Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

A Semiparametric Approach to Model Effect Modification

Muxuan Liang & Menggang Yu

To cite this article: Muxuan Liang & Menggang Yu (2022) A Semiparametric Approach to Model Effect Modification, Journal of the American Statistical Association, 117:538, 752-764, DOI: 10.1080/01621459.2020.1811099

To link to this article: https://doi.org/10.1080/01621459.2020.1811099

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 117, NO. 538, 752–764: Theory and Methods
https://doi.org/10.1080/01621459.2020.1811099

A Semiparametric Approach to Model Effect Modification

Muxuan Liang[a] and Menggang Yu[b]

[a]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA; [b]Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, Madison, WI

ABSTRACT

One fundamental statistical question for research areas such as precision medicine and health disparity is about discovering effect modification of treatment or exposure by observed covariates. We propose a semiparametric framework for identifying such effect modification. Instead of using the traditional outcome models, we directly posit semiparametric models on contrasts, or expected differences of the outcome under different treatment choices or exposures. Through semiparametric estimation theory, all valid estimating equations, including the efficient scores, are derived. Besides doubly robust loss functions, our approach also enables dimension reduction in presence of many covariates. The asymptotic and non-asymptotic properties of the proposed methods are explored via a unified statistical and algorithmic analysis. Comparison with existing methods in both simulation and real data analysis demonstrates the superiority of our estimators especially for an efficiency improved version. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received August 2018
Accepted August 2020

KEYWORDS
Dimension reduction; Interaction; Precision medicine; Semiparametric efficiency; Tangent space

1. Introduction

In many scientific investigations, estimation of the effect modification is a major goal. For example, in precision medicine research, recommending an appropriate treatment among many existing choices is a central question. Based on patient's characteristics, such recommendation amounts to estimating treatment effect modification (Kraemer 2013). Another example is health disparity research that focuses on measuring modification of the association between disparity categories (e.g., race and socioeconomic status) and health outcomes. The estimated effect modification can be used to improve the health system (Braveman 2006).

In the classical regression modeling framework, this amounts to estimating interactions between covariates and a certain interested variable. Take the precision medicine example, the goal is to find how the patient characteristics interact with the treatment indicator. If the interest focuses on treatment recommendation, then main effects of these characteristics do not directly contribute to it because they are the same for all treatment choices. Similarly for the health disparity example, the goal is to find how the modifiers interact with the disparity categories. If the interest focuses on elimination of disparity, then main effects of modifiers are of less importance because they are the same for all disparity categories.

Traditionally, effect modification or statistical interaction discovery is conducted mainly by testing or estimating product terms in outcome models. Such discovery is hard as it usually requires large sample sizes (Greenland 1993), especially when many covariates are present. Recent works in the area of precision medicine illustrate that when the goal is treatment recommendation, investigation on the product term in an outcome model may not be ideal as the outcome is also affected by covariate main effects (Zhang et al. 2012; Zhao et al. 2012; Lu, Zhang, and Zeng 2013; Tian et al. 2014; Xu et al. 2015; Chen et al. 2017). As we have discussed above, these main effects usually are not directly related to treatment recommendation. Therefore, these works focus on learning contrast functions which are differences of conditional expectations of the outcome under two treatment choices. Nonetheless, there is a lack of the literature on how the main effects or estimation of the main effects can contribute to the efficiency of learning such contrast functions.

Most of the existing works use either nonparametric (Zhao et al. 2012; Zhang et al. 2012) or parametric approaches (Kraemer 2013; Lu, Zhang, and Zeng 2013; Xu et al. 2015). The nonparametric approaches are flexible but may not be ideal when faced with a large number of covariates. The parametric approaches on the other hand can be sensitive to the underlying model assumptions. Song et al. (2017) considered a single index model for the contrast function to fill an important middle ground. Single index models are semiparametric models where the index is formed from a linear combination of covariates and a wrapper function that takes the index as argument is nonparametric. However, only an intuitive method of estimation was considered in Song et al. (2017). No systematic investigation was given to explore other possible estimating equations. Therefore, issues such as efficiency were left largely untackled.

CONTACT Menggang Yu *menggang.yu@gmail.com* Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison, 600 Highland Ave, Madison, WI 53792.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/JASA*.

More importantly, it is practical to provide more flexibility in the semiparametric framework by allowing more than one indices. That is multiple index models can better capture the heterogeneity in effect modification. As a simple example, a single index model with the linear index part depending on the product of two covariates is not a single index model any more, if this product is not included as a fitting covariate. However, multiple index models can easily capture this deviation from the linearity. When there are more than two treatments, it is also mathematically appealing to consider multiple index models. For example, single index models can be used to model the contrasts between treatments A and B and between B and C, respectively. But if the indices of these two models are different, the resulting contrast between A and C will be a double index model, not a single index model. This asymmetry, of assuming two single index models for two contrasts and one double index model for the other contrast, is easily avoided by assuming the multiple index models for all contrasts.

We therefore propose a more general semiparametric approach which is essentially a multiple index modeling framework for multiple treatments. We will also consider determination of the number of indices. Under our framework, we make the following new contributions. First, based on the well-established semiparametric estimation theory (Bickel et al. 1993; Tsiatis 2006), we characterize all valid estimating equations, including the efficient score under our framework. This leads to many possible choices of estimating equations, and efficiency consideration becomes very natural in our approach. Second, because multiple index models are intrinsically related to dimension reduction (Xia et al. 2002; Xia 2007), our method can also be used as a dimension reduction tool for interaction discovery with a specific variable. Third, we do not restrict the treatment or exposure variable to be binary. Literature for more than two treatment choices seem very sparse (Lou, Shao, and Yu 2018). Fourth, we also study the asymptotic and non-asymptotic properties of the resulting estimators based on a careful analysis of the computing algorithm. This enables inference and provides useful insights for using our approach in practice.

Estimating the effect modification is an important problem in causal inference (Imai and Ratkovic 2013; Abrevaya, Hsu, and Lieli 2015). Under the potential outcome framework (Rubin 1974, 2005), and the well-known assumptions of the Stable Unit Treatment Value Assumption (SUTVA), consistency, and treatment assignment ignorability (Imbens and Rubin 2015), the effect modification becomes the conditional average treatment effect (CATE). Under these assumptions, popular methods such as inverse probability weighting (IPW) and augmented inverse probability weighting (AIPW) (Robins, Rotnitzky, and Zhao 1994; Bang and Robins 2005; Cao, Tsiatis, and Davidian 2009; Tan 2010; Rotnitzky et al. 2012) were commonly used to estimate average treatment effect (ATE) (Hirano and Imbens 2001; Hirano, Imbens, and Ridder 2003) and the CATE (Imai and Ratkovic 2013; Abrevaya, Hsu, and Lieli 2015). On estimating the CATE, many literature also chose to directly work with outcome models (Green and Kern 2012; Xie, Brand, and Jann 2012; Lu et al. 2018; Wager and Athey 2018; Künzel et al. 2019). The well-known structural nested models and the corresponding G-estimation focused on parametric models for the CATE with relatively few covariates (Robins, Mark, and Newey 1992; Robins 1994; Vansteelandt and Joffe 2014). We posit a multiple index model on the contrast function or the CATE and show how the main effects contribute to the efficiency. Our proposed approach in some way extends these results on the CATE in a semiparametric modeling framework. In some literature (Huang and Chan 2017; Luo, Zhu, and Ghosh 2017; Persson et al. 2017), the effect modification appears to be used also as an important middle step to estimate the population level causal quantities such as the ATE. However, the methods proposed in these literature, including index models or dimension reduction, are for the outcomes, not for the contrast functions.

2. A Semiparametric Framework for Modeling Contrast Functions

Suppose $\boldsymbol{X} \in \mathcal{X}$ is a p-dimensional vector of covariates, Y is an outcome, and T is a discrete variable whose effect on Y and modification of this effect by $\boldsymbol{X}$ are of interest. We first consider the case when T has only two levels. We can also use $\{1, 2\}$, instead of $\{-1, 1\}$, to denote the levels of T and to conform with our notation below for the more general case. However, we keep $\{-1, 1\}$ as it leads to simpler nations in our presentation for the binary treatment setting.

The main goal is to learn the following contrast function based on observed data,

$$\Delta(\boldsymbol{X}) \equiv E[Y|T = 1, \boldsymbol{X}] - E[Y|T = -1, \boldsymbol{X}]. \tag{1}$$

We assume that a larger Y is better. Then when $\Delta(\boldsymbol{X}) > 0, T = 1$ rather than $T = -1$ leads to a better clinical outcome for given $\boldsymbol{X}$, and vice versa. Therefore, we consider the following model in this article

$$\Delta(\boldsymbol{X}) = g(\boldsymbol{B}_0^\top \boldsymbol{X}), \tag{2}$$

where g is an unknown function and $\boldsymbol{B}_0$ is a $p \times d$ matrix. Here d represents the number of indices. That is, $d = 1$ corresponds to a single index model and $d > 1$ to a multiple index model.

Note that there is an identifiability issue in Model (2) when both g and $\boldsymbol{B}_0$ are unrestricted. This is a known issue in both the index models and dimension reduction literature (Xia et al. 2002; Xia and Hardle 2006; Cook 2007; Xia 2007, 2008; Ma and Zhu 2012, 2013; Li 2018). To resolve this issue, we adopt the common strategy in the dimension reduction literature (Cook 2007; Ma and Zhu 2012; Li 2018) and assume that the columns of $\boldsymbol{B}_0$ form a Grassmann manifold. That is, $\boldsymbol{B}_0$ satisfies

$$\left(\boldsymbol{I}_{d\times d}, \boldsymbol{0}_{d\times(p-d)}\right) \boldsymbol{B}_0 = \boldsymbol{I}_{d\times d},$$

where $\boldsymbol{I}_{d\times d}$ is the identity matrix with rank d.

Model (2) is very flexible as the contrast function is defined in terms of the conditional means of the outcome, instead of its conditional distributions. The model is therefore semiparametric as it leaves the other parts of the distribution (e.g., variance) unspecified. This is similar to the well-known semiparametric conditional mean model commonly used in econometrics (Chamberlain 1987; Newey 2004).

Consequently, the outcome Y can be of many types as long as its mean function satisfies our model. For example, when Y is binary, the contrast function represents the difference of the success probabilities. Then Model (2) implies a single or multiple index model, depending on $d = 1$ or $d > 1$, for the difference of its success probabilities under the two treatment choices.

Now consider the case when T has K levels. To fully represent the effect modification, we need to use $K-1$ contrasts. For example when $K = 3$, we can use contrasts such as $E[Y|T = 1, \boldsymbol{X}] - E[Y|T = 2, \boldsymbol{X}]$ and $E[Y|T = 3, \boldsymbol{X}] - \frac{1}{2}(E[Y|T = 1, \boldsymbol{X}] + E[Y|T = 2, \boldsymbol{X}])$. In general, we extend the concept of the contrast function in (1) to a contrast vector function of length $K-1$ as follows

$$\boldsymbol{\Delta}(\boldsymbol{X}) \equiv \boldsymbol{\Omega} \begin{pmatrix} E[Y|T = 1, \boldsymbol{X}] \\ \vdots \\ E[Y|T = K, \boldsymbol{X}] \end{pmatrix}, \tag{3}$$

where $\boldsymbol{\Omega}$ is a prespecified $(K-1) \times K$ matrix. The $K-1$ rows of $\boldsymbol{\Omega}$ represent the interested contrasts. For $K = 2$, $\boldsymbol{\Omega} = (1, -1)$. For the above example of $K = 3$, we have

$$\boldsymbol{\Omega} = \begin{pmatrix} 1 & -1 & 0 \\ -\frac{1}{2} & -\frac{1}{2} & 1 \end{pmatrix}.$$

For the contrasts to be interpretable, we require the sum of ith row of $\boldsymbol{\Omega}$ to be 0, that is, $\sum_{j=1}^{K} \boldsymbol{\Omega}_{ij} = 0$ for $i = 1, \ldots, K-1$. Reasonably, we also require $\boldsymbol{\Omega}\boldsymbol{\Omega}^\top$ to be invertible.

In this setup, the corresponding model is

$$\boldsymbol{\Delta}(\boldsymbol{X}) = \boldsymbol{g}(\boldsymbol{B}_0^\top \boldsymbol{X}), \tag{4}$$

where $\boldsymbol{g}$ is a length $(K-1)$ vector function of $\boldsymbol{B}_0^\top \boldsymbol{X}$.

3. Tangent Spaces and Semiparametric Efficient Scores

Similar to the work in dimension reduction (Ma and Zhu 2012, 2013, 2014), we characterize the nuisance tangent space and its orthogonal complement for $\boldsymbol{B}_0$. The corresponding efficient score is also derived. We closely follow the notions and techniques of Tsiatis (2006). The derivation requires working with the full data likelihood even though we do not specify the form of the distribution of Y in Model (2) or (4). In other words, we need to convert these models into equivalent outcome models that involve $\boldsymbol{B}_0$, g, and the unspecified nonparametric parts.

In our supplementary materials, we show that Model (2) is equivalent to the following model for the outcome Y:

$$Y = \frac{1}{2} T g(\boldsymbol{B}_0^\top \boldsymbol{X}) + \epsilon, \tag{5}$$

where ϵ is some random variable satisfying the following conditional mean condition

$$E[\epsilon|T, \boldsymbol{X}] = E[\epsilon|\boldsymbol{X}]. \tag{6}$$

The equivalence can be shown by verifying that $\epsilon \equiv Y - \frac{1}{2} T g(\boldsymbol{B}_0^\top \boldsymbol{X})$ satisfies (6). This representation (5) enables us to directly work with the full data likelihood.

Similar to the binary setting, when T is multilevel, our model (4) is equivalent to the following model for the outcome Y:

$$Y = \boldsymbol{\Omega}_{\cdot T}^\top \left(\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\right)^{-1} \boldsymbol{g}(\boldsymbol{B}_0^\top \boldsymbol{X}) + \epsilon, \tag{7}$$

where $\boldsymbol{\Omega}_{\cdot T}$ is the column of $\boldsymbol{\Omega}$ that corresponds to the value of the treatment T. Similarly ϵ in (7) needs to satisfy the condition (6).

We first present results for the general multilevel T and assume that the function class of interest is the mean zero Hilbert space $\mathcal{H} = \{f(\epsilon, \boldsymbol{X}, T) : E(f) = 0\}$. These results will then be simplified for binary treatments.

The full data likelihood is

$$p_{\boldsymbol{X}}(\boldsymbol{X}) \pi_T(\boldsymbol{X}) p_\epsilon \left(Y - \boldsymbol{\Omega}_{\cdot T}^\top \left(\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\right)^{-1} \boldsymbol{g}(\boldsymbol{B}_0^\top \boldsymbol{X}), \boldsymbol{X}, T \right),$$

where $p_{\boldsymbol{X}}$ is the density of $\boldsymbol{X}$, $\pi_T(\boldsymbol{X})$ is the density of T conditional on $\boldsymbol{X}$, and p_ϵ is the density of ϵ conditional on $\boldsymbol{X}$ and T, with respect to some dominating measure. The density $\pi_T(\boldsymbol{X})$ is also known as propensity score (Rosenbaum and Rubin 1983). Note that $p_{\boldsymbol{X}}$, π_T, p_ϵ, and $\boldsymbol{g}$ are infinite-dimensional nuisance parameters. The tangent spaces correspond to $p_{\boldsymbol{X}}, p_\epsilon$, and π_T are

$$\begin{aligned} \Lambda_{\boldsymbol{X}} &= \{f(\boldsymbol{X}) \in \mathcal{H} : E[f] = 0\}, \\ \Lambda_\epsilon &= \Big\{ f(\epsilon, \boldsymbol{X}, T) \in \mathcal{H} : E(f|\boldsymbol{X}, T) = 0 \text{ and} \\ &\qquad E\big[f\epsilon|T, \boldsymbol{X}\big] = E\big[f\epsilon|\boldsymbol{X}\big] \Big\}, \\ \Lambda_\pi &= \{f(\boldsymbol{X}, T) \in \mathcal{H} : E[f \mid \boldsymbol{X}] = 0\}. \end{aligned}$$

Through some algebra, we can rewrite Λ_π as

$$\Lambda_\pi = \left\{ \boldsymbol{w}_T^\top \left(\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\right)^{-1} \boldsymbol{f}_\pi(\boldsymbol{X}), \forall \boldsymbol{f}_\pi(\boldsymbol{X}) : \mathcal{X} \mapsto R^{K-1} \right\},$$

where

$$\boldsymbol{w}_T = \frac{\boldsymbol{\Omega}_{\cdot T}}{\pi_T(\boldsymbol{X})}.$$

The tangent space of $\boldsymbol{g}$ is

$$\begin{aligned} \Lambda_{\boldsymbol{g}} = \Big\{ & \frac{p'_{\epsilon,1}(\epsilon, \boldsymbol{X}, T)}{p_\epsilon(\epsilon, \boldsymbol{X}, T)} \boldsymbol{\Omega}_{\cdot T}^\top \left(\boldsymbol{\Omega}\boldsymbol{\Omega}^\top\right)^{-1} \\ & \boldsymbol{f}_{\boldsymbol{g}}(\boldsymbol{B}_0^\top \boldsymbol{X}), \forall \boldsymbol{f}_{\boldsymbol{g}}(\boldsymbol{B}_0^\top \boldsymbol{X}) : \mathcal{X} \mapsto R^{K-1} \Big\}, \end{aligned}$$

where $p'_{\epsilon,1}(\cdot)$ is the derivative of $p_\epsilon(\epsilon, \boldsymbol{X}, T)$ w.r.t ϵ.

Let $\perp$ denote the orthogonal complement of a Hilbert space. Denote the nuisance tangent space $\Lambda \equiv \Lambda_{\boldsymbol{X}} + \Lambda_\epsilon + \Lambda_\pi + \Lambda_{\boldsymbol{g}}$. Then we have

Theorem 3.1. The orthogonal complement of the nuisance tangent space, $\Lambda^\perp$, is a subspace characterized by all functions with the form

$$\boldsymbol{w}_T^\top [\epsilon - E(\epsilon|\boldsymbol{X})] \left[\boldsymbol{\alpha}(\boldsymbol{X}) - E\{\boldsymbol{\alpha}(\boldsymbol{X})|\boldsymbol{B}_0^\top \boldsymbol{X}\} \right],$$

for any function $\boldsymbol{\alpha}(\boldsymbol{X}) : \mathcal{X} \mapsto R^{K-1}$.

Detailed proofs of this theorem and other theorems and corollaries are given in the supplementary materials. To obtain the efficient score, we need to project the score function onto $\Lambda^{\perp}$. The following theorem provides a formula to project any function onto $\Lambda^{\perp}$ and thus contains the efficient score as a special case.

Theorem 3.2. For any function $f(\epsilon, \boldsymbol{X}, T) \in \mathcal{H}$, its projection onto $\Lambda^{\perp}$ is given by

$$\boldsymbol{w}_T^{\top}\{\epsilon - E(\epsilon|\boldsymbol{X})\}\boldsymbol{C}(\boldsymbol{B}_0^{\top}\boldsymbol{X}),$$

where

$$\begin{aligned}\boldsymbol{C}(\boldsymbol{B}_0^{\top}\boldsymbol{X}) &= \boldsymbol{V}(\boldsymbol{X})\{\boldsymbol{D}(\boldsymbol{X}) - E[\boldsymbol{V}(\boldsymbol{X})|\boldsymbol{B}_0^{\top}\boldsymbol{X}]^{-1}E[\boldsymbol{V}(\boldsymbol{X})\boldsymbol{D}(\boldsymbol{X})|\boldsymbol{B}_0^{\top}\boldsymbol{X}]\},\\ \boldsymbol{V}(\boldsymbol{X})^{-1} &= E(\boldsymbol{w}_T\boldsymbol{w}_T^{\top}\epsilon^2|\boldsymbol{X}) - E(\boldsymbol{w}_T\boldsymbol{w}_T^{\top}|\boldsymbol{X})E(\epsilon|\boldsymbol{X})^2,\\ \boldsymbol{D}(\boldsymbol{X}) &= E(\boldsymbol{w}_T f\epsilon|\boldsymbol{X}) - E(\boldsymbol{w}_T f|\boldsymbol{X})E(\epsilon|\boldsymbol{X}).\end{aligned}$$

Note that $\boldsymbol{C}(\boldsymbol{B}_0^{\top}\boldsymbol{X})$ depends on $\boldsymbol{X}$, in addition to $\boldsymbol{B}_0^{\top}\boldsymbol{X}$. But we have suppressed it for notational simplicity. After setting f as the score function in Theorem 3.2, we obtain the efficient score in the following corollary.

Corollary 3.1. The efficient score of $\boldsymbol{B}_0$ is given by the vectorization of a $d \times p$ matrix whose (i,j) coordinate is given by

$$\boldsymbol{w}_T^{\top}\{\epsilon - E(\epsilon|\boldsymbol{X})\}\boldsymbol{C}^*_{i,j}(\boldsymbol{B}_0^{\top}\boldsymbol{X}),$$

where

$$\begin{aligned}\boldsymbol{C}^*_{i,j}(\boldsymbol{B}_0^{\top}\boldsymbol{X}) =& \boldsymbol{V}(\boldsymbol{X})\left\{X_j - E[\boldsymbol{V}(\boldsymbol{X})|\boldsymbol{B}_0^{\top}\boldsymbol{X}]^{-1}E[\boldsymbol{V}(\boldsymbol{X})X_j \,|\, \boldsymbol{B}_0^{\top}\boldsymbol{X}]\right\}\\ &\times \partial_i\boldsymbol{g}(\boldsymbol{B}_0^{\top}\boldsymbol{X}),\end{aligned}$$

X_j is the jth component of $\boldsymbol{X}$, and $\partial_i\boldsymbol{g}$ is the derivative of $\boldsymbol{g}$ with respect to its ith index.

In cases like clinical trials, $\pi_T(\boldsymbol{X})$ may be *known*. In this case, there is no corresponding tangent space Λ_π and the corresponding nuisance tangent space $\tilde{\Lambda} \equiv \Lambda_{\boldsymbol{X}} + \Lambda_\epsilon + \Lambda_{\boldsymbol{g}}$. Its orthogonal complement $\tilde{\Lambda}^{\perp}$ is then larger and can be shown to be the sum of $\Lambda^{\perp}$ and $\mathcal{S}_2$ defined in the supplementary materials. For any function $f(\epsilon, \boldsymbol{X}, T)$, its projection on $\tilde{\Lambda}^{\perp}$ is its projection on $\Lambda^{\perp}$ plus an additional term $\boldsymbol{w}_T^{\top}E(\boldsymbol{w}_T\boldsymbol{w}_T^{\top}|\boldsymbol{X})^{-1}E(\boldsymbol{w}_T f|\boldsymbol{X})$. However, the efficient score is unchanged as $E(\boldsymbol{w}_T f|\boldsymbol{X}) = 0$ when f is chosen as the score function.

As a special case of Theorem 3.2 and Corollary 3.1, when $K = 2$, we have the following corollaries, recognizing that $\boldsymbol{w}_T = \pi_T(\boldsymbol{X})^{-1}T$ now becomes a scalar.

Corollary 3.2. For $K = 2$ and $T \in \{-1, 1\}$,

$$\begin{aligned}\Lambda^{\perp} =& \left\{\pi_T(\boldsymbol{X})^{-1}T\left[\alpha(\boldsymbol{X}) - E\{\alpha(\boldsymbol{X})|\boldsymbol{B}_0^{\top}\boldsymbol{X}\}\right][\epsilon - E(\epsilon|\boldsymbol{X})],\right.\\ &\left.\forall\alpha(\boldsymbol{X}) : \mathcal{X} \mapsto R\right\}.\end{aligned}$$

Corollary 3.3. For $K = 2$ and $T \in \{-1, 1\}$, the projection of any function $f(\epsilon, \boldsymbol{X}, T) \in \mathcal{H}$ onto $\Lambda^{\perp}$ is given by

$$\pi_T(\boldsymbol{X})^{-1}T\,C(\boldsymbol{B}_0^{\top}\boldsymbol{X})\{\epsilon - E[\epsilon|\boldsymbol{X}]\},$$

where

$$\begin{aligned}C(\boldsymbol{B}_0^{\top}\boldsymbol{X}) &= V(\boldsymbol{X})\left\{D(\boldsymbol{X}) - \frac{E[V(\boldsymbol{X})D(\boldsymbol{X})|\boldsymbol{B}_0^{\top}\boldsymbol{X}]}{E[V(\boldsymbol{X})|\boldsymbol{B}_0^{\top}\boldsymbol{X}]}\right\},\\ V(\boldsymbol{X})^{-1} &= E[\pi_T(\boldsymbol{X})^{-2}\epsilon^2|\boldsymbol{X}] - E[\pi_T(\boldsymbol{X})^{-2}|\boldsymbol{X}]E(\epsilon|\boldsymbol{X})^2,\\ D(\boldsymbol{X}) &= E[\pi_T(\boldsymbol{X})^{-1}Tf\epsilon|\boldsymbol{X}] - E[\pi_T(\boldsymbol{X})^{-1}Tf|\boldsymbol{X}]E(\epsilon|\boldsymbol{X}).\end{aligned}$$

Therefore, the efficient score is

$$\pi_T(\boldsymbol{X})^{-1}T\,\boldsymbol{C}^*(\boldsymbol{B}_0^{\top}\boldsymbol{X})\{\epsilon - E(\epsilon|\boldsymbol{X})\},$$

where

$$\boldsymbol{C}^*(\boldsymbol{B}_0^{\top}\boldsymbol{X}) = V(\boldsymbol{X})\,\nabla g(\boldsymbol{B}_0^{\top}\boldsymbol{X}) \otimes \left\{\boldsymbol{X} - \frac{E[V(\boldsymbol{X})\boldsymbol{X}|\boldsymbol{B}_0^{\top}\boldsymbol{X}]}{E[V(\boldsymbol{X})|\boldsymbol{B}_0^{\top}\boldsymbol{X}]}\right\},$$

and $\otimes$ is the Kronecker product.

4. Estimation and Algorithm

We first consider estimation of $\boldsymbol{B}_0$ with a fixed d. Then we propose a method for determining d similar to Xia et al. (2002). For simplicity, we present our method with $K = 2$. Generalization to $K > 2$ is straightforward and relegated to the supplementary materials. From Corollary 3.3, the efficiency score can be written as

$$\begin{aligned}V(\boldsymbol{X})\frac{T}{\pi_T(\boldsymbol{X})}\nabla g(\boldsymbol{B}_0^{\top}\boldsymbol{X}) \otimes &\left\{\boldsymbol{X} - \frac{E[V(\boldsymbol{X})\boldsymbol{X}|\boldsymbol{B}_0^{\top}\boldsymbol{X}]}{E[V(\boldsymbol{X})|\boldsymbol{B}_0^{\top}\boldsymbol{X}]}\right\}\\ &\times\{\epsilon - E(\epsilon|\boldsymbol{X})\}. \qquad (8)\end{aligned}$$

We can see that the efficient score is hard to estimate directly due to many conditional expectations involved. We therefore use (8) to accomplish two tasks.

The first task is to construct more practical and simplified estimation procedures by exploring the robustness of the efficient score (8). In particular, (8) remains unbiased (for 0) by omitting the fraction $E[V(\boldsymbol{X})\boldsymbol{X}|\boldsymbol{B}_0^{\top}\boldsymbol{X}]/E[V(\boldsymbol{X})|\boldsymbol{B}_0^{\top}\boldsymbol{X}]$ and the leading term $V(\boldsymbol{X})$. In addition, $\pi_T(\boldsymbol{X})$ and $E[\epsilon|\boldsymbol{X}]$ form a pair for robustness in the sense that, if one is known or consistently estimated, the other can be misspecified. This is the well-known double robustness property in semiparametric estimation (Tsiatis 2006). Therefore, we propose the following class of estimating equations that are all unbiased for estimating $\boldsymbol{B}_0$ under Model (7),

$$\tilde{S} = \left\{\pi_T(\boldsymbol{X})^{-1}T\nabla g(\boldsymbol{B}_0^{\top}\boldsymbol{X}) \otimes \boldsymbol{X}(\epsilon - \eta(\boldsymbol{X})), \forall\eta(\boldsymbol{X}) : \mathcal{X} \mapsto R\right\}.$$

This will be our choice of estimating equations. The obvious benefit of using this function class $\tilde{S}$ is that solving the estimating equations is equivalent to minimizing the loss function $\pi_T(\boldsymbol{X})^{-1}\{Y - \frac{1}{2}Tg(\boldsymbol{B}_0^{\top}\boldsymbol{X}) - \eta(\boldsymbol{X})\}^2$. The corresponding sample version is

$$L_g(\boldsymbol{B}) = \frac{1}{n}\sum_{i=1}^{n}\frac{\{Y_i - \frac{1}{2}T_i g(\boldsymbol{B}^T\boldsymbol{X}_i) - \eta(\boldsymbol{X}_i)\}^2}{\pi_{T_i}(\boldsymbol{X}_i)}. \qquad (9)$$

The proposed loss function remains doubly robust in the sense that the minimizer of the proposed loss function is consistent if either $\pi_T(\boldsymbol{X})$ or $\eta(\boldsymbol{X}) = E[\epsilon|\boldsymbol{X}]$ is correctly specified. When $\pi_T(\boldsymbol{X})$ is known or can be consistently estimated, the

choice of $\eta(\boldsymbol{X})$ can be flexible. A convenient choice is $\eta(\boldsymbol{X}) = 0$ adopted in Chen et al. (2017) and Tian et al. (2014). Another choice is $\eta(\boldsymbol{X}) = \{1-2\pi(\boldsymbol{X})\}g(\boldsymbol{B}_0^\top \boldsymbol{X})$ used by Song et al. (2017). However, from the proof of Theorem 3.1 and Corollary 3.1,

$$\eta^*(\boldsymbol{X}) = E[\epsilon|\boldsymbol{X}]$$

leads to the most efficient estimator.

Because g is unknown, to estimate $\boldsymbol{B}_0$ through minimizing $L_g(\boldsymbol{B})$, we employ a minimum average variance estimation (MAVE) type of method as advocated in Xia et al. (2002). In particular, minimization is based on the following approximating loss function:

$$L(\boldsymbol{B}, \{a_j, \boldsymbol{b}_j\}_{j=1}^n) \tag{10}$$
$$= \sum_{j=1}^n \sum_{i=1}^n \frac{\{Y_i - \frac{1}{2}T_i[a_j + \boldsymbol{b}_j^\top(\boldsymbol{B}^\top \boldsymbol{X}_i - \boldsymbol{B}^\top \boldsymbol{X}_j)] - \eta(\boldsymbol{X}_i)\}^2}{n^2\, \pi_{T_i}(\boldsymbol{X}_i)} w_{ij},$$

where $w_{ij} = K_h(\boldsymbol{B}^\top \boldsymbol{X}_j - \boldsymbol{B}^\top \boldsymbol{X}_i)$ and $K_h(\cdot) = \frac{1}{h^d}K(\cdot/h)$ is a kernel function with bandwidth h. The extra parameters $a_j \in R$ and $\boldsymbol{b}_j \in R^d$ can be thought of as approximations to g and its gradient at each point $\boldsymbol{B}^\top \boldsymbol{X}_j$, and the kernel weight w_{ij} ensures the adequacy of the local linear approximation of g in its neighborhood. We can also normalize the weight w_{ij}'s by $\tilde{w}_{ij} = w_{ij}/\sum_j w_{ij}$. In the next two subsections, we will consider both the case of fixing $\eta(\boldsymbol{X})$ through a sensible or convenient choice and of estimating $\eta^*(\boldsymbol{X}) = E[\epsilon|\boldsymbol{X}]$. We term the two methods interaction MAVE (iMAVE) and iMAVE2, respectively.

The second task is to use the variance of the efficient score (8), or the efficiency bound, to evaluate our method. Obviously, our simplified method will lead to efficiency loss in general cases. However, if we further impose two assumptions

(a) $\epsilon \perp T|\boldsymbol{X}$, $\mathrm{var}(\epsilon|\boldsymbol{X})$ is a constant;
(b) $\pi_1(\boldsymbol{X}) \equiv \pi_1$, where π_1 is a constant.

Then the efficiency bound (based on the asymptotic variance of the efficient score) is exactly the same as the variance of our iMAVE2 method derived in Theorem 5.3. Therefore, iMAVE2 attains local efficiency under the above two assumptions.

4.1. The iMAVE Method With a Fixed $\eta(X)$

In this section, a weighted least square algorithm to minimize (10) is introduced that consists of the following steps.

1. An initial estimator, $\boldsymbol{B}_{(1)}$, is obtained. Please see our comments after the algorithm on how to obtain $\boldsymbol{B}_{(1)}$.
2. Let $\boldsymbol{B}_{(t)}$ be the estimator at the tth iteration. Calculate
$$w_{ij}^{(t)} = K_h(\boldsymbol{B}_{(t)}^\top \boldsymbol{X}_i - \boldsymbol{B}_{(t)}^\top \boldsymbol{X}_j).$$
3. Solve the following weighted least square problem to obtain
$$(a_j^{(t)}, \boldsymbol{b}_j^{(t)}) = \arg\min_{a_j, \boldsymbol{b}_j} L_1(a_j, \boldsymbol{b}_j),$$
for $j = 1, \ldots, n$, where
$$L_1(a_j, \boldsymbol{b}_j) = \frac{1}{n}\sum_{i=1}^n \frac{\{Y_i - \eta(\boldsymbol{X}_i) - \frac{1}{2}T_i[a_j + \boldsymbol{b}_j^\top(\boldsymbol{B}_{(t)}^\top \boldsymbol{X}_i - \boldsymbol{B}_{(t)}^\top \boldsymbol{X}_j)]\}^2}{\pi_{T_i}(\boldsymbol{X}_i)} w_{ij}^{(t)}.$$
4. Solve the following weighted least square problem to obtain
$$\tilde{\boldsymbol{B}}_{(t+1)} = \arg\min_{\boldsymbol{B}} L_2(\boldsymbol{B}),$$
where
$$L_2(\boldsymbol{B}) = \frac{1}{n^2}\sum_{j=1}^n \sum_{i=1}^n \frac{\{Y_i - \eta(\boldsymbol{X}_i) - \frac{1}{2}T_i[a_j^{(t)} + \boldsymbol{b}_j^{(t)\top}(\boldsymbol{B}^\top \boldsymbol{X}_i - \boldsymbol{B}^\top \boldsymbol{X}_j)]\}^2}{\pi_{T_i}(\boldsymbol{X}_i)} w_{ij}^{(t)}.$$
5. Normalize to obtain $\boldsymbol{B}_{(t+1)}$ by projecting $\tilde{\boldsymbol{B}}_{(t+1)}$ onto the Grassmann manifold.
6. If the discrepancy, $|\boldsymbol{B}_{(t+1)} - \boldsymbol{B}_{(t)}|$, is smaller than a prespecified tolerance, or a max number of iterations achieved, then output $\boldsymbol{B}_{(t+1)}$. If not, go back to Step 2 and start a new iteration.

The initial estimator $\boldsymbol{B}_{(1)}$ needs to be a consistent estimator for our theoretical analysis. To get a consistent $\boldsymbol{B}_{(1)}$, one choice is to solve a simplified version of (10) by only expanding g at $\boldsymbol{0}$,

$$L(\boldsymbol{B}) = \frac{1}{n}\sum_{i=1}^n \frac{\{Y_i - \frac{1}{2}T_i\boldsymbol{B}^\top \boldsymbol{X}_i\}^2}{\pi_{T_i}(\boldsymbol{X}_i)} \tilde{w}_{i0},$$

where $\tilde{w}_{i0} = K_h(\boldsymbol{B}^\top \boldsymbol{X}_i)$. For $d = 1$, one can also utilize the method of Song et al. (2017). In our simulation studies, we found that a simple choice of $\boldsymbol{B}_{(1)} = \boldsymbol{0}$ almost always led to stable convergent results.

4.2. The iMAVE2 Method With an Estimated $\eta^*(X)$

The following two-step procedure is proposed to estimate $\eta^*(\boldsymbol{X}) = E[\epsilon|\boldsymbol{X}]$. First, we obtain an estimate $\hat{\boldsymbol{B}}$ of $\boldsymbol{B}_0$ with a prefixed η. Then $g(\boldsymbol{B}^\top \boldsymbol{X})$ is estimated by

$$\hat{g}(\hat{\boldsymbol{B}}^\top \boldsymbol{X}) = \frac{\sum_{i=1}^n \pi_{T_i}(\boldsymbol{X}_i)^{-1} T_i Y_i K_h(\hat{\boldsymbol{B}}^\top(\boldsymbol{X}_i - \boldsymbol{X}))}{\sum_{i=1}^n K_h(\hat{\boldsymbol{B}}^\top(\boldsymbol{X}_i - \boldsymbol{X}))}, \tag{11}$$

where K_h is a kernel function with $K_h(\boldsymbol{X}) = h^{-d}K(\boldsymbol{X}/h)$. The kernel K and bandwidth h can be different from those used before in (10).

The estimated residual is $\hat{\epsilon}_i = Y_i - \frac{1}{2}T_i\hat{g}(\hat{\boldsymbol{B}}^\top \boldsymbol{X}_i)$. We can then estimate $E[\epsilon|\boldsymbol{X}]$, by

$$\frac{\sum_{i=1}^n \hat{\epsilon}_i K_h(\boldsymbol{X}_i - \boldsymbol{X})}{\sum_{i=1}^n K_h(\boldsymbol{X}_i - \boldsymbol{X})}, \tag{12}$$

where K_h is another kernel function with $K_h(\boldsymbol{X}) = h^{-p}K(\boldsymbol{X}/h)$. Again, the kernel K and bandwidth h can be different from those used before. On the other hand, noticing that $E[\pi_{T_i}(\boldsymbol{X}_i)^{-2}|\boldsymbol{X}]^{-1} = \pi_1(\boldsymbol{X})\pi_{-1}(\boldsymbol{X})$, η^* can also be estimated by

$$\hat{\eta}^*(\boldsymbol{X}) = \pi_1(\boldsymbol{X})\pi_{-1}(\boldsymbol{X})\frac{\sum_{i=1}^n \pi_{T_i}(\boldsymbol{X}_i)^{-2}\hat{\epsilon}_i K_h(\boldsymbol{X}_i - \boldsymbol{X})}{\sum_{i=1}^n K_h(\boldsymbol{X}_i - \boldsymbol{X})}. \tag{13}$$

With an estimated $\hat{\eta}^*$, a possibly improved estimator $\hat{\boldsymbol{B}}^*$ of $\boldsymbol{B}_0$ can be obtained. We call this efficiency improved estimation method iMAVE2.

Other approaches to obtain η^* can also be considered. For example, it may be estimated from an external independent dataset or given directly through prior knowledge. When η^* cannot be estimated reliably, especially when the dimensionality of $\boldsymbol{X}$ is high or when the sample size n is small, as long as the estimator is a function of $\boldsymbol{X}$, the resulting $\hat{\boldsymbol{B}}^*$ is still unbiased in principle. Therefore instead of nonparametric estimators, parametric models may also be used to estimate η^*.

4.3. Dimension Determination

There is a need to determine the dimension d, especially when p is large. Many methods proposed in the dimension reduction literature are applicable in our setting too (Schott 1994; Cook 1998; Koch and Naito 2007). In this article, we adopt the same procedure as Xia et al. (2002), which is a consistent procedure based on cross-validation. In particular, because

$$E\left[\frac{T}{\pi_T(\mathbf{X})}Y\middle|\mathbf{X}\right] = E\left[\frac{T}{\pi_T(\mathbf{X})}Y\middle|\mathbf{B}_0^\top\mathbf{X}\right],$$

consistency of the dimension determination procedure can be established by a direct application of Theorem 2 in Xia et al. (2002).

Given a dimension $d \in \{0, 1, \ldots, p\}$, the procedure goes through the following steps based on iMAVE.

1. Randomly split the dataset into five folds, and $\mathcal{I}_m, m = 1, \ldots, 5$ are the sets corresponding to these folds.
2. For $m = 1, \ldots, 5$, choose $\mathcal{I}_m$ as a testing set and the rest $\mathcal{I}_{-m}$ as a training dataset. Fit iMAVE on $\mathcal{I}_{-m}$ to obtain estimates of $\hat{\boldsymbol{B}}_{(-m)}$ and $\hat{g}_{(-m)}(\cdot)$. Then calculate the following score.

$$\mathrm{CV}(d, m) = \frac{1}{|\mathcal{I}_m|}\sum_{i\in\mathcal{I}_m}\left(\frac{1}{2}\frac{T_iY_i}{\pi_{T_i}(\boldsymbol{X}_i)} - \hat{g}_{(-m)}(\hat{\boldsymbol{B}}^\top\boldsymbol{X}_i)\right)^2,$$

   where $\hat{g}_{(-m)}(\cdot)$ is estimated using $\mathcal{I}_{-m}$.
3. The estimated dimension is $\hat{d} = \arg\min_{0\le d\le p}\sum_{m=1}^5 \mathrm{CV}(d, m)$.

These same steps can also be based on iMAVE2 to determine the dimension. It is intuitively clear that over-estimating the true dimension d to a slightly larger value is much less of a concern than under-estimating.

5. Theoretical Results

In this section, we analyze our estimator in a unified framework of statistical and algorithmic properties assuming a binary T for notational simplicity. We study both iMAVE and iMAVE2.

The non-convexity of (10) makes it intractable to obtain theoretical results for prediction or classification error by simply mimicking the usual analysis of empirical risk minimization (Vapnik 2000). It is also hard to analyze the convergence rate or asymptotic distribution of the proposed estimators due to a lack of characterization of the minimizers. On the other hand, because we carry out our optimization by iteratively solving a weighted least square problem, we can track the change of each iteration similar to Xia et al. (2002) and Xia (2007). This leads us to propose a unified framework of joint statistical and algorithmic analysis.

For any matrix $\boldsymbol{A}$, $|\boldsymbol{A}|$ represents the Frobenius norm of $\boldsymbol{A}$. For any random matrix $\boldsymbol{A}_n$, we say $\boldsymbol{A}_n = O_p(a_n)$ if each entry of $\boldsymbol{A}_n$ is $O_p(a_n)$. Let $\boldsymbol{B}_{(t)}$ be the estimator in the tth iteration of the iMAVE algorithm, and $\hat{\boldsymbol{B}}$ be the limit of $\boldsymbol{B}_{(t)}$ when $t \to +\infty$. The existence of the limit of $\boldsymbol{B}_{(t)}$ as well as the convergence of the algorithm, similar to Xia (2007), can be concluded from the proof. Denote $\delta_{\boldsymbol{B}}^{(t)} = |\boldsymbol{B}_{(t)} - \boldsymbol{B}_0|$. Our goal is to answer the following questions for both iMAVE and iMAVE2:

1. Suppose that $\delta_{\boldsymbol{B}}^{(1)}$ has some convergence rate to 0. After t iterations, what is the convergence rate of $\delta_{\boldsymbol{B}}^{(t)}$?
2. What is the convergence rate of $\delta_{\hat{\boldsymbol{B}}} \equiv |\hat{\boldsymbol{B}} - \boldsymbol{B}_0|$?
3. What are the answers for Questions 1 and 2 when iMAVE2 is used.
4. Whether there is asymptotic efficiency gain of iMAVE2 compared with iMAVE?

Questions 1 and 2 are answered by Theorems 5.1 and 5.2, respectively. Question 3 is answered by Theorem 5.5. Question 4 is answered by Theorems 5.3 and 5.5.

Theorem 5.1 is a new result beyond Xia et al. (2002) and Xia (2007). It essentially quantifies the nonasymptotic property of our estimators. It implies that under certain conditions, $\delta_{\boldsymbol{B}}^{(t)}$ converges to 0 with a rate of at least $(n/\log n)^{-1/2}$ almost surely when t is large enough and $d \le 5$. When $d > 5$, the convergence rate is bounded by a quantity related to bandwidth and d, and slower than $(n/\log n)^{-1/2}$. Theorem 5.2 implies that under certain conditions, $\delta_{\hat{\boldsymbol{B}}}$ converges to 0 in probability with the order of $n^{-1/2}$ when $d \le 5$. When $d > 5$, the convergence rate is slower than $n^{-1/2}$. The convergence rate in Theorem 5.2 is different than that in Theorem 5.1 by a factor of $\log n$ due to the difference of convergence modes. Theorem 5.1 provides deeper results with both statistical and algorithmic properties.

Theorems 5.3 and 5.5 provide the asymptotic distributions of iMAVE and iMAVE2 estimators, respectively. Theorem 5.4 provides the accuracy of estimating g based on $\hat{\boldsymbol{B}}$. Combining with the previous results in Section 2, we will see that difference of the asymptotic covariance matrices of iMAVE and iMAVE2 is always positive semidefinite. Thus, iMAVE2 is more efficient than iMAVE.

The conditions needed for our theorems are as follows. Let $\xi_{\boldsymbol{B}}(\boldsymbol{u}) = E(\boldsymbol{X}\boldsymbol{X}^\top|\boldsymbol{B}^\top\boldsymbol{X} = \boldsymbol{u})$ and $\mu_{\boldsymbol{B}}(\boldsymbol{u}) \equiv E(\boldsymbol{X}|\boldsymbol{B}^\top\boldsymbol{X} = \boldsymbol{u})$. We denote the distribution of $\boldsymbol{B}^\top\boldsymbol{X}$ as $p_{\boldsymbol{B}}(\boldsymbol{B}^\top\boldsymbol{x})$.

(C.1) The density of $\boldsymbol{X}$, $p_{\boldsymbol{X}}(\boldsymbol{x})$, has bounded 4th order derivatives and compact support. $\mu_{\boldsymbol{B}}(\boldsymbol{u})$ and $\xi_{\boldsymbol{B}}(\boldsymbol{u})$ have bounded derivatives with respect to $\boldsymbol{u}$ and $\boldsymbol{B}$ where $\boldsymbol{B}$ is in a small neighborhood of $\boldsymbol{B}_0 : |\boldsymbol{B} - \boldsymbol{B}_0| \le \delta$, for some $\delta > 0$.

(C.2) The matrix $\boldsymbol{M}_0 = \int \nabla g(\boldsymbol{B}_0^\top\boldsymbol{x})\nabla^\top g(\boldsymbol{B}_0^\top\boldsymbol{x}) \times p_{\boldsymbol{B}_0}(\boldsymbol{B}_0^\top\boldsymbol{x})p_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x}$ has full rank d.

(C.3) $K(\cdot)$ is a spherical symmetric univariate density function with a bounded 2nd order derivative and compact support.

(C.4) g has a bounded derivative. The error ϵ satisfies that there exists some M and $\nu_0 \in [0, +\infty)$ such that

$$E\left\{\exp\left[\frac{T\epsilon}{\pi_T(\boldsymbol{X})M}\right] - 1 - \frac{|T\epsilon|}{\pi_T(\boldsymbol{X})M} \,\Big|\, \boldsymbol{X}\right\} M^2 \le \nu_0/2.$$

(C.5) The bandwidth $h_1 = c_1 n^{-r_h}$, where $0 < r_h \le 1/\{\max(p,3)+6\}$. For $t \ge 2$, $h_t = \max\{n^{-r_h/2}h_{t-1}, \hbar\}$, where $\hbar = c_3 n^{-r_h'}$ with $0 < r_h' \le 1/(d+3)$. Here, c_1–c_4 are constants.

(C.6) $p_{\boldsymbol{B}}(\boldsymbol{B}^\top \boldsymbol{x})$ is bounded away from 0. In addition, $E[\pi_T(\boldsymbol{X})^{-1}TY|\boldsymbol{B}^\top\boldsymbol{X} = \boldsymbol{u}]$ is Lipschitz continuous and $\pi_T(\boldsymbol{X})$ is bounded away from 0 and 1.

Condition (C.6) is only needed for Theorem 5.4. Conditions (C.1)–(C.5) are similar to Xia (2007) except the requirement for compact support of covariates. This requirement is needed for iMAVE2 because g needs to be estimated to a certain rate for the asymptotic property of iMAVE2. For iMAVE, this requirement can be replaced by a finite moment condition. Epanechnikov and quadratic kernels satisfy Condition (C.3). The Gaussian kernel can also be used to guarantee our theoretical results with some modification to the proofs. According to Xia (2007), Condition (C.2) suggests that the dimension d cannot be further reduced. The bandwidth requirement in Condition (C.5) can be easily met. Condition (C.6) characterizes the smoothness of g as typically required for conditional expectation estimation.

Theorem 5.1. Under Conditions (C.1)–(C.5), suppose that the initial estimator for iMAVE, $\boldsymbol{B}_{(1)}$, satisfies $\delta_{\boldsymbol{B}}^{(1)}/h_1 \to 0$, then there exists a constant C_1 such that when the number of iterations t satisfies

$$t \ge 1 + \log\min\left\{\frac{3C_1\{\delta_n + \delta_{d\hbar}^2\hbar + \hbar^4\}}{\delta_{\boldsymbol{B}}^{(1)} + 2C_1 h_1^4}, 1\right\} \Big/ \log\frac{2}{3},$$

we have $\delta_{\boldsymbol{B}}^{(t)} \le (3C_1+1)\{\delta_n + \delta_{d\hbar}^2\hbar + \hbar^4\}$almost surely, where $\delta_n = (n/\log n)^{-1/2}$ and $\delta_{d\hbar} = (n\hbar^d/\log n)^{-1/2}$.

A simple observation from Theorem 5.1 implies that to reach the same accuracy when d increases, the number of iterations required is increasing linearly in d. This provides a useful guidance on the maximum number of iterations for the algorithm.

Theorem 5.2. Under the same conditions as Theorem 5.1, there exists a matrix $\boldsymbol{B}_0^\perp$ whose column space is the orthogonal complement of the column space of $\boldsymbol{B}_0$, such that the iMAVE estimator satisfies

$$\hat{\boldsymbol{B}} = \boldsymbol{B}_0\{\boldsymbol{I}_d + O_p(\hbar^4 + \delta_{d\hbar}^2 + n^{-1/2})\} + \boldsymbol{B}_0^\perp O_p(\hbar^4 + \delta_{d\hbar}^2 + n^{-1/2}).$$

Theorem 5.2 implies that when $\hat{\boldsymbol{B}}$ is decomposed based on the column space of $\boldsymbol{B}_0$ and its orthogonal complement, the component in the column space of $\boldsymbol{B}_0^\perp$ converges to 0, and the projection of $\hat{\boldsymbol{B}}$ on the column space of $\boldsymbol{B}_0$ converges to $\boldsymbol{B}_0$. To obtain the $n^{-1/2}$ convergence rate, we need $\hbar^4 + \delta_{d\hbar}^2 = O(n^{-1/2})$. In this case, d has to be no larger than 5.

Theorem 5.3. Assume the same conditions as Theorem 5.1 and $\hbar^4 + \delta_{d\hbar}^2 = o_p(n^{-1/2})$. Denote $\nu_{\boldsymbol{B}}(\boldsymbol{x}) \equiv \mu_{\boldsymbol{B}}(\boldsymbol{B}^\top\boldsymbol{x}) - \boldsymbol{x}$. Let $l(\hat{\boldsymbol{B}})$ and $l(\boldsymbol{B}_0)$ be vectorizations of the matrices $\hat{\boldsymbol{B}}$ and $\boldsymbol{B}_0$, respectively. Then

$$\sqrt{n}\{l(\hat{\boldsymbol{B}}) - l(\boldsymbol{B}_0)\} \to N(0, \boldsymbol{D}_0^+\boldsymbol{\Sigma}_0\boldsymbol{D}_0^+),$$

where $\boldsymbol{\Sigma}_0 = \mathrm{var}[\pi_{T_i}(\boldsymbol{X}_i)^{-1}T_i\nabla g(\boldsymbol{B}_0^\top\boldsymbol{X}_i)\otimes\nu_{\boldsymbol{B}_0}(\boldsymbol{X}_i)\{\epsilon_i - \eta(\boldsymbol{X}_i)\}]$. The expression of $\boldsymbol{D}_0^+$ can be found in our proof of this theorem from the supplementary materials.

Theorem 5.4. Suppose that Conditions (C.1)–(C.6) are satisfied and g is estimated by some kernel K_h of order m. Then h can be selected such that when n is large enough,

$$\|\hat{g}(\hat{\boldsymbol{B}}^\top\boldsymbol{X}) - g(\boldsymbol{B}_0^\top\boldsymbol{X})\|_\infty \le O\left\{(n/\log n)^{-\frac{m}{2m+d}}\right\}, \text{almost surely,}$$

where m can be any integer when $d \le 5$, but $m \le 4d/(d-5)$ when $d > 5$.

Theorem 5.5. Denote $\delta_{ph} \equiv (nh^p/\log n)^{-1/2}$. In iMAVE2, suppose $d \le 5$ and $\delta_{ph}^2 + h^{2m} = o(n^{-1/2})$ when estimating η^* by $\hat{\eta}^*$ using (12) or (13). Then, under Conditions (C.1)–(C.5), for iMAVE2, Theorems 5.1 and 5.2 still hold and Theorem 5.3 holds with the asymptotic variance, $\boldsymbol{D}_0^+\boldsymbol{\Sigma}_0^*\boldsymbol{D}_0^+$, where $\boldsymbol{\Sigma}_0^* = \mathrm{var}\left[\pi_{T_i}(\boldsymbol{X}_i)^{-1}T_i\nabla g(\boldsymbol{B}_0^\top\boldsymbol{X}_i)\otimes\nu_{\boldsymbol{B}_0}(\boldsymbol{X}_i)\{\epsilon_i - \eta^*(\boldsymbol{X}_i)\}\right]$, and $\boldsymbol{\Sigma}_0 - \boldsymbol{\Sigma}_0^*$ is positive semidefinite.

Detailed proofs for all theorems are given in the supplementary materials. Here, we consider construction of confidence intervals for $\mathbf{B}_0$ and possible improvement of empirical estimation with limited sample sizes. From Theorems 5.3 and 5.5, we know that the estimators are both $\sqrt{n}$-consistent and asymptotically normal under suitable conditions. This makes the inference of $\mathbf{B}_0$ possible if we have a stable way to estimate the asymptotic variances to form confidence intervals. In theory we just need to evaluate the variance formulas using observed data.

However, we found from our simulation studies that estimation of ∇g in the asymptotic variance formulas can be challenging. If we directly use all the data to estimate ∇g, the resulting confidence intervals often over cover. This is because estimation of ∇g is directly related to estimation of $\mathbf{B}_0$. Using data twice to first estimate $\mathbf{B}_0$ and then estimate ∇g leads to overfitting. Therefore, we propose a sample split procedure to alleviate this issue, similar to some recent works (Athey and Wager 2017; Chernozhukov et al. 2018; Zhao et al. 2019). Specifically, the whole dataset is split into halves randomly. On the first half, an iMAVE or iMAVE2 estimate of $\mathbf{B}_0$ is obtained. On the other half, we estimate g and ∇g using smoothing splines.

In addition, we found that a further one-step Newton–Raphson estimator for $\mathbf{B}_0$ can lead to some improvement, especially when the sample size is limited. In particular, we use the following step:

$$\hat{\mathbf{B}}_{\mathrm{NR}} = \hat{\mathbf{B}}_{\mathrm{MV}} - \left\{E^{(1)}\left[\frac{\partial\hat{\mathcal{S}}(\hat{\mathbf{B}}_{\mathrm{MV}};\boldsymbol{X},T,Y)}{\partial\hat{\mathbf{B}}_{\mathrm{MV}}}\right]\right\}^{-1}$$
$$E^{(1)}\left[\hat{\mathcal{S}}(\hat{\mathbf{B}}_{\mathrm{MV}};\boldsymbol{X},T,Y)\right],$$

where

$$\hat{\mathcal{S}}(\hat{\mathbf{B}}_{\mathrm{MV}}; \boldsymbol{X}, T, Y) = \pi_T(\boldsymbol{X})^{-1} T \nabla \hat{g}(\hat{\mathbf{B}}_{\mathrm{MV}}^{\top} \boldsymbol{X}) \otimes \hat{v}_{\hat{\mathbf{B}}_{\mathrm{MV}}}(\boldsymbol{X})(\epsilon - \eta(\boldsymbol{X})),$$

$\hat{\mathbf{B}}_{\mathrm{MV}}$ is the iMAVE or iMAVE2 estimator with corresponding choice of η or $\hat{\eta}^*(\mathbf{X})$ on the first half of the dataset, $\hat{v}_{\hat{\mathbf{B}}_{\mathrm{MV}}}$ is the estimator of $v_{\boldsymbol{B}_0}$ on the second half of the dataset, and $\nabla \hat{g}$ is the estimator of gradient on the second half of the dataset. $E^{(1)}[\cdot]$ represents expectation taken over the first half of the dataset. From the theory of one-step Newton–Raphson estimators, $\hat{\mathbf{B}}_{\mathrm{NR}}$ is still a $\sqrt{n}$-consistent estimator and its asymptotic variance can be estimated by

$$\left\{ E^{(1)} \left[\frac{\partial \hat{\mathcal{S}}(\hat{\mathbf{B}}_{\mathrm{MV}}; \boldsymbol{X}, T, Y)}{\partial \hat{\mathbf{B}}_{\mathrm{MV}}} \right] \right\}^{-1} \mathrm{var} \left[\hat{\mathcal{S}}(\hat{\mathbf{B}}_{\mathrm{MV}}; \boldsymbol{X}, T, Y) \right] \left\{ E^{(1)} \left[\frac{\partial \hat{\mathcal{S}}(\hat{\mathbf{B}}_{\mathrm{MV}}; \boldsymbol{X}, T, Y)}{\partial \hat{\mathbf{B}}_{\mathrm{MV}}} \right] \right\}^{-1}.$$

Due to the sample split procedure, the estimation error of $\hat{g}$ is not related to the first half of the data, which results in a more stable estimation of the asymptotic variance.

6. Simulation

Here, our method is evaluated and compared with existing methods. In particular, we compare with the outcome weighted learning method based on a logistic loss in Xu et al. (2015), the modified covariate method under the squared loss proposed in Tian et al. (2014), and residual weighted learning method (Zhou et al. 2017) based on a logistic loss. We also compare with Q-learning with linear basis functions as a parametric version of the proposed loss function (Qian and Murphy 2011). We first evaluate estimation results assuming d is known and then investigate dimension determination. Given the fact that Song et al. (2017) is a special case of iMAVE and their method can be applied only when $d = 1$, we do not include it as our comparison method.

We report part of the results for estimating effect modification and dimension determination in the main text. The rest of the simulation results is relegated to the supplementary materials. There we also report confidence interval coverage, and results for additional settings including more complex data generation models and correlated covariates.

6.1. Estimation Evaluation With Known d

Data are generated by the following model,

$$y = (\boldsymbol{\beta}^{\top} \boldsymbol{X})^2 + \frac{1}{2} T g(\boldsymbol{\beta}^{\top} \boldsymbol{X}) + \epsilon, \tag{14}$$

where $\epsilon \sim N(0, \sigma^2)$ and g is chosen as

1. Linear: $g(\boldsymbol{\beta}^{\top} \boldsymbol{X}) = \tau \boldsymbol{\beta}^{\top} \boldsymbol{X}$;
2. Logistic: $g(\boldsymbol{\beta}^{\top} \boldsymbol{X}) = \tau\{(1 + e^{-\boldsymbol{\beta}^{\top} \boldsymbol{X}})^{-1} - 0.5\}$;
3. Gaussian: $g(\boldsymbol{\beta}^{\top} \boldsymbol{X}) = \tau\{\Phi(\boldsymbol{\beta}^{\top} \boldsymbol{X}) - 0.5\}$, where $\Phi(\cdot)$ is the Gaussian distribution function.

We set $\sigma = 0.6$, $\tau = 7$, and T is generated to be -1 or 1 with equal probability and independent with all other variables. The true $\boldsymbol{\beta}_0$ is chosen to be $(1, 1, 1, 1)^{\top}$. $\boldsymbol{X}$ is generated from $N(0, \boldsymbol{I}_{4\times 4})$. The sample size n varies from 200, 500 to 1000. Results are summarized from 1000 simulated datasets.

Table 1 investigates the asymptotic bias of the iMAVE and iMAVE2 and the possible gain in efficiency from the latter. The ratios $\hat{\beta}_j/\hat{\beta}_1$, $j = 2, 3, 4$, are reported due to the Grassmann manifold assumption for identifiability. Whereas there are some empirical biases for nonlinear g under small sample sizes, as the sample size increases, the means of the ratios all approach 1, the true value. There is noticeable improvement from iMAVE2 over iMAVE in terms of MSE.

We further consider prediction results under the settings of known and estimated propensity scores. In particular, we

Table 1. Simulation results for coefficient estimation.

Size			Linear		Gaussian		Logistic	
			iMAVE	iMAVE2	iMAVE	iMAVE2	iMAVE	iMAVE2
200	Mean	$\hat{\beta}_2/\hat{\beta}_1$	0.9995	0.9986	0.8630	0.9161	0.7797	0.8611
		$\hat{\beta}_3/\hat{\beta}_1$	1.0021	1.0021	0.8960	0.9410	0.8192	0.8884
		$\hat{\beta}_4/\hat{\beta}_1$	1.0042	1.0035	0.8891	0.9408	0.8013	0.8802
	$\sqrt{\mathrm{MSE}}$	$\hat{\beta}_2/\hat{\beta}_1$	0.0563	0.0378	0.3122	0.2044	0.4106	0.2890
		$\hat{\beta}_3/\hat{\beta}_1$	0.0586	0.0386	0.2971	0.1977	0.4056	0.2837
		$\hat{\beta}_4/\hat{\beta}_1$	0.0540	0.0361	0.3075	0.2055	0.4191	0.2847
500	Mean	$\hat{\beta}_2/\hat{\beta}_1$	0.9978	0.9994	0.9526	0.9759	0.8995	0.9484
		$\hat{\beta}_3/\hat{\beta}_1$	1.0010	1.0004	0.9701	0.9854	0.9193	0.9625
		$\hat{\beta}_4/\hat{\beta}_1$	1.0020	1.0004	0.9452	0.9798	0.8994	0.9477
	$\sqrt{\mathrm{MSE}}$	$\hat{\beta}_2/\hat{\beta}_1$	0.0372	0.0207	0.1676	0.0975	0.2539	0.1558
		$\hat{\beta}_3/\hat{\beta}_1$	0.0329	0.0188	0.1663	0.0935	0.2587	0.1507
		$\hat{\beta}_4/\hat{\beta}_1$	0.0326	0.0184	0.1675	0.0925	0.2531	0.1505
1000	Mean	$\hat{\beta}_2/\hat{\beta}_1$	1.0015	1.0006	0.9994	1.0032	0.9728	0.9913
		$\hat{\beta}_3/\hat{\beta}_1$	1.0009	1.0007	1.0020	1.0026	0.9794	0.9946
		$\hat{\beta}_4/\hat{\beta}_1$	0.9993	1.0006	0.9980	1.0018	0.9756	0.9897
	$\sqrt{\mathrm{MSE}}$	$\hat{\beta}_2/\hat{\beta}_1$	0.0233	0.0124	0.1014	0.0515	0.1656	0.0905
		$\hat{\beta}_3/\hat{\beta}_1$	0.0247	0.0125	0.1017	0.0533	0.1672	0.0894
		$\hat{\beta}_4/\hat{\beta}_1$	0.0236	0.0123	0.1033	0.0520	0.1627	0.0885

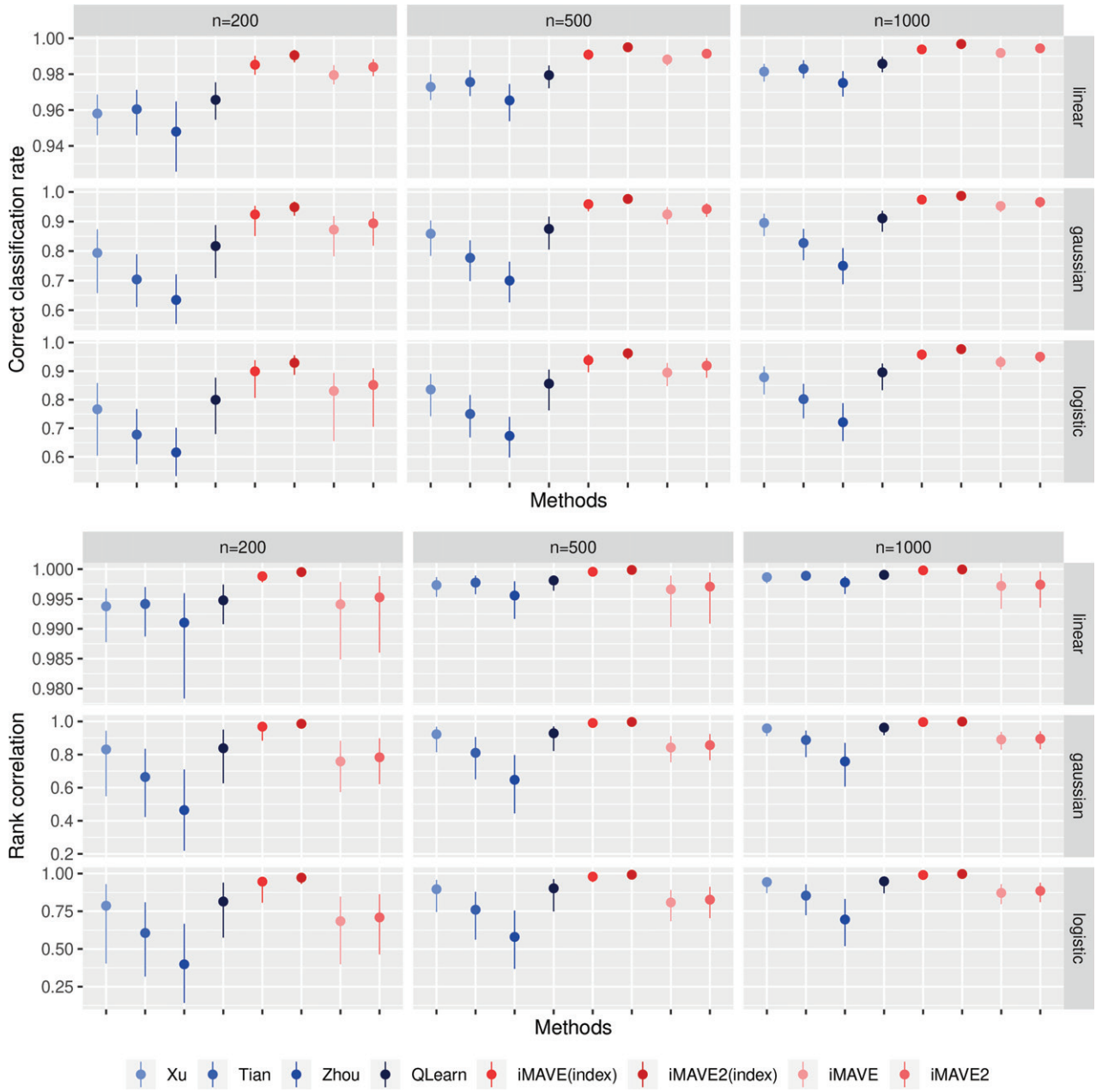

Figure 1. Simulation results for rank correlation and classification rate with *known* $\pi_T(\boldsymbol{X})$. The point represents the median, and the vertical line represents the range from the 0.25 to the 0.75 quantiles, of the results from 1000 simulations.

investigate the estimated effect modification in terms of correct classification rate and rank correlation over test datasets generated independently according to the true simulation model above but with sample sizes of 10,000. The rank correlation is determined by the fitted classifier and the true $g(\boldsymbol{\beta}_0^\top \boldsymbol{X})$ and the classification rate by their corresponding signs. For example, for iMAVE and iMAVE2, we evaluate the rank correlation between $\hat{g}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X})$ and $g(\boldsymbol{\beta}_0^\top \boldsymbol{X})$ and the concordance between $\hat{g}(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}) > 0$ and $g(\boldsymbol{\beta}_0^\top \boldsymbol{X}) > 0$ to determine the correct classification rate.

In our simulation setting where g is monotone and $g(0) = 0$, the sign of $g(\boldsymbol{\beta}_0^\top \boldsymbol{X})$ is also identical to that of $\boldsymbol{\beta}_0^\top \boldsymbol{X}$. In addition, the rank correlation between $g(\hat{\boldsymbol{\beta}}^\top \boldsymbol{X})$ and $g(\boldsymbol{\beta}_0^\top \boldsymbol{X})$ is also identical to that between $\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}$ and $\boldsymbol{\beta}_0^\top \boldsymbol{X}$. Because the resulting estimators of Tian et al. (2014), Xu et al. (2015), and Zhou et al. (2017) are parametric and target at the decision boundary $\boldsymbol{\beta}_0^\top \boldsymbol{X}$, we also include results of iMAVE(index) and iMAVE2(index) which compare the concordance between $\hat{\boldsymbol{\beta}}^\top \boldsymbol{X} > 0$ and $\boldsymbol{\beta}_0^\top \boldsymbol{X} > 0$ and the rank correlation between $\hat{\boldsymbol{\beta}}^\top \boldsymbol{X}$ and $\boldsymbol{\beta}_0^\top \boldsymbol{X}$ when g is monotone and $g(0) = 0$. This represents a more fair comparison with the parametric methods. Again, the index comparison only makes sense when g is monotone which is the case in our simulation setting.

From Figure 1, our methods have the best correct classification rates for the test datasets in all settings with known propensity score. When g is monotone and $g(0) = 0$, in terms of rank correlation, iMAVE2(index) is the best followed by iMAVE(index). The performances of iMAVE and iMAVE2 sacrifice slightly due to the estimation of g.

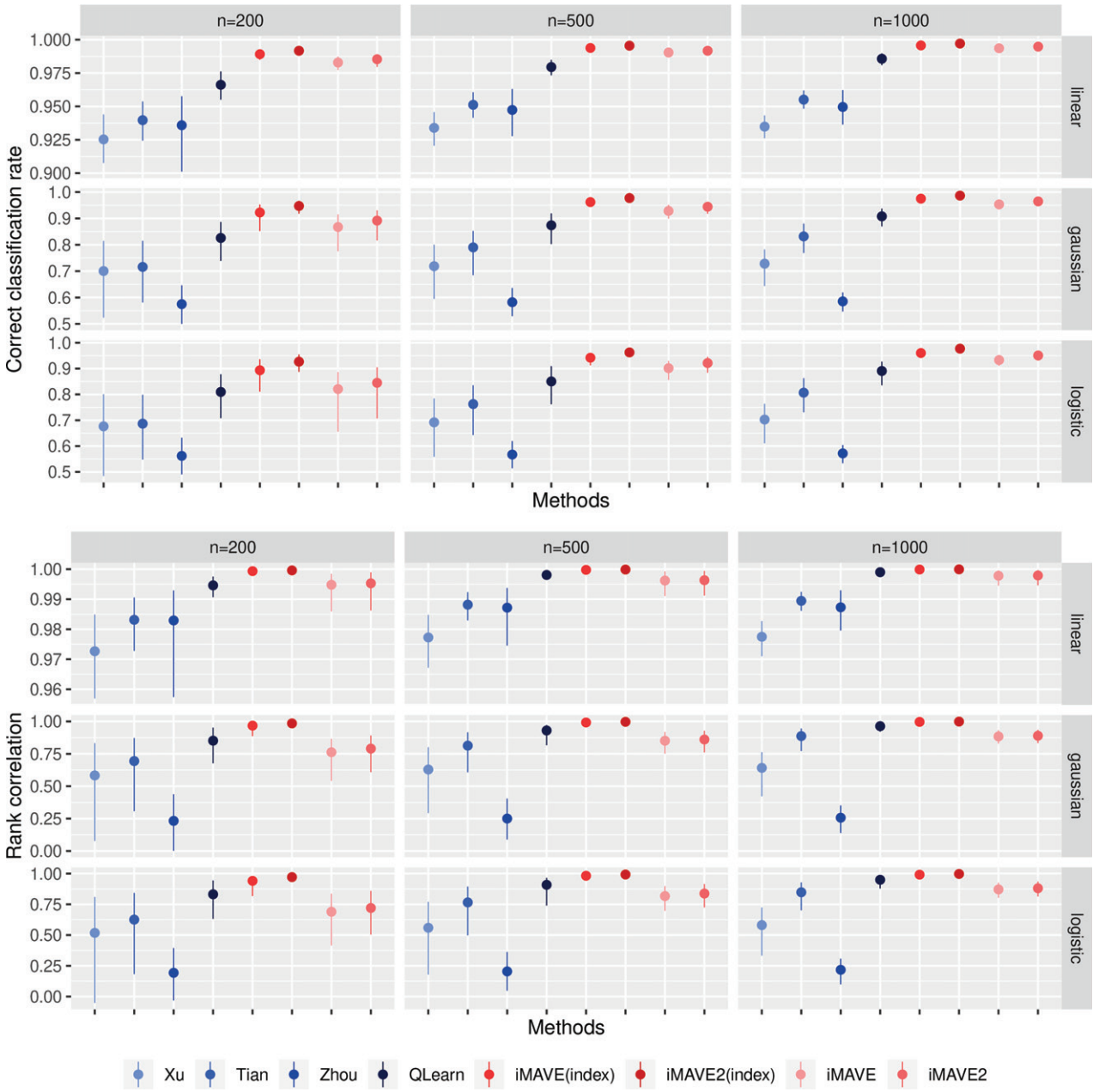

Figure 2. Simulation results for rank correlation and classification rate with *estimated* $\pi_T(\boldsymbol{X})$. The point represents the median, and the vertical line represents the range from the 0.25 to the 0.75 quantiles, of the results from 1000 simulations.

We further investigate the setting when $\pi_T(\boldsymbol{X})$ needs to be estimated. In this case, we generate T from a logistic model with coefficients $\tilde{\boldsymbol{\beta}} = (0.2, -0.2, 0.2, -0.2)^\top$ and then fit a logistic regression for $\pi_T(\boldsymbol{X})$. After estimating $\pi_T(\boldsymbol{X})$, all methods are implemented with the estimated $\pi_T(\boldsymbol{X})$. From Figure 2, our methods have the best correct classification rate and rank correlation than all other methods in all settings.

6.2. Dimension Determination

Here, we evaluate our dimension determination procedure through simulation. We follow Section 6.1 mostly except that we set $p = 10$ and the true $d = 2$. Consequently, the function g is

$$g(\boldsymbol{B}^\top \boldsymbol{X}) = \tau\{\Phi(\boldsymbol{\beta}_1^\top \boldsymbol{X}) - 0.5\} + \tau\{\Phi(\boldsymbol{\beta}_2^\top \boldsymbol{X}) - 0.5\},$$

where $\boldsymbol{\beta}_1 = (1, 1, 1, 1, 1, 1, 1, 1, 1, 1)^\top$ and $\boldsymbol{\beta}_2 = (1, -1, 1, -1, 1, -1, 1, -1, 1, -1)^\top$. We set $\gamma = 0.1$ and the sample size n is fixed at 500. Over 100 simulated datasets, our procedure was able to choose the correct dimension 2 for 72 times, 3 for 26 times, and 4 for 2 times. As we mentioned before, over-estimating the dimension slightly is not a big issue. There is no under-estimation of d, but slight over-estimation in some datasets.

7. Application to a Mammography Screening Study

This is a randomized study that included female subjects who were non-adherent to mammography screening guidelines at baseline (i.e., no mammogram in the year prior to baseline) (Champion et al. 2007). One primary interest of the study

was to compare the intervention effect of phone counseling on mammography screening (phone intervention) versus usual care at 21 months post-baseline. The outcome is whether a subject took mammography screening during this time period. There are 530 subjects with 259 in the phone intervention group and 271 in the usual care group. Baseline covariates include socio-demographics, health belief variables, stage of readiness to undertake mammography screening, and number of years had a mammogram in past 2–5 years in the study. In total, there are 211 covariates including second-order interactions among the covariates.

Our methods, together with our comparator methods (Tian et al. 2014; Xu et al. 2015; Zhou et al. 2017), were applied to this dataset. To compare the results of the estimated treatment assignment rules, we used the following metrics. An assignment rule $T(\boldsymbol{X})$ refers to a mapping from $\boldsymbol{X}$ to $\{1, -1\}$. For example, in our model set up with $\Delta(\boldsymbol{X}) = E[Y|T = 1, \boldsymbol{X}] - E[Y|T = -1, \boldsymbol{X}] = g(\boldsymbol{\beta}^\top \boldsymbol{X})$, the assignment rule that maximizes the expected value of the outcome is $T(\boldsymbol{X}) = 1\{g(\boldsymbol{\beta}^\top \boldsymbol{X}) > 0\}$. For a fitted assignment rule, say $\hat{T}(\boldsymbol{X})$, the following two quantities are used to evaluate the performances.

$$E[\Delta_1] = E[Y|\hat{T}(\boldsymbol{X}) = 1, T = 1] - E[Y|\hat{T}(\boldsymbol{X}) = 1, T = -1],$$

and,

$$E[\Delta_{-1}] = E[Y|\hat{T}(\boldsymbol{X}) = -1, T = -1] - E[Y|\hat{T}(\boldsymbol{X}) = -1, T = 1].$$

They represent gains in the outcome expectations between the recommendation agreeing and disagreeing subgroups. If both $E[\Delta_{-1}]$ and $E[\Delta_1]$ are positive, then the estimated treatment decision rule can improve the outcome.

The actual evaluation was based on cross-validation. First, 80% of subjects were randomly selected into a training set and the rest into a testing set. Apparently, due to this further reduction of sample size, we had to reduce the number of covariates for fitting. We performed screening procedures for all methods in a uniform fashion. In particular, the method of Tian et al. (2014) with lasso penalty was fitted on the training sets for variable selection. After variable selection, the selected covariates were fitted by each method. For iMAVE and iMAVE2, dimension selection from $d = 1, 2, 3$ was also implemented. Then, the benefit quantities defined above were calculated on the testing set. The cross-validation was based on 100 splits. The SDs in Table 2 refer to the standard deviations of $\hat{E}[\Delta_1]$ and $\hat{E}[\Delta_{-1}]$ from these 100 repeats. In Table 2, our methods seem to have advantages as they lead to larger $\hat{E}[\Delta_1]$ and $\hat{E}[\Delta_{-1}]$. The average percentages of subjects assigned to $T = 1$ and -1 in the test sets are also given in the table. A list of the top selected variables by the screening method is provided in the supplementary materials.

Table 2. Results for the mammography screening study from 100 cross-validations.

	$\hat{E}[\Delta_1]$		$\hat{E}[\Delta_{-1}]$	
Method	Mean (SD)	Avg % of subj in $T = 1$	Mean (SD)	Avg % of subj to $T = -1$
iMAVE	0.032(0.014)	42%	0.052(0.012)	58%
iMAVE2	0.036(0.014)	42%	0.054(0.012)	58%
Tian	0.022(0.013)	44%	0.043(0.011)	56%
Xu	0.026(0.012)	43%	0.044(0.012)	57%
Zhou	0.020(0.013)	41%	0.041(0.011)	59%
QLearn	0.018(0.012)	33%	0.022(0.011)	67%

8. Discussion

In this article, we have proposed a very general semiparametric modeling framework for effect modification estimation. Whereas our main motivational setting is from precision medicine, the framework is generally applicable to statistical interaction discovery with interested variables in many other settings. For example in health disparities research, a complex and interrelated set of individual, provider, health system, societal, and environmental factors contribute to disparities in health and health care. Federal efforts to reduce disparities often include a focus on designated priority populations who are particularly vulnerable to health and health care disparities. Our approach seems ideal for data analysis in this setting.

When there are many covariates, we have focused on dimension reduction. In high-dimensional settings, variable screening may be needed to reduce the number of covariates. Various methods can be applied in our framework. For example, because $E[TY/\pi_T|\boldsymbol{X}] = g(\boldsymbol{\beta}^\top \boldsymbol{X})$, we can implement a nonparametric variable screening method such as the distance correlation based approach (Li, Zhong, and Zhu 2012). Alternatively, regression with penalty for variable selection such as lasso can be used (Tian et al. 2014; Xu et al. 2015). Ideally, one could also incorporate variable selection into our framework when the dimension d is fixed. In particular, lasso type of regularization can be used together with our estimating equations. This can be a fruitful path for future work as variable selection is an important practical issue.

Supplementary Materials

Estimation with multiple level treatments or exposures, proofs of Theorems 3.1–5.5, additional simulation results, and supplemental results for the mammography screening study are contained in the supplementary materials.

Funding

Research reported in this article was partially funded through a Patient-Centered Outcomes Research Institute (PCORI) Award (ME-1409-21219). The views in this publication are solely the responsibility of the authors and do not necessarily represent the views of the PCORI, its Board of Governors or Methodology Committee.

References

Abrevaya, J., Hsu, Y. C., and Lieli, R. P. (2015), "Estimating Conditional Average Treatment Effects," *Journal of Business & Economic Statistics*, 33, 485–505. [753]

Athey, S., and Wager, S. (2017), "Efficient Policy Learning," Stanford Institute for Economic Policy Research Working Paper 17-031, available at *https://siepr.stanford.edu/sites/default/files/publications/17-031.pdf*. [758]

Bang, H., and Robins, J. M. (2005), "Doubly Robust Estimation in Missing Data and Causal Inference Models," *Biometrics*, 61, 962–972. [753]

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, Baltimore, MD: The Johns Hopkins University Press. [753]

Braveman, P. (2006), "Health Disparities and Health Equity: Concepts and Measurement," *Annual Review of Public Health*, 27, 167–194. [752]

Cao, W., Tsiatis, A. A., and Davidian, M. (2009), "Improving Efficiency and Robustness of the Doubly Robust Estimator for a Population Mean With Incomplete Data," *Biometrika*, 96, 723–734. [753]

Chamberlain, G. (1987), "Asymptotic Efficiency in Estimation With Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334. [753]

Champion, V., Skinner, C. S., Hui, S., Monahan, P., Juliar, B., Daggy, J., and Menon, U. (2007), "The Effect of Telephone v. Print Tailoring for Mammography Adherence," *Patient Education and Counseling*, 65, 416. [761]

Chen, S., Tian, L., Cai, T., and Yu, M. (2017), "A General Statistical Framework for Subgroup Identification and Comparative Treatment Scoring," *Biometrics*, 73, 1199–1209. [752,756]

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., and Robins, J. (2018), "Double/Debiased Machine Learning for Treatment and Structural Parameters," *The Econometrics Journal*, 21, C1–C68. [758]

Cook, R. D. (1998), "Principal Hessian Directions Revisited," *Journal of the American Statistical Association*, 93, 84–94. [757]

——— (2007), "Fisher Lecture: Dimension Reduction in Regression," *Statistical Science*, 22, 1–26. [753]

Green, D. P., and Kern, H. L. (2012), "Modeling Heterogeneous Treatment Effects in Survey Experiments With Bayesian Additive Regression Trees," *Public Opinion Quarterly*, 76, 491–511. [753]

Greenland, S. (1993), "Basic Problems in Interaction Assessment," *Environmental Health Perspectives*, 101, 59–66. [752]

Hirano, K., and Imbens, G. W. (2001), "Estimation of Causal Effects Using Propensity Score Weighting: An Application to Data on Right Heart Catheterization," *Health Services and Outcomes Research Methodology*, 2, 259–278. [753]

Hirano, K., Imbens, G. W., and Ridder, G. (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71, 1161–1189. [753]

Huang, M.-Y., and Chan, K. C. G. (2017), "Joint Sufficient Dimension Reduction and Estimation of Conditional and Average Treatment Effects," *Biometrika*, 104, 583–596. [753]

Imai, K., and Ratkovic, M. (2013), "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation," *The Annals of Applied Statistics*, 7, 443–470. [753]

Imbens, G. W., and Rubin, D. B. (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, New York: Cambridge University Press. [753]

Koch, I., and Naito, K. (2007), "Dimension Selection for Feature Selection and Dimension Reduction With Principal and Independent Component Analysis," *Neural Computation*, 19, 513–545. [757]

Kraemer, H. C. (2013), "Discovering, Comparing, and Combining Moderators of Treatment on Outcome After Randomized Clinical Trials: A Parametric Approach," *Statistics in Medicine*, 32, 1964–1973. [752]

Künzel, S. R., Sekhon, J. S., Bickel, P. J., and Yu, B. (2019), "Metalearners for Estimating Heterogeneous Treatment Effects Using Machine Learning," *Proceedings of the National Academy of Sciences of the United States of America*, 116, 4156–4165. [753]

Li, B. (2018), *Sufficient Dimension Reduction: Methods and Applications With R*, Monographs on Statistics and Applied Probability, Boca Raton, FL: Chapman & Hall/CRC. [753]

Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening via Distance Correlation Learning," *Journal of the American Statistical Association*, 107, 1129–1139. [762]

Lou, Z., Shao, J., and Yu, M. (2018), "Optimal Treatment Assignment to Maximize Expected Outcome With Multiple Treatments," *Biometrics*, 74, 506–516. [753]

Lu, M., Sadiq, S., Feaster, D. J., and Ishwaran, H. (2018), "Estimating Individual Treatment Effect in Observational Data Using Random Forest Methods," *Journal of Computational and Graphical Statistics*, 27, 209–219. [753]

Lu, W., Zhang, H. H., and Zeng, D. (2013), "Variable Selection for Optimal Treatment Decision," *Statistical Methods in Medical Research*, 22, 493–504. [752]

Luo, W., Zhu, Y., and Ghosh, D. (2017), "On Estimating Regression-Based Causal Effects Using Sufficient Dimension Reduction," *Biometrika*, 104, 51–65. [753]

Ma, Y., and Zhu, L. (2012), "A Semiparametric Approach to Dimension Reduction," *Journal of the American Statistical Association*, 107, 168–179. [753,754]

——— (2013), "Efficient Estimation in Sufficient Dimension Reduction," *The Annals of Statistics*, 41, 250–268. [753,754]

——— (2014), "On Estimation Efficiency of the Central Mean Subspace," *Journal of the Royal Statistical Society*, Series B, 76, 885–901. [754]

Newey, W. K. (2004), "Efficient Semiparametric Estimation via Moment Restrictions," *Econometrica*, 72, 1877–1897. [753]

Persson, E., Hggstrm, J., Waernbaum, I., and de Luna, X. (2017), "Data-Driven Algorithms for Dimension Reduction in Causal Inference," *Computational Statistics and Data Analysis*, 105, 280–292. [753]

Qian, M., and Murphy, S. A. (2011), "Performance Guarantees for Individualized Treatment Rules," *The Annals of Statistics*, 39, 1180. [759]

Robins, J. M. (1994), "Correcting for Non-Compliance in Randomized Trials Using Structural Nested Mean Models," *Communications in Statistics—Theory and Methods*, 23, 2379–2412. [753]

Robins, J. M., Mark, S. D., and Newey, W. K. (1992), "Estimating Exposure Effects by Modelling the Expectation of Exposure Conditional on Confounders," *Biometrics*, 48, 479–495. [753]

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994), "Estimation of Regression Coefficients When Some Regressors Are Not Always Observed," *Journal of the American Statistical Association*, 89, 846–866. [753]

Rosenbaum, P. R., and Rubin, D. B. (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–55. [754]

Rotnitzky, A., Lei, Q., Sued, M., and Robins, J. M. (2012), "Improved Double-Robust Estimation in Missing Data and Causal Inference Models," *Biometrika*, 99, 439–456. [753]

Rubin, D. B. (1974), "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies," *Journal of Educational Psychology*, 66, 688. [753]

——— (2005), "Causal Inference Using Potential Outcomes," *Journal of the American Statistical Association*, 100, 322–331. [753]

Schott, J. R. (1994), "Determining the Dimensionality in Sliced Inverse Regression," *Journal of the American Statistical Association*, 89, 141–148. [757]

Song, R., Luo, S., Zeng, D., Zhang, H. H., Lu, W., and Li, Z. (2017), "Semiparametric Single-Index Model for Estimating Optimal Individualized Treatment Strategy," *Electronic Journal of Statistics*, 11, 364–384. [752,756,759]

Tan, Z. (2010), "Bounded, Efficient and Doubly Robust Estimation With Inverse Weighting," *Biometrika*, 97, 661–682. [753]

Tian, L., Alizadeh, A. A., Gentles, A. J., and Tibshirani, R. (2014), "A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates," *Journal of the American Statistical Association*, 109, 1517–1532. [752,756,759,760,762]

Tsiatis, A. A. (2006), *Semiparametric Theory and Missing Data* (1st ed.), New York: Springer. [753,754,755]

Vansteelandt, S., and Joffe, M. (2014), "Structural Nested Models and g-Estimation: The Partially Realized Promise," *Statistical Science*, 29, 707–731. [753]

Vapnik, V. (2000), *The Nature of Statistical Learning Theory*, New York: Springer. [757]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. [753]

Xia, Y. (2007), "A Constructive Approach to the Estimation of Dimension Reduction Directions," *The Annals of Statistics*, 35, 2654–2690. [753,757,758]

——— (2008), "A Multiple-Index Model and Dimension Reduction," *Journal of the American Statistical Association*, 103, 1631–1640. [753]

Xia, Y., and Hardle, W. (2006), "Semi-Parametric Estimation of Partially Linear Single-Index Models," *Journal of Multivariate Analysis*, 97, 1162–1184. [753]

Xia, Y., Tong, H., Li, W. K., and Zhu, L.-X. (2002), "An Adaptive Estimation of Dimension Reduction Space," *Journal of the Royal Statistical Society*, Series B, 64, 363–410. [753,755,756,757]

Xie, Y., Brand, J. E., and Jann, B. (2012), "Estimating Heterogeneous Treatment Effects With Observational Data," *Sociological Methodology*, 42, 314–347. [753]

Xu, Y., Yu, M., Zhao, Y. Q., Li, Q., Wang, S., and Shao, J. (2015), "Regularized Outcome Weighted Subgroup Identification for Differential Treatment Effects," *Biometrics*, 71, 645–653. [752,759,760,762]

Zhang, B., Tsiatis, A. A., Davidian, M., Zhang, M., and Laber, E. (2012), "Estimating Optimal Treatment Regimes From a Classification Perspective," *Stat*, 1, 103–114. [752]

Zhao, Y., Laber, E. B., Ning, Y., Saha, S., and Sands, B. E. (2019), "Efficient Augmentation and Relaxation Learning for Individualized Treatment Rules Using Observational Data," *Journal of Machine Learning Research*, 20, 1–23. [758]

Zhao, Y., Zeng, D., Rush, A. J., and Kosorok, M. R. (2012), "Estimating Individualized Treatment Rules Using Outcome Weighted Learning," *Journal of the American Statistical Association*, 107, 1106–1118. [752]

Zhou, X., Mayer-Hamblett, N., Khan, U., and Kosorok, M. R. (2017), "Residual Weighted Learning for Estimating Individualized Treatment Rules," *Journal of the American Statistical Association*, 112, 169–187. [759,760,762]