{q}_t^t, \hat{\gamma}_t^t)} = 0.$$

Because of the nonnegativity of $\hat{w}_i$, we need $1 + \hat{\rho}_i^T g_i(X_i, \hat{\beta}, \hat{q}_t^t, \hat{\gamma}_t^t) > 0$ for all $i$. The modified Newton–Raphson algorithm discussed in [7] can be applied.

**Remark 2.3:** The proposed calibration MR estimators, which include all the information of candidate models, have the same structure as the IPW estimator, with $T_i/\hat{\pi}(X_i)$ replaced by the weights $\hat{w}_{i \in S_t}$ and $(1 - T_i)/(1 - \hat{\pi}(X_i))$ replaced by the weights $\hat{w}_{i \in S_0}$, respectively.

### 2.2. Multiple robustness and asymptotic normality

Under some regularity conditions [21], Theorem 2.1 shows that $\hat{q}_{mr}$ is multiply robust, in the sense of $\hat{q}_{mr}$ is a consistent estimator as long as one working model is correctly specified, either for $\pi(X)$ or $f(X)$. This is a significant improvement over the IPW, MI and AIPW estimators. Its proof is given in the Supplementary Material.

(C1) Parameter spaces $Q_1$ for $q_1$ and $Q_0$ for $q_0$ are compact, true values $q_0$ and $q_0$ are in the interior of $Q_1$ and $Q_0$, respectively.
Without loss of generality, assume that $\pi$ contains a correctly specified vector model for $f$.

**Theorem 2.1:** Under (C1)–(C5), when $W$ contains a correctly specified model for $\pi(X)$ or $F$ contains a correctly specified vector model for $f(X)$, $\hat{q}_{mr}$ is a consistent estimator of $q_0$ as $n \to \infty$, where $q_0$ is the true value.

**Theorem 2.2:** Under (C1)–(C7), $\sqrt{n}(\hat{q}_{mr} - q_0)$ has an asymptotic normal distribution with mean 0 and variance var, where

$$Z = f_1^{-1}(q_0^1)(Q_1(\beta_{1,0}) - E(Q_1 \Phi_f^T)E^{-1}(\Phi_f^{\otimes 2})\Phi_f) - f_0^{-1}(q_0^0)(Q_0(\beta_{1,0})$$

$$- E(Q_0 \Phi_f^T)E^{-1}(\Phi_f^{\otimes 2})\Phi_f],$$

with $Q_1(\beta_1) = T \{\psi_t(Y - q_0^1) - A_1(G_1)^{-1}g(\beta_1, q_1^0, \gamma_1^0)\}/\pi_1(\beta_1),$ $Q_0(\beta_1) = (1 - T)\{\psi_t(Y - q_0^0) - A_0(G_0)^{-1}g(\beta_1, q_0^0, \gamma_0^0)\}/(1 - \pi_1(\beta_1))$.
and
\[ A_1 = E(\psi_1(Y_1 - q_1^1)g(\beta_*, q_1^0, \gamma_*)^{T}/\pi_1(\beta_{1,0})), \]
\[ A_0 = E(\psi_1(Y_0 - q_0^1)g(\beta_*, q_0^0, \gamma_*)^{T}/(1 - \pi_1(\beta_{1,0}))). \]

Remark 2.4: A closed-form expression for the asymptotic variance is difficult to derive, because it depends on not only the number of models but also the particular functional form of those models. As in [18,22,31] and many others, we propose to use the bootstrap approach to provide a reasonable standard error (SE), i.e. the square roots of bootstrap variance estimators based on 200 bootstrap replications, to avoid tedious formula-based implementation.

3. Ensemble QTE estimation

3.1. Compression by refitting

In theory, the number of models postulated in \( \mathcal{W} \) and \( \mathcal{F} \) has no effect on the multiple robustness, as long as \( M \) and \( K \) are fixed. However, when \( M \) and \( K \) are large, the computational cost of obtaining \( \hat{w}_i \) will increase. Moreover, the convex hulls of \( g_i(X_i, \hat{\beta}, \hat{q}_i, \hat{\gamma}_i) \) may not contain the vector 0 such that the set for \( \hat{w}_i \) are empty. To solve this issue, we apply the following ensemble approach [12] to reduce the computational burden. Denote
\[ \hat{U}_{\pi_1} = \{ \hat{\pi}_{1i}, \ldots, \hat{\pi}_{Mi} \}, \quad \hat{U}_{i} = \{ \hat{Y}_i(\hat{\gamma}_1^l), \ldots, \hat{Y}_i(\hat{\gamma}_K^l) \}, \quad \text{for } l = 1, \ldots, L, \]
where \( \hat{\pi}_{mi} = \pi_m(X_i, \hat{\beta}_m) \) and \( \{ \hat{Y}_i(\hat{\gamma}_i^l) \}_{l=1}^L \) are independently random draws given \( X_i \) based on \( f_k(X, \hat{\gamma}_k^l) \) obtained in Section 2.1. The key step for defining our ensemble estimators is to perform twice least-squares model refitting to summarize the respective working model’s information. Specifically, by regressing \( T_i \) on \( \hat{U}_{\pi_1} \) for \( i = 1, \ldots, n \), we have
\[ \hat{\alpha}_\pi = \left( \sum_{i=1}^{n} \hat{U}_{\pi_1} \hat{U}_{\pi_1}^T \right)^{-1} \left( \sum_{i=1}^{n} \hat{U}_{\pi_1} T_i \right) \]
and define a weighted value as
\[ \tilde{\pi}_i = \hat{U}_{\pi_1} \times \{ \hat{\alpha}_\pi^2 / (\hat{\alpha}_\pi^T \hat{\alpha}_\pi) \}, \]
where \( a^2 \doteq (a_1^2, \ldots, a_d^2)^T \) for a vector \( a = (a_1, \ldots, a_d)^T \). In order to increase the weights of important models and reduce the weights of unimportant ones, we square the weights instead of using \( \hat{\alpha}_\pi \) directly, which makes the candidate models more distinguishable in their final contributions. Since the final weighted value \( \tilde{\pi}_i \) is refitted based on all working models, the consistency of \( \tilde{\pi}_i \) guarantees if the correct model is contained in \( \mathcal{W} \). On the other hand, by regressing \( Y_i \) on \( \hat{U}_{i}^l \) based on the complete subsamples, we have
\[ \hat{\alpha}_1^l = \left( \sum_{i=1}^{n} T_i \hat{U}_{1i} (\hat{U}_{1i}^l)^T \right)^{-1} \left( \sum_{i=1}^{n} T_i \hat{U}_{1i} Y_i \right) \]
\[ \hat{\alpha}_0^l = \left( \sum_{i=1}^{n} (1 - T_i) \hat{U}_{0i} (\hat{U}_{0i}^l)^T \right)^{-1} \left( \sum_{i=1}^{n} (1 - T_i) \hat{U}_{0i} Y_i \right). \]
Define the corresponding weighted values as
\[ \tilde{y}_{1i} = \tilde{U}_{1i} \times \{(\hat{\alpha}_1^1)^2/(\hat{\alpha}_1^1)^T \hat{\alpha}_1^1)\}, \quad \tilde{y}_{0i} = \tilde{U}_{0i} \times \{(\hat{\alpha}_0^1)^2/(\hat{\alpha}_0^1)^T \hat{\alpha}_0^1)\}. \]

The consistency of \( \tilde{y}_{1i} \) and \( \tilde{y}_{0i} \) is guaranteed if the correct models for conditional densities are contained in \( \mathcal{F} \). Define \( \tilde{m}^t_i = \sum_{l=1}^L \psi_T (\tilde{y}_{1i} - \tilde{q}^l)/L, \) where \( \tilde{q}^1 \) and \( \tilde{q}^0 \) are imputation estimators calculated by solving the following equation:
\[
\frac{1}{n} \sum_{i=1}^n \left[ T_i \psi_T (Y_i - \tilde{q}^1) + (1 - T_i) \frac{1}{L} \sum_{l=1}^L \psi_T (\tilde{y}_{1i} - \tilde{q}^l) \right] = 0, \\
\frac{1}{n} \sum_{i=1}^n \left[ (1 - T_i) \psi_T (Y_i - \tilde{q}^0) + T_i \frac{1}{L} \sum_{l=1}^L \psi_T (\tilde{y}_{0i} - \tilde{q}^l) \right] = 0,
\]
and then define
\[ \tilde{\theta} = \frac{1}{n} \sum_{i=1}^n \tilde{m}^t_i, \quad \tilde{\eta}^t = \frac{1}{n} \sum_{i=1}^n \tilde{m}^t_i. \]

### 3.2. Ensemble MR estimator

The ensemble MR estimators are defined as
\[ \hat{q}_{es} = \hat{q}_{es} - \hat{q}_{es}^0, \]
where \( \hat{q}_{es} \) is the solution to
\[ \sum_{i \in S_t} \hat{w}^t_i \psi_T (Y_i - \hat{q}^t) = 0, \quad (9) \]
and the weights \( w^t_i \)s are obtained by maximizing \( \prod_{i \in S_t} w_i \) under the following constraints:
\[ \sum_{i \in S_t} w_i = 1, \quad \sum_{i \in S_t} w_i \tilde{\pi}_i = \tilde{\theta}, \quad \sum_{i \in S_t} w_i \tilde{m}_i^t = \tilde{\eta}^t. \]

Using the Lagrange multipliers method,
\[ \hat{w}^t_i = \frac{1}{n_t \{1 + ((\tilde{\pi}_i - \tilde{\theta})(-1)^{i+1}, \tilde{m}_i^t - \tilde{\eta}^t)\hat{\rho}_t^{es}\}}, \text{ for } i \in S_t, \]
where \( \hat{\rho}_t^{es} \) satisfies two-dimensional estimating equations
\[ \sum_{i \in S_t} \frac{((\tilde{\pi}_i - \tilde{\theta})(-1)^{i+1}, \tilde{m}_i^t - \tilde{\eta}^t)^T}{1 + ((\tilde{\pi}_i - \tilde{\theta})(-1)^{i+1}, \tilde{m}_i^t - \tilde{\eta}^t)\hat{\rho}_t^{es}} = 0. \]

Under the same regularity conditions in Theorem 2.1, the consistency of ensemble estimator \( \hat{q}_{es} \) still holds.

**Theorem 3.1:** Under (C1)–(C5), when \( \mathcal{W} \) contains a correctly specified model for \( \pi(X) \) or \( \mathcal{F} \) contains a correctly specified vector model for \( f(Y|X) \), \( \hat{q}_{es} \) is a consistent estimator of \( q_0 \) as \( n \to \infty \).
4. Simulation studies

The simulation setting is the combination of [25,31] with some modifications. In specific, covariate $X_t$ is generated from a 8-dimensional normal distribution with mean zero and identity covariance matrix. Outcome models for $Y_t$ have two choices:

(1) $Y_t = \gamma_{11}^t + \gamma_{12}^t W_1 + \gamma_{13}^t W_2 + \gamma_{14}^t W_3 + \gamma_{15}^t W_4 + \gamma_{16}^t W_5 + \gamma_{17}^t W_6 + \epsilon_t,$
with transformation $W_k = X_k^2$, $k = 1, 2, 3, 4, 5, 6,$

(2) $Y_t = \gamma_{21}^t + \gamma_{22}^t W_1 + \gamma_{23}^t W_2 + \gamma_{24}^t W_3 + \gamma_{25}^t W_4 + \gamma_{26}^t W_5 + \gamma_{27}^t W_6 + \epsilon_t,$
with transformation $W_k = \exp(X_k/2), \ k = 1, 2, 3, 4, 5, 6,$

where $\epsilon_t$’s are independently from $\mathcal{N}(0, 1)$ for $t = 0, 1$. We generate $T$ from a Bernoulli distribution with probability $\pi(X)$ and consider two choices for $\pi(X)$:

(1) $\pi(X) = 1/[1 + \exp(\beta_{11}X_1 + \beta_{12}X_2 + \beta_{13}X_3 + \beta_{14}X_4 + \beta_{15}X_5 + \beta_{16}X_6 + \beta_{17}X_7 + \beta_{18}X_8)],$

(2) $\pi(X) = 1/[1 + \exp(\beta_{21}X_1 + \beta_{22}X_2 + \beta_{23}X_3 + \beta_{24}X_4 + \beta_{25}X_5 + \beta_{26}X_6 + \beta_{27}X_7 + \beta_{28}X_8)],$
\(\text{with transformation } W_k = \exp(X_k/2), \ k = 1, 2, 4, 5, 7, 8.\)

The true parameters are set as

$\gamma_{10}^1 = (0, 1.5, 1.5, 1.5, 3.5, 3.5, 3.5)^T, \quad \gamma_{10}^0 = (-2.5, 1.5, 1.5, 1.5, 3.5, 3.5, 3.5)^T,$

$\gamma_{20}^1 = (0, 1.5, 1.5, 1.5, 3.5, 3.5, 3.5)^T, \quad \gamma_{20}^0 = (-2.5, 1.5, 1.5, 1.5, 3.5, 3.5, 3.5)^T,$

$\beta_{10} = (-0.7, 0.7, 0.3, 0.7, 0.7, 0.7, 0.7)^T, \quad \beta_{20} = (-0.7, 0.7, 0.3, 0.7, 0.3, 0.3)^T.$

It can be verified that the true value of QTE is 2.5 in these cases.

We evaluate the finite-sample performance of the following estimators: (a) the naive estimator $\hat{\psi}_{\text{naive}}^1 = \hat{\psi}_{\text{naive}}^0 - \hat{\psi}_{\text{naive}}^0$ based on the observed $Y_t$’s, where $\hat{\psi}_{\text{naive}}^1$ and $\hat{\psi}_{\text{naive}}^0$ are the solutions to $\sum_{t=1}^n T(Y_t - q^1)/n = 0$ and $\sum_{t=1}^n (1 - T_i)\psi_t(Y_t - q^0)/n = 0$; (b) the IPW estimator $\hat{\psi}_{\text{ipw}}^1$ defined by (1); (c) the MI estimator $\hat{\psi}_{\text{mi}}$ defined by (2) and (3); (d) the AIPW estimator $\hat{\psi}_{\text{aipw}}$ defined by (4) and (5); (e) the MR estimators $\hat{\psi}_{\text{mr}}$ defined by (6); (f) the ensemble MR estimators $\hat{\psi}_{\text{es}}$ defined by (7). The MR estimators of [31] are not included since it is hard to specify the conditional distribution functions in these cases. To implement our proposed methods, we postulate two models for $\pi(X)$ as follows:

$\pi_1(X, \beta_1) = 1/[1 + \exp(\beta_{10}X_1 + \beta_{11}X_1 + \ldots + \beta_{18}X_8)],$

$\pi_2(X, \beta_2) = 1/[1 + \exp(\beta_{10}X_1 + \beta_{11}X_1 + \exp(X_1/2) + \ldots + \beta_{18}X_8/2)],$

and two functions for $f(Y_t | X)$ as follows:

$f_1^1(X, \gamma_1^1) = \frac{1}{\sqrt{2}\pi} \exp \left\{ -\frac{1}{2} (Y - (\gamma_1^1 + \gamma_1^1 X_1^2 + \ldots + \gamma_1^1 X_8^2))^2 \right\},$

$f_1^2(X, \gamma_2^1) = \frac{1}{\sqrt{2}\pi} \exp \left\{ -\frac{1}{2} (Y - (\gamma_2^1 + \gamma_2^1 X_1 + \exp(X_1/2) + \ldots + \gamma_2^1 X_8/2))^2 \right\}.$

Here, $\beta_1 = (\beta_{10}, \ldots, \beta_{18})^T, \beta_2 = (\beta_{20}, \ldots, \beta_{28})^T, \gamma_1^1 = (\gamma_1^1, \ldots, \gamma_1^1)^T$ and $\gamma_2^1 = (\gamma_2^1, \ldots, \gamma_2^1)^T$ are unknown parameter vectors. Following the notations of [22], a four-digit
subscript is used to distinguish estimators constructed using different candidate models. Each digit indicates the use of \( \pi_1(X, \beta_1) \), \( \pi_2(X, \beta_2) \), \( f_1(X, \gamma_1) \) and \( f_2(X, \gamma_2) \) in order. For example, \( \hat{\pi}_{\text{ipw}(1000)} \) denotes the IPW estimator constructed using \( \pi_1(X, \beta_1) \), \( \hat{\pi}_{\text{mr}(1101)} \) denotes the MR estimator constructed using \( \pi_1(X, \beta_1) \), \( \pi_2(X, \beta_2) \) and \( f_2(X, \gamma_2) \).

To investigate the effect of different sample sizes \( n \) and different quantile levels \( \tau \), six scenarios are considered, i.e. \((n, \tau) = (200, 0.25), (200, 0.50), (200, 0.75), (500, 0.25), (500, 0.50), (500, 0.75)\). Simulation results based on 1000 replications are presented in Tables 1–6. The performance of six estimators is assessed by the bias enlarged 100 times, standard deviation (SD) and mean square error (MSE).

A few conclusions can be drawn from the simulation results:

(i) When either the propensity score model \( \pi_1(X) \) or the conditional density vector model \( f_k(X, \gamma_k) \) is correctly specified, the AIPW estimators show small values of biases suggesting that they are doubly robust, while the proposed calibration MR and ensemble MR estimators are also unbiased. On the contrary, when the wrong model is used, it can be seen that the biases of the IPW or MI estimators are much larger. These findings agree with the theoretical results.

(ii) When both models are misspecified, the AIPW estimators exhibit large biases. For example, for \( \tau = 0.25, 0.5 \) and 0.75 with \( n = 200 \), the biases of \( \hat{\pi}_{\text{ipw}(101)} \) are about 1.39, 1.98 and 2.32 when the underlying \( \pi_1(X) \) and \( f_1(X, \gamma_1) \) are true; the biases of \( \hat{\pi}_{\text{ipw}(100)} \) are about \(-1.22, -1.49 \) and \(-1.42 \) when the underlying \( \pi_2(X) \) and \( f_1(X, \gamma_1) \) are true. Turning to the calibration MR and ensemble MR estimators, the multiple robustness is well demonstrated by inspecting their biases, which are consistent under these four data generating scenarios according to our theory. Neither the AIPW estimator nor any existing doubly robust estimator can achieve such robustness.

(iii) To assess the efficiency of our proposed estimators, we use the MSEs of \( \hat{\pi}_{\text{ipw}(\text{opt})} \) as the benchmark, where \( \hat{\pi}_{\text{ipw}(\text{opt})} \) denotes the corresponding AIPW estimator based on two correctly specified models; that is, \( \hat{\pi}_{\text{ipw}(\text{opt})} = \hat{\pi}_{\text{ipw}(1011)} \) under the first scenario (both \( \pi_1(X) \) and \( f_1(X, \gamma_1) \) are true), \( \hat{\pi}_{\text{ipw}(\text{opt})} = \hat{\pi}_{\text{ipw}(1001)} \) under the second scenario (both \( \pi_1(X) \) and \( f_2(X, \gamma_2) \) are true), \( \hat{\pi}_{\text{ipw}(\text{opt})} = \hat{\pi}_{\text{ipw}(0110)} \) under the third scenario (both \( \pi_2(X) \) and \( f_1(X, \gamma_1) \) are true), \( \hat{\pi}_{\text{ipw}(\text{opt})} = \hat{\pi}_{\text{ipw}(0101)} \) under the fourth scenario (both \( \pi_2(X) \) and \( f_2(X, \gamma_2) \) are true). It can be seen that the MSEs of the corresponding \( \hat{\pi}_{\text{dr}(1011)} \) and \( \hat{\pi}_{\text{es}(1011)} \) are identical to that of \( \hat{\pi}_{\text{ipw}(\text{opt})} \) under the first two scenarios, and the MSEs of \( \hat{\pi}_{\text{dr}(0111)} \) and \( \hat{\pi}_{\text{es}(0111)} \) are identical to that of \( \hat{\pi}_{\text{ipw}(\text{opt})} \) under the last two scenarios. Further, as long as the two correctly specified models are contained, we find that the MSEs of the calibration MR and ensemble MR estimators are smaller than or identical to that of \( \hat{\pi}_{\text{ipw}(\text{opt})} \).

(iv) According to our theory, the estimators \( \hat{\pi}_{\text{mr}(1111)} \) and \( \hat{\pi}_{\text{es}(1111)} \) are always unbiased under all four data-generating models. In addition, compared with \( \hat{\pi}_{\text{ipw}(\text{opt})} \), the MSEs of \( \hat{\pi}_{\text{mr}(1111)} \) and \( \hat{\pi}_{\text{es}(1111)} \) do not increase noticeably by adding more models, consistent with what the theoretical results suggested.

(v) Compared with the calibration MR estimators, the ensemble MR estimators have the similar performance. It can be concluded that the ensemble estimators are also multiply robust with good estimation efficiency. Moreover, we calculate the average computational time of the estimators \( \hat{\pi}_{\text{mr}(1111)} \) and \( \hat{\pi}_{\text{es}(1111)} \) for 50th percentile QTE.
non-normal outcome variables are given in Table S1 in the Supplementary Material, and we

To see the robustness of our proposed methods, the simulation results based on

| Method | Bias × 100 | SD | MSE |
|--------|------------|----|-----|
| $\hat{\pi}_{\text{naive}}$ | -24.8255 | 1.2007 | 1.5032 |
| $\hat{\pi}_{\text{qpw(1000)}}$ | 10.4358 | 1.4871 | 2.2224 |
| $\hat{\pi}_{\text{qpw(0100)}}$ | 103.5603 | 1.3826 | 2.9842 |
| $\hat{\pi}_{\text{qmi(0010)}}$ | 0.3186 | 0.2208 | 0.0487 |
| $\hat{\pi}_{\text{qmi(0001)}}$ | 333.2928 | 1.7858 | 14.2974 |
| $\hat{\pi}_{\text{qpw(1010)}}$ | -1.6007 | 0.6321 | 0.3999 |
| $\hat{\pi}_{\text{qpw(1001)}}$ | 35.9457 | 1.7630 | 3.2375 |
| $\hat{\pi}_{\text{qpw(0110)}}$ | -1.2020 | 0.5973 | 0.3569 |
| $\hat{\pi}_{\text{qpw(0101)}}$ | 139.8683 | 1.5865 | 4.4733 |
| $\hat{\pi}_{\text{qmi(1110)}}$ | -1.8295 | 0.6367 | 0.4058 |
| $\hat{\pi}_{\text{qmi(1101)}}$ | 24.2532 | 1.5823 | 2.5624 |
| $\hat{\pi}_{\text{qmi(0111)}}$ | 1.0850 | 0.6135 | 0.3765 |
| $\hat{\pi}_{\text{qmi(1011)}}$ | -0.9743 | 0.6489 | 0.4211 |
| $\hat{\pi}_{\text{qes(1111)}}$ | -2.1589 | 0.6608 | 0.4369 |
| $\hat{\pi}_{\text{qmr(1111)}}$ | 0.4411 | 0.6286 | 0.3952 |
| $\hat{\pi}_{\text{qes(1110)}}$ | 31.4157 | 1.5912 | 2.6305 |
| $\hat{\pi}_{\text{qes(1011)}}$ | 2.5627 | 0.6007 | 0.3615 |
| $\hat{\pi}_{\text{qes(0111)}}$ | 0.5112 | 0.6371 | 0.4059 |
| $\hat{\pi}_{\text{qmr(1011)}}$ | -1.0290 | 0.6280 | 0.3944 |

and

with different sample sizes $n$ and $\tau$, the above conclusions can also be established.

To see the robustness of our proposed methods, the simulation results based on

non-normal outcome variables are given in Table S1 in the Supplementary Material, and we
have the similar conclusions. In addition, Table 7 shows the simulation results on the estimated standard errors, i.e. the square roots of bootstrap variance estimators based on 200 bootstrap replications, of the QTE estimators on 50th percentile with \( n = 200 \). It can be seen that the SEs are similar to the SDs, which indicates the bootstrap approach works well. In summary, the simulation results suggest that the calibration MR and ensemble MR estimators all provide extra protection against model misspecification. Besides, the proposed ensemble estimators can reduce the computational burden efficiently.
### Table 3. Bias (× 100), standard deviation (SD) and mean square error (MSE) of the QTE estimators on 75th percentile with \( n = 200 \).

| Method   | \( \pi_1 \) is true | \( \pi_2 \) is true |
|----------|----------------------|----------------------|
|          | Bias × 100 | SD     | MSE   | Bias × 100 | SD     | MSE   |
| \( \hat{q}_{naive} \) | -55.0355 | 2.2135 | 5.2023 | -249.0107 | 0.9265 | 7.0591 |
| \( \hat{q}_{ipw(1000)} \) | 26.7685 | 3.0912 | 9.6274 | -8.5436 | 0.8703 | 0.7648 |
| \( \hat{q}_{ipw(0100)} \) | 243.0359 | 2.7452 | 13.4430 | 6.9864 | 0.8019 | 0.6479 |
| \( \hat{q}_{mi(0010)} \) | -0.1862 | 0.2376 | 0.0565 | -258.2188 | 0.7105 | 7.1725 |
| \( \hat{q}_{ipw(1010)} \) | -0.3718 | 0.8008 | 0.6412 | -55.6024 | 1.0504 | 1.4125 |
| \( \hat{q}_{ipw(0101)} \) | 100.8858 | 1.5930 | 3.5556 | 2.5897 | 0.5065 | 0.2573 |
| \( \hat{q}_{qaipw(1010)} \) | 1.0535 | 0.7578 | 0.5744 | -78.0753 | 1.0425 | 1.6964 |
| \( \hat{q}_{qaipw(1001)} \) | 231.6853 | 2.8135 | 13.2838 | 2.7956 | 0.4939 | 0.2448 |
| \( \hat{q}_{mi(1110)} \) | -1.5686 | 0.8378 | 0.7022 | -18.3669 | 0.9196 | 0.8862 |
| \( \hat{q}_{mi(1101)} \) | 57.5476 | 2.7253 | 7.7582 | 1.9227 | 0.5172 | 0.2678 |
| \( \hat{q}_{mi(0111)} \) | -0.2764 | 0.8380 | 0.7023 | 0.3220 | 0.5378 | 0.2892 |
| \( \hat{q}_{mi(1011)} \) | -1.2237 | 0.8294 | 0.6880 | 1.4108 | 0.5441 | 0.2963 |
| \( \hat{q}_{mi(0111)} \) | -1.2273 | 0.8423 | 0.7097 | 1.3067 | 0.5462 | 0.2985 |
| \( \hat{q}_{es(1110)} \) | -0.9500 | 0.8373 | 0.7011 | -20.1202 | 0.9196 | 0.8862 |
| \( \hat{q}_{es(1101)} \) | 65.4680 | 2.8296 | 8.4352 | 2.4070 | 0.5264 | 0.2777 |
| \( \hat{q}_{es(0111)} \) | -2.8079 | 0.8238 | 0.6795 | 2.3712 | 0.5200 | 0.2709 |
| \( \hat{q}_{es(1011)} \) | 0.8640 | 0.8382 | 0.7027 | 2.9441 | 0.5079 | 0.2589 |
| \( \hat{q}_{es(1111)} \) | -0.3277 | 0.8283 | 0.6861 | -54.5747 | 2.7989 | 8.1315 |

5. Two real data applications

5.1. Birthweight data

We apply the proposed methods to the birthweight data discussed in Section 1. There are a total of 50,000 observations. The outcome response is birthweight of infants recorded in
grams with five covariates: mother’s marriage status \( X_1 \), mother’s race \( X_2 \), gender of the infant \( X_3 \), mother’s age \( X_4 \) and mother’s education level \( X_5 \). The QTE is defined as the quantile difference of birthweight distributions between the treatment (mother smokes; \( T = 1 \)) and control (mother does not smoke; \( T = 0 \)) groups.
For the propensity score, we consider two candidate models,

(1) \[ \pi_1(X, \beta_1) = 1 / \{1 + \exp(\beta_{10} + \beta_{11}X_1 + \cdots + \beta_{15}X_5)\} \],

(2) \[ \pi_2(X, \beta_2) = 1 / \{1 + \exp(\beta_{20} + \beta_{21}\exp(X_1/2) + \cdots + \beta_{25}\exp(X_5/2))\} \].
For conditional density, we also consider two functions,

\[(A) \quad f_1^t(X, \gamma_1^t) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left[ Y - (\gamma_{10}^t + \gamma_{11}^t X_1 + \ldots + \gamma_{15}^t X_5) \right]^2 \right\}, \]

\[(B) \quad f_2^t(X, \gamma_2^t) = \frac{1}{\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left[ Y - (\gamma_{20}^t + \gamma_{21}^t \exp(X_1/2) + \ldots + \gamma_{25}^t \exp(X_5/2)) \right]^2 \right\}. \]
Table 7. Standard deviation (SD) and bootstrapped standard error (SE) of the QTE estimators on 50th percentile with \( n = 500 \).

| Method        | \( \pi_1 \) is true | \( \pi_2 \) is true | \( \pi_1 \) is true | \( \pi_2 \) is true |
|---------------|----------------------|----------------------|----------------------|----------------------|
|               | SD                   | SE                   | SD                   | SE                   |
| \( q_{\text{naive}} \) | 0.9431               | 1.0168               | 0.4550               | 0.4593               |
| \( q_{\text{pw(1000)}} \) | 1.2005               | 1.2821               | 0.3975               | 0.4326               |
| \( q_{\text{pw(0100)}} \) | 1.1229               | 1.6444               | 0.3634               | 0.4022               |
| \( q_{\text{mi(0100)}} \) | 0.1215               | 0.1242               | 0.3346               | 0.3255               |
| \( q_{\text{mi(0001)}} \) | 0.9336               | 0.9412               | 0.1191               | 0.1193               |
| \( q_{\text{ipw(1010)}} \) | 0.4099               | 0.4463               | 0.5190               | 0.5683               |
| \( q_{\text{ipw(0101)}} \) | 1.2832               | 1.4721               | 0.2614               | 0.2928               |
| \( q_{\text{ipw(0011)}} \) | 0.3825               | 0.4261               | 0.5415               | 0.5851               |
| \( q_{\text{ipw(0110)}} \) | 1.3428               | 1.4280               | 0.2506               | 0.2845               |
| \( q_{\text{mr(1110)}} \) | 0.4193               | 0.4601               | 0.4072               | 0.4691               |
| \( q_{\text{mr(1101)}} \) | 1.1731               | 1.2210               | 0.2716               | 0.3003               |
| \( q_{\text{mr(1011)}} \) | 0.3993               | 0.4519               | 0.2586               | 0.2941               |
| \( q_{\text{mr(0111)}} \) | 0.4243               | 0.4606               | 0.2714               | 0.3019               |
| \( q_{\text{mr(0110)}} \) | 0.4132               | 0.4597               | 0.4026               | 0.4670               |
| \( q_{\text{mr(1100)}} \) | 1.2071               | 1.2574               | 0.2721               | 0.2998               |
| \( q_{\text{mr(1001)}} \) | 0.3905               | 0.4431               | 0.2532               | 0.2870               |
| \( q_{\text{mr(0101)}} \) | 0.4165               | 0.4605               | 0.2706               | 0.2990               |
| \( q_{\text{mr(0011)}} \) | 0.4111               | 0.4600               | 0.2710               | 0.2991               |

Table 8. Estimates of the QTE (with standard errors) for the birthweight data.

| Method        | \( \tau = 0.25 \) | \( \tau = 0.50 \) | \( \tau = 0.75 \) |
|---------------|-------------------|-------------------|-------------------|
|               | QTE    | SE     | QTE    | SE     | QTE    | SE     |
| \( q_{\text{naive}} \) | -229.9581 | 7.0490 | -230.8787 | 9.3462 | -225.0790 | 9.6223 |
| \( q_{\text{pw(1000)}} \) | -238.6364 | 11.7369 | -240.5241 | 13.2462 | -223.1643 | 11.3477 |
| \( q_{\text{pw(0100)}} \) | -243.2168 | 13.9244 | -244.6566 | 13.9895 | -223.6857 | 11.3544 |
| \( q_{\text{mr(0101)}} \) | -259.6074 | 12.6416 | -254.3605 | 10.3636 | -253.3463 | 12.7102 |
| \( q_{\text{mr(0001)}} \) | -175.4767 | 33.7354 | -260.4832 | 31.5832 | -254.5129 | 32.3834 |
| \( q_{\text{ipw(1010)}} \) | -226.8475 | 12.0008 | -255.9659 | 27.9477 | -222.1215 | 11.6819 |
| \( q_{\text{ipw(0101)}} \) | -242.0030 | 18.9663 | -284.4943 | 27.3348 | -215.7643 | 14.2382 |
| \( q_{\text{mr(1010)}} \) | -233.9161 | 11.0169 | -245.2229 | 12.9668 | -229.6715 | 9.6017 |
| \( q_{\text{mr(1001)}} \) | -237.8042 | 12.2734 | -247.1024 | 12.2976 | -231.9572 | 10.2591 |
| \( q_{\text{mr(0110)}} \) | -234.8951 | 11.0675 | -245.6265 | 12.7009 | -231.1429 | 9.7476 |
| \( q_{\text{mr(0111)}} \) | -235.6574 | 11.7102 | -241.0843 | 13.6694 | -230.5143 | 10.4754 |
| \( q_{\text{mr(1101)}} \) | -233.5665 | 10.9342 | -240.8855 | 13.2931 | -228.5643 | 9.5173 |
| \( q_{\text{mr(1100)}} \) | -230.8400 | 10.8118 | -240.5903 | 13.2683 | -229.4000 | 9.4980 |
| \( q_{\text{mr(1011)}} \) | -237.4965 | 12.1330 | -247.0180 | 12.3448 | -231.8572 | 10.1878 |
| \( q_{\text{mr(0111)}} \) | -236.0630 | 11.9501 | -247.3012 | 12.3791 | -237.8042 | 12.2734 |
| \( q_{\text{mr(0011)}} \) | -232.9441 | 10.2920 | -243.2168 | 13.6694 | -234.8951 | 12.7009 |
| \( q_{\text{mr(0010)}} \) | -229.6084 | 7.8184 | -249.8735 | 11.6672 | -231.6643 | 11.2899 |

for \( t = 0, 1 \). Here, \( \beta_1 = (\beta_{10}, \ldots, \beta_{15})^T \), \( \beta_2 = (\beta_{20}, \ldots, \beta_{25})^T \), \( \gamma_1^t = (\gamma_{10}^t, \ldots, \gamma_{15}^t)^T \) and \( \gamma_2^t = (\gamma_{20}^t, \ldots, \gamma_{25}^t)^T \) are unknown parameter vectors. The resulting QTE estimates and their standard errors (SEs) by the square roots of bootstrap variance estimators based on 200 bootstrap replicates are presented in Table 8.
Table 9. Estimates of the QTE (with standard errors) for ACTG175.

|       | \( \tau = 0.25 \) |       | \( \tau = 0.50 \) |       | \( \tau = 0.75 \) |
|-------|-----------------|-------|-----------------|-------|-----------------|
|       | QTE             | SE    | \( \Delta_1 \) | QTE   | SE   | \( \Delta_1 \) |
| \( \hat{q}_{naive} \) | 71.3459 | 20.0358 | 42.2058 | 14.6791 | 53.9581 | 13.8785 |
| \( q_{gwp(1000)} \)  | 69.8850 | 18.8603 | 42.9600 | 11.1684 | 54.6550 | 13.5096 |
| \( q_{gwp(0100)} \)  | 69.9400 | 18.8220 | 42.9550 | 11.2094 | 54.6450 | 13.4456 |
| \( q_{mi(0010)} \)   | 66.3811 | 18.4286 | 65.5699 | 12.6397 | 61.5123 | 14.0202 |
| \( q_{mi(0001)} \)   | 37.6686 | 14.5264 | 62.3850 | 61.9249 | 75.7147 | 249.2752 |
| \( q_{aipw(1010)} \) | 69.7351 | 18.3929 | 44.7665 | 11.2651 | 56.2639 | 13.2649 |
| \( q_{aipw(1001)} \) | 70.1526 | 18.2191 | 44.6825 | 11.1181 | 56.4743 | 13.6682 |
| \( q_{aipw(0101)} \) | 69.7833 | 18.3909 | 44.7746 | 11.2319 | 56.2490 | 13.2747 |
| \( q_{aipw(0100)} \) | 70.1626 | 18.2040 | 44.6965 | 11.1304 | 56.5009 | 13.6574 |
| \( q_{mr(1110)} \)   | 69.7450 | 18.2671 | 44.5150 | 11.0271 | 55.3900 | 13.3250 |
| \( q_{mr(1101)} \)   | 70.2400 | 18.2027 | 44.2350 | 11.0749 | 56.0250 | 13.3606 |
| \( q_{mr(1111)} \)   | 69.8600 | 17.9539 | 44.4950 | 10.7717 | 56.0950 | 13.0452 |
| \( q_{mr(1011)} \)   | 69.6100 | 18.0420 | 44.4600 | 10.8627 | 56.0500 | 13.0645 |
| \( q_{mr(1111)} \)   | 69.7800 | 18.2151 | 44.5600 | 10.7215 | 56.0150 | 13.2447 |
| \( q_{es(1110)} \)   | 69.2650 | 18.5958 | 44.4500 | 11.0161 | 55.8200 | 13.4988 |
| \( q_{es(1101)} \)   | 69.8800 | 18.2103 | 44.3350 | 11.1591 | 56.0950 | 13.3230 |
| \( q_{es(1011)} \)   | 69.3950 | 18.2880 | 44.3950 | 11.0662 | 55.6800 | 13.0472 |
| \( q_{es(1010)} \)   | 69.2850 | 18.4310 | 44.3150 | 11.1159 | 55.5850 | 12.9666 |
| \( q_{es(1111)} \)   | 69.3050 | 18.2488 | 44.3700 | 11.1126 | 55.6200 | 13.0902 |

|       | \( \Delta_2 \) |
|-------|-----------------|
| \( \hat{q}_{naive} \) | 72.2414 | 22.1438 |
| \( q_{gwp(1000)} \)  | 71.3650 | 18.8688 |
| \( q_{gwp(0100)} \)  | 71.5750 | 19.0032 |
| \( q_{mi(0010)} \)   | 81.8023 | 18.7827 |
| \( q_{mi(0001)} \)   | 61.0501 | 14.7372 |
| \( q_{aipw(1010)} \) | 70.3742 | 18.1412 |
| \( q_{aipw(1001)} \) | 71.6532 | 18.5652 |
| \( q_{aipw(0101)} \) | 70.4202 | 18.1531 |
| \( q_{aipw(0100)} \) | 71.7254 | 18.5994 |
| \( q_{mr(1110)} \)   | 70.7150 | 18.2312 |
| \( q_{mr(1101)} \)   | 71.4850 | 18.6324 |
| \( q_{mr(1111)} \)   | 70.8100 | 18.0375 |
| \( q_{es(1110)} \)   | 70.7300 | 17.9711 |
| \( q_{es(1101)} \)   | 70.9250 | 18.3402 |
| \( q_{es(1011)} \)   | 70.5700 | 18.4841 |
| \( q_{es(1010)} \)   | 71.2350 | 18.5383 |
| \( q_{es(1111)} \)   | 70.4350 | 18.3266 |
| \( q_{es(1011)} \)   | 70.5000 | 18.1565 |
| \( q_{es(1111)} \)   | 70.4650 | 18.1007 |

|       | \( \Delta_3 \) |
|-------|-----------------|
| \( \hat{q}_{naive} \) | 44.4620 | 15.6223 |
| \( q_{gwp(1000)} \)  | 44.0300 | 13.2824 |
| \( q_{gwp(0100)} \)  | 44.0450 | 13.2712 |
| \( q_{mi(0010)} \)   | 54.4630 | 16.6472 |
| \( q_{mi(0001)} \)   | 23.4713 | 12.9447 |
| \( q_{aipw(1010)} \) | 46.0368 | 13.0046 |
| \( q_{aipw(0101)} \) | 47.5850 | 13.0701 |
| \( q_{aipw(0100)} \) | 46.1016 | 13.0753 |
| \( q_{aipw(0101)} \) | 47.6370 | 13.1076 |
| \( q_{mr(1110)} \)   | 45.4850 | 13.0661 |
| \( q_{mr(1101)} \)   | 46.1150 | 13.1822 |
| \( q_{mr(1111)} \)   | 45.7800 | 13.1287 |
| \( q_{mr(1111)} \)   | 45.7900 | 13.1484 |

(continued)
| τ = 0.25 | QTE | SE  | QTE | SE  | QTE | SE  |
| --- | --- | --- | --- | --- | --- | --- |
| $\hat{q}_{mr(1111)}$ | 45.8300 | 13.1539 | 41.0850 | 12.7195 | 39.7400 | 16.6526 |
| $\hat{q}_{mr(1110)}$ | 45.5100 | 13.2198 | 40.6450 | 12.9688 | 39.5400 | 17.0612 |
| $\hat{q}_{mr(1101)}$ | 46.2200 | 13.2465 | 40.2150 | 12.7889 | 39.2000 | 16.7071 |
| $\hat{q}_{mr(0111)}$ | 45.3400 | 13.0755 | 40.8400 | 12.9197 | 39.7450 | 16.5748 |
| $\hat{q}_{mr(1011)}$ | 45.2900 | 13.0825 | 40.8700 | 12.9527 | 39.7600 | 16.6701 |
| $\hat{q}_{mr(1111)}$ | 45.3500 | 13.0175 | 40.7100 | 12.9726 | 39.6850 | 16.7009 |

It can be shown that (i) the proposed MR estimates and ensemble MR estimates are close, and smaller than naive estimates in all cases. The estimates $\hat{q}_{mi(0001)}$ have significantly different values from other estimates and their SEs are much larger than those of $\hat{q}_{mi(0010)}$, which indicates model (A) is more reasonable and the response may follow a normal distribution; (ii) we look into the estimates using model (A) as a candidate model and find that their values are all close; (iii) the negative values of the QTE estimates indicate that the 25th, 50th and 75th percentiles of the infants’ birthweights under $T = 1$ (mother smokes) are smaller than those under $T = 0$ (mother does not smoke); (iv) the SEs of $\hat{q}_{ipw(0110)}$ are larger than those of $\hat{q}_{ipw(1010)}$, which indicates that model (1) may be more suitable for the treatment assignment; (v) compared to the IPW, MI and AIPW estimates, our proposed estimates have smaller SEs, which shows that our estimates are more efficient due to the inclusion of more propensity score models and conditional density models.

### 5.2. ACTG 175 data

We further consider the ACTG 175 data to investigate the treatment effects of different antiretroviral regimens. In specific, 1340 HIV positive patients were divided into 4 arms, i.e. 321 subjects with zidovudine or ZDV, 333 subjects with didanosine or ddi, 336 subjects with ZDV and ddi, 350 subjects with ZDV and zalcitabine. The outcome response $Y$ is CD4 count at 96 ± 5 weeks with six continuous covariates: age ($X_1$), weight ($X_2$), CD4 cell counts at baseline and 20 ± 5 weeks ($X_3$ and $X_4$), and CD8 cell counts at baseline and 20 ± 5 weeks ($X_5$ and $X_6$). In this study, the traditional regimen, i.e. ZDC, is the control arm, and the regimens ddi, ZDV and ddi, ZDV and zalcitabine are three new treatments. To see whether the new three regimens work or not, we compute the quantile differences of the CD4 count at 96 ± 5 weeks between the three new treatments and control arm, respectively, which are denoted as $\Delta_1$, $\Delta_2$ and $\Delta_3$. It is reasonable to assume that, given the six baseline covariates, the treatment assignment does not depend on the CD4 count at 96 ± 5 weeks. Thus, the unconfoundedness assumption holds.

For the propensity score models, two functions are considered

\begin{align*}
(1) \quad \pi_1(X, \beta_1) &= \frac{1}{1 + \exp(\beta_{10} + \beta_{11}X_1 + \beta_{12}X_2 + \cdots + \beta_{16}X_6)}, \\
(2) \quad \pi_2(X, \beta_2) &= 1 - \exp(-\exp(\beta_{20} + \beta_{21}X_1 + \cdots + \beta_{26}X_6)).
\end{align*}

Since the four outcome distributions are skewed, the log-transformation is considered. In specific, we assume that $Y$ has a normal distribution left truncated at 0 or log($Y$) has a normal distribution right truncated at the maximum of observed $Y$ such that we consider
the following two conditional density functions:

\[ f_1^t(X, \gamma_t^1) = \frac{\phi\left(\frac{Y - (\gamma_{10}^t + \gamma_{11}^t X_1 + \ldots + \gamma_{16}^t X_6)}{\sigma}\right)}{\sigma [1 - \Phi(- (\gamma_{10}^t + \gamma_{11}^t X_1 + \ldots + \gamma_{16}^t X_6)/\sigma)]}, \]

\[ f_2^t(X, \gamma_t^2) = \frac{\phi\left(\frac{Z - (\gamma_{20}^t + \gamma_{21}^t X_1 + \ldots + \gamma_{26}^t X_6)}{\sigma}\right)}{\sigma \Phi\left(\frac{b - (\gamma_{20}^t + \gamma_{21}^t X_1 + \ldots + \gamma_{26}^t X_6)}{\sigma}\right)}, \]

with transformation \( Z = \log(Y) \) and \( b = \max(Z_i) \),

for \( t = 0, 1 \). Here, \( \beta_1 = (\beta_{10}, \ldots, \beta_{16})^T, \beta_2 = (\beta_{20}, \ldots, \beta_{26})^T, \gamma_1^t = (\gamma_{10}^t, \ldots, \gamma_{16}^t)^T, \gamma_2^t = (\gamma_{20}^t, \ldots, \gamma_{26}^t)^T \) and \( \sigma \) are unknown parameters, \( \phi(x) = \exp(-x^2/2)/\sqrt{2\pi} \) is a standard normal distribution and \( \Phi(x) \) is the cumulative distribution function of \( \phi(x) \). Both models (A) and (B) can be fitted by R package ‘truncreg’. The resulting QTE estimates and their standard errors (SEs) by the square roots of bootstrap variance estimators based on 200 bootstrap replications are presented in Table 9.

It can be shown that (i) the proposed MR estimates and ensemble MR estimates are close; (ii) the estimates \( \hat{q}_{mi(0001)} \) are different from other estimates, which indicates a normal distribution with truncation is more reasonable; (iii) the positive values of the QTE estimates indicate that the 25th, 50th and 75th percentiles of the new regimens have a positive influence than the traditional ZDV; (iv) compared with other estimates, our proposed estimates have slightly smaller SEs by including the multiple models.

6. Discussion

In this paper, we consider the QTE estimation and propose two classes of MR estimators. The first class of MR estimators allows more than one candidate model for both propensity score and conditional density of potential outcomes. The second class of MR estimators is based on the ensemble approach to reduce computational burden efficiently while ensuring multiple robustness. Both of these two classes of estimators are based on the calibration approaches that have the same structure as the IPW estimator and the weights are calculated by the constrained EL method. Throughout this paper, we assume that the unconfoundedness condition holds. However, in observational studies, there often exist unmeasured confounding variables, which may cause some problems for evaluating the QTE, i.e. the estimates may have large biases or population parameters are not identifiable. In addition, the working models considered are all parametric, and the robustness could be further improved by using nonparametric working models.

Acknowledgments

We are grateful to the Editor, the Associate Editor and two anonymous referees for their insightful comments and suggestions on this article, which have led to significant improvements.

Disclosure statement

No potential conflict of interest was reported by the author(s).
Funding

This paper was supported by the National Natural Science Foundation of China [Grant Nos. 11871287, 11771144, 11801359], the Natural Science Foundation of Tianjin [Grant No. 18JCBC41100], Fundamental Research Funds for the Central Universities and the Key Laboratory for Medical Data Analysis and Statistical Research of Tianjin.

References

[1] A. Abadie, *Semiparametric instrumental variable estimation of treatment response models*, J. Econom. 113 (2003), pp. 231–263.

[2] A. Abadie, J.D. Angrist, and G.W. Imbens, *Instrumental variables estimates of the effect of subsidized training on the quantile of trainee earnings*, Econometrica 70 (2002), pp. 91–117.

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

[4] M.D. Cattaneo, *Efficient semiparametric estimation of multi-valued treatment effects under ignorability*, J. Econom. 155 (2010), pp. 138–154.

[5] K.C.G. Chan and S.C.P. Yam, *Oracle, multiple robust and multipurpose calibration in a missing response problem*, Statist. Sci. 29 (2014), pp. 380–396.

[6] S. Chen and D. Haziza, *Multiple robust imputation procedures for the treatment of item nonresponse in surveys*, Biometrika 102 (2017), pp. 439–453.

[7] J. Chen, R.R. Sitter, and C. Wu, *Using empirical likelihood methods to obtain range restricted weights in regression estimators for surveys*, Biometrika 89 (2002), pp. 230–237.

[8] X. Chen, A.T.K. Wan, and Y. Zhou, *Efficient quantile regression analysis with missing observations*, J. Am. Stat. Assoc. 110 (2015), pp. 723–741.

[9] S. Chen and C.L. Yu, *Parameter estimation through semiparametric quantile regression imputation*, Electron. J. Stat. 10 (2016), pp. 3621–3647.

[10] V. Chernozhukov and C. Hansen, *An IV model of quantile treatment effects*, Econometrica 73 (2005), pp. 245–261.

[11] K. Doksum, *Empirical probability plots and statistical inference for nonlinear models in the two-sample case*, Ann. Statist. 2 (1974), pp. 267–277.

[12] X. Duan and G. Yin, *Ensemble approaches to estimating the population mean with missing response*, Scand. J. Statist. 44 (2017), pp. 899–917.

[13] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties*, J. Am. Stat. Assoc. 96 (2001), pp. 1348–1360.

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

[15] M. Frölich and B. Melly, *Estimation of quantile treatment effects with Stata*, Stata J. 10 (2010), pp. 423–457.

[16] M. Frölich and B. Melly, *Unconditional quantile treatment effects under endogeneity*, J. Bus. Econom. Statist. 31 (2013), pp. 346–357.

[17] S.M. Hammer, D.A. Katzenstein, M.D. Hughes, H. Gundacker, R.T. Schooley, R.H. Haubrich, W.K. Henry, M.M. Lederman, J.P. Phair, M. Niu, M.S. Hirsch, and T.C. Merigan, *A trial comparing nucleoside monotherapy with combination therapy in HIV-infected adults with CD4 cell counts from 200 to 500 per cubic millimeter*, New Engl. J. Med. 335 (1996), pp. 1081–1089.

[18] P. Han, *Multiply robust estimation in regression analysis with missing data*, J. Am. Stat. Assoc. 109 (2014), pp. 1159–1173.

[19] P. Han, *Combining inverse probability weighting and multiple imputation to improve robustness of estimation*, Scand. J. Statist. 43 (2016), pp. 246–260.

[20] P. Han, *Intrinsic efficiency and multiple robustness in longitudinal studies with dropout*, Biometrika 103 (2016), pp. 683–700.

[21] P. Han, L. Kong, J.W. Zhao, and X.C. Zhou, *A general framework for quantile estimation with incomplete data*, J. R. Statist. Soc. Ser. B (Statist. Methodol.) 81 (2019), pp. 305–333.
[22] P. Han and L. Wang, *Estimation with missing data: Beyond double robustness*, Biometrika 100 (2013), pp. 417–430.
[23] D.G. Horvitz and D.J. Thompson, *A generalization of sampling without replacement from a finite universe*, J. Am. Stat. Assoc. 47 (1952), pp. 663–685.
[24] G.W. Imbens and J.D. Angrist, *Identification and estimation of local average treatment effects*, Econometrica 62 (1994), pp. 467–475.
[25] J.D.Y. Kang and J.L. Schafer, *Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data (with discussion)*, Statist. Sci. 22 (2007), pp. 523–539.
[26] E.L. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, San Francisco, 1974.
[27] J. Qin and J. Lawless, *Empirical likelihood and general estimating equations*, Ann. Statist. 22 (1994), pp. 300–325.
[28] J.M. Robins, A. Rotnitzky, and L.P. Zhao, *Estimation of regression coefficients when some regressors are not always observed*, J. Am. Stat. Assoc. 89 (1994), pp. 846–866.
[29] D.B. Rubin, *Multiple imputation after 18+ years*, J. Am. Stat. Assoc. 91 (1996), pp. 473–489.
[30] C.Y. Tang and Y. Qin, *An efficient empirical likelihood approach for estimating equations with missing data*, Biometrika 99 (2012), pp. 1001–1007.
[31] Y. Xie, C. Cotton, and Y. Zhu, *Multiply robust estimation of causal quantile treatment effects*, Stat. Med. 39 (2020), pp. 4238–4251.