Statistica Sinica Preprint No: SS-2021-0339	
Title	Learning Non-monotone Optimal Individualized Treatment Regimes
Manuscript ID	SS-2021-0339
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0339
Complete List of Authors	Trinetri Ghosh, Yanyuan Ma, Wensheng Zhu and Yuanjia Wang
Corresponding Authors	Trinetri Ghosh
E-mails	tghosh3@wisc.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica

LEARNING NON-MONOTONE OPTIMAL INDIVIDUALIZED TREATMENT REGIMES

Trinetri Ghosh[1], Yanyuan Ma[2], Wensheng Zhu[3] and Yuanjia Wang[4]

[1] *University of Wisconsin-Madison,* [2] *Pennsylvania State University,*

[3] *Northeast Normal University and* [4] *Columbia University*

Abstract: We propose a new modeling and estimation approach to select the optimal treatment regime from different options through constructing a robust estimating equation. The method is protected against misspecification of the propensity score model, the outcome regression model for the non-treated group, or the potential non-monotonic treatment difference model. Our method also allows residual errors to depend on covariates. A single index structure is incorporated to facilitate the nonparametric estimation of the treatment difference. We then identify the optimal treatment through maximizing the value function. Theoretical properties of the treatment assignment strategy are established. We illustrate the performance and effectiveness of our proposed estimators through extensive simulation studies and a real dataset on the effect of maternal smoking on baby birth weight.

Key words and phrases: Double- and multi-robust, optimal treatment regimes, propensity score, value function.

1. Introduction

Due to the between-person heterogeneity, it is common to observe different responses to the same treatment in different individuals. Various factors can contribute to the heterogeneity from one individual to another, such as genetic risk factors, age and individual-specific environmental exposures. Thus, when different treatment options are available, it is desirable to select the best treatment regime that is specific to a particular individual, which is one of the goals of precision medicine. Precision medicine aims to determine a strategic assignment of treatments to patients according to their characteristics and medical history. This goal can be achieved by using an individualized treatment rule (ITR), i.e., a deterministic function of subject-specific factors that are responsible for patients' heterogeneous responses to treatment. The optimal ITR maximizes the expected clinical outcome of interest under the ITR. Furthermore, an optimal dynamic treatment regime usually consists of a set of sequential decision rules applied at a set of decision points. There has been significant research developments on estimating the optimal treatment regimes in recent years, as will be elaborated subsequently. In this work, we will focus on estimating the individualized treatment regimes at a single decision time point.

Two popular model-based methods to derive the optimal dynamic treat-

ment regimes are Q-learning and A-learning, standing for quality-learning and advantage-learning, respectively. Q-learning (Watkins 1989, Watkins & Dayan 1992, Nahum-Shani et al. 2012, Zhao et al. 2009, 2011, Murphy 2005, Qian & Murphy 2011, Song et al. 2015, Goldberg & Kosorok 2012, Chakraborty et al. 2010) is built on a postulated regression outcome model for the outcome of interest and is implemented through a backward induction fitting procedure. This approach was initially proposed by Watkins (1989), and later a detailed proof of convergence was provided by Watkins & Dayan (1992). The performance of the optimal treatment decision rule obtained by Q-learning depends on the correctly-specified outcome model. On the other hand, A-learning (Murphy 2003, Blatt et al. 2004, Robins 2004, Orellana et al. 2010, Liang et al. 2018) maximizes estimating equations to estimate the contrast functions, using the estimated probability of observed treatment assignment given patient information at each decision point (i.e., treatment propensity scores). The performance of the optimal treatment decision rule obtained by A-learning thus relies on the suitable treatment assignment model.

Another approach, known as model free or policy (value) search method (Zhang, Tsiatis, Laber & Davidian 2012, Zhang, Tsiatis, Davidian, Zhang & Laber 2012, Zhao et al. 2012, Jiang, Lu, Song & Davidian 2017, Jiang,

Lu, Song, Hudgens & Naprvavnik 2017), directly derives and maximizes a consistent estimator for the value function over a prespecified class of treatment regimes indexed by a finite dimensional parameter or over a class of nonparametric treatment regimes. For example, Zhang, Tsiatis, Laber & Davidian (2012) formulated the inverse propensity score weighted (IPW) estimator and the double-robust augmented IPW estimator for the value function with a single decision time point. Later, Zhang et al. (2013) extended this idea to more than one decision point. Zhang, Tsiatis, Davidian, Zhang & Laber (2012) and Zhao et al. (2012) recast the original problem of finding the optimal treatment regime as a weighted classification problem. The former obtains the optimal treatment regime by minimizing the expected weighted misclassification error, whereas the later used outcome-weighted support vector machine. Other relevant work includes Robins (2004), Foster et al. (2011), Zhao et al. (2013), Matsouaka et al. (2014), Song et al. (2017), Bai et al. (2017), Fan et al. (2017), Shi et al. (2018), and Huang & Yang (2020).

In this paper, we propose a new modeling and estimation method to determine the optimal treatment regimes at a decision time point, combining the advantages of Q-learning, A-learning and the model-free approach. In addition, our model has the advantage that it only assumes the treat-

ment difference as a smooth function of an index of the covariates, without requiring the smooth function to be monotonic. This is practically important. For example, for a patient with heart disease, low blood pressure and high blood pressure both can increase the risk of heart attack, hence resulting in a possible non-monotonic treatment difference model. Another example is the relationship between BMI and health risks, where both underweight and obese individuals have increased risks of a range of health measures. Our model also has the flexibility of allowing the model error to be dependent on the covariates. In practice, this is also an important feature. Further, we consider a multi-robust estimating equation as protection against misspecified propensity score function, treatment difference model, and outcome regression model for the non-treated group. Benefiting from the smoothness of the treatment difference function, our treatment regime identification rate is $O_p(n^{-2/5})$, faster than the existing rate of $O_p(n^{-1/3})$ (Fan et al. 2017), where treatment difference function is assumed to be monotonic.

The remainder of the paper is organized as follows. In Section 2, we introduce the estimation procedure and the algorithm for our proposed method. Section 3 provides the asymptotic properties of the proposed estimators for $\boldsymbol{\beta}$ and the treatment difference function, $Q(\cdot)$. In Section 4,

we summarize the finite-sample performance of the estimators for different designs, including well-specified and misspecified models. We implemented our method on baby birth weight dataset, where the research interest is to investigate whether the maternal smoking during pregnancy has any effect on birth weight in Section 5. Section 6 concludes the paper.

2. Model and Estimation

We consider the following treatment difference model

$$Y_{i1} - Y_{i0} = Q(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) + \epsilon_i, \tag{2.1}$$

where Y_{i1} is the potential outcome for individual i if treatment is received, Y_{i0} is the potential outcome for individual i if no treatment is received, $\mathbf{X}_i \in \mathcal{R}^{d_{\boldsymbol{\beta}}}$ is the set of covariates, the treatment difference function $Q(\cdot)$ is an unknown smooth function and $E(\epsilon \mid \mathbf{X}) = 0$, where ϵ is the model error. Here $\boldsymbol{\beta} \in \mathcal{R}^{d_{\boldsymbol{\beta}}}$ is a vector of unknown parameters and $d_{\boldsymbol{\beta}}$ is the dimension of $\boldsymbol{\beta}$. Let A_i be the treatment indicator. Our estimation is performed under the following two assumptions commonly assumed in the literature.

Assumption 1. (Stable unit treatment value assumption) $Y_i = Y_{i1}A_i + Y_{i0}(1 - A_i)$.

Assumption 2. (No-unmeasured-confounders assumption) $A_i \perp (Y_{i1}, Y_{i0}) \mid \mathbf{X}_i$.

For identifiability of $\boldsymbol{\beta}$, we require $\boldsymbol{\beta}$ to have the form $\boldsymbol{\beta} = (1, \boldsymbol{\beta}_L^{\mathrm{T}})^{\mathrm{T}}$, where the lower sub-vector is an arbitrary vector of length $d_{\boldsymbol{\beta}} - 1$. If Y_{i1} and Y_{i0} had been both available, we could have estimated $\boldsymbol{\beta}$ through simultaneously solving $\sum_{i=1}^{n} \left\{ Y_{i1} - Y_{i0} - \widetilde{Q}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \right\} \left\{ \mathbf{X}_{Li} - E(\mathbf{X}_{Li} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \right\} = \mathbf{0}$ and $\sum_{j=1}^{n} K_h \left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i \right) (Y_{j1} - Y_{j0} - c_i) = \mathbf{0}$ for $i = 1, \ldots, n$. Here $\mathbf{X}_L$ represents the sub-vector of $\mathbf{X}$ formed by its lower $d_{\boldsymbol{\beta}} - 1$ components and $\widetilde{Q}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) = c_i$. Here we use $\widetilde{Q}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)$ instead of c_i in the first equation to emphasize that it is an estimate of the function $Q(\cdot)$ evaluated at $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i$, for $i = 1, \ldots, n$. Note that $K_h(\cdot) = K(\cdot/h)/h$, where $K(\cdot)$ is a kernel function and h is a bandwidth. However, since we only observe Y_i, we may consider an inverse probability weighting based estimator (Robins et al. 1994) and modify the above equations to $\sum_{i=1}^{n} \left[A_i Y_i / \pi(\mathbf{X}_i) - (1 - A_i) Y_i / \{1 - \pi(\mathbf{X}_i)\} - \widetilde{Q}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \right] \times \left\{ \mathbf{X}_{Li} - E(\mathbf{X}_{Li} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \right\} = \mathbf{0}$ and $\sum_{i=1}^{n} K_h \left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j \right) \left[A_i Y_i / \pi(\mathbf{X}_i) - (1 - A_i) Y_i / \{1 - \pi(\mathbf{X}_i)\} - c_j \right] = \mathbf{0}$ for $j = 1, \ldots, n$, where $\pi(\mathbf{X}_i)$ is a known propensity score model. To gain protection against a misspecified $\pi(\mathbf{X}_i)$, we adopt models $\mu(\mathbf{X}_i, \boldsymbol{\alpha}) = E(Y_{i0} \mid \mathbf{X}_i)$ and $\pi(\mathbf{X}_i, \boldsymbol{\gamma}) = P(A_i = 1 \mid \mathbf{X}_i)$, and obtain the estimate of $\boldsymbol{\beta}$ by modifying the above

equations to a double-robust augmented version (Robins et al. 1994)

$$\sum_{i=1}^{n}\left[\frac{\{A_i-\pi(\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\}\{Y_i-\mu(\mathbf{X}_i,\widehat{\boldsymbol{\alpha}})\}}{\pi(\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\{1-\pi(\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\}}+\left\{1-\frac{A_i}{\pi(\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})}\right\}\widetilde{Q}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\right] \tag{2.2}$$
$$\times\left\{\mathbf{X}_{Li}-E(\mathbf{X}_{Li}\mid\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i)\right\} \;=\; \mathbf{0},$$

and

$$\sum_{i=1}^{n}K_h\left(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j\right)\left[\frac{\{A_i-\pi(\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\}\{Y_i-\mu(\mathbf{X}_i,\widehat{\boldsymbol{\alpha}})\}}{\pi(\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\{1-\pi(\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\}}-\frac{A_i}{\pi(\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})}c_j\right]=\mathbf{0}$$

for $j=1,\ldots,n$. Here, the above relation can be equivalently written as

$$\begin{aligned}&\widetilde{Q}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j,\boldsymbol{\beta},\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\gamma}}) \qquad (2.3)\\ =\;&\frac{\sum_{i=1}^{n}K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j)\{A_i-\pi(\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\}\{Y_i-\mu(\mathbf{X}_i,\widehat{\boldsymbol{\alpha}})\}/[\pi(\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\{1-\pi(\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})\}]}{\sum_{i=1}^{n}K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i-\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j)A_i/\pi(\mathbf{X}_i,\widehat{\boldsymbol{\gamma}})}.\end{aligned}$$

The $\boldsymbol{\beta}$ estimator based on (2.2) is double robust with respect to $\pi(\mathbf{X},\boldsymbol{\gamma})$ and $\mu(\mathbf{X},\boldsymbol{\alpha})$. Following the literature, we consider parametric models $\pi(\mathbf{X},\boldsymbol{\gamma})$ and $\mu(\mathbf{X},\boldsymbol{\alpha})$ for simplicity.

Proposition 1. *Under the model in (2.1), as long as one of the two models $\pi(\mathbf{x},\boldsymbol{\gamma})$ and $\mu(\mathbf{x},\boldsymbol{\alpha})$ is correct, then estimator for $\boldsymbol{\beta}$ is consistent. In addition, for the estimation of $\boldsymbol{\beta}$, we can further use a working model for the $Q(\cdot)$ function, which may be different from the true treatment difference*

function if both $\pi(\mathbf{x}, \boldsymbol{\gamma})$ *and* $\mu(\mathbf{x}, \boldsymbol{\alpha})$ *are correctly specified.*

Note that in Proposition 1, we do not require known values of $\boldsymbol{\gamma}$ or $\boldsymbol{\alpha}$. Instead, $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$ are unknown parameters. As long as one of the models in $\pi(\mathbf{x}, \boldsymbol{\gamma})$ and $\mu(\mathbf{x}, \boldsymbol{\alpha})$ is correctly specified, the conclusion of Proposition 1 holds. Note also that $\widehat{\boldsymbol{\gamma}}$ can be obtained based on data $(\mathbf{X}_i, A_i), i = 1, \ldots, n$ via, for example, maximum likelihood estimator (MLE). Likewise, $\mu(\mathbf{X}, \boldsymbol{\alpha}) = E(Y_i \mid \mathbf{X}_i, A_i = 0)$ and hence $\widehat{\boldsymbol{\alpha}}$ can be obtained based on data $(\mathbf{X}_i, Y_i)$ for the i where $A_i = 0$, via, for example, solving generalized estimating equations (GEEs). In solving (2.2) to obtain $\widehat{\boldsymbol{\beta}}$, the choice of bandwidth h is flexible and can be any positive number as long as $n^{-1/2} << h << n^{-1/4}$. However, once we obtain $\widehat{\boldsymbol{\beta}}$, we estimate $Q(\cdot)$ using an optimal bandwidth of order $n^{-1/5}$, which can be obtained from cross-validation. We now describe the algorithm of the estimation procedure in detail.

Algorithm:

Step 1. Obtain the estimate of $\boldsymbol{\gamma}$, $\widehat{\boldsymbol{\gamma}}$ by MLE based on the data $(\mathbf{X}_i, A_i)$, $i = 1, \ldots, n$.

Step 2. Extract the observations with $A_i = 0$. Denote the sub-dataset corresponding to $A_i = 0$ as $(\mathbf{X}_i, Y_i^0), i = 1, \ldots, n_0$. Using the subset of the observations to compute the estimator of $\boldsymbol{\alpha}$, $\widehat{\boldsymbol{\alpha}}$ by solving the

GEEs, $\sum_{i=1}^{n_0} \mathbf{W}(\mathbf{X}_i, \boldsymbol{\alpha})\{Y_i^0 - \mu(\mathbf{X}_i, \boldsymbol{\alpha})\} = \mathbf{0}$, where $\mathbf{W}(\mathbf{X}_i, \boldsymbol{\alpha})$ is an arbitrary $d_{\boldsymbol{\alpha}} \times 1$ matrix of functions of covariates $\mathbf{X}_i$, the parameter $\boldsymbol{\alpha} \in \mathcal{R}^{d_{\boldsymbol{\alpha}}}$ and $d_{\boldsymbol{\alpha}}$ is the dimension of $\boldsymbol{\alpha}$.

Step 3. Plug $\widehat{\boldsymbol{\gamma}}$ and $\widehat{\boldsymbol{\alpha}}$ in (2.3) and obtain $\widetilde{Q}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j, \boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})$.

Step 4. Plug $\widehat{\boldsymbol{\gamma}}$, $\widehat{\boldsymbol{\alpha}}$ and $\widetilde{Q}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i, \boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})$ in (2.2) and solve (2.2) to get $\widehat{\boldsymbol{\beta}}_L$.

Step 5. Select a bandwidth h_{opt}.

Step 6. Obtain $\widehat{Q}(\cdot, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})$ by (2.3), while plugging in $\widehat{\boldsymbol{\gamma}}$, $\widehat{\boldsymbol{\alpha}}$, $\widehat{\boldsymbol{\beta}}$ and h_{opt}.

In step 5, to estimate $Q(\cdot)$, we need a suitable bandwidth. We adopt the leave-one-out cross-validation method to select the bandwidth. Specifically, we estimate $Q(\cdot)$ by

$$\begin{aligned} &\widetilde{Q}_{-j}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_j, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) \\ =\ &\frac{\sum_{i=1, i\neq j}^{n} K_h(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_j)\{A_i - \pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}})\}\{Y_i - \mu(\mathbf{X}_i, \widehat{\boldsymbol{\alpha}})\}/[\pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}})\{1 - \pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}})\}]}{\sum_{i=1, i\neq j}^{n} K_h(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i - \widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_j) A_i/\pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}})}, \end{aligned}$$

where $\widetilde{Q}_{-j}(\cdot)$ denotes the estimator with the j^{th} observation left out. Then we calculate the leave-one-out cross-validated prediction MSE as $\mathrm{CV}(h) = n^{-1}\sum_{i=1}^{n}\big[\{A_i - \pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}})\}\{Y_i - \mu(\mathbf{X}_i, \widehat{\boldsymbol{\alpha}})\}/\pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}})\{1 - \pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}})\}$ $-A_i\widetilde{Q}_{-i}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})/\pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}})\big]^2$ and we choose h to be the minimizer of $\mathrm{CV}(h)$.

Considering that $Q(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ may be a non-monotonic function, we detect all the regions where $Q(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) > 0$ as the treatment region, i.e., we assign treatment 1 if and only if $Q(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) > 0$ to an individual. Obviously, this maximizes the value function, hence leads to the optimal treatment regime. Specifically, the value function $V\{Q(\cdot), \boldsymbol{\beta}\} = E[Y_{i1}I\{Q(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) > 0\} + Y_{i0}I\{Q(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_i) \leqslant 0\}]$ under our identification strategy. Therefore, even if $Q(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ has multiple roots, we can still identify the optimal treatment regimes. We can also observe that $Q(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) > 0$ simplifies to $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X} > 0$, if $Q(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ is monotone. Therefore, our strategy $Q(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}) > 0$ can accommodate both monotone and non-monotone functions. Thus, once we obtain $\widehat{Q}(\cdot, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})$ and $\widehat{\boldsymbol{\beta}}$, we directly identify the optimal treatment regime by assigning treatment 1 if and only if $\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}) > 0$. We can further estimate the subsequent maximum value function as

$$
\begin{aligned}
& \widehat{V}\{\widehat{Q}(\cdot), \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}\} \\
= \; & n^{-1}\sum_{i=1}^{n} \frac{[A_i I\{\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) > 0\} + (1 - A_i) I\{\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) \leqslant 0\}]Y_i}{\pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}}) I\{\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) > 0\} + \{1 - \pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}})\} I\{\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) \leqslant 0\}} \\
& + n^{-1}\sum_{i=1}^{n} \{\pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}}) - A_i\} \left[\frac{\mu(\mathbf{X}_i, \widehat{\boldsymbol{\alpha}}) + \widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})}{\pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}})} I\{\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) > 0\} \right. \\
& \left. - \frac{\mu(\mathbf{X}_i, \widehat{\boldsymbol{\alpha}})}{1 - \pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}})} I\{\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) \leqslant 0\} \right] \\
= \; & n^{-1}\sum_{i=1}^{n} \left(\frac{[A_i + (1 - 2A_i) I\{\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) \leqslant 0\}]Y_i}{\pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}}) + \{1 - 2\pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}})\} I\{\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) \leqslant 0\}} + \{\pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}}) - A_i\} \right. \qquad (2.4) \\
\times \; & \left. \frac{[\mu(\mathbf{X}_i, \widehat{\boldsymbol{\alpha}}) + \widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) - \{2\mu(\mathbf{X}_i, \widehat{\boldsymbol{\alpha}}) + \widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})\} I\{\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) \leqslant 0\}]}{\pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}}) + \{1 - 2\pi(\mathbf{X}_i, \widehat{\boldsymbol{\gamma}})\} I\{\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X}_i, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}) \leqslant 0\}} \right),
\end{aligned}
$$

which is a consistent estimator of the true value function $V\{Q(\cdot), \boldsymbol{\beta}\}$.

3. Theoretical Properties

We now study the theoretical properties of the proposed estimators. For notational simplicity, define $\mathbf{W}(\boldsymbol{\gamma}) \equiv E\left(\partial^2 \log[\pi(\mathbf{X},\boldsymbol{\gamma})^A\{1-\pi(\mathbf{X},\boldsymbol{\gamma})\}^{1-A}]/\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}^{\mathrm{T}}\right)$, $\boldsymbol{\phi}_{\boldsymbol{\gamma}}(\mathbf{X}_i, A_i, \boldsymbol{\gamma}) \equiv \mathbf{W}(\boldsymbol{\gamma})^{-1}\partial \log\left[\pi(\mathbf{x}_i,\boldsymbol{\gamma})^{A_i}\{1-\pi(\mathbf{x}_i,\boldsymbol{\gamma})\}^{1-A_i}\right]/\partial\boldsymbol{\gamma}$, and $\boldsymbol{\phi}_{\boldsymbol{\alpha}}(\mathbf{X}_i, A_i, Y_i, \boldsymbol{\alpha}) \equiv [E\{\mathbf{W}(\mathbf{X},\boldsymbol{\alpha})\mathbf{D}(\mathbf{X},\boldsymbol{\alpha})\}]^{-1}\mathbf{W}(\mathbf{X}_i,\boldsymbol{\alpha})(1-A_i)\{Y_i-\mu(\mathbf{X}_i,\boldsymbol{\alpha})\}$, where $\mathbf{W}(\mathbf{X}_i,\boldsymbol{\alpha})$ is an arbitrary weight matrix and $\mathbf{D}(\mathbf{X},\boldsymbol{\alpha}) = \partial\mu(\mathbf{X},\boldsymbol{\alpha})/\partial\boldsymbol{\alpha}^{\mathrm{T}}$. Through out the paper, $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^{\mathrm{T}}$.

Proposition 2. *Write the conditional distribution of A given $\mathbf{X}$ as $\pi(\mathbf{x},\boldsymbol{\gamma})^a\{1-\pi(\mathbf{x},\boldsymbol{\gamma})\}^{1-a}$. Regardless if $\pi(\mathbf{X},\boldsymbol{\gamma})$ is the true propensity score model or not, there exists $\boldsymbol{\gamma}_0$ so that the MLE, $\widehat{\boldsymbol{\gamma}}$ satisfies $\sqrt{n}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \boldsymbol{\phi}_{\boldsymbol{\gamma}}(\mathbf{X}_i, A_i, \boldsymbol{\gamma}) + o_p(1)$, hence $\sqrt{n}(\widehat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0) \to N\{\mathbf{0}, \mathbf{W}(\boldsymbol{\gamma}_0)^{-1}\mathbf{B}(\boldsymbol{\gamma}_0)\mathbf{W}(\boldsymbol{\gamma}_0)^{-1}\}$ in distribution when $n\to\infty$, where $\mathbf{B}(\boldsymbol{\gamma}_0) \equiv E\left\{\boldsymbol{\phi}_{\boldsymbol{\gamma}}(\mathbf{X}_i, A_i, \boldsymbol{\gamma}_0)^{\otimes 2}\right\}$.*

Specifically, when the $\pi(\mathbf{x},\boldsymbol{\gamma})$ model is correct, $\boldsymbol{\gamma}_0$ is the true parameter value which yields $\pi_0(\mathbf{x}) = \pi(\mathbf{x},\boldsymbol{\gamma}_0)$, and the covariance matrix simplifies to inverse of Fisher's Information matrix $\mathbf{I}(\boldsymbol{\gamma}_0)$, which is $\mathbf{I}(\boldsymbol{\gamma}_0) = -\mathbf{W}(\boldsymbol{\gamma}_0)^{-1}$. When the $\pi(\mathbf{x},\boldsymbol{\gamma})$ model is incorrect, $\boldsymbol{\gamma}_0$ is the parameter vector which minimizes the Kullback-Leibler distance $E\left\{\log\left([\pi_0(\mathbf{X})^A\{1-\pi_0(\mathbf{X})\}^{1-A}]/[\pi(\mathbf{X},\boldsymbol{\gamma})^A\{1-\pi(\mathbf{X},\boldsymbol{\gamma})\}^{1-A}]\right)\right\}$.

Proposition 3. *Whether* $\mu(\mathbf{X}, \boldsymbol{\alpha})$ *is the true model or not, let* $\widehat{\boldsymbol{\alpha}}$ *be estimator which solves the estimating equation* $\sum_{i=1}^{n_0} \mathbf{W}(\mathbf{X}_i, \boldsymbol{\alpha})\{Y_i^0 - \mu(\mathbf{X}_i, \boldsymbol{\alpha})\} = \mathbf{0}$. *Then,* $\sqrt{n_0}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) = n_0^{-1/2} \sum_{i=1}^{n_0} \boldsymbol{\phi}_{\boldsymbol{\alpha}}(\mathbf{X}_i, A_i, Y_i, \boldsymbol{\alpha}_0) + o_p(1) = n_0^{-1/2} \sum_{i=1}^{n} \boldsymbol{\phi}_{\boldsymbol{\alpha}}(\mathbf{X}_i, A_i, Y_i, \boldsymbol{\alpha}_0) + o_p(1)$, *hence* $\sqrt{n_0}(\widehat{\boldsymbol{\alpha}} - \boldsymbol{\alpha}_0) \to N(\mathbf{0}, \mathbf{V}_{\boldsymbol{\alpha}})$ *in distribution when* $n \to \infty$, *where* $\mathbf{V}_{\boldsymbol{\alpha}} = [E\{\mathbf{W}(\mathbf{X}, \boldsymbol{\alpha}_0)\mathbf{D}(\mathbf{X}, \boldsymbol{\alpha}_0)\}]^{-1} \times E\{\mathbf{W}(\mathbf{X}, \boldsymbol{\alpha}_0) v(\mathbf{X}) \mathbf{W}(\mathbf{X}, \boldsymbol{\alpha}_0)^{\mathrm{T}}\}([E\{\mathbf{W}(\mathbf{X}, \boldsymbol{\alpha}_0)\mathbf{D}(\mathbf{X}, \boldsymbol{\alpha}_0)\}]^{-1})^{\mathrm{T}}$, *and* $v(\mathbf{X})$ *is the conditional variance of* Y *given* $\mathbf{X}$. *When the model* $\mu(\mathbf{X}, \boldsymbol{\alpha})$ *is correctly specified,* $\boldsymbol{\alpha}_0$ *satisfies* $\mu_0(\mathbf{X}) = \mu(\mathbf{X}, \boldsymbol{\alpha}_0)$, *where* $\mu_0(\mathbf{x})$ *is the true mean outcome under* $A = 0$, *while when the model* $\mu(\mathbf{X}, \boldsymbol{\alpha})$ *is incorrectly specified,* $\boldsymbol{\alpha}_0$ *satisfies* $E[\mathbf{W}(\mathbf{X}_i, \boldsymbol{\alpha}_0)\{Y_i^0 - \mu(\mathbf{X}_i, \boldsymbol{\alpha}_0)\}] = 0$.

The results in Proposition 2 and 3 are direct consequence from White (1982) and Yi & Reid (2010), hence we omit the detailed proofs. The asymptotic properties for the estimators $\widehat{\boldsymbol{\beta}}$, $\widehat{Q}(\cdot, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})$, the root of $\widehat{Q}(\cdot, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})$, and $\widehat{V}\{\widehat{Q}(\cdot), \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}\}$ are developed under the following conditions.

Regularity Conditions:

(C1) The true parameter value $\boldsymbol{\beta}_0$ belongs to a compact set Ω.

(C2) The univariate kernel $K(\cdot)$ is symmetric, has compact support and is Lipschitz continuous on its support. It satisfies $\int K(u)du = 1$, $\int uK(u)du = 0$, $0 \neq \int u^2 K(u)du < \infty$.

(C3) The probability density function of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, which is denoted by $f(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$, is bounded away from 0 and ∞.

(C4) $E(\mathbf{X} \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})f(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ and $Q(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ are twice differentiable and their second derivatives, as well as $f(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})$ are locally Lipschitz-continuous and bounded.

(C5) The bandwidth $h = O(n^{-\kappa})$ for $1/8 < \kappa < 1/2$.

(C6) The treatment assignment probability satisfies $c < \pi_0(x) < 1 - c$, where c is a small positive constant.

(C7) The true treatment responses, $\mu_0(\mathbf{x})$ for non-treated group and $\mu_1(\mathbf{x})$ for treated group are bounded by a constant C.

(C8) The true treatment effect function $Q(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$ has roots $r_1, \ldots, r_K$, $K < \infty$. In addition, $Q'(r_k) \neq 0$ for all $k = 1, \ldots, K$.

Conditions (C1)–(C5) are standard conditions and they ensure sufficient convergence rate of the nonparametric estimators. Condition (C6) is also routinely assumed to exclude weights near 0 and 1. Condition (C7) is very mild and is usually satisfied in practice. Condition (C8) allows roots for the function $Q(\cdot)$ and is also very mild.

Lemma 1. *Under Conditions* (C2), (C3) *and* (C4), *at any* $\boldsymbol{\beta} \in \Omega$ *and for any function* $H(\mathbf{X}_j, A_j, Y_j)$ *such that* $E\left\{H(\mathbf{X}_j, A_j, Y_j) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j\right\}$ *is twice differentiable, we have*

$$\begin{aligned} & E\left\{H(\mathbf{X}_j, A_j, Y_j)K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\right\} - E\left\{H(\mathbf{X}_j, A_j, Y_j) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right\} f(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) \\ = & \frac{\partial^2}{(\partial\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})^2}[E\left\{H(\mathbf{X}_j, A_j, Y_j) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\right\} f(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})]\frac{h^2}{2}\int z^2K(z)dz + o(h^2), \\ & var\{n^{-1}\sum_{j=1}^{n} H(\mathbf{X}_j, A_j, Y_j)K_h(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}_j - \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\} \\ = & (nh)^{-1}E\{H^2(\mathbf{X}_j, A_j, Y_j) \mid \boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}\}f(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x})\int K^2(z)dz + O(n^{-1}). \end{aligned}$$

Lemma 2. *Assume the regularity Conditions* (C1)-(C5) *hold. Then at any* $\boldsymbol{\beta} \in \Omega$ *the kernel estimator* $\widetilde{Q}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0)$ *satisfies* $\widetilde{Q}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0) - Q(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}) = O_p\{h^2 + (nh)^{-1/2}\}$.

Note that the convergence rate of $\widetilde{Q}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \boldsymbol{\beta}, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0)$ is slower than root-n while $\widehat{\boldsymbol{\gamma}}, \widehat{\boldsymbol{\alpha}}$ have root-n convergence rate. Thus, estimating $\widetilde{Q}(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}, \boldsymbol{\beta}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})$, which is based on $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\boldsymbol{\gamma}}$ instead of based on $\boldsymbol{\alpha}_0$ and $\boldsymbol{\gamma}_0$, does not change the results in Lemma 2.

Theorem 1. *Assume* $\widehat{\boldsymbol{\beta}}_L$ *solves (2.2). Then, under the regularity conditions* (C1)-(C5), $\widehat{\boldsymbol{\beta}}_L$ *satisfies* $\sqrt{n}(\widehat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}_{L0}) \to N\{\mathbf{0}, \mathbf{B}^{-1}\mathbf{V}_1(\mathbf{B}^{-1})^{\mathrm{T}}\}$ *in distribution as* $n \to \infty$, *where* $\mathbf{V}_1 \equiv E\left[\left\{\boldsymbol{\phi}_{\boldsymbol{\beta}}(\mathbf{X}_i, Y_i, A_i, \boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0) + \mathbf{B}_{\boldsymbol{\gamma}}\boldsymbol{\phi}_{\boldsymbol{\gamma}}(\mathbf{X}_i, A_i, \boldsymbol{\gamma}_0) + \mathbf{B}_{\boldsymbol{\alpha}}\boldsymbol{\phi}_{\boldsymbol{\alpha}}(\mathbf{X}_i, A_i, Y_i, \boldsymbol{\alpha}_0)\right\}^{\otimes 2}\right]$. *The detailed expression for*

$\mathbf{B}$, $\mathbf{B}_{\boldsymbol{\gamma}}$, $\mathbf{B}_{\boldsymbol{\alpha}}$ *and* $\boldsymbol{\phi}_{\boldsymbol{\beta}}(\mathbf{X}_i, Y_i, A_i, \boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0)$ *are provided in Section* **??** *of the Supplementary material.*

The first term in the variance expression $\mathbf{V}_1$ captures the variability of estimating different functions, the second term captures the variability in estimating $\boldsymbol{\beta}$ due to the estimation of $\boldsymbol{\gamma}$ and the third term captures the same induced by $\widehat{\boldsymbol{\alpha}}$.

Lemma 3. *Assume the regularity Conditions* (C1)-(C5) *hold. Then the kernel estimator obtained from Step 6,* $\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})$ *satisfies*

$$\begin{aligned} & bias\{\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})\} \\ = \; & h_{\mathrm{opt}}^2 \left\{ \frac{Q'(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) d[E\{\pi_0(\mathbf{X}_j)/\pi(\mathbf{X}_j, \boldsymbol{\gamma}_0) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\} f(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})]/d(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})}{f(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) E\{\pi_0(\mathbf{X}_j)/\pi(\mathbf{X}_j, \boldsymbol{\gamma}_0) \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\}} + \frac{Q''(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})}{2} \right\} \\ & \times \int z^2 K(z) dz + o(h_{\mathrm{opt}}^2 + n^{-1/2} h_{\mathrm{opt}}^{-1/2}), \end{aligned}$$

and

$$\begin{aligned} var\{\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}, \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}})\} \; = \; & \frac{1}{n h_{\mathrm{opt}}} \Bigg(E\left[\frac{\pi_0(\mathbf{X}_j)}{\pi^2(\mathbf{X}_j, \boldsymbol{\gamma}_0)} \{Y_{1j} - \mu(\mathbf{X}_j, \boldsymbol{\alpha}_0)\}^2 \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x} \right] \\ & + E\left[\frac{1 - \pi_0(\mathbf{X}_j)}{\{1 - \pi(\mathbf{X}_j, \boldsymbol{\gamma}_0)\}^2} \{Y_{0j} - \mu(\mathbf{X}_j, \boldsymbol{\alpha}_0)\}^2 \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x} \right] \\ & - Q^2(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) E\left\{ \frac{\pi_0(\mathbf{X}_j)}{\pi^2(\mathbf{X}_j, \boldsymbol{\gamma}_0)} \mid \boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x} \right\} \end{aligned}$$

$$-2Q(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})E\left[\frac{\pi_0(\mathbf{X}_j)}{\pi^2(\mathbf{X}_j,\boldsymbol{\gamma}_0)}\{\mu_0(\mathbf{X}_j)-\mu(\mathbf{X}_j,\boldsymbol{\alpha}_0)\}\mid\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right]\Bigg)$$
$$\times\frac{1}{f(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})}\left[E\left\{\frac{\pi_0(\mathbf{X}_j)}{\pi(\mathbf{X}_j,\boldsymbol{\gamma}_0)}\mid\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}\right\}\right]^{-2}\int K^2(z)dz+O(n^{-1}).$$

Here, for a generic function $r(\cdot)$, $r'(\cdot)$ and $r''(\cdot)$ are respectively its first and second derivatives.

Theorem 2. *Let $Q(z_0)=0$ and $\widehat{Q}(\widehat{z},\widehat{\boldsymbol{\beta}},\widehat{\boldsymbol{\alpha}},\widehat{\boldsymbol{\gamma}})=0$. Then as $n\to\infty$, under the regularity Conditions* (C1)-(C5)*, $\widehat{z}\to z_0$ at the rate $n^{-2/5}$. Specifically, the leading term of the bias of $\widehat{z}$ is*

$$-h_{\mathrm{opt}}^2\left\{\frac{d[E\{\pi_0(\mathbf{X})/\pi(\mathbf{X},\boldsymbol{\gamma}_0)\mid\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}=z_0\}f(z_0)]/d(z_0)}{E\{\pi_0(\mathbf{X})/\pi(\mathbf{X},\boldsymbol{\gamma}_0)\mid\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}=z_0\}f(z_0)}+\frac{Q''(z_0)}{2Q'(z_0)}\right\}\int z^2K(z)dz,$$

and the leading term of the variance of $\widehat{z}$ is

$$\frac{1}{nh_{\mathrm{opt}}}\Bigg(E\left[\frac{\pi_0(\mathbf{X})}{\pi^2(\mathbf{X},\boldsymbol{\gamma}_0)}\{Y_1-\mu(\mathbf{X},\boldsymbol{\alpha}_0)\}^2\mid\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}=z_0\right]$$
$$+E\left[\frac{1-\pi_0(\mathbf{X})}{\{1-\pi(\mathbf{X},\boldsymbol{\gamma}_0)\}^2}\{Y_0-\mu(\mathbf{X},\boldsymbol{\alpha}_0)\}^2\mid\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}=z_0\right]\Bigg)$$
$$\times\frac{1}{f(z_0)Q'(z_0)^2}\left[E\left\{\frac{\pi_0(\mathbf{X})}{\pi(\mathbf{X},\boldsymbol{\gamma}_0)}\mid\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}=z_0\right\}\right]^{-2}\int K^2(z)dz.$$

From Theorem 2, we can observe that our treatment region identification rate is $O_p(n^{-2/5})$, better than the classical rate $O_p(n^{-1/3})$ (Fan et al. 2017). This is due to the smoothness assumption made in Condition (C4).

Theorem 3. *Under the regularity Conditions* (C1)-(C8)*, the optimal value*

function estimator given in (2.4) satisfies $n^{1/2}[\widehat{V}\{\widehat{Q}(\cdot), \widehat{\boldsymbol{\beta}}, \widehat{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\gamma}}\} - V\{Q(\cdot), \boldsymbol{\beta}_0\}] \to N(0, \sigma^2)$ *in distribution when* $n \to \infty$, *where* $\sigma^2 = E\big[-\mathbf{U}_{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{B}^{-1}$ $\{\boldsymbol{\phi}_{\boldsymbol{\beta}}(\mathbf{X}_i, Y_i, A_i, \boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0) + \mathbf{B}_{\boldsymbol{\gamma}}\boldsymbol{\phi}_{\boldsymbol{\gamma}}(\mathbf{X}_i, A_i, \boldsymbol{\gamma}_0) + \mathbf{B}_{\boldsymbol{\alpha}}\boldsymbol{\phi}_{\boldsymbol{\alpha}}(\mathbf{X}_i, A_i, Y_i, \boldsymbol{\alpha}_0)\}$ $+\mathbf{U}_{\boldsymbol{\alpha}}^{\mathrm{T}}\boldsymbol{\phi}_{\boldsymbol{\alpha}}(\mathbf{X}_i, A_i, Y_i, \boldsymbol{\alpha}_0) + \mathbf{U}_{\boldsymbol{\gamma}}^{\mathrm{T}}\boldsymbol{\phi}_{\boldsymbol{\gamma}}(\mathbf{X}_i, A_i, \boldsymbol{\gamma}_0) + v_Q\{\mathbf{X}_i, A_i, Y_i, \boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, Q(\cdot)\}$ $+v_0(\mathbf{X}_i, A_i, Y_i)\big]^2$. *The detailed expression for* $\mathbf{B}$, $\mathbf{B}_{\boldsymbol{\gamma}}$, $\mathbf{B}_{\boldsymbol{\alpha}}$, $\mathbf{U}_{\boldsymbol{\alpha}}$, $\mathbf{U}_{\boldsymbol{\gamma}}$, $\boldsymbol{\phi}_{\boldsymbol{\beta}}(\mathbf{X}_i, Y_i, A_i, \boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0)$, $v_Q\{\mathbf{X}_i, A_i, Y_i, \boldsymbol{\beta}_0, \boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0, Q(\cdot)\}$ *and* $v_0(\mathbf{X}_i, A_i, Y_i)$ *are provided in Section* **??** *of the Supplementary material.*

We can understand the first term in σ^2 as the variability in value function due to $\boldsymbol{\beta}$. The second term is related to the variability induced by $\boldsymbol{\alpha}$. The third term captures the variability due to the $\boldsymbol{\gamma}$ estimation. The fourth term measures the variability in the value function induced by estimating the treatment effect function. Lastly, the fifth term captures the variability in the value function inherited from the variability of the covariates.

4. Simulations

We conducted simulation studies to compare the performance of the estimators discussed in Section 2. We considered different scenarios, depending misspecification of either $\pi(\mathbf{X}, \boldsymbol{\gamma})$ or $\mu(\mathbf{X}, \boldsymbol{\alpha})$ to show the robustness property of our estimators. We used sample size $n = 500$ with 1000 replicates.

4.1 Simulation 1

Our first simulation follows similar designs as those in Fan et al. (2017), where the monotonicity of $Q(\cdot)$ is required. We set $d_{\boldsymbol{\beta}} = 4$ and generated the covariate vector $\mathbf{X}_i$ from the multivariate normal distribution with zero mean and identity covariance matrix. We generated the treatment indicator A_i from a Bernoulli distribution with probability $\pi_0(\mathbf{X}_i) = 0.5$. The response variables are formed from $Y_i = \mu_0(\mathbf{X}_i) + A_i Q_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{X}_i) + \epsilon_i$, where ϵ_i is generated from a centered normal distribution with variance 0.25. Here, $Q_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) = 2\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}$ and $\mu_0(\mathbf{x}) = 1 + \boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}$, where $\boldsymbol{\alpha}_0 = (1, -1, 1, 1)^{\mathrm{T}}$ and $\boldsymbol{\beta}_0 = (1, 1, -1, 1)^{\mathrm{T}}$.

To illustrate the robustness of our method, we considered four cases in estimation. In Case I, we used the constant treatment probability model and a linear model for $\mu(\mathbf{x}, \boldsymbol{\alpha})$ in the implementation. Note that these two models are both correctly specified. In Case II, we used a constant model for $\mu(\mathbf{x}, \boldsymbol{\alpha})$, which is a misspecified model, while keeping the treatment probability π unchanged. In Case III, we fixed π at 0.4 and used the same μ model as in Case I. Thus, μ is correctly specified, whereas π is misspecified. Lastly, in Case IV, both models are misspecified by using the same model for μ as in Case II and setting π as in Case III.

We followed the algorithm described in Section 2, where we used the

Epanechnikov kernel in the nonparametric implementation, and used the bandwidth $c\sigma n^{-1/3}$ to estimate $\boldsymbol{\beta}$, where σ^2 is the estimated variance of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{x}$ and c is a constant between 7 and 7.5 in step 3.

From the results summarized in Table 1, we observe that in the first three cases, our estimation for $\boldsymbol{\beta}$ yields small bias. In contrast, for Case IV, when both $\pi(\mathbf{x})$ and $\mu(\mathbf{x}, \boldsymbol{\alpha})$ are misspecified, the estimation for $\boldsymbol{\beta}$ is biased. In terms of inference, the estimated standard deviations based on the asymptotic properties match closely with the empirical variability of the estimators and the 95% confidence intervals have coverage close to the nominal level in the first three cases. Interestingly, the value function estimator and the root of $Q(\cdot)$ perform well in all cases. From Figure 1, we can also see that the 95% confidence interval for $Q(\cdot)$ includes the true function $Q_0(\cdot)$.

4.2 Simulation 2

In our second simulation study, we examine the performance of our estimators in the presence of non-monotonic $Q_0(\cdot)$ function and heteroscedastic error variance. In generating data, the true treatment difference function is $Q_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) = (\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)^2 - 2$, $\mu_0(\mathbf{x}) = 1 + \sin(\boldsymbol{\alpha}_{10}^{\mathrm{T}}\mathbf{x}) + 0.5(\boldsymbol{\alpha}_{20}^{\mathrm{T}}\mathbf{x})^2$, and the errors satisfy $\epsilon_i \sim \mathcal{N}(0, \log\{(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)^2 + 1\})$. Here $\boldsymbol{\beta}_0 = (1, 1, -1, 1)^{\mathrm{T}}$,

Table 1: Simulation 1. $Q_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) = 2\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}$ and $\mu_0(\mathbf{x}) = 1+\boldsymbol{\alpha}_0^{\mathrm{T}}\mathbf{x}$. Case I: $\mu(\cdot)$ and $\pi(\cdot)$ are correctly-specified, Case II: $\mu(\cdot)$ is misspecified, Case III: $\pi(\cdot)$ is misspecified, Case IV: both models are misspecified. For the different cases, we also compute the mean of the estimated sd based on asymptotics ($\hat{\text{sd}}$), empirical coverage obtained with 95% confidence intervals based on these estimated sd (cvg), and mean squared error (mse).

Results for $\boldsymbol{\beta}$ and value function V							
Case	parameters	True	Estimate	sd	$\hat{\text{sd}}$	cvg	MSE
I	β_2	1	0.9960	0.0567	0.0563	95.1%	0.0032
	β_3	-1	-0.9953	0.0563	0.0559	95.3%	0.0032
	β_4	1	0.9941	0.0559	0.0549	94.4%	0.0032
	V	2.5958	2.5955	0.1453	0.1458	97.2%	0.0211
II	β_2	1	1.0256	0.1598	0.2033	93.4%	0.0262
	β_3	-1	-1.0252	0.1561	0.1982	93.7%	0.0250
	β_4	1	1.0068	0.1361	0.1913	94.6%	0.0186
	V	2.5958	2.6083	0.1926	0.1702	96.2%	0.0373
III	β_2	1	1.0049	0.0468	0.0470	95.3%	0.0022
	β_3	-1	-1.0044	0.0470	0.0473	95.2%	0.0022
	β_4	1	1.0032	0.0482	0.0464	94.7%	0.0023
	V	2.5958	2.6176	0.1476	0.1471	96.9%	0.0223
IV	β_2	1	0.7494	0.1106	0.0940	41.3%	0.0751
	β_3	-1	-0.7485	0.1073	0.0919	42.1%	0.0748
	β_4	1	1.0243	0.0901	0.0939	95.8%	0.0087
	V	2.5958	2.6401	0.1665	0.1655	96.4%	0.0297

Results for the root of $Q_0(t) = 2t$								
Case	true	mean	bias	$\hat{\text{bias}}$	sd	$\hat{\text{sd}}$	cvg	MSE
I	0	0.0023	0.0023	-0.0013	0.0641	0.0663	96.6%	0.0041
II	0	-0.0004	-0.0004	-0.0001	0.1703	0.1693	93.7%	0.0290
III	0	0.0015	0.0015	-0.0004	0.0589	0.0594	95.7%	0.0035
IV	0	-0.0080	-0.0080	0.0044	0.1242	0.1225	93.2%	0.0155

$\boldsymbol{\alpha}_{10} = (1, -1, 1, 1)^{\mathrm{T}}$ and $\boldsymbol{\alpha}_{20} = (1, 0, -1, 0)^{\mathrm{T}}$. All other aspects of the simulation setting are identical to those in Simulation 1.

Despite the heteroscedasticity and non-monotone treatment difference function, similar to Simulation 1, we considered four cases to demonstrate the robustness property of our estimator. In Case I, we used correctly

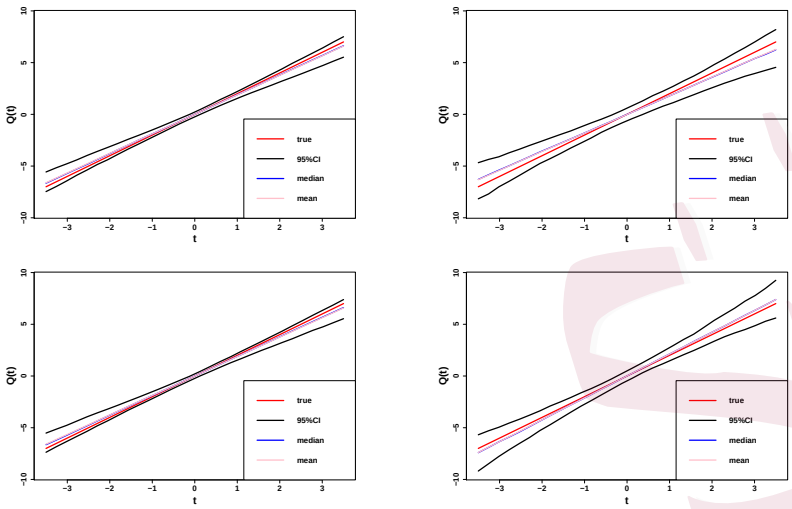

Figure 1: Simulation 1. Mean, median and 95% confidence band of the estimators of $Q_0(t) = 2t$, when (I) $\mu(\cdot)$ and $\pi(\cdot)$ are both correct (top-left), (II) $\mu(\cdot)$ is misspecified and $\pi(\cdot)$ is correct (top-right), (III) $\pi(\cdot)$ is misspecified and $\mu(\cdot)$ is correct (bottom-left), (IV) both $\mu(\cdot)$ and $\pi(\cdot)$ are misspecified (bottom-right).

specified models for both π and $\mu(\mathbf{x}, \boldsymbol{\alpha})$. In Case II, we misspecify the $\mu(\mathbf{x}, \boldsymbol{\alpha})$ model to a linear one. In Case III, we misspecified π to 0.4. Finally, in Case IV, we used the misspecified models for both $\mu(\mathbf{x}, \boldsymbol{\alpha})$ and π.

We used the same nonparametric estimation procedures as in Simulation 1 to implement the algorithm in Section 2. From the results summarized in Table 2, we observe that despite of both a non-monotonic $Q_0(\cdot)$ function and a heteroscedastic error variance, the estimation for parameters $\boldsymbol{\beta}$, the value function and the two roots of $Q_0(\cdot)$ yields very small bias in the first three cases. Also, the estimated standard deviations are close

to the empirical standard deviations and the confidence intervals are close to nominal coverage levels. As expected, the estimation of $\boldsymbol{\beta}$ in Case IV does not perform well, although the performance of the value function and the two roots of $Q(\cdot)$ show certain robustness even in Case IV. In Figure 2, we note that the 95% confidence interval for $Q(\cdot)$ includes the true $Q_0(\cdot)$ function in all four cases.

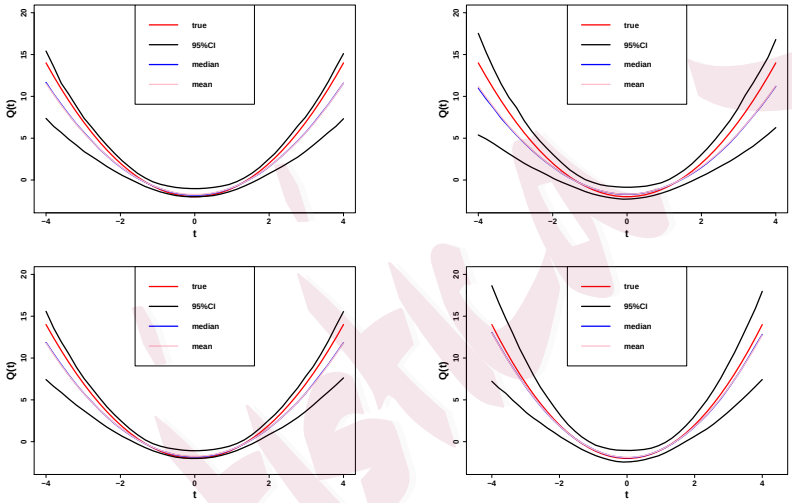

Figure 2: Simulation 2. Mean, median and 95% confidence band of the estimators of $Q_0(t) = t^2 - 2$ with heteroscedastic error variance. See also the caption of Figure 1.

4.3 Simulation 3

In the previous simulation settings, we only considered the situation where the true propensity score is a constant. We now consider non-constant propensity score case, which better reflects the situation in observational

Table 2: Simulation 2. $Q_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) = (\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})^2 - 2$ and $\mu_0(\mathbf{x}) = 1 + \sin(\boldsymbol{\alpha}_{10}^{\mathrm{T}}\mathbf{x}) + 0.5(\boldsymbol{\alpha}_{20}^{\mathrm{T}}\mathbf{x})^2$, when error variance is heteroscedastic. See also the caption of Table 1.

Results for $\boldsymbol{\beta}$ and value function V.							
Case	parameters	True	Estimate	sd	$\hat{\text{sd}}$	cvg	MSE
I	β_2	1	1.0364	0.0646	0.0911	93.5%	0.0055
	β_3	-1	-1.0368	0.0653	0.0911	93.1%	0.0056
	β_4	1	1.0330	0.0646	0.0893	92.9%	0.0053
	V	4.7166	4.6415	0.2843	0.2970	94.9%	0.0864
II	β_2	1	1.0398	0.1124	0.1384	92.8%	0.0142
	β_3	-1	-1.0341	0.1062	0.1320	94.0%	0.0124
	β_4	1	1.0411	0.1166	0.1385	93.2%	0.0153
	V	4.7166	4.6160	0.3095	0.3053	94.4%	0.1059
III	β_2	1	1.0295	0.0618	0.0833	92.6%	0.0047
	β_3	-1	-1.0300	0.0640	0.0830	93.6%	0.0050
	β_4	1	1.0262	0.0629	0.0826	94.2%	0.0046
	V	4.7166	4.7024	0.2967	0.3052	95.9%	0.0882
IV	β_2	1	0.9549	0.0894	0.0990	86.4%	0.0100
	β_3	-1	-1.0283	0.0865	0.1020	92.0%	0.0083
	β_4	1	0.9563	0.0910	0.1007	89.4%	0.0102
	V	4.7166	4.7518	0.3122	0.3174	96.1%	0.0987

Results for the two roots of $Q_0(t) = t^2 - 2$								
Case	true	mean	bias	$\hat{\text{bias}}$	sd	$\hat{\text{sd}}$	cvg	MSE
I	-1.4142	-1.4469	-0.0327	-0.0109	0.1498	0.1120	93.2%	0.0235
	1.4142	1.4480	0.0338	0.0047	0.1259	0.1108	93.7%	0.0170
II	-1.4142	-1.4482	-0.0340	-0.0144	0.2096	0.1943	94.7%	0.0451
	1.4142	1.4408	0.0266	0.1191	0.1854	0.1830	95.2%	0.0351
III	-1.4142	-1.4423	-0.0281	-0.0170	0.1104	0.1033	93.1%	0.0130
	1.4142	1.4436	0.0294	0.0113	0.1112	0.1019	93.3%	0.0132
IV	-1.4142	-1.4087	0.0056	-0.0180	0.1585	0.1745	97.4%	0.0252
	1.4142	1.4363	0.0221	-0.0145	0.1497	0.1662	97.5%	0.0229

studies. Specifically, we let $\pi(\mathbf{X}_i) = \exp(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{X}_i)/\{1 + \exp(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{X}_i)\}$, where $\boldsymbol{\gamma}_0 = (0.1, 0, -0.1, 0)^{\mathrm{T}}$. Further, we consider $Q_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) = (\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i) + \sin(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}_i)$ and keep other data generation identical to Simulation 2.

To show the robustness of our method, we consider four cases similar to Simulation 2. In Case I, we used correctly specified models for both $\pi(\mathbf{x}, \boldsymbol{\gamma})$

and $\mu(\mathbf{x}, \boldsymbol{\alpha})$. In Case II, we misspecify the $\mu(\mathbf{x}, \boldsymbol{\alpha})$ model to a linear one, while keeping the treatment probability model $\pi(\mathbf{X}, \boldsymbol{\gamma})$ unchanged. In Case III, we use a constant model for $\pi(\mathbf{x}, \boldsymbol{\gamma})$, which is misspecified and used the same model for $\mu(\mathbf{X}, \boldsymbol{\alpha})$ as in Case I. Finally, in Case IV, both models are misspecified by using the same model for $\mu(\mathbf{X}, \boldsymbol{\alpha})$ as in Case II and considering $\pi(\mathbf{X}, \boldsymbol{\gamma})$ as in Case III.

We followed the algorithm described in Section 2 and summarized the results in Table 3. From the Table 3, we observe that despite the heteroscedastic error and non-constant propensity score model, in the first three cases, the estimations for parameters $\boldsymbol{\beta}$, the value function, V and the root of the treatment difference function yield small bias. In addition, the estimated standard deviations are still close to the empirical version of the estimators, and the confidence intervals have coverage close to the nominal levels. Interestingly, the estimator of the root of $Q(\cdot)$ performs well even in Case IV. In Figure 3, we note that the 95% confidence interval for $Q(\cdot)$ includes the true $Q_0(\cdot)$ function in all four cases.

4.4 Simulation 4

Here, we consider a non-constant propensity score model similar to Simulation 3. Other data generation procedure is identical to Simulation 2. Thus,

Table 3: Simulation 3. $Q_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) = (\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) + \sin(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})$, $\pi_0(\mathbf{x}) = \exp(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{X})/\{1+\exp(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{X})\}$ and $\mu_0(\mathbf{x}) = 1 + \sin(\boldsymbol{\alpha}_{10}^{\mathrm{T}}\mathbf{x}) + 0.5(\boldsymbol{\alpha}_{20}^{\mathrm{T}}\mathbf{x})^2$. See also the caption of Table 1.

Results for $\boldsymbol{\beta}$ and value function V.							
Case	parameters	True	Estimate	sd	$\hat{\text{sd}}$	cvg	MSE
I	β_2	1	1.0100	0.1197	0.1603	93.2%	0.0144
	β_3	-1	-1.0093	0.1177	0.1625	93.5%	0.0139
	β_4	1	1.0120	0.1227	0.1640	93.3%	0.0152
	V	3.0533	3.0345	0.1254	0.1287	95.7%	0.0161
II	β_2	1	1.0418	0.1742	0.1896	93.9%	0.0321
	β_3	-1	-0.9983	0.1696	0.2087	93.0%	0.0287
	β_4	1	1.0467	0.1749	0.1866	93.1%	0.0327
	V	3.0533	2.9907	0.1157	0.1359	95.5%	0.0173
III	β_2	1	1.0220	0.1278	0.1757	92.4%	0.0168
	β_3	-1	-1.0170	0.1243	0.1755	93.2%	0.0157
	β_4	1	1.0168	0.1276	0.1765	93.7%	0.0165
	V	3.0533	3.0440	0.1440	0.1203	95.0%	0.0208
IV	β_2	1	1.0512	0.2019	0.1727	80.5%	0.0434
	β_3	-1	-1.0083	0.1976	0.1929	82.7%	0.0391
	β_4	1	1.0575	0.2094	0.1837	82.6%	0.0472
	V	3.0533	2.9647	0.1402	0.1406	90.6%	0.0275

Results for the root of $Q_0(t) = t + \sin(t)$								
Case	true	mean	bias	$\hat{\text{bias}}$	sd	$\hat{\text{sd}}$	cvg	MSE
I	0	0.0271	0.0271	0.0147	0.0795	0.0767	93.8%	0.0070
II	0	0.0105	0.0105	0.0179	0.1564	0.1669	97.2%	0.0245
III	0	0.0078	0.0078	-0.0013	0.0769	0.0823	96.2%	0.0059
IV	0	0.0089	0.0089	0.0186	0.1726	0.1650	97.3%	0.0299

we consider a non-constant propensity score model with a non-monotonic treatment difference function and heteroscedastic error variance in this simulation.

Similar to Simulation 3, we consider four cases and implemented our estimation procedures to illustrate the robustness property of estimators. From the results summarized in Table 4, we observe that the estimations

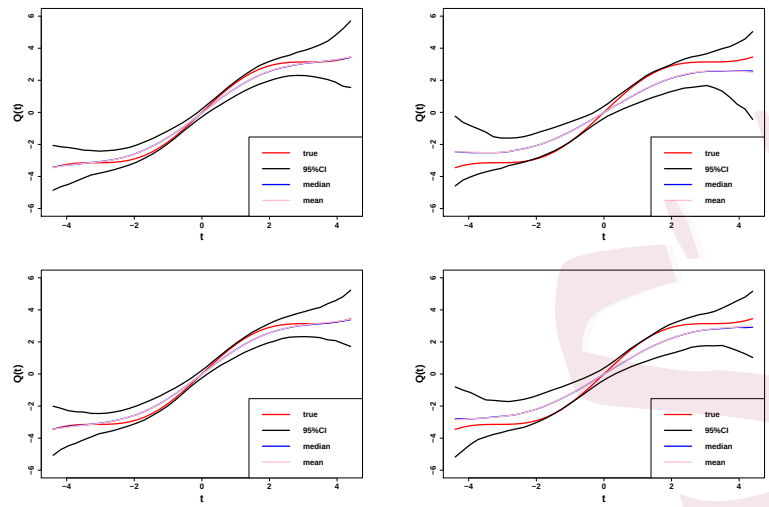

Figure 3: Simulation 3. Mean, median and 95% confidence band of the estimators of $Q_0(t) = t + \sin(t)$ with non-constant propensity score model and heteroscedastic error variance. See also the caption of Figure 1.

for parameters $\boldsymbol{\beta}$, the value function, V, and the root of the treatment difference function result in small bias in the first three cases, as expected. Also, the estimated standard deviations are close to the empirical standard deviations and the confidence intervals are close to nominal coverage levels. Interestingly, the estimation and inference of $\boldsymbol{\beta}$, V, and the two roots perform well in Case IV. In Figure 4, we note that the 95% confidence interval for $Q(\cdot)$ includes the true $Q_0(\cdot)$ function in all four cases.

For comparison, we implemented the methods in Fan et al. (2017) for all simulations. These results are summarized in Tables **??** to **??** in the Supplementary Material. As we can see, when the monotonicity assumption

is violated, Fan et al. (2017) deteriorates and performs worse than our proposed method. We also provided additional simulation studies in the Supplementary Material. Further, we implemented two machine learning methods (Zhang et al. 2015, Zhao et al. 2012) and reported the results in Tables **??** and **??** in the Supplementary Material.

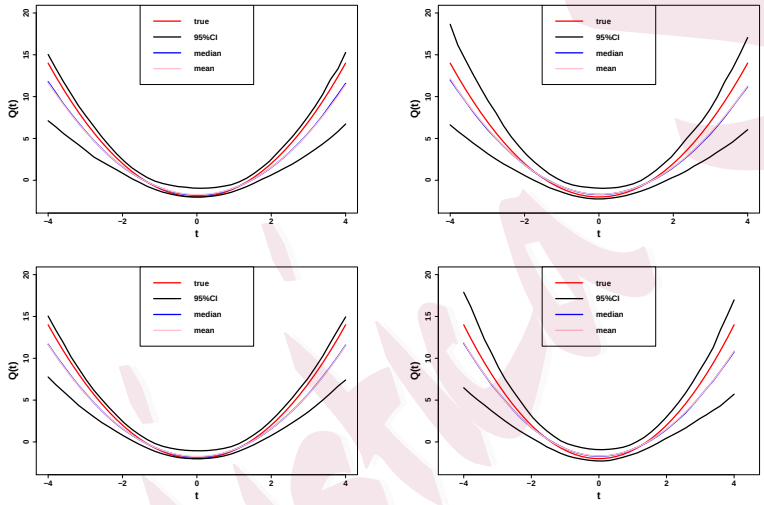

Figure 4: Simulation 4. Mean, median and 95% confidence band of the estimators of $Q_0(t) = t^2 - 2$ with non-constant propensity score model and heteroscedastic error variance. See also the caption of Figure 1.

5. Real Data Application

We now apply our proposed method to a study of the effect of smoking during pregnancy on baby's birth weight. The primary outcomes is birth weight (in grams) of singleton births in Pennsylvania, USA (Almond et al.

Table 4: Simulation 4. $Q_0(\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x}) = (\boldsymbol{\beta}_0^{\mathrm{T}}\mathbf{x})^2 - 2$, $\pi_0(\mathbf{x}) = \exp(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{X})/\{1 + \exp(\boldsymbol{\gamma}_0^{\mathrm{T}}\mathbf{X})\}$ and $\mu_0(\mathbf{x}) = 1 + \sin(\boldsymbol{\alpha}_{10}^{\mathrm{T}}\mathbf{x}) + 0.5(\boldsymbol{\alpha}_{20}^{\mathrm{T}}\mathbf{x})^2$. See also the caption of Table 1.

Results for $\boldsymbol{\beta}$ and value function V.							
Case	parameters	True	Estimate	sd	$\hat{\text{sd}}$	cvg	MSE
I	β_2	1	1.0316	0.0659	0.0870	92.7%	0.0053
	β_3	-1	-1.0346	0.0644	0.0890	92.8%	0.0053
	β_4	1	1.0346	0.0635	0.0874	93.9%	0.0052
	V	4.7166	4.6549	0.3035	0.3294	96.3%	0.0959
II	β_2	1	1.0236	0.1117	0.1377	93.4%	0.0130
	β_3	-1	-1.0186	0.1111	0.1362	93.6%	0.0127
	β_4	1	1.0173	0.1143	0.1349	92.9%	0.0133
	V	4.7166	4.6134	0.2952	0.3269	96.3%	0.0977
III	β_2	1	1.0361	0.0637	0.0889	93.0%	0.0054
	β_3	-1	-1.0343	0.0637	0.0895	95.0%	0.0052
	β_4	1	1.0327	0.0642	0.0877	94.0%	0.0052
	V	4.7166	4.6381	0.2930	0.3020	95.8%	0.0920
IV	β_2	1	1.0396	0.1219	0.1368	93.5%	0.0164
	β_3	-1	-1.0245	0.1103	0.1321	94.0%	0.0128
	β_4	1	1.0343	0.1215	0.1371	92.3%	0.0159
	V	4.7166	4.6004	0.2905	0.3056	94.2%	0.0979

Results for the two roots of $Q_0(t) = t^2 - 2$								
Case	true	mean	bias	$\hat{\text{bias}}$	sd	$\hat{\text{sd}}$	cvg	MSE
I	-1.4142	-1.4268	-0.0126	-0.0199	0.1201	0.1085	95.3%	0.0146
	1.4142	1.4556	0.0414	0.0238	0.1115	0.1172	94.8%	0.0142
II	-1.4142	-1.3810	0.0331	0.0368	0.1811	0.1885	94.4%	0.0339
	1.4142	1.4553	0.0411	0.0086	0.1792	0.1835	93.2%	0.0338
III	-1.4142	-1.4410	-0.0268	-0.0159	0.1221	0.1169	93.2%	0.0156
	1.4142	1.4554	0.0412	0.0137	0.1160	0.1206	93.6%	0.0151
IV	-1.4142	-1.3941	0.0201	-0.0065	0.1905	0.1950	94.9%	0.0367
	1.4142	1.4508	0.0366	-0.0080	0.1769	0.1885	95.9%	0.0326

2005). This study aims to determine whether pregnant women should quit smoking to ensure healthy birth in terms of baby's birth-weight. For illustration purpose, we consider a subset of 1394 unmarried mothers. The data set contains the maternal smoking habit during pregnancy which is treated

as treatment A_i (1 =Non-smoking, 0 =Smoking). The covariates observed are mother's age (mage), an indicator variable for alcohol consumption during pregnancy (alcohol), an indicator variable of previous birth in which the infant died (deadkids), mother's education (medu), father's education (fedu), number of prenatal care visits (nprenatal), months since last birth (monthslb), mother's race (mrace) and an indicator variable of first born child (fbaby).

To estimate the propensity score, mean outcome model for non-treated group and treatment difference function, first we normalize all the continuous covariates. We use the expit model $\pi(\mathbf{X}, \boldsymbol{\gamma})$ to describe the propensity score, and use MLE to estimate $\boldsymbol{\gamma}$. In addition, we consider a linear model for the mean outcome model for the non-treatment group $\mu(\mathbf{X}, \boldsymbol{\alpha})$, and solved GEE to obtain $\widehat{\boldsymbol{\alpha}}$. Lastly, we estimate the treatment difference model $Q(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X})$ using the proposed method.

In implementing the algorithm described in Section 2, we used the quartic kernel in the nonparametric implementation to estimate $\boldsymbol{\beta}$ with bandwidth $c\sigma n^{-1/3}$, where σ^2 is the estimated variance of $\boldsymbol{\beta}^{\mathrm{T}}\mathbf{X}$, and $c = 0.05$.

For identifiability purpose, we fix the coefficient of the first covariate (here, mage) to be 1 and estimate the remaining eight coefficients. The estimated parameters in $\boldsymbol{\beta}$, their standard errors, and the value function

are summarized in Table 5. From the 95% confidence interval for $\boldsymbol{\beta}$, we can conclude that all covariates are significant, except the indicator variable for the alcohol consumption. We provide the estimated treatment difference model, $\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X})$ in Figure 5. Here the covariate alcohol is included in estimating $\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X})$. It shows a higher baby birth weight for those mothers who did not smoke during their pregnancy, once $\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X})$ is greater than 0. We further construct 95% confidence band for the difference function $Q(t)$ based on 500 bootstrap samples by resampling the residuals. The results obtained from CAL and CAL-DR method Fan et al. (2017) are summarized in Table 6. It can be seen that no significant covariate is detected by CAL or CAL-DR. Also the variability in estimating the value function using CAL or CAL-DR is higher than that from the proposed method. Further, we observe that the 95% confidence intervals computed by CAL and CAL-DR methods include the estimated value function obtained by our method.

Remark 1. We further performed analysis after excluding the covariate alcohol and observed that the estimated function $\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{X})$ does not change much. However, excluding alcohol influences the estimated confidence band. This suggests that more careful consideration on variable selection and on how to perform inference post variable selection is needed, which is beyond the scope of our work. This will be an interesting future research topic.

Table 5: Birth-weight study analysis: Results for $\boldsymbol{\beta}$ and value function V with 95% CI.

parameters	Estimate	$\widehat{\text{sd}}$	Confidence interval
β_2 (alcohol)	0.2965	0.4842	(-0.6525,1.2455)
β_3 (deadkids)	0.3406	0.0233	(0.2950,0.3862)
β_4 (medu)	-0.1972	0.0073	(-0.2116,-0.1828)
β_5 (fedu)	-0.0947	0.0005	(-0.0957,-0.0938)
β_6 (nprenatal)	0.2822	0.0061	(0.2703,0.2941)
β_7 (monthslb)	0.0183	0.0002	(0.0178,0.0188)
β_8 (mrace)	3.0882	0.3753	(2.3527,3.8237)
β_9 (fbaby)	-1.8505	0.4267	(-2.6868,-1.0143)
Value function	3274.9	25.439	(3225.1,3324.8)

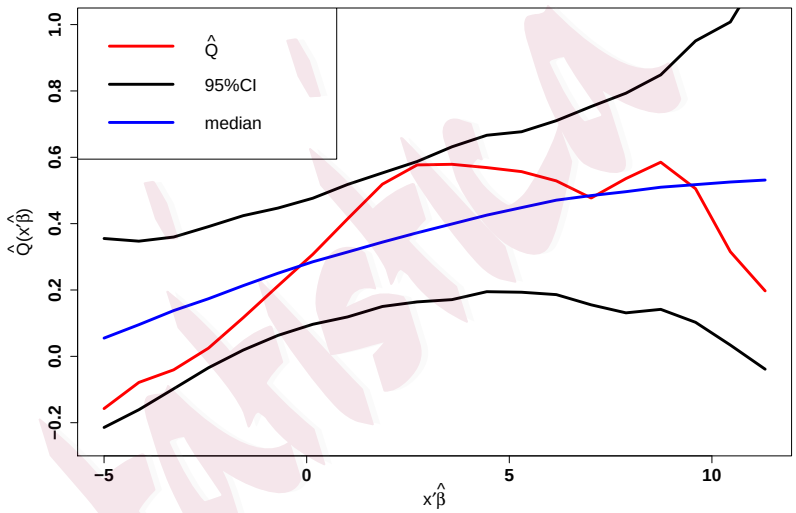

Figure 5: Data analysis. The estimated treatment difference model, $\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x})$, its median, 95% confidence bands based on low baby birth weight data set.

Table 6: Application to birth-weight study using CAL and CAL-DR methods.

parameters	Estimate	$\widehat{\text{sd}}$	Confidence interval
CAL estimates for $\boldsymbol{\beta}$ and value function V for the real data analysis.			
β_2 (alcohol)	-0.0223	7.0207	(-13.783,13.738)
β_3 (deadkids)	0.2088	5.9428	(-11.439,11.857)
β_4 (medu)	-0.1998	0.6850	(-1.5424,1.1428)
β_5 (fedu)	-0.0902	0.2929	(-0.6643,0.4838)
β_6 (nprenatal)	0.2966	0.5499	(-0.7812,1.3743)
β_7 (monthslb)	0.0192	0.1653	(-0.3048,0.3431)
β_8 (mrace)	3.1975	7.1149	(-10.747,17.142)
β_9 (fbaby)	-2.0682	3.8728	(-9.6587,5.5222)
Value function	3244.4	30.345	(3185.0,3303.9)
CAL-DR estimates for $\boldsymbol{\beta}$ and value function V for the real data analysis.			
β_2 (alcohol)	-0.1111	1.7382	(-3.5179,3.2958)
β_3 (deadkids)	0.2589	0.9554	(-1.6136,2.1314)
β_4 (medu)	-0.2058	0.3147	(-0.8226,0.4109)
β_5 (fedu)	-0.0815	0.1545	(-0.3843,0.2213)
β_6 (nprenatal)	0.2759	0.2782	(-0.2693,0.8212)
β_7 (monthslb)	0.0183	0.0383	(-0.0568,0.0933)
β_8 (mrace)	3.2234	2.4674	(-1.6127,8.0594)
β_9 (fbaby)	-2.0459	1.7858	(-5.5460,1.4542)
Value function	3241.9	30.243	(3182.6,3301.2)

6. Discussion

We proposed a new multi-robust estimation to estimate the optimal treatment regimes for a single decision time point under weak conditions, i.e., our treatment difference model $Q(\cdot)$ does not need to be monotonic and we only require $E(\epsilon \mid \mathbf{X}) = 0$. Our method enjoys the protection against misspecification of either the propensity score model or the outcome regression model for the non-treated group or the non-monotonic treatment difference model. We use nonparametric kernel-based estimator to obtain

the treatment difference model and show that the treatment identification rate is $O_p(n^{-2/5})$. The extensive simulation studies illustrate superior performance under various scenarios.

Regardless the true treatment difference function $Q(\cdot)$ has single or multiple roots, our procedure always identifies the region $\{\mathbf{x} : \widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}) > 0\}$ as the treatment region. When $\widehat{Q}(\cdot)$ has multiple root, the corresponding treatment region is then the union of several intervals for $\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}$. In practice, this does not cause any issue, since when a new patient enters with covariate $\mathbf{x}_0$, we simply evaluate $\widehat{Q}(\widehat{\boldsymbol{\beta}}^{\mathrm{T}}\mathbf{x}_0)$ to determine whether treatment should be given. In this paper, we consider parametric models for the propensity score function and the mean outcome model for the non-treated group. Alternatively, one can use the semiparametric or nonparametric method to obtain these two functions. For example, one can use the semiparametric estimation procedure in Ma & Zhu (2013) to estimate propensity score and the mean outcome model for the non-treated group to obtain $n^{1/2}$-consistent estimators for $\boldsymbol{\gamma}$ and $\boldsymbol{\alpha}$. The treatment identification rate will remain unchanged.

Many extensions following this work can be interesting and worth pursuing. For example, we may consider multiple treatment decision times while incorporating the usual backward induction to obtain the optimal dynamic

treatment regimes. We may also consider multiple treatment choices sharing the same index. Further studies along these lines can be challenging and fruitful.

Supplementary Material

The online Supplementary Material contains proofs for Proposition 1, Lemma 1–3, and Theorems 1–3.

Acknowledgements

The research is supported by grants from NSF, NIH, National Natural Science Foundation of China (No. 12171077), and the National Institute of Neurological Disorders and Stroke (No. NS073671). The authors would also like to thank the associate editor and reviewers for their helpful comments and suggestions.

References

Almond, D., Chay, K. Y. & Lee, D. S. (2005), 'The costs of low birth weight', *The Quarterly Journal of Economics* **120**, 1031–1083.

Bai, X., Tsiatis, A. A., Lu, W. & Song, R. (2017), 'Optimal treatment regimes for survival end-

points using a locally-efficient doubly-robust estimator from a classification perspective', *Lifetime data analysis* **23**(4), 585–604.

Blatt, D., Murphy, S. A. & Zhu, J. (2004), 'A-learning for approximate planning', *Ann Arbor* **1001**, 48109–2122.

Chakraborty, B., Murphy, S. & Strecher, V. (2010), 'Inference for non-regular parameters in optimal dynamic treatment regimes', *Statistical methods in medical research* **19**(3), 317–343.

Fan, C., Lu, W., Song, R. & Zhou, Y. (2017), 'Concordance-assisted learning for estimating optimal individualized treatment regimes', *Journal of the royal statistical society, Series B.* **79**, 1565–1582.

Foster, J. C., Taylor, J. M. & Ruberg, S. J. (2011), 'Subgroup identification from randomized clinical trial data', *Statistics in medicine* **30**(24), 2867–2880.

Goldberg, Y. & Kosorok, M. R. (2012), 'Q-learning with censored data', *Annals of statistics* **40**(1), 529–560.

Huang, M.-Y. & Yang, S. (2020), 'Robust inference of conditional average treatment effects using dimension reduction', *arXiv preprint arXiv:2008.13137* .

Jiang, R., Lu, W., Song, R. & Davidian, M. (2017), 'On estimation of optimal treatment regimes for maximizing t-year survival probability', *Journal of the Royal Statistical Society. Series B, Statistical methodology* **79**(4), 1165–1185.

REFERENCES

Jiang, R., Lu, W., Song, R., Hudgens, M. G. & Naprvavnik, S. (2017), 'Doubly robust estimation of optimal treatment regimes for survival data—with application to an hiv/aids study', *The annals of applied statistics* **11**(3), 1763–1786.

Liang, S., Lu, W. & Song, R. (2018), 'Deep advantage learning for optimal dynamic treatment regime', *Statistical theory and related fields* **2**(1), 80–88.

Ma, Y. & Zhu, L. (2013), 'Efficient estimation in sufficient dimension reduction', *The Annals of Statistics* **41**, 250–268.

Matsouaka, R. A., Li, J. & Cai, T. (2014), 'Evaluating marker-guided treatment selection strategies', *Biometrics* **70**(3), 489–499.

Murphy, S. A. (2003), 'Optimal dynamic treatment regimes', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **65**(2), 331–355.

Murphy, S. A. (2005), 'A generalization error for q-learning', *Journal of Machine Learning Research* **6**(Jul), 1073–1097.

Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W. E., Gnagy, B., Fabiano, G. A., Waxmonsky, J. G., Yu, J. & Murphy, S. A. (2012), 'Q-learning: A data analysis method for constructing adaptive interventions.', *Psychological methods* **17**(4), 478–494.

Orellana, L., Rotnitzky, A. & Robins, J. M. (2010), 'Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, part i: main content', *The international journal of biostatistics* **6**(2), Article 8.

Qian, M. & Murphy, S. A. (2011), 'Performance guarantees for individualized treatment rules', *Annals of statistics* **39**(2), 1180–1210.

Robins, J. M. (2004), 'Optimal structural nested models for optimal sequential decisions', *Proceedings of the second seattle Symposium in Biostatistics* pp. 189–326.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994), 'Estimation of regression coefficients when some regressors are not always observed', *Journal of the American Statistical Association* **89**, 846–866.

Shi, C., Song, R., Lu, W. & Fu, B. (2018), 'Maximin projection learning for optimal treatment decision with heterogeneous individualized treatment effects', *Journal of the Royal Statistical Society. Series B, Statistical methodology* **80**(4), 681–702.

Song, R., Luo, S., Zeng, D., Zhang, H. H., Lu, W. & Li, Z. (2017), 'Semiparametric single-index model for estimating optimal individualized treatment strategy', *Electron. J. Statist.* **11**(1), 364–384.

Song, R., Wang, W., Zeng, D. & Kosorok, M. R. (2015), 'Penalized q-learning for dynamic treatment regimens', *Statistica Sinica* **25**(3), 901–920.

Watkins, C. J. C. H. (1989), 'Learning from delayed rewards', *Ph.D. thesis, University of Cambridge, England* .

Watkins, C. J. & Dayan, P. (1992), 'Q-learning', *Machine learning* **8**(3-4), 279–292.

White, H. (1982), 'Maximum likelihood estimation of misspecified models', *Econometrica*

50(1), 1–25.

Yi, G. Y. & Reid, N. (2010), 'A note on mis-specified estimating functions', *Statistica Sinica* pp. 1749–1769.

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M. & Laber, E. (2012), 'Estimating optimal treatment regimes from a classification perspective', *Stat* **1**(1), 103–114.

Zhang, B., Tsiatis, A. A., Laber, E. B. & Davidian, M. (2012), 'A robust method for estimating optimal treatment regimes', *Biometrics* **68**(4), 1010–1018.

Zhang, B., Tsiatis, A. A., Laber, E. B. & Davidian, M. (2013), 'Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions', *Biometrika* **100**(3), 681–694.

Zhang, Y., Laber, E. B., Tsiatis, A. & Davidian, M. (2015), 'Using decision lists to construct interpretable and parsimonious treatment regimes', *Biometrics* **71**(4), 895–904.

Zhao, L., Tian, L., Cai, T., Claggett, B. & Wei, L.-J. (2013), 'Effectively selecting a target population for a future comparative study', *Journal of the American Statistical Association* **108**(502), 527–539.

Zhao, Y., Kosorok, M. R. & Zeng, D. (2009), 'Reinforcement learning design for cancer clinical trials', *Statistics in medicine* **28**(26), 3294–3315.

Zhao, Y., Zeng, D., Rush, A. J. & Kosorok, M. R. (2012), 'Estimating individualized treatment rules using outcome weighted learning', *Journal of the American Statistical Association*

107(499), 1106–1118.

Zhao, Y., Zeng, D., Socinski, M. A. & Kosorok, M. R. (2011), 'Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer', *Biometrics* **67**(4), 1422–1433.

University of Wisconsin-Madison

E-mail: (tghosh3@wisc.edu)

Pennsylvania State University

E-mail: (yzm63@psu.edu)

Northeast Normal University

E-mail: (wszhu@nenu.edu.cn)

Columbia University

E-mail: (yw2016@cumc.columbia.edu)