Submitted to the Annals of Statistics

A ROBUST AND EFFICIENT APPROACH TO CAUSAL INFERENCE BASED ON SPARSE SUFFICIENT DIMENSION REDUCTION

BY SHUJIE MA[†,*], LIPING ZHU[‡,†], ZHIWEI ZHANG[§], CHIH-LING TSAI[¶] AND RAYMOND J. CARROLL[‖]

University of California at Riverside[†] , Renmin University of China[‡] , University of California at Riverside[§] , University of California at Davis[¶] , and Texas A&M University[‖]

A fundamental assumption used in causal inference with observational data is that treatment assignment is ignorable given measured confounding variables. This assumption of no missing confounders is plausible if a large number of baseline covariates are included in the analysis, as we often have no prior knowledge of which variables can be important confounders. Thus, estimation of treatment effects with a large number of covariates has received considerable attention in recent years. Most existing methods require specifying certain parametric models involving the outcome, treatment and confounding variables, and employ a variable selection procedure to identify confounders. However, selection of a proper set of confounders depends on correct specification of the working models. The bias due to model misspecification and incorrect selection of confounding variables can yield misleading results. We propose a robust and efficient approach for inference about the average treatment effect via a flexible modeling strategy incorporating penalized variable selection. Specifically, we consider an estimator constructed based on an efficient influence function that involves a propensity score and an outcome regression. We then propose a new sparse sufficient dimension reduction method to estimate these two functions without making restrictive parametric modeling assumptions. The proposed estimator of the average treatment effect is asymptotically normal and semiparametrically efficient without the need for variable selection consistency. The proposed methods are illustrated via simulation studies and a biomedical application.

[*]Ma's research was supported by NSF grant DMS-1712558 and NIH grant R01 ES024732-03. Zhu's work was supported by NNSFC grant 11731011 and National Youth Top-notch Talent Support Program, China. Carroll's research was supported by a grant from the National Cancer Institute (U01-CA057030).

MSC 2010 subject classifications: Primary 62G08; secondary 62G10, 62G20, 62J07

Keywords and phrases: Average treatment effect, Dimension reduction, High dimensional data, Multiple-index model, Outcome regression, Semiparametric efficiency

1. Introduction. Causal inference in observational studies is challenged by the fact that treatment assignment may depend on some baseline covariates known as confounding variables that are also associated with the outcome of interest. Most existing methods for causal inference can be cast in terms of potential outcomes under Rubin's causal model (Rubin, 1974). A fundamental assumption is that treatment assignment is strongly ignorable, i.e., conditionally independent of potential outcomes given measured confounders; see Rosenbaum and Rubin (1983). A common approach to understanding causality is to adjust for confounding in a regression model that relates the outcome to the treatment under investigation. This outcome regression (OR) approach is straightforward to implement and its validity depends on correct specification of the OR model. In contrast, many alternative methods require a model for the propensity score (PS), that is, the conditional probability of being treated given the covariates (Rosenbaum and Rubin, 1983). The estimated PS can be used to match each member of the treated group with one or more subjects in the untreated group, stratify the sample so that the resulting two groups are more comparable in each stratum, or weight each observation by the inverse of the estimated PS, or one minus it, depending on the actual treatment (Rosenbaum and Rubin, 1984, 1985; Heckman et al., 1998; Robins et al., 2000; Hirano et al., 2003; Abadie and Imbens, 2006). It is of interest to note that much of the recent research has focused on doubly robust (DR) estimation that encompasses both OR and PS models so that the resulting estimators are consistent and asymptotically normal if either model is correctly specified (e.g., van der Laan and Robins, 2003; Bang and Robins, 2005; Tan, 2006; Freedman and Berk, 2008; Cao et al., 2009; Tan, 2010; van der Laan and Rose, 2011; Rotnitzky et al., 2012; Chan and Yam, 2014).

For the sake of simplicity, it is often assumed that the PS and OR models are parametric. However, parametric models may be misspecified, resulting in asymptotically biased estimators with poor finite sample performance. On the other hand, the DR estimators are relatively robust against model misspecification. Yet they would not be efficient if one of the two models is misspecified, and they could perform rather poorly when both models are misspecified (Kang and Schafer, 2007; Freedman and Berk, 2008). Hence, it is desirable to work with less restrictive PS and OR models. In practice, there can be a large collection of potential confounding variables, of which only a few have to be adjusted. This leads to variable selection in regression for causal inference. To this end, Belloni et al. (2014) and Farrell (2015) proposed penalized estimation procedures for estimating linear PS and OR models for high-dimensional data. Selection of a proper set of confounding

variables depends on correct specification of the working models. With a large sample size, the bias due to model misspecification and incorrect selection of confounding variables becomes pronounced in comparison to sampling variability, and may lead to statistically significant false findings. Thus, to conduct robust and efficient causal inference, it is essential to employ a flexible modeling strategy that incorporates variable selection. The need for such a strategy is particularly crucial in analyzing big data, which frequently involve a large number of variables measured on a large number of subjects. On the other hand, big data, which often involve a large sample size, present an opportunity to employ state-of-the-art methods for dimension reduction and nonparametric regression to achieve a good balance between flexibility and parsimony of statistical modeling.

In this article, we propose a sparse sufficient dimension reduction (SSDR) method to estimate the PS and OR models. It is known that sufficient dimension reduction (Li, 1991; Cook and Li, 2002) provides a general and effective way to reduce the dimension of covariates while preserving information on regression. We employ multiple-index models with a small number of linear combinations of relevant covariates to estimate PS and OR. Multiple-index models are flexible and contain various parametric and semiparametric models as special cases (Yin et al., 2008; Xia, 2008), yet their estimation is challenging, especially in the high-dimensional setting. To accomplish this difficult task, we show that estimation of the directions in a multiple-index model is equivalent to finding vectors that span the same subspace as the left-singular vectors of the low-rank coefficient matrix in a sparse reduced-rank regression problem. We then use sparsity-inducing penalization to select relevant covariates with group Lasso penalties (Yuan and Lin, 2006), and employ an Iterative Shrinkage and Thresholding algorithm for parameter estimation. Our proposed method is able to identify important confounders from a large number of candidate variables and characterize their roles in treatment assignments and outcome predictions, without making more restrictive parametric modeling assumptions.

To relax the assumptions on parametric forms, Hahn (1998) proposed a nonparametric estimator of the average treatment effect (ATE) for low-dimensional covariates. The resulting estimator attains the semiparametric information bound. It, however, suffers from the "curse of dimensionality" with increasing dimension. To alleviate the problem of dimensionality, Luo et al. (2017) applied minimum average variance estimation (Xia, et al., 2002, MAVE) to recover the OR function, and Ghosh (2011) employed a single-index model, together with sufficient dimension and partial least squares methods, to estimate the PS function. These two methods allow us to es-

timate ATE via a flexible modeling strategy, yet their computational algorithms and the associated theories are developed for fixed dimensions. Nevertheless, our proposed estimator can be used for data with both moderate and high dimensions. Specifically, we develop a DR estimation method for ATE by making use of the SSDR estimates for the PS and OR functions. The resulting estimator of ATE is shown to be root-n consistent, asymptotically normal and efficient with high-dimensional covariates. These properties hold without the requirement of variable selection consistency and restrictive parametric modeling assumptions, which is remarkable because post-selection inference is known to be a challenging task in general (Wasserman and Roeder, 2009; Berk et al., 2013; Lockhart et al., 2014; van de Geer et al., 2014; Zhang and Zhang, 2014). In the context of causal inference, Belloni et al. (2014) demonstrated that standard penalized methods such as Lasso can lead to the biased estimation of ATE. The construction of the DR estimator, based on the efficient influence function, can mitigate the bias of the Lasso estimator and allows for imperfect variable selection.

The rest of the paper is organized as follows. Section 2 introduces the proposed sparse sufficient dimension reduction method. Section 3 describes PS and OR, and introduces the DR estimation. Section 4 presents the proposed estimator of ATE. Section 5 establishes the theoretical properties of the proposed estimator and Section 6 provides the computational algorithm. In Section 7, we evaluate the finite sample performance via simulation studies. An empirical example is reported in Section 8. Concluding remarks are given in Section 9, and all technical proofs are provided in the Appendix and on-line supplemental materials Ma et al. (2018).

2. Sparse sufficient dimension reduction. We first introduce the following notation which will be used frequently in this paper. For positive a_n and b_n, let $b_n \ll a_n$ denote $a_n^{-1} b_n = o(1)$ and $b_n \asymp a_n$ denote $\lim_{n\to\infty} a_n^{-1} b_n = c$ for a positive constant c. Moreover, let $a_n \vee b_n = \max(a_n, b_n)$. For a vector $\mathbf{a} = (a_1, \ldots, a_p)^\top$, define $||\mathbf{a}||_\infty = \max(|a_i|)$. For a matrix $\mathbf{A} = (A_{jk}) = (\mathbf{A}_1, \ldots, \mathbf{A}_p)^\top = (\mathbf{A}_{\cdot 1}, \ldots, \mathbf{A}_{\cdot q}) \in \mathbb{R}^{p\times q}$, let $||\mathbf{A}||^2 = \sum\sum A_{jk}^2$, $\mathrm{vec}(\mathbf{A}) = (\mathbf{A}_{\cdot 1}^\top, \ldots, \mathbf{A}_{\cdot q}^\top)^\top$, and $\mathrm{span}(\mathbf{A})$ be the subspace of $\mathbb{R}^q$ spanned by the columns of $\mathbf{A}$. For a subset $\mathcal{S} \subseteq \{1, \ldots, p\}$, let $\mathbf{A}_\mathcal{S}$ be the submatrix of $\mathbf{A}$ associated with the row indices $\mathcal{S}$. For a subset $\mathcal{B} \subseteq \{1, \ldots, q\}$, let $\mathbf{A}_{\cdot\mathcal{B}}$ be the submatrix of $\mathbf{A}$ associated with the column indices $\mathcal{B}$. For a symmetric matrix $\mathbf{A}$, let $\lambda_{\min}(\mathbf{A})$ denote the smallest eigenvalue of $\mathbf{A}$. Denote $|\mathcal{S}|$ as the cardinality of a set $\mathcal{S}$.

For DR estimation, it is essential to obtain good estimates for the PS and OR functions. Assuming a restrictive parametric form on these two

functions can lead to large biases due to possible model misspecification. On the other hand, directly estimating them via classical nonparametric regression is difficult when the dimension of covariates is high. To achieve modeling flexibility with high-dimensional covariates, we propose a SSDR method to estimate them. We denote by Z a generic response of interest and X a vector of p-dimensional covariates. Let $X_i = (X_{i1}, \ldots, X_{ip})^\top$ be a vector of covariates and let $(Z_i, X_i^\top)^\top$, $i = 1, \ldots, n$, be independent and identically distributed (i.i.d.) samples from $(Z, X^\top)^\top$. Our interest is to estimate the conditional expectation $E(Z_i \mid X_i)$. To facilitate subsequent illustrations, we assume that $E(X_{ik}) = 0$ and $\operatorname{var}(X_{ik}) = 1$ for $1 \le k \le p$. Denote $X_{i,\mathcal{S}} = (X_{ik}, k \in \mathcal{S})^\top$ and $\mathbf{X}_{\cdot\mathcal{S}} = (X_{1,\mathcal{S}}, \ldots, X_{n,\mathcal{S}})^\top$. Without loss of generality, let $X_i = (X_{i,\mathcal{R}}^\top, X_{i,\mathcal{I}}^\top)^\top$, where $\mathcal{R}$ and $\mathcal{I}$ are the sets of indices of relevant and irrelevant covariates, respectively, for $E(Z_i \mid X_i)$.

We consider a SSDR model in which the conditional mean $E(Z_i \mid X_i)$ depends on r linear combinations of the relevant covariates, so that we have the sparse multiple-index model:

$$E(Z_i \mid X_i) = E(Z_i \mid X_{i,\mathcal{R}}) = E(Z_i \mid \mathbf{B}_{\mathcal{R}}^\top X_{i,\mathcal{R}}), \tag{1}$$

where $\mathbf{B}_{\mathcal{R}}$ is an $|\mathcal{R}| \times r$ matrix of unknown parameters with $r \le |\mathcal{R}|$. Model (1) implies that the $|\mathcal{R}|$-dimensional vector of relevant covariates can be replaced by the r-dimensional vector $\mathbf{B}_{\mathcal{R}}^\top X_{i,\mathcal{R}}$ without loss of information in the mean regression. Let $\mathbf{B} = (\mathbf{B}_{\mathcal{R}}^\top, \mathbf{B}_{\mathcal{I}}^\top)^\top = (\mathbf{B}_{\mathcal{R}}^\top, \mathbf{0}_{(p-|\mathcal{R}|)\times r}^\top)^\top$, indicating that the coefficients of irrelevant covariates are zero. Thus model (1) can be written as

$$E(Z_i \mid X_i) = E(Z_i \mid \mathbf{B}_{\mathcal{R}}^\top X_{i,\mathcal{R}}) = E(Z_i \mid \mathbf{B}^\top X_i).$$

We next assume that

(A1) X_i, $1 \le i \le n$, are i.i.d. observations from the multivariate normal distribution $\mathcal{N}(\mathbf{0},\ \mathbf{\Sigma})$.

For the sake of shortening proofs, we make the above assumption on the distribution of covariates, and it can be relaxed to the linearity condition jointly with the constant variance condition (Cook and Lee, 1999; Duan and Li, 1991; Li, 1992). The same assumption as (A1) is also given in van de Geer et al. (2014) for studying the de-biased Lasso estimators.

Let $\widetilde{Z}_i = Z_i - E(Z_i)$. Under Assumption (A1), Duan and Li (1991) showed that $\operatorname{span}\{\mathbf{\Sigma}^{-1}E(\widetilde{Z}_i X_i)\} \subseteq \operatorname{span}(\mathbf{B})$. Subsequently, Li (1992) and Cook and Li (2002) employed principal Hessian directions (pHd) to further demonstrate that $\operatorname{span}\{\mathbf{\Sigma}^{-1}E(\widetilde{Z}_i X_i X_i^\top)\} \subseteq \operatorname{span}(\mathbf{B})$. These two results imply that $\operatorname{span}(\mathbf{\Sigma}^{-1}\mathbf{\Lambda}) \subseteq \operatorname{span}(\mathbf{B})$, where $\mathbf{\Lambda} \equiv E(\widetilde{Z}_i X_i \widetilde{X}_i^\top) \in \mathbb{R}^{p\times(p+1)}$ and

$\widetilde{X}_i = (1, X_i^\top)^\top$. With the coverage assumption that $\text{span}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}) = \text{span}(\mathbf{B})$ (Li, 1992; Cook and Li, 2002), $\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda}$ is a matrix with rank r that can be written as $\mathbf{V}^0\mathbf{A}^{0\top}$, where $\mathbf{V}^0$ and $\mathbf{A}^0$ are $p \times r$ and $(p+1) \times r$ matrices, and $\mathbf{A}^0$ satisfies $\mathbf{A}^{0\top}\mathbf{A}^0 = \mathbf{I}$. In addition, $\text{span}(\mathbf{V}^0) = \text{span}(\mathbf{B})$ implying $\mathbf{V}^0_{\mathcal{I}} = \mathbf{0}$. Thus, we propose to recover $\text{span}(\mathbf{B})$ using $\mathbf{V}^0$.

Note that $\mathbf{V}^0$ and $\mathbf{A}^0$ can be obtained through minimizing

$$E||\widetilde{\mathbf{W}} - \mathbf{X}\mathbf{V}\mathbf{A}^\top||^2$$

subject to $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$ and $\mathbf{V}_{\mathcal{I}} = \mathbf{0}_{(p-|\mathcal{R}|)\times r}$, where $\widetilde{W}_i = \widetilde{X}_i\widetilde{Z}_i$, $\widetilde{\mathbf{W}} = (\widetilde{W}_1, \ldots, \widetilde{W}_n)^\top$, $\mathbf{X} = (X_1, \ldots, X_n)^\top$, and $\mathbf{A}$ is a $(p+1) \times r$ matrix. In practice, we use the empirical version of $\widetilde{Z}_i$ for estimation. Estimation of $\mathbf{V}^0$ and $\mathbf{A}^0$ is a sparse reduced-rank regression (SRRR) problem (Chen and Huang, 2012). For our purpose, however, we only need to obtain an estimate of $\mathbf{V}$, which satisfies $\text{span}(\mathbf{V}) = \text{span}(\mathbf{V}^0)$. Indeed, for any given $\mathbf{A}$ satisfying $\mathbf{A}^\top\mathbf{A} = \mathbf{I}$, there is a matrix $\mathbf{A}^\perp$ with orthonormal columns such that $(\mathbf{A}, \mathbf{A}^\perp)$ is an orthogonal matrix. Accordingly, we have

$$||\widetilde{\mathbf{W}} - \mathbf{X}\mathbf{V}\mathbf{A}^\top||^2 = ||\widetilde{\mathbf{W}}\mathbf{A} - \mathbf{X}\mathbf{V}||^2 + ||\widetilde{\mathbf{W}}\mathbf{A}^\perp||^2.$$

Thus, for any given $\mathbf{A}^*$ satisfying $\mathbf{A}^{*\top}\mathbf{A}^* = \mathbf{I}$,

$$\begin{aligned}
\mathbf{V}^* &= \arg \min_{\mathbf{V}\in\mathbb{R}^{p\times r}, \mathbf{V}_{\mathcal{I}}=\mathbf{0}_{(p-|\mathcal{R}|)\times r}} E||\widetilde{\mathbf{W}} - \mathbf{X}\mathbf{V}\mathbf{A}^{*\top}||^2 \\
&= \arg \min_{\mathbf{V}\in\mathbb{R}^{p\times r}, \mathbf{V}_{\mathcal{I}}=\mathbf{0}_{(p-|\mathcal{R}|)\times r}} E||\widetilde{\mathbf{W}}\mathbf{A}^* - \mathbf{X}\mathbf{V}||^2 \\
&= \arg \min_{\mathbf{V}\in\mathbb{R}^{p\times r}, \mathbf{V}_{\mathcal{I}}=\mathbf{0}_{(p-|\mathcal{R}|)\times r}} E||\mathbf{X}\mathbf{V}^0\mathbf{A}^{0\top}\mathbf{A}^* - \mathbf{X}\mathbf{V}||^2 = \mathbf{V}^0\mathbf{A}^{0\top}\mathbf{A}^*.
\end{aligned}$$

The above equation indicates that $\text{span}(\mathbf{V}^*) = \text{span}(\mathbf{V}^0)$ as long as $\mathbf{A}^{0\top}\mathbf{A}^*$ is a full rank matrix. Thus, $\text{span}(\mathbf{V}^*) = \text{span}(\mathbf{V}^0) = \text{span}(\mathbf{B})$, and

$$E(Z_i \mid X_i) = E(Z_i \mid \mathbf{B}^\top X_i) = E(Z_i \mid \mathbf{V}^{*\top} X_i). \tag{2}$$

Based on the above discussion, we make the following assumption for a given $\mathbf{A}^*$.

(A2) (i) $\mathbf{A}^{*\top}\mathbf{A}^* = \mathbf{I}$ and (ii) $\mathbf{A}^{0\top}\mathbf{A}^*$ is a full rank matrix.

Assumption (A2) on $\mathbf{A}^*$ is needed for model identification as explained above. Without (A2), the column space $\text{span}(\mathbf{V}^*)$ is not identifiable. Since

$$\mathbf{V}^* = \arg \min_{\mathbf{V}\in\mathbb{R}^{p\times r}, \mathbf{V}_{\mathcal{I}}=\mathbf{0}_{(p-|\mathcal{R}|)\times r}} E||\widetilde{\mathbf{W}}\mathbf{A}^* - \mathbf{X}\mathbf{V}||^2, \tag{3}$$

the estimator of $\mathbf{V}^*$ can be obtained by adopting a group Lasso penalized approach (Yuan and Lin, 2006). Specifically, for the given $\mathbf{A}^*$, we can obtain the estimator $\widehat{\mathbf{V}}$ of $\mathbf{V}^*$ by minimizing

$$(1/2)||\widetilde{\mathbf{W}}\mathbf{A}^* - \mathbf{X}\mathbf{V}||^2 + \lambda \sum_{k=1}^{p} ||\mathbf{V}_k||, \tag{4}$$

where λ is a tuning parameter and $\mathbf{V}_k$ is the k^{th} component of $\mathbf{V} = (\mathbf{V}_1, \cdots, \mathbf{V}_p)^\top$ with the dimension $r \times 1$. Let $\widehat{\mathcal{R}} = \{k : \widehat{\mathbf{V}}_k \neq \mathbf{0}\}$ be the set of indices of the nonzero estimated coefficients, and denote $\widehat{s} = |\widehat{\mathcal{R}}|$ and $\widehat{\mathcal{I}} = \widehat{\mathcal{R}}^c$. To ameliorate the bias caused by the penalties, we subsequently use the selected variables to obtain the refitted unpenalized estimator of $\mathbf{V}^*$, which is

$$\widetilde{\mathbf{V}} = \arg \min_{\mathbf{V} \in \mathbb{R}^{p \times r}, \mathbf{V}_{\widehat{\mathcal{I}}} = \mathbf{0}_{(p-\widehat{s}) \times r}} ||\widetilde{\mathbf{W}}\mathbf{A}^* - \mathbf{X}\mathbf{V}||^2.$$

The choice of $\mathbf{A}^*$ will be discussed in Section 6.

3. Propensity score, outcome regression, and doubly robust method. In this section we introduce the DR estimator, which depends on the PS and OR functions. Let D_i denote a dummy variable such that $D_i = 1$ when the treatment is given to the i^{th} individual, and $D_i = 0$ otherwise. Let Y_{0i} and Y_{1i} be potential outcomes corresponding to $D_i = 0$ and $D_i = 1$, respectively. Then $Y_{1i} - Y_{0i}$ is the treatment effect for the i^{th} individual. However, individual treatment effects are not observed. Instead, we observe D_i and $Y_i \equiv D_i Y_{1i} + (1 - D_i)Y_{0i}$. Then the data set consists of (D_i, Y_i, X_i), $i = 1, \ldots, n$. Our main interest is to estimate the ATE:

$$\tau \equiv E(Y_{1i} - Y_{0i}).$$

The major challenge in estimating ATE is that, for each i, we only observe either Y_{1i} or Y_{0i}, but not both. The PS, defined as

$$\pi(x) \equiv P(D_i = 1 \mid X_i = x), \tag{5}$$

plays an important role in adjusting for confounding. Following Rosenbaum and Rubin (1983, 1984), we make the following assumption about confounding.

(A3) (i) D_i and (Y_{0i}, Y_{1i}) are independent of each other given X_i and (ii) $0 < \pi(X_i) < 1$ for all X_i.

Assumption A3 (i) implies that

$$\tau_j(X_i) = E(Y_{ji} \mid X_i) = E(Y_{ji} \mid D_i = j, X_i) = E(Y_i \mid D_i = j, X_i) \tag{6}$$

for $j = 0, 1$, which is called the OR function (Tan, 2006). Assumption A3 (ii) further ensures identifiability of (6).

For observational data, the PS based method and the OR approach are two common procedures used for reducing selection bias. Alternatively, one might consider a DR estimator that makes use of both $\pi(X_i)$ and $\tau_j(X_i)$ given in (5) and (6). The DR estimator can be constructed based on the efficient influence function (Hahn, 1998) given as:

$$\frac{D_i\{Y_{1i} - \tau_1(X_i)\}}{\pi(X_i)} - \frac{(1 - D_i)\{Y_{0i} - \tau_0(X_i)\}}{1 - \pi(X_i)} + \tau_1(X_i) - \tau_0(X_i). \tag{7}$$

Let $\tau_j^*(X_i)$ and $\pi^*(X_i)$ be the postulated models of $\tau_j(X_i)$ and $\pi(X_i)$, respectively, for $j = 0, 1$. By the facts that

$$E(D_iY_i \mid X_i) = E(D_iY_{1i} \mid X_i) = E(D_i \mid X_i)E(Y_{1i} \mid X_i) = \pi(X_i)\tau_1(X_i);$$
$$E\{(1 - D_i)Y_i \mid X_i\} = \{1 - \pi(X_i)\}\tau_0(X_i),$$

it can be seen that the expected value of (7) equals τ when either $\tau_j^*(X_i) = \tau_j(X_i)$ or $\pi^*(X_i) = \pi(X_i)$. Then, the DR estimator that is the sample average of (7) is asymptotically unbiased if either the PS model or the OR model is correctly specified. However, the DR estimator is not semiparametrically efficient when one of them is misspecified. Moreover, it can perform poorly when both models are misspecified (Kang and Schafer, 2007).

To solve the problem of model misspecification, Hahn (1998) employed nonparametric techniques to estimate $\tau_j(X_i)$ and $\pi(X_i)$ consistently without assuming any specific model structure. Accordingly, the resulting estimator is root-n consistent and efficient. However, this nonparametric approach is only applicable in practice for data with low dimensional covariates (generally one to three). When p becomes large, it is known that the nonparametric regression method suffers from the "curse of dimensionality". For high dimensional data, Belloni et al. (2014) and Farrell (2015) proposed penalized estimation under the postulated parametric PS and OR models. They showed that their estimators perform well when the parametric models are correctly specified or have negligible approximation errors to the true models. Analogous to the low dimensional case, however, those estimators can be less efficient and more biased if the postulated models are misspecified. In addition, model selection procedures used in high dimensional data analysis may fail to identify the key confounders under misspecified models. To resolve these problems, we consider the SSDR model for the PS and OR functions and estimate the index parameters in the SSDR model by the method given in Section 2. In the next section, we present the estimator of the ATE τ.

4. Estimation of the average treatment effect. To estimate the ATE τ, we first obtain the estimator of $E(Z_i \mid X_i)$ given in (2) by multivariate kernel smoothing. Consider a multivariate kernel density function $K(u_1, \ldots, u_r)$ and a bandwidth vector $\mathbf{h} = (h_1, \ldots, h_r)^\top$. For ease of implementation, we let $h_1 = \cdots = h_r = h$. Denote $K_h(\mathbf{u}) = h^{-r}K(u_1/h, \ldots, u_r/h)$, where $\mathbf{u} = (u_1, \ldots, u_r)^\top$. For given $x = (x_1, \ldots, x_p)^\top$, the conditional mean $E(Z_i \mid X_i = x) = E(Z_i \mid \mathbf{V}^{*\top} X_i = \mathbf{V}^{*\top} x)$ is estimated by

$$\begin{aligned} \widehat{E}(Z_i \mid X_i = x) &= \widehat{E}(Z_i \mid \widetilde{\mathbf{V}}^\top X_i = \widetilde{\mathbf{V}}^\top x) \\ &= \sum\nolimits_{i=1}^n K_h(\widetilde{\mathbf{V}}^\top X_i - \widetilde{\mathbf{V}}^\top x) Z_i \Big/ \sum\nolimits_{i=1}^n K_h(\widetilde{\mathbf{V}}^\top X_i - \widetilde{\mathbf{V}}^\top x). \end{aligned} \tag{8}$$

We let $Z_i = D_i$, and obtain the estimator $\widehat{\pi}(x) = \widehat{E}(D_i \mid x)$ by (8). Moreover, by the derivation in (6), we obtain the estimator $\widehat{\tau}_j(x) = \widehat{E}(Y_{ji} \mid x) = \widehat{E}(Y_{ji} \mid D_i = j, x)$ of $\tau_j(x)$ by letting $Z_i = Y_{ji}$ and using the observations in the control and treatment groups, respectively, for $j = 0, 1$. Thus, for $Z_i = Y_{1i}$ and $Z_i = Y_{0i}$, their corresponding sample sizes used for estimating $E(Z_i \mid X_i)$ are n_1 (the sample size of the treatment group) and n_0 (the sample size of the control group). Note $n_1/n \to E(D_i)$, $n_0/n \to 1 - E(D_i)$, and $E(D_i) \in (0, 1)$. Hence, $n_1 \asymp n$ and $n_0 \asymp n$. Since using either n_j $(j = 0, 1)$ or n does not affect the asymptotic order, we suppress the subscription j in n_j for notational simplicity. Furthermore, $\widetilde{Z}_i = Y_{ji} - E(Y_{ji} \mid D_i = j)$ for $Z_i = Y_{ji}$, and thus we replace $E(Y_{ji} \mid D_i = j)$ with the corresponding sample analog within the control and treatment groups, respectively, in estimation.

Next, we replace $\tau_1(X_i)$, $\tau_0(X_i)$ and $\pi(X_i)$ in the influence function (7) by the corresponding estimators given above. Then τ is estimated by

$$\widehat{\tau} = n^{-1} \sum_{i=1}^n \left[\frac{D_i\{Y_i - \widehat{\tau}_1(X_i)\}}{\widehat{\pi}(X_i)} - \frac{(1 - D_i)\{Y_i - \widehat{\tau}_0(X_i)\}}{1 - \widehat{\pi}(X_i)} + \widehat{\tau}_1(X_i) - \widehat{\tau}_0(X_i) \right].$$

In the next section, we present the theoretical properties of the proposed estimators. Specifically, we first establish estimation consistency for $\widehat{\mathbf{V}}$ and $\widetilde{\mathbf{V}}$. We then derive the asymptotic normality of $\widehat{\tau}$, based on which we can conduct statistical inference for ATE. We also show that $\widehat{\tau}$ achieves the semiparametric efficiency bound. It is worth noting that $\widehat{\tau}$ enjoys these properties without the need for variable selection consistency.

5. Inference for the average treatment effect. We first establish the estimation error bounds for the group Lasso estimator $\widehat{\mathbf{V}}$ and the refitted unpenalized estimator $\widetilde{\mathbf{V}}$. Under Assumption (A2) (i), $\mathbf{V}^* = \mathbf{V}^0 \mathbf{A}^{0\top} \mathbf{A}^*$ and it is a $p \times r$ matrix with $|\mathcal{R}|$ nonzero rows, where $|\mathcal{R}| \le s$. Here, s is an upper

bound on the row sparsity of $\mathbf{V}^*$. Both s and p can depend on the sample size n such that $s \equiv s_n$ and $p \equiv p_n$. For notational convenience, we suppress n in their expressions. We assume that $s \ll n$, $p \geq 2$ and $\log p = O(n^{\varpi})$ for some $\varpi \in (0, 1)$.

For a matrix $\mathbf{\Delta} = (\mathbf{\Delta}_1, \ldots, \mathbf{\Delta}_p)^\top \in \mathbb{R}^{p \times r}$, let $\mathcal{R}'$ be the subset of indices in $\mathcal{I}$ corresponding to the s largest values of $||\mathbf{\Delta}_k||$. Denote $\mathcal{R}_{2s} = \mathcal{R}' \cup \mathcal{R}$. For $\mathbf{X}$ satisfying (A1), we make the following assumption on $\mathbf{\Sigma}$.

(R) Let $\kappa(2s) \equiv \min\{\frac{||\mathbf{\Sigma}^{1/2}\mathbf{\Delta}||}{||\mathbf{\Delta}_{\mathcal{R}_{2s}}||} : \mathbf{\Delta} \in \mathbb{R}^{p \times r} \backslash \{\mathbf{0}\}, \sum_{k \in \mathcal{I}} ||\mathbf{\Delta}_k|| \leq 3 \sum_{k \in \mathcal{R}} ||\mathbf{\Delta}_k||\}$. Assume $0 < \kappa(2s) < \infty$.

In addition, we assume that

(A4) (i) for any $\mathbf{a} \in \mathbb{R}^p$, there exists a constant $0 < \rho < \infty$ such that $\mathbf{a}^\top \mathbf{\Sigma} \mathbf{a} \leq \rho ||\mathbf{a}||^2$, and (ii) for each $\ell = 1, \ldots, r$, $\mathbf{V}_{\cdot\ell}^{*\top} \mathbf{V}_{\cdot\ell}^* \leq c_\ell$ for some constant $0 < c_\ell < \infty$.

It is worth noting that (R) is the Restricted Eigenvalue (RE) assumption for random design matrices satisfying (A1) (Zhou et al., 2009). The RE assumption is needed and commonly used for establishing the estimation error bound of the Lasso estimators (e.g., Zhang and Huang, 2008; Bickel et al., 2009; Raskutti et al., 2010). For high-dimensional settings with $p \geq n$, the matrix $\mathbf{X}^\top \mathbf{X}/n$ is degenerate, i.e., $\lim_{\mathbf{\Delta} \in \mathbb{R}^{p \times r} \backslash \{\mathbf{0}\}} \frac{||\mathbf{X}\mathbf{\Delta}||}{\sqrt{n}||\mathbf{\Delta}||} = 0$. As a consequence, ordinary least squares estimation does not work in this case, since it requires $\lim_{\mathbf{\Delta} \in \mathbb{R}^{p \times r} \backslash \{\mathbf{0}\}} \frac{||\mathbf{X}\mathbf{\Delta}||}{\sqrt{n}||\mathbf{\Delta}||} > 0$. Thus, the Lasso estimator requires a much weaker assumption. Under Assumption (R), we have $\lambda_{\min}(\mathbf{\Sigma}_{\mathcal{R},\mathcal{R}}) \geq \kappa(2s) > 0$, where $\mathbf{\Sigma}_{\mathcal{R},\mathcal{R}}$ is the submatrix of $\mathbf{\Sigma}$ with rows and columns both indexed by the indices in $\mathcal{R}$, so that the parameters in the sparse regression are uniquely defined. It has been proven in Zhou et al. (2009) that (A1) and (R) together imply $\lambda_{\min}(\boldsymbol{X}_{\cdot\mathcal{R}}^\top \boldsymbol{X}_{\cdot\mathcal{R}}/n) > 0$ and the random design matrix $\mathbf{X}$ behaves nicely with high probability. Moreover, Assumption (A4) (i) is given below (4.5) of Zhang and Huang (2008). This, in conjunction with Assumption (A4) (ii), ensures that $\mathbf{V}_{\cdot\ell}^{*\top} X_i$ follows a normal distribution with finite variance.

Denote $\epsilon_i = Z_i - E(Z_i \mid \mathbf{V}^{*\top} X_i)$ and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$. We assume that

(A5) Z_i is bounded, or its error ϵ_i satisfies

(i) the noise vector $\boldsymbol{\epsilon}$ has sub-Gaussian tails such that $P(|\mathbf{a}^\top \boldsymbol{\epsilon}| > ||\mathbf{a}||x) \leq \gamma \exp(-Cx^2)$ for any vector $\mathbf{a} \in R^n$ and $x \geq 0$, and for some positive finite constants C and γ, (ii) ϵ_i and X_i are independent for each i, and (iii) $\sup_{X_i} |E(Z_i \mid \mathbf{V}^{*\top} X_i)| \leq \widetilde{C}$ for some positive finite constant $\widetilde{C}$.

Let $\phi_{\max}$ be the maximum eigenvalue of the matrix $\mathbf{X}^\top\mathbf{X}/n$. For a set $\mathcal{S} \subseteq \{1,\ldots,p\}$, denote $\underline{\phi}(Q_\mathcal{S},\mathcal{S}) = \min_{\delta\in R^{|\mathcal{S}|}} \delta^\top Q_\mathcal{S}\delta/||\delta||^2$, where $Q_\mathcal{S} = \mathbf{X}_{\cdot\mathcal{S}}^\top\mathbf{X}_{\cdot\mathcal{S}}/n$. The following theorem provides estimation error bounds for the estimators $\widehat{\mathbf{V}}$ and $\widetilde{\mathbf{V}}$ given in Section 2.

Theorem 1. *Under Assumptions (A1), (A2), (A4), (A5) and (R),* $\lambda \asymp \sqrt{rn\log(p\vee n)}$ *and* $s = o(\sqrt{n/\log(p\vee n)})$*, for sufficiently large* n*, we have that, with probability at least* $1-3(p\vee n)^{-1}$*,*

$$\begin{aligned} ||\mathbf{X}(\widehat{\mathbf{V}}-\mathbf{V}^*)|| &\le 4\sqrt{2}\lambda\sqrt{s}/(\kappa(2s)\sqrt{n}); \\ \textstyle\sum_{k=1}^p ||\widehat{\mathbf{V}}_k-\mathbf{V}_k^*|| &\le 32\lambda s/(\kappa(2s)^2 n); \\ \widehat{s} &\le 128\kappa(2s)^{-2}\phi_{\max}s; \\ ||\widehat{\mathbf{V}}-\mathbf{V}^*|| &\le 4\sqrt{2}\lambda\sqrt{s}/(\kappa(2s)\sqrt{n})^2. \end{aligned}$$

We further obtain that, with probability at least $1-3(p\vee n)^{-1}$*,* $||\widetilde{\mathbf{V}}-\mathbf{V}^*|| \le c^*\lambda\sqrt{s}/n$*, where* $c^* = \min(8\sqrt{2}\underline{\phi}(Q_{\widehat{\mathcal{R}}\cup\mathcal{R}},\widehat{\mathcal{R}}\cup\mathcal{R})^{-1/2}\kappa(2s)^{-1}, 2\{128\kappa(2s)^{-2}\phi_{\max}+1\}^{1/2}\underline{\phi}(Q_{\widehat{\mathcal{R}}\cup\mathcal{R}},\widehat{\mathcal{R}}\cup\mathcal{R})^{-1})$.

We subsequently explore the convergence rate of $\widehat{\mathbf{V}}$ and $\widetilde{\mathbf{V}}$ and an upper bound of $\widehat{s}$. To this end, we introduce the following assumption.

(A6) (i) Assume that r is a fixed number. (ii) With probability approaching one, $\phi_{\max} \le C_\phi$ for some constant $C_\phi \in (0,\infty)$, and $\underline{\phi}(Q_\mathcal{S},\mathcal{S}) \ge c_{\underline{\phi}} > 0$ uniformly in $\mathcal{S} \subseteq \{1,\ldots,p\}$ with $|\mathcal{S}| \le \{128\kappa(2s)^{-2}\phi_{\max}+1\}s$.

Corollary 1. *Suppose Assumptions (A1), (A2), (A4)-(A6) and (R) hold. For* $\lambda \asymp \sqrt{rn\log(p\vee n)}$ *and* $s = o(\sqrt{n/\log(p\vee n)})$*, we have, as* $n\to\infty$*,* $P(\widehat{s}\le C^*s)\to 1$*, where* $C^* = 128\kappa(2s)^{-2}C_\phi$*. In addition,*

$$||\widehat{\mathbf{V}}-\mathbf{V}^*|| = O_p(\sqrt{s\log(p\vee n)/n}), \text{ and } ||\widetilde{\mathbf{V}}-\mathbf{V}^*|| = O_p(\sqrt{s\log(p\vee n)/n}).$$

The results in Corollary 1 follow immediately from Theorem 1, and they are required for establishing the asymptotic distribution of the ATE estimator. For this purpose, we also consider the following conditions.

(C1) The r-dimensional kernel function is a product of r univariate kernel functions, i.e., $K_h(\mathbf{u}) = h^{-r}K(u_1/h)\cdots K(u_r/h)$, where h is a bandwidth and $\mathbf{u} = (u_1,\ldots,u_r)^\top$. The univariate kernel function $K(\cdot)$ is symmetric, has compact support and is Lipschitz continuous on its

support. Furthermore, it satisfies

$$\begin{aligned} \int K(u)du &= 1, \int u^i K(u)du = 0 \ (i = 1, \ldots, m-1), \text{ and} \\ 0 &\neq \int |u|^m K(u)du < \infty. \end{aligned}$$

Accordingly, K is a m^{th} order kernel.

(C2) The $(m-1)^{\text{th}}$ derivative of $E(Z \mid \mathbf{V}^\top X)$ is a locally Lipschitz continuous function of $\mathbf{V}^\top X$ for $\mathbf{V}$ in a neighborhood of $\mathbf{V}^*$.

(C3) (i) $\max\{n^{-1/(2r)}(\log n)^{1/r}, n^{-1/(r+2)}(\log n)^{1/(r+2)}\} \ll h \ll n^{-1/(4m)}$, where $r < 2m$ and $m > 1$; (ii) $s\log(p \vee n) = o(n^{1/4} + h^{-m+1} + \sqrt{nh^{r+2}/\log(n)})$.

Conditions (C1) and (C2) are commonly used in the kernel nonparametric smoothing literature; see, for example, Ma and Zhu (2012). Condition (C3) states the order requirements for the bandwidth h, the dimension of the covariates p, and the upper bound of the number of relevant covariates s. They are needed in order to have the root-n consistency of the ATE estimator $\widehat{\tau}$. Suppose that $h \asymp n^{-1/(2m+r)}$. Then h achieves the optimal order in kernel estimation. By Assumption (C3) (ii), s and p need to satisfy $s\log(p \vee n) = o(n^{1/4} + n^{(m-1)/(2m+r)}/\sqrt{\log(n)})$. Let τ^0 be the true ATE.

Theorem 2. *Under Assumptions (A1)-(A6) and (R), and Conditions (C1)-(C3), we have that, for* $\lambda \asymp \sqrt{rn\log(p \vee n)}$, $\widehat{\tau} - \tau^0 = O_p(n^{-1/2})$, *and* $\sigma^{-1}\sqrt{n}(\widehat{\tau} - \tau^0) \to N(0,1)$, *where*

$$\sigma^2 = E\left[\frac{\sigma_1^2(X_i)}{\pi(X_i)} + \frac{\sigma_0^2(X_i)}{1-\pi(X_i)} + (\tau(X_i) - \tau^0)^2\right], \tag{9}$$

$\sigma_1^2(X_i) = var(Y_{1i} \mid X_i)$, $\sigma_0^2(X_i) = var(Y_{0i} \mid X_i)$ *and* $\tau(X_i) = \tau_1(X_i) - \tau_0(X_i)$.

Remark 1. In Theorem 2, we obtain the root-n consistency and asymptotic normality of the estimator $\widehat{\tau}$ without the need for variable selection consistency, i.e., that $P(\widehat{\mathcal{R}} = \mathcal{R}) \to 1$. It is worth noting that achieving selection consistency typically requires a uniform signal strength condition (Zhang and Zhang, 2014) under which all non-zero regression coefficients should be greater in magnitude than a threshold value. However, this condition can be easily violated when weak signals may exist.

Remark 2. The asymptotic variance σ^2 given in (9) reaches the semiparametric efficiency bound in Theorem 1 of Hahn (1998). Thus, $\widehat{\tau}$ is semiparametrically efficient.

Remark 3. The asymptotic variance σ^2 given in (9) equals

$$E\left[\frac{D_i\{Y_i-\tau_1(X_i)\}}{\pi(X_i)}-\frac{(1-D_i)\{Y_i-\tau_0(X_i)\}}{1-\pi(X_i)}+\tau(X_i)-\tau^0\right]^2.$$

Hence, we estimate it by

(10)
$$\sigma_n^2=n^{-1}\sum_{i=1}^n\left[\frac{D_i\{Y_i-\widehat{\tau}_1(X_i)\}}{\widehat{\pi}(X_i)}-\frac{(1-D_i)\{Y_i-\widehat{\tau}_0(X_i)\}}{1-\widehat{\pi}(X_i)}+\widehat{\tau}(X_i)-\widehat{\tau}\right]^2,$$

where $\widehat{\tau}(X_i)=\widehat{\tau}_1(X_i)-\widehat{\tau}_0(X_i)$.

We next show that σ_n^2 is a consistent estimator of σ^2.

Theorem 3. *Under Assumptions (A1)-(A6) and (R), and Conditions (C1)-(C3), we have that, for* $\lambda\asymp\sqrt{rn\log(p\vee n)}$, $\sigma_n^2-\sigma^2=o_p(1)$.

Using the results of Theorems 2 and 3, we obtain the distribution of $\sigma_n^{-1}(\widehat{\tau}-\tau^0)$ below.

Corollary 2. *Under Assumptions (A1)-(A6) and (R), and Conditions (C1)-(C3), we have that, for* $\lambda\asymp\sqrt{rn\log(p\vee n)}$, $\sigma_n^{-1}\sqrt{n}(\widehat{\tau}-\tau^0)\to N(0,1)$.

Remark 4. By Corollary 2, we are able to construct a $(1-\alpha)100\%$ confidence interval for the true ATE, τ^0, given as $\widehat{\tau}\pm z_{\alpha/2}\sigma_n/\sqrt{n}$, where $z_{\alpha/2}$ is the $(1-\alpha/2)$ quantile of the standard normal.

6. Computational algorithm. After studying the theoretical properties of the proposed estimators, this section focuses on the computation of the primary estimator $\widehat{\mathbf{V}}$ of $\mathbf{V}^*$. As stated in (3), this estimator can be obtained by minimizing $Q_n(\mathbf{V};\mathbf{A}^*)=L_n(\mathbf{V};\mathbf{A}^*)+\lambda\sum_{k=1}^p||\mathbf{V}_k||$, where $L_n(\mathbf{V};\mathbf{A}^*)=(1/2)||\widetilde{\mathbf{W}}\mathbf{A}^*-\mathbf{X}\mathbf{V}||^2$. This is a convex optimization problem with group Lasso penalties. We employ an Iterative Shrinkage and Thresholding (IST) algorithm, which converges quickly for finding the parameter estimator with convex penalties (Beck and Teboulle, 2009).

Specifically, for given $\mathbf{V}^{(m-1)}$, the estimator $\mathbf{V}^{(m)}$ in the IST algorithm is obtained by solving the proximal operator problem (Gong et al., 2013):

$$\mathbf{V}^{(m)}=\arg\min_{\mathbf{V}}(1/2)||\mathbf{V}-\mathbf{U}^{(m)}||^2+t^{(m)}\lambda\sum_{k=1}^p||\mathbf{V}_k||, \tag{11}$$

where $\mathbf{U}^{(m)}=\mathbf{V}^{(m-1)}-\nabla L_n(\mathbf{V};\mathbf{A}^*)t^{(m)}$, $\nabla L_n(\mathbf{V};\mathbf{A}^*)=-\mathbf{X}^\top(\widetilde{\mathbf{W}}\mathbf{A}^*-\mathbf{X}\mathbf{V}^{(m)})$, and $t^{(m)}$ is the step size in the m^{th} step. Then the minimizer

in (11) has a closed form solution $\mathbf{V}_k^{(m)} = (1 - \lambda t^{(m)}/||\mathbf{U}_k^{(m)}||)_+ \mathbf{U}_k^{(m)}$, for $k = 1, \ldots, p$, where $(x)_+ = x$ if $x > 0$ and 0, otherwise. We use a line search criterion considered in Gong et al. (2013) to find the step size at step m. The step size $t^{(m)}$ is acceptable if the following monotone line search criterion is satisfied:

$$Q_n(\mathbf{V}^{(m)}; \mathbf{A}^*) \leq Q_n(\mathbf{V}^{(m-1)}; \mathbf{A}^*) - (\zeta/2)||\mathbf{V}^{(m)} - \mathbf{V}^{(m-1)}||^2/t^{(m)},$$

where ζ is a constant in the interval $(0, 1)$. We let $t^{(m)} = 0.5^\rho$, where ρ is the minimal value that satisfies the above criterion. Following Gong et al. (2013), we use $\zeta = 0.01$ in our implementation.

In the computational algorithm, we need an $\mathbf{A}^*$ that satisfies Assumption (A2). We use the convergent value of the sequence $\mathbf{A}^{(m)}$ as $\mathbf{A}^*$, where $\mathbf{A}^{(m)}$ and $\mathbf{V}^{(m)}$ are obtained by minimizing $Q_n(\mathbf{V}; \mathbf{A})$ iteratively until convergence. For the given $\mathbf{V}^{(m)}$, the minimizer of $Q_n(\mathbf{V}^{(m)}; \mathbf{A})$ is $\mathbf{A}^{(m)} = \mathbf{U}_L \mathbf{U}_R^\top$, where $\mathbf{U}_L$ and $\mathbf{U}_R$ are the left-singular vectors and right-singular vectors of $\widetilde{\mathbf{W}}^\top \mathbf{X} \mathbf{V}^{(m)}$, respectively. In the process, we use the following strategy to find an initial value $\mathbf{V}^{(0)}$ of $\mathbf{V}$. We fit the Lasso regression for each column of $\widetilde{\mathbf{W}}$ on $\mathbf{X}$, and obtain the union set of all selected variables, denoted by $\widehat{\mathcal{R}}^{(0)}$. Let $\widehat{\mathbf{b}}^{(0)} = \arg\min_{\mathbf{b} \in \mathbb{R}^{p \times (p+1)}, \mathbf{b}_{(\widehat{\mathcal{R}}^{(0)})^c} = \mathbf{0}_{(p - |\widehat{\mathcal{R}}^{(0)}|) \times (p+1)}} ||\widetilde{\mathbf{W}} - \mathbf{X}\mathbf{b}||^2$. The initial value $\mathbf{V}^{(0)}$ is the r left-singular vectors of $\widehat{\mathbf{b}}^{(0)}$ multiplied by the corresponding singular values.

From the penalized estimator $\widehat{\mathbf{V}}$, we are able to compute the refitted unpenalized estimator $\widetilde{\mathbf{V}}$. Then we obtain the estimator $\widehat{E}(Z_i \mid \widetilde{\mathbf{V}}^\top x)$ in (8) by using the Gaussian kernel for estimation and employing the leave-one-out cross validation approach for the selection of bandwidth h. Finally, we apply the five-fold cross validation (CV) method to choose the tuning parameter λ and the order r. It is worth noting that different methods have been proposed for the determination of r. Some popular approaches with good statistical properties include the sequential test methods (Li, 1991; Bura and Cook, 2001), the BIC-type methods (Feng et al., 2013) and the cross-validation type approaches (Xia, et al., 2002; Xia, 2008). Furthermore, Luo and Li (2016) proposed a new procedure through exploiting a special eigenvalue-eigenvector pattern to assist order determination. In our framework, the estimation of parameters is essentially a SRRR problem, so we adopt the same method as given in Chen and Huang (2012) by using the five-fold CV to select r.

7. Simulation studies.

7.1. *Background and methods used.* In this section, we illustrate the finite sample performance of our proposed method via simulations in which

we generate data from different PS and OR models.

We call our proposed estimator of ATE as the sparse sufficient dimension reduction (sparse_dim) estimator. We compare it with six other estimators. Three are feasible estimators, (a) the "sparse_linear" estimator from fitting a sparse logistic linear model and a sparse linear regression model to PS and OR, respectively, where the variables are selected by Lasso and the estimated coefficients are obtained by refitting the models with the selected variables; (b) the "full_dim" estimator from fitting PS and OR with all covariates using the pHd method for sufficient dimension reduction without variable selection; (c) the "kernel" estimator from fitting PS and OR with all covariates using the nonparametric kernel regression. For comparison purposes, we also consider three infeasible estimators obtained by using the true covariates with nonzero coefficients, namely (d) the "oracle_linear" estimator from correspondingly fitting the linear models with the true covariates to PS and OR; (e) the "oracle_dim" estimator from fitting PS and OR with the true covariates using the sufficient dimension reduction approach; (f) the "oracle" estimator from fitting the data with the true PS and OR models. For methods involving kernel estimation, we use the leave-one-out cross validation to select the bandwidth. It is expected that the oracle estimate should perform the best.

7.2. *Data generation mechanism and settings.* We consider three models, namely

$$\begin{aligned}
&\text{Model 1:} && \operatorname{logit}\{E(D_i \mid X_i)\} = (X_{i1} + X_{i2})(X_{i3} + 1)/2,\\
&&& E(Y_i \mid D_i, X_i) = D_i + X_{i1}^2 + X_{i2}^2;\\
&\text{Model 2:} && \operatorname{logit}\{E(D_i \mid X_i)\} = (X_{i1} + X_{i2} + X_{i3})/2,\\
&&& E(Y_i \mid D_i, X_i) = D_i + (X_{i1} + 2)(X_{i2} + X_{i3} + 2);\\
&\text{Model 3:} && \operatorname{logit}\{E(D_i \mid X_i)\} = (X_{i1} + 2X_{i2} - X_{i3})/2,\\
&&& E(Y_i \mid D_i, X_i) = D_i(X_{i1} + X_{i2} + 1) + X_{i1} + X_{i2} + X_{i3} + X_{i4},
\end{aligned}$$

where $Y_i = E(Y_i \mid D_i, X_i) + \epsilon_i$, X_i are generated from $\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, $\mathbf{\Sigma} = \{\sigma_{jj'}\}$, $\sigma_{jj'} = 0.5^{|j-j'|}$ for $1 \le j, j' \le p$, and ϵ_i are independently generated from the standard normal distribution for $i = 1, \cdots, n$.

In Model 1, both PS and OR are nonlinear models with $r = 2$. In Model 2, PS is a linear model with $r = 1$, while OR is a nonlinear model with $r = 2$. In Model 3, both PS and OR are linear models with $r = 1$.

We consider $p = 20, 40, 100$ and $n = 1500, 3000, 5000$. All simulation results are based on 500 realizations. Observational studies often have large sample sizes, so we focus on studying the performance of the proposed estimator with moderately large p and large n in different model settings.

This consideration is consistent with the data setting in our empirical applications. For the sake of illustration, we also provide simulations for ultra high-dimensional data in Section S.1 of the supplemental materials Ma et al. (2018).

TABLE 1
The empirical coverage rates (rate), and the absolute values of biases (bias) and the average of the estimated standard deviations (est_sd) of the estimated ATE for $p = 20$.

		Model 1			Model 2			Model 3		
n		rate	bias	est_sd	rate	bias	est_sd	rate	bias	est_sd
1500	sparse_linear	0.000	0.882	0.187	0.958	0.019	0.274	0.982	0.001	0.078
	oracle_linear	0.000	0.858	0.188	0.958	0.019	0.273	0.984	0.002	0.077
	full_dim	0.643	0.066	0.084	0.624	0.181	0.107	0.944	0.060	0.108
	oracle_dim	0.914	0.013	0.065	0.912	0.019	0.062	0.976	0.030	0.096
	sparse_dim	0.912	0.013	0.064	0.910	0.019	0.062	0.960	0.029	0.095
	kernel	0.020	0.331	0.086	0.000	1.530	0.102	0.000	0.837	0.109
	oracle	0.960	0.001	0.067	0.960	0.001	0.063	0.978	0.001	0.077
3000	sparse_linear	0.000	0.858	0.130	0.950	0.009	0.203	0.996	0.003	0.055
	oracle_linear	0.000	0.843	0.130	0.952	0.009	0.203	0.992	0.003	0.055
	full_dim	0.738	0.038	0.063	0.790	0.071	0.066	0.968	0.047	0.072
	oracle_dim	0.920	0.010	0.050	0.924	0.006	0.043	0.992	0.025	0.070
	sparse_dim	0.918	0.010	0.050	0.924	0.008	0.044	0.994	0.025	0.069
	kernel	0.000	0.318	0.054	0.000	1.458	0.064	0.000	0.810	0.070
	oracle	0.950	0.003	0.048	0.956	0.004	0.044	0.992	0.004	0.055
5000	sparse_linear	0.000	0.874	0.103	0.960	0.014	0.161	0.980	0.001	0.043
	oracle_linear	0.000	0.846	0.102	0.964	0.013	0.161	0.978	0.001	0.042
	full_dim	0.766	0.020	0.037	0.726	0.052	0.045	0.974	0.030	0.054
	oracle_dim	0.938	0.004	0.038	0.940	0.001	0.033	0.982	0.022	0.050
	sparse_dim	0.936	0.001	0.038	0.940	0.001	0.033	0.962	0.025	0.050
	kernel	0.000	0.260	0.037	0.000	1.365	0.047	0.000	0.763	0.052
	oracle	0.948	0.004	0.041	0.948	0.002	0.034	0.976	0.001	0.042

7.3. *Results.* Tables 1-3 report the empirical coverage rates (rate) of the 95% confidence intervals, and the absolute values of biases (bias) and the average values of the estimated standard deviations (est_sd) of the seven estimated ATE for $p = 20$, 40 and 100, respectively, based on 500 simulation realizations. For Model 1 and Model 2, we observe that, as n increases, the coverage rates of the sparse_dim estimate and the oracle_dim estimate become closer to the nominal rate 95%, the est_sd values of these two estimates are similar to that of the oracle estimate, and their estimation biases are close to zero. These findings indicate that the sparse_dim estimate performs similarly to the oracle_dim estimate by knowing the true covariates and the oracle estimate by knowing the true models. In contrast, the sparse_linear and oracle_linear estimates for Model 1 have zero coverage rates and yield large estimation biases and est_sd values. This implies that when both PS

TABLE 2

The empirical coverage rates (rate), and the absolute values of biases (bias) and the average of the estimated standard deviations (est_sd) of the estimated ATE for $p = 40$.

		Model 1			Model 2			Model 3		
n		rate	bias	est_sd	rate	bias	est_sd	rate	bias	est_sd
1500	sparse_linear	0.000	0.886	0.187	0.960	0.025	0.274	0.980	0.005	0.078
	oracle_linear	0.000	0.858	0.188	0.958	0.022	0.273	0.984	0.002	0.077
	full_dim	0.522	0.159	0.091	0.012	0.966	0.121	0.928	0.002	0.126
	oracle_dim	0.914	0.013	0.065	0.912	0.019	0.062	0.976	0.030	0.096
	sparse_dim	0.904	0.015	0.064	0.906	0.020	0.069	0.930	0.027	0.090
	kernel	0.000	0.398	0.094	0.000	2.261	0.112	0.000	1.231	0.112
	oracle	0.960	0.001	0.067	0.960	0.001	0.063	0.978	0.005	0.077
3000	sparse_linear	0.000	0.865	0.130	0.976	0.009	0.203	0.992	0.001	0.055
	oracle_linear	0.000	0.843	0.130	0.952	0.009	0.203	0.992	0.003	0.055
	full_dim	0.648	0.066	0.055	0.578	0.177	0.086	0.946	0.055	0.078
	oracle_dim	0.920	0.010	0.050	0.924	0.010	0.044	0.992	0.025	0.069
	sparse_dim	0.918	0.014	0.050	0.920	0.009	0.050	0.978	0.026	0.069
	kernel	0.000	0.320	0.054	0.000	2.118	0.087	0.000	1.138	0.074
	oracle	0.950	0.003	0.048	0.956	0.004	0.044	0.992	0.004	0.055
5000	sparse_linear	0.000	0.863	0.103	0.968	0.005	0.156	0.984	0.001	0.043
	oracle_linear	0.000	0.846	0.102	0.964	0.013	0.161	0.978	0.001	0.042
	full_dim	0.714	0.036	0.039	0.624	0.077	0.061	0.964	0.038	0.057
	oracle_dim	0.938	0.004	0.038	0.940	0.001	0.033	0.982	0.022	0.050
	sparse_dim	0.932	0.001	0.037	0.934	0.003	0.033	0.974	0.023	0.050
	kernel	0.000	0.268	0.040	0.000	2.055	0.059	0.000	1.085	0.053
	oracle	0.948	0.002	0.041	0.948	0.001	0.034	0.976	0.001	0.042

and OR models are nonlinear, the estimates obtained from the parametric linear model fittings can be very biased and inefficient due to the model misspecification. Although both the sparse_linear and oracle_linear estimates for Model 2 have better coverage rates, their est_sd values are quite large. This indicates that, for the nonlinear OR model, the linear estimates are inefficient even though they are unbiased. In Model 3, both models are linear. The sparse_dim estimate and the oracle_dim estimate perform reasonably well, and they are slightly inferior to the linear estimates as expected. Moreover, we find that the nonparametric kernel estimate has very small coverage rates that are close to zero for all cases and it has large biases. The performance of both the full_dim and nonparametric kernel estimates deteriorates sharply as the dimension p becomes larger. This suggests that using all covariates with the sufficient dimension reduction approach or nonparametric kernel estimation may not yield a reliable estimate of ATE. In sum, the proposed sparse_dim estimate performs well in estimating ATE with a large set of covariates even when the true model structure is not known *a priori*.

To further illustrate the bias and variance of the estimated ATE, $\widehat{\tau}$, cal-

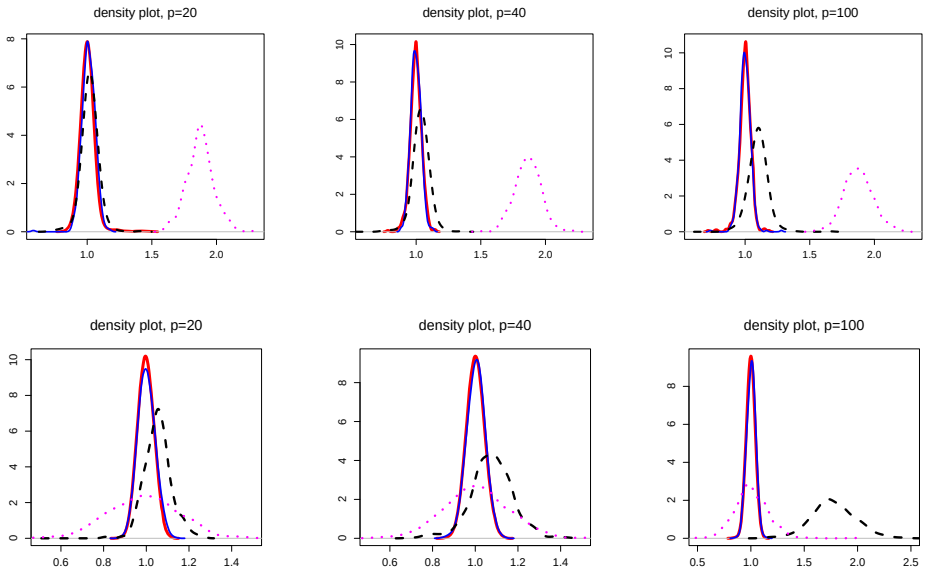
TABLE 3
The empirical coverage rates (rate), and the absolute values of biases (bias) and the average of the estimated standard deviations (est_sd) of the estimated ATE for $p = 100$.

		Model 1			Model 2			Model 3		
n		rate	bias	est_sd	rate	bias	est_sd	rate	bias	est_sd
1500	sparse_linear	0.000	0.890	0.186	0.958	0.036	0.276	0.970	0.003	0.079
	oracle_linear	0.000	0.858	0.188	0.958	0.019	0.273	0.984	0.002	0.077
	full_dim	0.192	0.309	0.098	0.000	2.404	0.164	0.002	0.939	0.149
	oracle_dim	0.914	0.013	0.065	0.912	0.019	0.062	0.976	0.030	0.096
	sparse_dim	0.902	0.014	0.065	0.904	0.016	0.066	0.920	0.034	0.096
	kernel	0.004	0.423	0.104	0.000	2.518	0.168	0.000	1.427	0.150
	oracle	0.960	0.001	0.067	0.960	0.001	0.063	0.978	0.001	0.077
3000	sparse_linear	0.000	0.878	0.131	0.952	0.007	0.202	0.990	0.001	0.055
	oracle_linear	0.000	0.843	0.130	0.952	0.009	0.203	0.992	0.003	0.055
	full_dim	0.192	0.193	0.059	0.000	1.655	0.095	0.786	0.053	0.089
	oracle_dim	0.920	0.010	0.050	0.924	0.008	0.044	0.992	0.025	0.070
	sparse_dim	0.908	0.016	0.050	0.904	0.012	0.046	0.972	0.022	0.073
	kernel	0.000	0.322	0.058	0.000	2.447	0.094	0.000	1.330	0.082
	oracle	0.960	0.002	0.048	0.956	0.004	0.044	0.992	0.004	0.055
5000	sparse_linear	0.000	0.876	0.102	0.942	0.006	0.156	0.978	0.001	0.043
	oracle_linear	0.000	0.846	0.102	0.964	0.013	0.161	0.978	0.001	0.042
	full_dim	0.358	0.097	0.045	0.000	0.775	0.060	0.894	0.042	0.067
	oracle_dim	0.938	0.004	0.038	0.940	0.001	0.033	0.982	0.022	0.050
	sparse_dim	0.918	0.006	0.038	0.930	0.001	0.034	0.976	0.022	0.050
	kernel	0.000	0.272	0.050	0.000	2.365	0.064	0.000	1.302	0.068
	oracle	0.948	0.004	0.041	0.948	0.002	0.034	0.976	0.001	0.042

culated from the oracle, sparse_dim, full_dim, and sparse_linear estimates, Figure 1 depicts the kernel density plots of $\widehat{\tau}$ for Model 1 and Model 2 when $p = 20, 40, 100$ and $n = 5000$. Figure 1 demonstrates that both sparse_dim and oracle estimates are symmetrically distributed around 1, which is the true ATE. However, the sparse_linear estimate shows a large bias in Model 1, and exhibits large variances in Model 2. As for the full_dim estimate, it becomes a more biased and less efficient estimate as p increases. This implies that the redundant variables included in the model can significantly affect the estimation accuracy of ATE when p is large.

We next demonstrate the impact of different methods on the test statistic $\vartheta_n \equiv \sigma_n^{-1}\sqrt{n}(\widehat{\tau} - \tau^0)$. Accordingly, Figure 2 depicts the kernel density plots of ϑ_n with four different ATE estimates in Model 1 discussed above for $p = 20, 40, 100$ and $n = 5000$. It shows that the density plots of ϑ_n calculated from the oracle and sparse_dim estimates exhibit a similar pattern, being symmetric around zero. This indicates that these two estimates yield a reliable test statistic. In contrast, the density plot of ϑ_n computed from the sparse_linear estimate exhibits a large bias due to the misspecification of

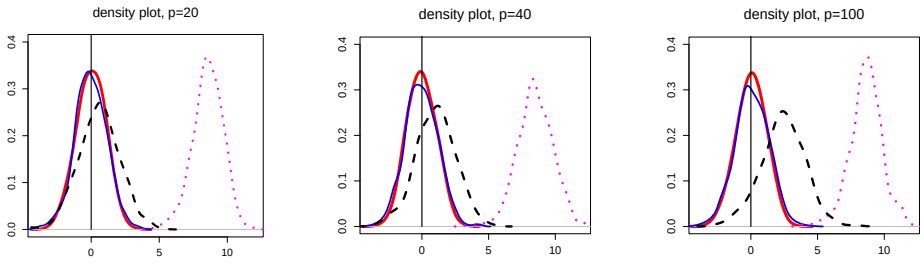
FIG 1. *The kernel density plots of the estimated average treatments effects, $\widehat{\tau}$, calculated using the oracle estimate (red thick curves), the sparse_dim estimate (blue thin curves), the full_dim estimate (black dashed curves) and the sparse_linear estimate (magenta dotted curves); the upper and lower panels correspond to Model 1 and Model 2.*



both PS and OR models. As for the plot of ϑ_n calculated from the full_dim estimate, it becomes more biased and less efficient as p increases. Based on our Monte Carlo studies, we finally conclude that the proposed sparse_dim estimate performs well in estimating and testing the average treatment effect when the true model is not known *a priori*.

Lastly, we compare our proposed method "sparse_dim" with several other popular methods for estimating ATE. These methods include the "MAVE" estimator proposed in Luo et al. (2017) using MAVE to recover the OR function, the "IPW" estimator which is an inverse probability weighting estimator with the propensity score estimated by the method given in Imai and Ratkovic (2014), the "Matching" estimator obtained based on one-to-one matching using the R package Matching (Sekhon, 2008), the "TMLE" estimator which is the targeted maximum likelihood estimator proposed in van der Laan and Rubin (2006), and the "RF" and "GAM" estimators from applying random forest and the generalized additive model (GAM), respectively, in G-computation (Robins, 1986; Snowden, 2011). Random forest (van der Laan et al., 2007) and GAM (Hastie and Tibshirani, 1986) are two popular nonparametric methods for estimating regression models. We refer to Luo et al. (2017) for the detailed descriptions of the above methods. Table

FIG 2. *The kernel density plots of the statistic, $\vartheta_n \equiv \sigma_n^{-1}\sqrt{n}(\widehat{\tau} - \tau^0)$, for Model 1, where the estimated average treatments effects are calculated using the oracle estimate (red thick curves), the sparse_dim estimate (blue thin curves), the full_dim estimate (black dashed curves) and the sparse_linear estimate (magenta dotted curves).*



4 reports the biases and the empirical standard deviations (emp_sd) of the estimated ATE by the seven methods for $p = 20, 40, 100$ and $n = 1500$ when the data are generated from Model 2. We exclude the "TMLE" estimate for $p = 100$ due to its computational burden for large p. We observe that our proposed sparse_dim estimator has the smallest bias and emp_sd values among all the estimators. It is of interest to note that the MAVE estimator performs better than the other five estimates, whereas its performance deteriorates as p becomes larger.

TABLE 4
The biases ("bias") and empirical standard deviations (emp_sd) of the estimated ATE by the seven methods for $p = 20, 40, 100$ when data are generated from Model 2.

		sparse_dim	MAVE	IPW	Matching	TMLE	RF	GAM
$p = 20$	bias	0.019	0.072	1.572	1.059	0.040	-0.108	0.664
	emp_sd	0.065	0.087	0.184	0.196	0.070	0.080	0.125
$p = 40$	bias	0.020	0.092	1.582	1.074	0.455	-0.128	0.664
	emp_sd	0.072	0.093	0.188	0.211	0.088	0.082	0.115
$p = 100$	bias	0.016	0.215	1.603	1.152	—	-0.182	0.653
	emp_sd	0.068	0.105	0.178	0.197	—	0.081	0.125

8. Application. In this section, we consider the NIH-AARP Study of Diet and Health (Schatzkin et al., 2001). We employ our proposed method to investigate the causal effect of smoking on body mass index (BMI). The confounding variables are dietary pattern scores for nutritional intakes, which were calculated by using the U.S. Department of Agriculture's (USDA's) Healthy Eating Index-2005 (HEI-2005, http://www.cnpp.usda.gov/Healthy

EatingIndex.htm). The HEI-2005 comprises 12 distinct component scores. Intakes of each food or nutrient, represented by one of the 12 components and adjusted for caloric intake (energy), are assessed and given a score. A higher score represents a better dietary quality. All confounding variables are centered and standardized in the analysis.

The data consist of 7432 African American women aged 55-70 who had not been diagnosed with any cancer at baseline and who did not have missing BMI. In our analysis, let Y_i =BMI, D_i =indicator for smoking, and $X_{i1}, \ldots, X_{i12}$ be the dietary scores of Total Fruit (TF), Whole Fruit (WF), Total Grains (TG), Whole Grains (WG), Total Vegetables (TV), DOL Vegetables (DV), Dairy (D), Meat and Beans (MB), Oils, Sodium (S), Saturated Fat (SF) and Empty Calories (EC), respectively, for $i = 1, \cdots, 7432$.

We apply our proposed sparse sufficient dimension reduction (sparse_dim) method to estimate PS and OR, respectively. By employing the five-fold CV method, we obtain the estimated number of indices, which is $\widehat{r} = 1$, in model (1) for PS and OR, respectively. For comparison, we also consider the sparse_linear method discussed in simulation studies.

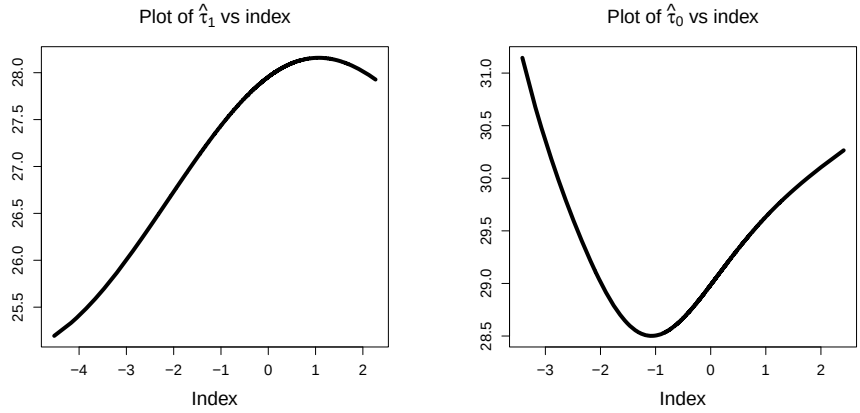
TABLE 5

The selected variables among the 12 dietary intakes by the sparse_linear and sparse_dim methods for PS, OR in the smoking group (OR_smoke), and OR in the non-smoking group (OR_non-smoke). "√" means that the variable is selected.

	PS		OR_smoke		OR_non-smoke	
	sparse_linear	sparse_dim	sparse_linear	sparse_dim	sparse_linear	sparse_dim
TF	√	√		√	√	√
WF	√	√				
TG		√		√		
WG				√		√
TV						
DV				√	√	√
D				√		√
MB					√	√
Oils						
S						
SF			√	√	√	√
EC	√	√				

Table 5 reports the variables selected by these two methods for estimating PS, OR in the smoking group, and OR in the non-smoking group. The results show that our approach captures the variables that would be missed by the sparse_linear method. For example, it is evident in other studies that fruit, vegetable and whole grain intakes influence BMI (Steffen et al., 2003; Heo et al., 2011; Charlton et al., 2014). However, fruit and vegetable intakes are not

FIG 3. *The plots of $\widehat{\tau}_1$ for the smoking group and $\widehat{\tau}_0$ for the non-smoking group, respectively, versus the estimated index value $\widehat{\mathbf{V}}^\top X_i$.*



selected by the sparse_linear method for OR of the smoking group and whole grain intake is not selected by the sparse_linear method for both smoking and non-smoking groups.

To examine the relationship between BMI and dietary intakes, Figure 3 depicts the estimated conditional means $\widehat{\tau}_1(\cdot)$ and $\widehat{\tau}_0(\cdot)$ versus the estimated index value $\widehat{\mathbf{V}}^\top X_i$ for the smoking and non-smoking groups, respectively. It is of interest to note that the estimated conditional mean in the smoking group is smaller than that in the non-smoking group at the same index value. Both plots in Figure 3 clearly show a nonlinear relationship between BMI and the estimated index value. Specifically, the plot for the smoking group exhibits a nonlinear increasing pattern along with index value and the slope becomes flatter as the index value becomes larger. The plot for the non-smoking group displays a quadratic pattern. It shows that the BMI of non-smokers decreases along with the index value in the beginning and then it increases after the index exceeds certain value.

To further illustrate the relationship between BMI and the dietary score of each nutrient intake, Figure 4 depicts $\widehat{\tau}_1(\cdot)$ and $\widehat{\tau}_0(\cdot)$ versus the dietary score for Total Fruit, respectively, by fixing the dietary scores of other nutrient intakes at their means. We use this dietary score for illustration because it is selected as relevant dietary intakes for OR by the sparse_dim method. In the smoking group, it shows a positive relationship between BMI and the Total Fruit score, and the slope becomes flatter as the score increases. In the non-smoking group, the plot shows a quadratic pattern with the Total Fruit score. Overall, Figure 4 indicates that a better dietary score of Total Fruit can increase BMI for smokers, and the Total Fruit score is inversely

FIG 4. *The plots of $\widehat{\tau}_1$ for the smoking group and $\widehat{\tau}_0$ for the non-smoking group, respectively, versus the dietary score of total fruit.*

associated with BMI when the score is less than 3 and their association becomes positive as the score becomes larger.

Next we compare our proposed sparse_dim estimator of ATE with three other estimators: (a) the "sparse_linear" estimator; (b) the "linear_dim" estimator from fitting the sparse logistic linear model to PS and using the proposed sparse dimension reduction method to estimate OR; (c) the "dim_linear" estimator from using the proposed sparse dimension reduction method to estimate PS and fitting the sparse linear model to OR.

Table 6 reports the estimated values ("est.") of ATE and their associated standard errors ("s.e.") by these four different methods. It shows that all four methods have negative values for the estimated ATE. This result confirms the earlier finding that current smokers have significantly lower BMI than non-smokers (see Kaufman et al., 2012). Furthermore, the linear_dim and dim_linear methods yield the estimates of ATE that are close to that obtained from the sparse_dim method, but they produce larger standard errors. This is because both linear_dim and dim_linear methods can lead to asymptotically unbiased but not efficient estimates due to possible misspecification of either the PS or the OR model. Moreover, we compare our sparse_dim estimator with the six estimators, MAVE, IPW, Matching, TMLE, RF and GAM, given in Section 7. Table 7 reports the "est." and "s.e." values of these estimators. The sparse_dim estimator has the smallest standard error value. We also observe that the MAVE and TMLE methods have estimated values close to that obtained from the sparse_dim estimator.

TABLE 6
The estimates ("est.") and standard errors ("s.e.") of ATE obtained by four different methods: sparse_dim, sparse_linear, linear_dim, and dim_linear.

	sparse_dim	sparse_linear	linear_dim	dim_linear
est.	-1.218	-1.147	-1.195	-1.236
s.e.	0.179	0.189	0.188	0.184

TABLE 7
The estimates ("est.") and standard errors ("s.e.") of ATE obtained by seven different methods: sparse_dim, MAVE, IPW, Matching, TMLE, RF and GAM.

	sparse_dim	MAVE	IPW	Matching	TMLE	RF	GAM
est.	-1.218	-1.173	-1.154	-1.082	-1.189	-0.946	-1.086
s.e.	0.179	0.197	0.185	0.207	0.186	0.193	0.195

9. Discussion. In this paper, we consider an estimator of ATE constructed based on an efficient influence function, which involves a PS function and an OR function. We propose a sparse sufficient dimension reduction method to estimate these two functions, without making restrictive parametric modeling assumptions. Theoretically, we show that the proposed estimator is asymptotically normal and semiparametric efficient without the need for variable selection consistency. Practically, we illustrate the proposed method through a number of simulation studies and an empirical example. The numerical studies support our theoretical findings. Our method provides a new flexible strategy for efficient inference of ATE with big data which often involve a large number of variables measured on a large number of subjects. Our proposed method can be extended to estimate quantile, heterogeneous and longitudinal treatment effects in observational studies. In sum, these three avenues can shed light on areas of future research that deserve a thorough study. It is worth noting that in practice one can also apply other popular approaches such as MAVE (Luo et al., 2017) and machine learning methods (van der Laan et al., 2007; van der Laan and Rose, 2011) to estimate the working models without imposing restrictive modeling assumptions. However, careful and thorough investigations are needed to develop the computational algorithms and establish the theoretical properties of the resulting estimators in high-dimensional settings.

10. Acknowledgements. The authors sincerely thank the Editor, the Associate Editor and the two anonymous referees for their insightful comments and suggestions that lead to a substantial improvement of the paper.

Appendix. The Appendix contains the technical proofs of Theorems 2 and 3. The proof of Theorem 1 is given in the on-line supplemental materials Ma et al. (2018).

A.1. Proof of Theorem 2. Let $\tau_j(x) = E(Y_{ji}|X_i = x)$ and $g_j(x) = E(D_{ji}|X_i = x)$ for $j = 0$ and 1, where $D_{1i} = D_i$ and $D_{0i} = 1 - D_i$. In addition, let $\widehat{\tau}_j(x)$ and $\widehat{g}_j(x)$ be the estimators of $\tau_j(x)$ and $g_j(x)$, respectively, and let $T_i = (D_i, Y_i, X_i)$ be the i^{th} individual observation. Denote

$$m_j(T_i, \tau_j, g_j) = \frac{D_{ji}\{Y_i - \tau_j(X_i)\}}{E(D_{ji}|X_i)} + \tau_j(X_i) = \frac{D_{ji}\{Y_i - \tau_j(X_i)\}}{g_j(X_i)} + \tau_j(X_i).$$

To prove this theorem, we will show that

$$n^{-1}\sum_{i=1}^n m_1(T_i, \widehat{\tau}_1, \widehat{g}_1) = n^{-1}\sum_{i=1}^n m_1(T_i, \tau_1, g_1) + o_p(n^{-1/2}). \tag{A.1}$$

Employing the same techniques as those for obtaining the above result, we can demonstrate that

$$n^{-1}\sum_{i=1}^n m_0(T_i, \widehat{\tau}_0, \widehat{g}_0) = n^{-1}\sum_{i=1}^n m_0(T_i, \tau_0, g_0) + o_p(n^{-1/2}).$$

By Central Limit Theorem, we have

$$\sigma^{-1}\sqrt{n}[n^{-1}\sum_{i=1}^n\{m_1(T_i, \tau_1, g_1) - m_0(T_i, \tau_0, g_0)\} - \tau^0] \to N(0, 1),$$

where σ^2 is defined in (9). This, together with the Slutsky's Theorem and $\widehat{\tau} = n^{-1}\sum_{i=1}^n\{m_1(T_i, \widehat{\tau}_1, \widehat{g}_1) - m_0(T_i, \widehat{\tau}_0, \widehat{g}_0)\}$, yields the asymptotic result of $\widehat{\tau}$ in Theorem 2. Furthermore, by the weak law of large numbers, we obtain that

$$n^{-1}\sum_{i=1}^n\{m_1(T_i, \tau_1, g_1) - m_0(T_i, \tau_0, g_0)\} - \tau^0 = O_p(n^{-1/2}),$$

which implies $\widehat{\tau} - \tau^0 = O_p(n^{-1/2})$.

To complete the proof, we demonstrate (A.1) below. By the Taylor series expansion, we have

$$\begin{aligned} & m_1(T_i, \widehat{\tau}_1, \widehat{g}_1) - m_1(T_i, \tau_1, g_1) \\ = \; & -g_1^{-2}D_{1i}(Y_i - \tau_1)(\widehat{g}_1 - g_1) + (-g_1^{-1}D_{1i} + 1)(\widehat{\tau}_1 - \tau_1) \\ & +\widetilde{g}_1^{-2}D_{1i}(\widehat{g}_1 - g_1)(\widehat{\tau}_1 - \tau_1) + \widetilde{g}_1^{-3}D_{1i}(Y_i - \widetilde{\tau}_1)(\widehat{g}_1 - g_1)^2 \end{aligned} \tag{A.2}$$

for some $\widetilde{g}_1$ between g_1 and $\widehat{g}_1$ and $\widetilde{\tau}_1$ between τ_1 and $\widehat{\tau}_1$. Then

$$\begin{aligned} & n^{-1}\sum_{i=1}^n m_1(T_i, \widehat{\tau}_1, \widehat{g}_1) - n^{-1}\sum_{i=1}^n m_1(T_i, \tau_1, g_1) \\ = \; & \varphi_{n1} + \varphi_{n2} + \varphi_{n3} + \varphi_{n4}, \end{aligned} \tag{A.3}$$

where

$$\begin{aligned}
\varphi_{n1} &= n^{-1}\sum_{i=1}^{n}[-g_1(X_i)^{-2}D_{1i}\{Y_i-\tau_1(X_i)\}]\{\widehat{g}_1(X_i)-g_1(X_i)\},\\
\varphi_{n2} &= n^{-1}\sum_{i=1}^{n}(-g_1(X_i)^{-1}D_{1i}+1)\{\widehat{\tau}_1(X_i)-\tau_1(X_i)\},\\
\varphi_{n3} &= n^{-1}\sum_{i=1}^{n}\widetilde{g}_1(X_i)^{-2}D_{1i}\{\widehat{\tau}_1(X_i)-\tau_1(X_i)\}\{\widehat{g}_1(X_i)-g_1(X_i)\} \quad \text{and}\\
\varphi_{n4} &= -n^{-1}\sum_{i=1}^{n}\widetilde{g}_1(X_i)^{-3}D_{i1}\{Y_i-\widetilde{\tau}_1(X_i)\}\{\widehat{g}_1(X_i)-g_1(X_i)\}^2.
\end{aligned}$$

It is worth noting that, by definitions of $\widehat{g}_1(X_i)$ and $g_1(X_i)$, we have

$$\begin{aligned}
\widehat{g}_1(X_i)-g_1(X_i) &= \widehat{E}(D_{i1}|\widetilde{\mathbf{V}}^\top X_i)-E(D_{i1}|\mathbf{V}^{*\top}X_i)\\
&= \{\widehat{E}(D_{i1}|\widetilde{\mathbf{V}}^\top X_i)-\widehat{E}(D_{i1}|\mathbf{V}^{*\top}X_i)\}\\
&+ \{\widehat{E}(D_{i1}|\mathbf{V}^{*\top}X_i)-E(D_{i1}|\mathbf{V}^{*\top}X_i)\}. \qquad \text{(A.4)}
\end{aligned}$$

Furthermore, let $\xi_i=-g_1(X_i)^{-2}D_{1i}\{Y_i-\tau_1(X_i)\}$. Then, under Assumption (A3), $E(\xi_i|X_i)=0$.

Applying (A.4) and Lemmas **S.1** and **S.2** presented in the supplemental materials Ma et al. (2018), we have $\varphi_{n1}=o_p(n^{-1/2})$. Employing the same approach, we can show that $\varphi_{n2}=o_p(n^{-1/2})$.

By Condition (C2) and the results of (S.17) and (S.18) in the proof of Lemma **S.2**, we obtain that

$$\begin{aligned}
&\sup_{X_i}|\widehat{E}(D_{i1}|\widetilde{\mathbf{V}}^\top X_i)-\widehat{E}(D_{i1}|\mathbf{V}^{*\top}X_i)|\\
\le\ &\sup_{X_i}|E^{(1)}(D_{i1}|\mathbf{V}^{*\top}X_i)|\times||X_i||_\infty\sqrt{|\mathcal{R}^*|}||\widetilde{\mathbf{V}}_{\mathcal{R}^*}-\mathbf{V}^*_{\mathcal{R}^*}||\{1+o(1)\}\\
=\ &O_p(\sqrt{\log(p\vee n)})O_p(\sqrt{s})O_p(\sqrt{s\log(p\vee n)/n})=O_p(\log(p\vee n)sn^{-1/2}).
\end{aligned}$$

Then employing the uniform convergence rate in Mack and Silverman (1982), we have

$$\sup_{X_i}|\widehat{E}(D_{i1}|\mathbf{V}^{*\top}X_i)-E(D_{i1}|\mathbf{V}^{*\top}X_i)|=O_p\{h^m+(nh^r)^{-1/2}\sqrt{\log n}\}.$$

The above two results, together with (A.4), imply that

$$\sup_{X_i}|\widehat{g}_1(X_i)-g_1(X_i)|=O_p\{h^m+(nh^r)^{-1/2}\sqrt{\log n}+\log(p\vee n)sn^{-1/2}\}. \qquad \text{(A.5)}$$

Analogously, we can show that

$$\sup_{X_i}|\widehat{\tau}_1(X_i)-\tau_1(X_i)|=O_p\{h^m+(nh^r)^{-1/2}\sqrt{\log n}+\log(p\vee n)sn^{-1/2}\}. \qquad \text{(A.6)}$$

As a result, (A.5), (A.6), and Condition (C3) imply that there exist constants $\widetilde{c}$ and $\widetilde{\widetilde{c}} \in (0, \infty)$ such that, with probability approaching 1,

$$\begin{aligned} |\varphi_{n3}| &\leq \widetilde{c} \sup_{X_i} |\widehat{g}_1(X_i) - g_1(X_i)| \sup_{X_i} |\widehat{\tau}_1(X_i) - \tau_1(X_i)| \\ &= O\{h^{2m} + (nh^r)^{-1}(\log n) + \{\log(p \vee n)\}^2 s^2 n^{-1}\} = o(n^{-1/2}), \end{aligned}$$

$$\begin{aligned} |\varphi_{n4}| &\leq \widetilde{\widetilde{c}} \sup_{X_i} |\widehat{g}_1(X_i) - g_1(X_i)|^2 \\ &= O\{h^{2m} + (nh^r)^{-1}(\log n) + \{\log(p \vee n)\}^2 s^2 n^{-1}\} = o(n^{-1/2}). \end{aligned}$$

The above results, in conjunction with (A.3), lead to (A.1), which completes the proof.

A.2. Proof of Theorem 3. Let $\widehat{m}_{ji} = m_j(T_i, \widehat{\tau}_j, \widehat{g}_j)$ and $m_{ji} = m_j(T_i, \tau_j, g_j)$ for $j = 0$ and 1. Based on the results of (A.1) and Theorem 2, we have $n^{-1} \sum_{i=1}^n (\widehat{m}_{1i} - m_{1i}) = o_p(n^{-1/2})$ and $\widehat{\tau} - \tau_0 = O_p(n^{-1/2})$, respectively. The above results, together with the definitions of σ^2 and σ_n^2 in (9) and (10), imply that

$$\begin{aligned} \sigma_n^2 - \sigma^2 &= n^{-1} \sum\nolimits_{i=1}^n (\widehat{m}_{1i} - \widehat{m}_{0i} - \widehat{\tau})^2 - n^{-1} \sum\nolimits_{i=1}^n (m_{1i} - m_{0i} - \tau_0)^2 \\ &= \sum\nolimits_{j,j'=0,1} n^{-1} \sum\nolimits_{i=1}^n (\widehat{m}_{ji} + m_{ji})(\widehat{m}_{j'i} - m_{j'i}) + o_p(1). \end{aligned}$$

Below we will show that

$$n^{-1} \sum\nolimits_{i=1}^n (\widehat{m}_{1i} + m_{1i})(\widehat{m}_{1i} - m_{1i}) = o_p(1).$$

Employing the same techniques, we can also demonstrate that $n^{-1} \sum_{i=1}^n (\widehat{m}_{ji} + m_{ji})(\widehat{m}_{j'i} - m_{j'i}) = o_p(1)$ for $j = j' = 0$ and $j \neq j'$. Accordingly, we have $\sigma_n^2 - \sigma^2 = o_p(1)$, which completes the proof of Theorem 3.

Note that

$$\begin{aligned} & n^{-1} \sum\nolimits_{i=1}^n (\widehat{m}_{1i} + m_{1i})(\widehat{m}_{1i} - m_{1i}) \\ = \; & n^{-1} \sum\nolimits_{i=1}^n (\widehat{m}_{1i} - m_{1i})^2 + 2n^{-1} \sum\nolimits_{i=1}^n (\widehat{m}_{1i} - m_{1i}) m_{1i}. \end{aligned}$$

Hence, we will show that

$$n^{-1} \sum\nolimits_{i=1}^n (\widehat{m}_{1i} - m_{1i})^2 = o_p(1), \tag{A.7}$$

$$n^{-1} \sum\nolimits_{i=1}^n (\widehat{m}_{1i} - m_{1i}) m_{1i} = o_p(1). \tag{A.8}$$

By (A.2), $\widehat{m}_{1i} - m_{1i} = \varphi_{1i} + \varphi_{2i} + \varphi_{3i} + \varphi_{4i}$, where

$$\begin{aligned}
\varphi_{1i} &= -g_1^{-2} D_{1i}(Y_i - \tau_1)(\widehat{g}_1 - g_1), \\
\varphi_{2i} &= (-g_1^{-1} D_{1i} + 1)(\widehat{\tau}_1 - \tau_1), \\
\varphi_{3i} &= \widetilde{g}_1^{-2} D_{1i}(\widehat{\tau}_1 - \tau_1)(\widehat{g}_1 - g_1), \\
\varphi_{4i} &= \widetilde{g}_1^{-3} D_{i1}(Y_i - \widetilde{\tau}_1)(\widehat{g}_1 - g_1)^2.
\end{aligned}$$

Using the fact that $n^{-1}\sum_{i=1}^n (\widehat{m}_{1i} - m_{1i})^2 \le 4\sum_{k=1}^4 n^{-1}\sum_{i=1}^n \varphi_{ki}^2$. We only need to demonstrate that $n^{-1}\sum_{i=1}^n \varphi_{ki}^2 = o_p(1)$ for $k = 1, ..., 4$.

By (A.5), (A.6), $|D_{i1}| \le 1$, and $0 < g_1(X_i) < 1$, it immediately follows that $n^{-1}\sum_{i=1}^n \varphi_{ki}^2 = o_p(1)$ for $k = 2, 3$. From Assumption (A3), we have

$$\begin{aligned}
& n^{-1}\sum_{i=1}^n \{-g_1^{-2} D_{1i}(Y_i - \tau_1)\}^2 \\
= \;& n^{-1}\sum_{i=1}^n \{-g_1^{-2} D_{1i}(Y_{1i} - \tau_1)\}^2 \\
\le \;& c' n^{-1}\sum_{i=1}^n (Y_{1i} - \tau_1)^2 = c' n^{-1}\sum_{i=1}^n \epsilon_i^2
\end{aligned}$$

for some constant $c' \in (0, \infty)$. By Assumption (A5) (i), $n^{-1}\sum_{i=1}^n \epsilon_i^2 = O_p(1)$, and hence

$$n^{-1}\sum_{i=1}^n \{-g_1^{-2} D_{1i}(Y_i - \tau_1)\}^2 = O_p(1).$$

The above result, in conjunction with (A.5), implies that

$$\begin{aligned}
n^{-1}\sum_{i=1}^n \varphi_{1i}^2 &\le \sup_{X_i} |\widehat{g}_1(X_i) - g_1(X_i)|^2 n^{-1}\sum_{i=1}^n \{-g_1^{-2} D_{1i}(Y_i - \tau_1)\}^2 \\
&= o_p(1) O_p(1) = o_p(1). \qquad \text{(A.9)}
\end{aligned}$$

Analogously, we can show that

$$\begin{aligned}
& n^{-1}\sum_{i=1}^n \{\widetilde{g}_1^{-3} D_{i1}(Y_i - \widetilde{\tau}_1)\}^2 \le c'' n^{-1}\sum_{i=1}^n (Y_i - \widetilde{\tau}_1)^2 \\
& \le c'' 2 n^{-1}\sum_{i=1}^n \epsilon_i^2 + c'' 2 \sup_{X_i} |\widetilde{\tau}_1(X_i) - \tau_1(X_i)|^2 = O_p(1),
\end{aligned}$$

for some constant $c'' \in (0, \infty)$. Accordingly,

$$\begin{aligned}
n^{-1}\sum_{i=1}^n \varphi_{4i}^2 &\le \sup_{X_i} |\widehat{g}_1(X_i) - g_1(X_i)|^4 n^{-1}\sum_{i=1}^n \{\widetilde{g}_1^{-3} D_{i1}(Y_i - \widetilde{\tau}_1)\}^2 \\
&= o_p(1) O_p(1) = o_p(1).
\end{aligned}$$

This, together with (A.9), completes the proof of (A.7).

It is worth noting that

$$\begin{aligned} & n^{-1}\sum_{i=1}^{n}(\widehat{m}_{1i}-m_{1i})m_{1i} \\ = \; & n^{-1}\sum_{i=1}^{n}\varphi_{1i}m_{1i}+n^{-1}\sum_{i=1}^{n}\varphi_{2i}m_{1i}+n^{-1}\sum_{i=1}^{n}\varphi_{3i}m_{1i}+n^{-1}\sum_{i=1}^{n}\varphi_{4i}m_{1i}. \end{aligned}$$

To verify (A.8), we only need to show that $n^{-1}\sum_{i=1}^{n}\varphi_{1i}m_{1i}=o_p(1)$. This is because the proofs of $n^{-1}\sum_{i=1}^{n}\varphi_{ki}m_{1i}=o_p(1)$ for $k=2,3,4$ follow the same arguments. By Assumption (A3)(ii), there exist constants $c_1,c_2\in(0,\infty)$ such that $|-g_1(X_i)^{-2}D_{1i}|\le c_1$ and $|g_1(X_i)^{-1}D_{1i}|\le c_2$. Then,

$$\begin{aligned} & |n^{-1}\sum_{i=1}^{n}\varphi_{1i}m_{1i}| \\ \le \; & \sup_{X_i}|\widehat{g}_1(X_i)-g_1(X_i)|n^{-1}\sum_{i=1}^{n}|-g_1^{-2}D_{1i}(Y_i-\tau_1)||\frac{D_{1i}(Y_i-\tau_1)}{g_1}+\tau_1| \\ \le \; & c_1\sup_{X_i}|\widehat{g}_1(X_i)-g_1(X_i)|n^{-1}\sum_{i=1}^{n}|\epsilon_i|(c_2|\epsilon_i|+\widetilde{C}) \\ = \; & c_1\sup_{X_i}|\widehat{g}_1(X_i)-g_1(X_i)|\{c_2n^{-1}\sum_{i=1}^{n}\epsilon_i^2+\widetilde{C}n^{-1}\sum_{i=1}^{n}|\epsilon_i|\}=o_p(1), \end{aligned}$$

where $\widetilde{C}$ is defined in Assumption (A5) (iii). This completes the whole proof.

SUPPLEMENTARY MATERIAL

Supplement to "A robust and efficient approach to causal inference based on sparse sufficient dimension reduction"
(). The supplement contains the technical proof of Theorem 1, two lemmas that will be used in the proof of Theorem 2, and additional simulation studies.

References.

Abadie, A. and Imbens, G. W. (2006). Large sample properties of matching estimators for average treatment effects. *Econometrica,* 74, 235-267.

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

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *Journal on Imaging Sciences*, 2, 183-202.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, 81, 608-650.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics*, 41, 802-837.

Bickel, P. J., Ritov, Y., Buja, A., and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37, 1705-1732.

Bura, E., and Cook, R. D. (2001). Extending sliced inverse regression: the weighted chi-squared test. *Journal of the American Statistical Association*, 96, 996-1003.

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.

Chan, K. C. G. and Yam, S. C. P. (2014). Oracle, multiple robust and multipurpose calibration in a missing response problem. *Statistical Science*, 29, 380-396.

Charlton, K., Kowal, P., Soriano, M. M., Williams, S., Banks, E., Vo, K., Byles ,J. (1014). Fruit and vegetable intake and body mass index in a large sample of middle-aged Australian men and women. *Nutrients*, 6, 2305-2319.

Chen, L. and Huang, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *Journal of the American Statistical Association*, 107, 1533-1545.

Cook, R. D. and Lee, H. (1999). Dimension reduction in binary response regression. *Journal of the American Statistical Association*, 94, 1187-1200.

Cook, D. R. and Li, B. (2002). Dimension reduction for conditional mean in regression. *Annals of Statistics*, 30, 455-474.

Duan, N. H. and Li, K. C. (1991). Slicing regression: A link-free regression method *The Annals of Statistics*, 19, 505-530.

Farrell, M. H. (2015). Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189, 1-23.

Feng, Z. H., Wen, X. R., Yu, Z., Zhu, Li. (2013). On partial sufficient dimension reduction with applications to partially linear multi-index models, *Journal of the American Statistical Association*, 108, 237-246.

Freedman, D. A. and Berk, R. A. (2008). Weighting regression by propensity scores. *Evaluation Review*, 32, 392-409.

Ghosh, D. (2011). Propensity score modelling in observational studies using dimension reduction methods. *Statistics & Probability Letters*, 81, 813-820.

Gong, P., Zhang, C., Lu, Z., Huang, J. Z., and Ye, J. (2013). A general iterative shrinkage and thresholding algorithm for non-convex regularized optimization problems. *In Proceedings of the 30th International Conference on Machine Learning (ICML)*, 28, 37-45.

Hahn, J. (1998). On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, 66, 315-331.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1, 297-310.

Heckman, J. J., Ichimura, H., and Todd, P. (1998). Matching as an econometric evaluation estimator. *The Review of Economic Studies*, 65, 261-294.

Heo, M., Kim, R. S., Wylie-Rosett, J., Allison, D. B., Heymsfield, S. B., Faith, M. S. (2011). Inverse association between fruit and vegetable intake and BMI even after controlling for demographic, socioeconomic and lifestyle factors. *Obesity Facts*, 4, 449-455.

Hirano, K., Imbens, G. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica,* 71, 1161-1189.

Imai, K. and Ratkovic, M. (2014). Covariate balancing propensity socre. *Journal of the Royal Statistical Society, Series B,* 76, 243-263.

Kang, J. D.Y. and Schafer, J. L. (2007). Demystifying double robustness: a comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22, 523-539.

Kaufman, A., Auguston, E. M. and Patrick, H. (2012). Unraveling the relationship between smoking and weight: the role of sedentary behavior. *Journal of Obesity*,

doi:10.1155/2012/735465.

Li, K. C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86, 316-342.

Li, K. C. (1992). On principal Hessian directions for data visualization and dimension reduction: another application of Stein's lemma. *Journal of the American Statistical Association*, 87, 1025-1039.

Lockhart, R., Taylor, J., Tibshirani, R., and Tibshirani, R. (2014). A significance test for the lasso. *Annals of Statistics*, 42, 413-468.

Luo, W. and Li, B. (2016). Combining eigenvalues and variation of eigenvectors for order determination. *Biometrika, 103,* 875-887.

Luo, W., Zhu, Y., and Ghosh, D. (2017). On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika*, 104, 51-65.

Ma, Y. and Zhu, L. (2012). A semiparametric approach to dimension reduction. *Journal of the American Statistical Association,* 107, 168-179.

Ma, Y. and Zhu, L. (2013). Efficiency loss and the linearity condition in dimension reduction. *Biometrika,* 100, 371-383.

Ma, S., Zhu, L., Zhang, Z., Tsai, C.L., and Caroll, R. J. (2018). Supplement to "A robust and efficient approach to causal inference based on sparse sufficient dimension reduction".

Mack, Y. P. and Silverman, B. W. (1982). Weak and strong uniform consistency of kernel regression estimates. *Probability Theory and Related Fields*, 61, 405-415.

Raskutti, G., Wainwright, M. J., and Yu, B. (2010). Restricted eigenvalue properties for correlated Gaussian designs. *Journal of Machine Learning Research*, 11, 2241-2259.

Robins, J. M. (1986). A new approach to causal inference in mortality studies with a sustained exposure period – application to control of the healthy worker survivor effect. *Mathematical Modelling*, 7, 1393-1512.

Robins, J. M., Hernan, M. A. and Brumback, B. (2000). Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11, 550-560.

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.

Rosenbaum, P. R. and Rubin D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, 79, 516-524.

Rosenbaum, P. R. and Rubin D. B. (1985). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, 39, 33-38.

Rotnitzky, A., Lei, Q., Sued, M. Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99, 439-456.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66, 688–701.

Schatzkin, A., Subar, A. F., Thompson, F. E., Harlan, L. C., Tangrea, J., Hollenbeck, A. R., Hurwitz, P. E., Coyle, L., Schussler, N., Michaud, D. S., Freedman, L. S., Brown, C. C., Midthune, D., and Kipnis, V. (2001). Design and serendipity in establishing a large cohort with wide dietary intake distributions: the national institutes of health-aarp diet and health study. *American Journal of Epidemiology*, 154, 1119–1125.

Sekhon, J. S. (2008). Multivariate and propensity score matching software with automated balance optimization: The matching package for R. *Journal of Statistical Software*, 42, 1-52.

Snowden, J. M., Rose, S., and Mortimer, K. M. (2011). Implementation of G-computation on a simulated data set: Demonstration of a causal inference technique. *American Jour-*

nal of Epidemiology, 173, 731-738.
Steffen, L. M., Jacobs, D.R., Murtaugh, M. A., Moran, A., Steinberger, J., Hong, C. P., and Sinaiko, A. R. (2003). Whole grain intake is associated with lower body mass and greater insulin sensitivity among adolescents. *American Journal of Epidemiology*, 158, 243-250.
Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101, 1619–1637.
Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97, 661–682.
van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high dimensional models. *Annals of Statistics*, 42, 1166-1202.
van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super learner. *Statistical Applications in Genetics and Molecular Biology*, 6, 1-21.
van der Laan, M. J. and Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. Springer, New York.
van der Laan, M. J. and Rose, S. (2011). *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer, New York.
van der Laan, M. J. and Rubin, D. (2006). Targeted maximum likelihood learning. *The International Journal of Biostatistics,* 2, article no. 11.
Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Annals of Statistics*, 37, 2178-2201.
Xia, Y., Tong, H., Li, W., and Zhu, L. X. (2002). An adaptive estimation of dimension reduction space. *Journal of the Royal Statistical Society, Series B,* 64, 363-410.
Xia, Y. (2008). A multiple-index model and dimension reduction. *Journal of the American Statistical Association*, 103, 1631-1640.
Yin, X., Li, B., and Cook, R. D. (2008). Successive direction extraction for estimating the central subspace in a multiple-index regression. *Journal of Multivariate Analysis*, 99, 1733-1757.
Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B*, 68, 49-67.
Zhang, C. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36, 1567-1594.
Zhang, C. and Zhang, S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society, Series B,* 76, 217-242.
Zhou, S., van de Geer, S., and Bühlmann, P. (2009). Adaptive Lasso for high dimensional regression and gaussian graphical modeling. *Available at arxiv:0903.2515.*

Shujie Ma
Department of Statistics
University of California at Riverside
Riverside, CA 92521, USA
E-mail: shujie.ma@ucr.edu

Liping Zhu
Center for Applied Statistics
Institute of Statistics and Big Data
Renmin University of China
Beijing 100872, China
E-mail: zhu.liping@ruc.edu.cn

Zhiwei Zhang
Department of Statistics
University of California at Riverside
Riverside, CA 92521, USA
E-mail: zhiwei.zhang@ucr.edu

Chih-Ling Tsai
Graduate School of Management
University of California at Davis
Davis, CA 95616, USA
E-mail: cltsai@ucdavis.edu

Raymond J. Carroll
Department of Statistics
Texas A&M University
College Station, TX 77843, USA
and School of Mathematical Sciences
University of Technology
Sydney, Broadway NSW 2007
E-mail: carroll@stat.tamu.edu