Appl. Math. J. Chinese Univ.
2022, 37(1): 131-146

Recent advances in statistical methodologies in evaluating program for high-dimensional data

ZHAN Ming-feng[1] CAI Zong-wu[2] FANG Ying[1,3] LIN Ming[1,3,*]

Abstract. The era of big data brings opportunities and challenges to developing new statistical methods and models to evaluate social programs or economic policies or interventions. This paper provides a comprehensive review on some recent advances in statistical methodologies and models to evaluate programs with high-dimensional data. In particular, four kinds of methods for making valid statistical inferences for treatment effects in high dimensions are addressed. The first one is the so-called doubly robust type estimation, which models the outcome regression and propensity score functions simultaneously. The second one is the covariate balance method to construct the treatment effect estimators. The third one is the sufficient dimension reduction approach for causal inferences. The last one is the machine learning procedure directly or indirectly to make statistical inferences to treatment effect. In such a way, some of these methods and models are closely related to the de-biased Lasso type methods for the regression model with high dimensions in the statistical literature. Finally, some future research topics are also discussed.

§1 Introduction

With the availability and development of statistical methodologies and models, it has become more and more important to evaluate the effect of social and economic policies. Indeed, an accurate policy evaluation can improve the pre-design, implementation, and post-adjustment of the policies. Modern policy evaluation methods basically rely on the potential outcome framework developed by Rubin (1973, 1974, 1977). The difficulty of the treatment effect estimation under Rubin's framework lies in the fact that one can not observe the two potential

Received: 2021-06-21. Revised: 2021-08-20.
MR Subject Classification: 62F02, 62G02.
Keywords: causal inference, covariate balance, de-biased Lasso, dimension reduction, doubly robust, high dimensions, machine learning, treatment effect.
Digital Object Identifier (DOI): https://doi.org/10.1007/s11766-022-4489-3.
Supported by the National Natural Science Foundation of China(71631004, 72033008), National Science Foundation for Distinguished Young Scholars(71625001), and Science Foundation of Ministry of Education of China(19YJA910003).
*Correspondence author.

outcomes simultaneously. To solve this problem, one would generally follow the unconfoundedness assumption (Rosenbaum and Rubin, 1983) stating that given the observed covariates, whether the individuals participate in the treated or control group is not affected by the values of potential outcomes. So far, many estimation methods for treatment effects have been proposed by researchers, including but not limited to, outcome regression (OR), inverse probability weighting (IPW), doubly robust (DR) estimation, matching, covariate balance weighting (CBW), difference-in-differences and regression discontinuity design. To have an overview of the methodology, the reader is referred to the survey paper by Imbens and Wooldrige (2009), Liu et al. (2020), and the book by Cerulli (2015), among many others.

Benefiting from the advances of modern statistics and computing technology, it is easier than ever for researchers to obtain massive amounts of data, which brings new opportunities and challenges to developing new policy evaluation methods. On the one hand, the emergence and utilization of rich data sets provide new sources and materials to obtain more accurate policy evaluation, for example, rich sets of covariates make the unconfoundedness assumption more credible and help to improve the accuracy of the identification. On the other hand, the emergence of big data has not fundamentally solved the dilemma that counterfactual outcomes of the individuals can not be directly observed in policy evaluation and raises more challenges to the statistical inference of treatment effect, for example, which covariate (or which functional form of the covariate) should be included in the underlying models.

It is well known that making valid statistical inference under high-dimensional settings, where there are possibly more covariates than the sample size, is a nontrivial task in both the statistics and econometrics literature. Due to the high dimensionality effect, the traditional estimators would perform poorly when the dimension of the covariates is greater than or/and increasing with the sample size. Recently, there have been some breakthroughs made by researchers in statistics and econometrics for high-dimensional inference. For example, Javanmard and Montanari (2014), Van de Geer et al. (2014), and Zhang and Zhang (2014) propose the de-biased least absolute shrinkage and selection operator (Lasso) methods from different perspectives to construct confidence intervals for the coefficients in regression models with high-dimensional data. The main idea within the de-biased Lasso methods is to add a compensation part to the classical Lasso estimator as in Tibshirani (1996) to remove the bias caused by the regularization. Meanwhile, Belloni, Chernozhukov and Hansen (2014), Farrell (2015), and Belloni et al. (2017) construct valid inference of treatment effects relying on the estimation of both the outcome regression (capturing the relationship between the outcome variables and the covariates) and propensity score (capturing the relationship between the treatment variable and the covariates) models in high-dimensional settings.

In addition to the unconfoundedness assumption, sparsity assumption, permitting to apply model selection methods, is needed in high dimensions. Generally, researchers would adopt the approximate sparsity assumptions for the regression models. The approximate sparsity assumptions allow for imperfect model selection that the true regression functions can be represented by a linear combination of a set of covariates (or functional transforms of the covariates), whose

number is much smaller than the sample size, with a small nonzero approximate error. In contrast to perfect model selection, the approximate sparsity is less restrictive in the sense that it contains the situation where there are some moderate but nonzero covariates.

Typically, the existing methods of making causal inference in high dimensions can be divided into four classes. The first class is to make inference of treatment effect based on the doubly robust type estimation, which needs to model the outcome regression and propensity score (PS) functions simultaneously. For example, Belloni, Chernozhukov, and Hansen (2014) propose the double selection method, applying the Lasso method to two equations, respectively, to make valid inference of the treatment effect under the framework of partially linear model, Farrell (2015) focuses on the doubly robust average treatment effect (ATE) estimators with multivalued treatments and applies the group Lasso method as in Yuan and Lin (2006) to both the outcome regression and propensity score models to make robust inference on ATE, and Belloni et al. (2017) provide inferential procedures for a variety of treatment effect parameters via the Neyman orthogonality assumption, implicitly containing the doubly robust ATE estimator as a special case. The second class is to involve the idea of covariate balance in the procedure of making causal inferences. For example, Athey, Imbens and Wager (2018) propose the approximate residual balancing method to make valid inference by assigning the stable balance weight as in Zubizarreta (2015) to the regression residuals, Ning, Peng and Imai (2020) propose the high-dimensional covariate balancing propensity score (CBPS) method to combine the model selection method with the CBPS method as in Imai and Ratkovic (2014), and Tan (2020b) proposes the model-assisted inference for treatment effects by applying regularization to the calibrated loss function. The third class lies in the utilization of the sufficient dimension reduction (SDR) method. For example, Ma et al. (2019) propose the sparse SDR method to estimate the outcome regression and propensity score models and use the DR estimator to make causal inferences. The last one is to make causal inferences by ingeniously exploiting machine learning methods directly or indirectly. For example, Wager and Athey (2018) adopt the random forest algorithm to estimate the treatment effect by thinking of the individuals in the same leaf of a causal tree as having come from randomization experiment and derive the large sample theories, and Athey et al. (2019) propose the generalized random forest method to estimate any quantity that can be identified via local moment conditions and derive the asymptotic consistency and normality results. Finally, high-dimensionality settings within the panel data framework proposed by Hsiao, Ching and Wan (2012) are also considered by Carvalho, Masini and Medeiros (2018), and Shi and Huang (2021). The reader is referred to Cai (2021) and the references therein to have a comprehensive review on recent developments in estimating treatment effects for panel data.

The rest of the paper is organized as follows. Section 2 presents the notations of the treatment effect and introduces the DR estimators which are popular for both the low-dimensional and high-dimensional settings. Section 3 gives a brief review on causal inference methods based on the outcome regression and propensity score models in high dimensions. Section 4 is devoted

to introducing inferential methods of adopting the covariate balance methodology to estimate treatment effect with high-dimensional data. Section 5 investigates the utilization of SDR methods for causal inferences. Section 6 describes the machine learning methods that are directly or indirectly used to estimate treatment effects. Section 7 concludes the paper.

§2 Treatment Effect Estimators with Unconfoundedness

Suppose that we have a binary treatment variable D taking the value of 1 if the individual is in the treated group and taking the value of 0 if the individual is in the control group. Let $Y(1)$ and $Y(0)$ be two potential outcomes corresponding to the treated and untreated status of the individual, respectively. The observed outcome of interest Y is defined as

$$Y = DY(1) + (1 - D)Y(0).$$

The ATE is defined as the mean difference between the two potential outcomes, $\Delta = E[Y(1)] - E[Y(0)]$. It is supposed that there is also a $p \times 1$ vector of covariates X representing the characteristics of the individual. The propensity score determining the conditional probability of receiving treatment is defined as $\pi(x) = P(D = 1|X = x)$. To identify the ATE, the so-called unconfoundedness and overlap assumptions (Rosenbaum and Rubin, 1983) are commonly adopted; that is,

(a) (Unconfoundedness) $(Y(0), Y(1)) \perp D \mid X$;

(b) (Overlap) for some $\epsilon > 0$ and all $x \in \mathbb{X}, \epsilon \leq \pi(x) \leq 1 - \epsilon$,

where $\mathbb{X} \subset \mathbb{R}^p$ is the support of X. Assume that $\{(X_i, D_i, Y_i)\}_{i=1}^n$ is a random sample from the population (X, D, Y), where n is the sample size.

In this section, it is devoted to briefly introducing the three types of widely-used ATE estimators; that is, the OR estimator, the IPW estimator and the DR estimator. The OR estimator is based on the relationship between the outcome variable and the covariates. Under the assumption of unconfoundedness, the ATE can be identified by

$$\Delta = E\big[\mu_1(X) - \mu_0(X)\big],$$

where $\mu_j(X) = E\big[Y(j)|X\big] = E\big[Y|X, D = j\big]$ for $j = 0$ and 1, and the corresponding OR estimator for ATE is given as

$$\hat{\Delta}_{or} = \frac{1}{n}\sum_{i=1}^{n}\big[\hat{\mu}_1(X_i) - \hat{\mu}_0(X_i)\big], \tag{1}$$

where $\hat{\mu}_j(X_i)$ is a consistent estimate of $\mu_j(X_i)$ for $j = 0$ and 1.

It is clear that the IPW estimator is based on the relationship between the treatment variable and the covariates. Under the identification assumptions, the ATE can be identified by

$$\Delta = E\left[\frac{DY}{\pi(X)}\right] - E\left[\frac{(1-D)Y}{1-\pi(X)}\right],$$

and the corresponding IPW estimator for ATE is defined as

$$\hat{\Delta}_{ipw} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i}{\hat{\pi}\left(X_i\right)}Y_i\right) - \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1-D_i}{1-\hat{\pi}\left(X_i\right)}Y_i\right), \tag{2}$$

where $\hat{\pi}(X_i)$ is a consistent estimate of $\pi(X_i)$. Furthermore, the ATE can also be identified by

$$\begin{aligned}\Delta &= E\left[\frac{DY}{\pi(X)}+\left(1-\frac{D}{\pi(X)}\right)\mu_1(X)\right]-E\left[\frac{(1-D)Y}{1-\pi(X)}+\left(1-\frac{1-D}{1-\pi(X)}\right)\mu_0(X)\right]\\ &= E\left[\mu_1(X)+\frac{D}{\pi(X)}(Y_i-\mu_1(X))\right]-E\left[\mu_0(X)+\frac{1-D}{1-\pi(X)}(Y-\mu_0(X_i))\right].\end{aligned} \tag{3}$$

Note that when $\pi(x)$ is correctly specified, $E\left[\left(1-\frac{D}{\pi(X)}\right)\mu_1(X)\right]=E\left[\left(1-\frac{1-D}{1-\pi(X)}\right)\mu_0(X)\right]=0$ even if $\mu_j(X)$, $j=0,1$, are misspecified. Meanwhile, when $\mu_j(X)$, $j=0,1$, are correctly specified, $E\left[\frac{D}{\pi(X)}(Y-\mu_1(X))\right]=E\left[\frac{1-D}{1-\pi(X)}(Y-\mu_0(X))\right]=0$ even if $\pi(x)$ is misspecified. Therefore, (3) holds if at least one of the OR and PS models is correctly specified. By combining the OR model $\mu_j(X)$, $j=0,1$ and the PS model $\pi(X)$, the DR estimator for ATE is constructed as

$$\begin{aligned}&\hat{\Delta}_{dr}\\ &=\frac{1}{n}\sum_{i=1}^{n}\left[\frac{D_i}{\hat{\pi}(X_i)}Y_i+\left(1-\frac{D_i}{\hat{\pi}(X_i)}\right)\hat{\mu}_1(X_i)\right]-\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1-D_i}{1-\hat{\pi}(X_i)}Y_i+\left(1-\frac{1-D_i}{1-\hat{\pi}(X_i)}\right)\hat{\mu}_0(X_i)\right]\\ &=\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_1(X_i)+\frac{D_i}{\hat{\pi}(X_i)}(Y_i-\hat{\mu}_1(X_i))\right]-\frac{1}{n}\sum_{i=1}^{n}\left[\hat{\mu}_0(X_i)+\frac{1-D_i}{1-\hat{\pi}(X_i)}(Y_i-\hat{\mu}_0(X_i))\right].\end{aligned} \tag{4}$$

The DR estimator is robust in the sense that if at least one of the OR and PS models is correctly specified, the resulting estimator is consistent. Among all the three types of treatment effect estimators, one can see that the DR estimator plays an important role in the statistical inferences of the ATE in high-dimensional settings in the next section.

§3 Inference of Treatment Effects Based on OR and PS Models

3.1 Double Selection Method

Belloni, Chernozhukov and Hansen (2014) propose the so-called double selection method under the framework of partially linear model

$$Y_i=\alpha D_i+f(X_i)+u_i,$$

with $E(u_i|X_i,D_i)=0$, $D_i=h(X_i)+v_i$, and $E(v_i|X_i)=0$, where $f(\cdot)$ and $h(\cdot)$ are two unknown functions, and u_i and v_i are unobserved disturbances. To combine the high-dimensional settings with the partially linear model, the approximate sparsity assumptions are made to the two functions $f(\cdot)$ and $h(\cdot)$; that is,

$$f(X_i)=\sum_{j=1}^{p}X_{ij}\beta_{or,j}+e_{or,i},\qquad\text{and}\qquad h(X_i)=\sum_{j=1}^{q}X_{ij}\beta_{ps,j}+e_{ps,i},$$

where the numbers of non-zero components in β_{or} and β_{ps} are less than s, an integer much smaller than the sample size n, and $e_{or,i}$ and $e_{ps,i}$ are small non-zero approximation error terms. For this structural model, Belloni, Chernozhukov and Hansen (2014) introduce the "post-double-selection" (Post-DS) procedure to make valid inference of the treatment effect α. Three steps are included in the Post-DS procedure: (i) apply the feasible Lasso model selection

method to regress Y_i on X_i in the first step; (ii) apply the feasible Lasso model selection method to regress D_i on X_i in the second step; (iii) run the OLS procedure by regressing Y_i on D_i and the combination of the selected covariates X_i in the previous two steps. More generally, Belloni, Chernozhukov and Hansen (2014) apply the Post-DS procedure to study the ATE and average treatment effect on the treated (ATT) in the data-rich environment based on the DR estimators in Section 5 of their paper, with the main idea similar to the paper by Belloni et al. (2017).

Remark 1: *Note that at the second step above, it is assumed that D_i is a linear probability model of X_i, which might be inappropriate. Also, other forms of $f(x)$ can be considered.*

3.2 Robust Inference Based on DR Estimators

Under the unconfoundedness and overlap assumptions, Farrell (2015) focuses on the DR ATE estimators in the framework of multivalued treatments and proposes a procedure for making robust inference arguing that the DR estimators are not only robust to model misspecification but also robust to model selection. The robustness property in high-dimensional settings permits the existence of the selection errors in modeling both the outcome regression functions and propensity scores without affecting the valid inference of the ATE estimators. Similar to Belloni, Chernozhukov and Hansen (2014), the approximate sparsity assumptions, allowing for imperfect model section, are made in the model selection stage. Robust inference on the ATE is then obtained via a model selection procedure similar to the Post-DS as in Belloni, Chernozhukov and Hansen (2014). First, apply the group Lasso method[1] to the regression functions and propensity scores, respectively, and then, refit the generalized linear regression model for the regression functions and propensity scores with the union of the selected covariates in the previous step.

3.3 Valid Inference Based on Neyman Orthogonality Condition

Belloni et al. (2017) provide a more general framework to make causal inference for a variety of treatment effect parameters such as (local) ATE and (local) quantile treatment effects. As argued by Belloni et al. (2017), perfect model selection may be too restrictive to be satisfied in some real applications, and in order to provide valid inference for the treatment effect, the approximate sparsity assumptions allowing for small nonzero approximation error are made for the outcome regression and propensity score functions. With these assumptions, machine learning techniques, such as Lasso and Post-Lasso[2] (Belloni and Chernozhukov, 2013) methods, can be used to conduct the model selection. Furthermore, two conditions are required to proceed. One is the so-called Neyman orthogonality condition, which ensures that the estimating equations are first-order insensitive to the perturbations in the nuisance components, for instance, the outcome regression and propensity score functions. The other one is no overfitting condition

[1]Group Lasso is a method that generalizes the Lasso procedure to jointly select groups of covariates into or out of a model, see Yuan and Lin (2006) for more details.

[2]Apply the Lasso model section method first, and then, run the OLS regression to the selected variables.

requiring that the nuisance components should be estimated at a relatively slower $o(n^{-1/4})$ rate to insure small estimation bias. As a concrete example of this general framework, it can be verified that the DR ATE estimator (4) satisfies the Neyman orthogonality condition, and the estimated outcome regression and propensity score functions via Lasso or Post-Lasso model selection method satisfy the no overfitting condition automatically. In other words, to estimate and infer the ATE in data-rich environment, one can use the DR estimator with the nuisance components estimated via Lasso or Post-Lasso method and rely on the multiplier bootstrap procedure as described in Section 3.2 of Belloni et al. (2017) to make inferences.

The causal inference method based on DR estimators needs to model both the outcome regression and propensity score functions and the consistency assumptions for the model selection stage of both models are required. To relax these restrictions, several methods have been proposed recently, described in Sections 4 and 5, respectively.

§4 Inference of Treatment Effects Via Covariate Balance with High-dimensional Data

4.1 CBW Methods

In recent years, some CBW methods have been proposed by researchers to improve the finite sample size performance of the traditional IPW estimators. Broadly speaking, there are two ways to achieve the covariate balance among different groups: one depends on modeling the propensity score; see, e.g., the papers by Imai and Ratkovic (2014) and Sant'Anna, Song and Xu (2020), and the other directly aims at the balance weights via optimization designs; see, e.g., the papers by Zubizarreta (2015) and Chan, Yam and Zhang (2016). To have a general idea of the CBW method, by the definition of the propensity score and the law of iterated expectations, it can be shown that

$$E\left[\frac{Dg(X)}{\pi(X)}\right] = E\left[\frac{(1-D)g(X)}{1-\pi(X)}\right] = E[g(X)], \tag{5}$$

where $g(\cdot) : R^p \to R^m$ for some $m \geq 1$ is a continuous and integrable function. It can be seen from (5) that the propensity score has the property of balancing the moment of any functional form of the covariates among the treated, control and combined groups. Imai and Ratkovic (2014) exploit the finite moment conditions of (5) and propose the CBPS method, while Sant'Anna, Song and Xu (2020) make full use of (5) to consider the infinite moment conditions and propose the integrated propensity score procedure. For example, Imai and Ratkovic (2014) first assume a logistic model for the propensity score as

$$\pi_\beta(X) = \pi(X^\top\beta) = \frac{\exp(X^\top\beta)}{1+\exp(X^\top\beta)},$$

where $\beta \in \Theta \subseteq \mathbb{R}^p$ is a p-dimensional vector of parameters. Then, the unknown parameter β is estimated by

$$\hat{\beta} = \arg\min_\beta \ \bar{g}_\beta(D,X)' H \bar{g}_\beta(D,X),$$

where H is an $m \times m$ positive definite matrix and

$$\bar{g}_\beta(D, X) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{D_i}{\pi_\beta(X_i)} - \frac{1-D_i}{1-\pi_\beta(X_i)}\right) g(X_i).$$

Once the propensity score is estimated, the ATE estimator based on the CBPS method can be obtained via (2).

Viewing the inverse propensity score of (5) as a weight, through careful optimization designs, Zubizarreta (2015) proposes the stable balance weighting (SBW) method to choose the minimum sample variance of the weights that balance the covariates, while Chan, Yam and Zhang (2016) propose the calibration balance weighting method to choose the weights that minimize the calibration distance and balance the covariates as well. Specifically, the SBW weights in Zubizarreta (2015) can be obtained by solving the following optimization problem for $j = 0$ and 1,

$$\min_{w_{ji}} \sum_{i:D_i=j} (w_{ji} - 1/n_j)^2$$

subject to $w_{ji} > 0$, $\sum_{i:D_i=j} w_{ji} = 1$ and $\left|\sum_{i:D_i=j} w_{ji}X_{ik} - \bar{X}_k\right| \leq \delta_{jk}$, $1 \leq k \leq p$, where n_0 and n_1 are the numbers of units in the control and treated groups, respectively, X_{ik} is the k-th covariate of unit i, $\bar{X}_k$ is the mean of the k-th covariate in the combined group, and δ_{jk}, $j = 0, 1$; $k = 1, \cdots, p$, are pre-specified small positive numbers. Once the stable balance weights w_{1i} and w_{0i} are obtained, the weighted estimator of the ATE based on SBW method is given by

$$\hat{\Delta}_{SBW} = \sum_{i:D_i=1} w_{1i}Y_i - \sum_{i:D_i=0} w_{0i}Y_i.$$

In high-dimensional settings, another line of making valid inferences of treatment effects is to take the idea of covariate balancing into the construction of the estimators.

4.2 Approximate Residual Balancing Method

For high-dimensional linear model under unconfoundedness setup, Athey, Imbens and Wager (2018) propose the approximate residual balancing method which combines the penalized linear outcome regression model, Lasso or elastic net (Zou and Hastie, 2005), with the stable balance weights (Zubizarreta, 2015) assigned to the regression residuals. Particularly, suppose that $\mu_j(X) = X^\top\alpha_j$, $j = 0, 1$, the ATE estimator based on the approximate residual balancing method is constructed as

$$\hat{\Delta}_{\mathrm{ARB}} = \bar{X}^T(\hat{\alpha}_1 - \hat{\alpha}_0) + \sum_{i:D_i=1}\gamma_{1i}\left(Y_i - X_i^\top\hat{\alpha}_1\right) - \sum_{i:D_i=0}\gamma_{0i}\left(Y_i - X_i^\top\hat{\alpha}_0\right),$$

where

$$\hat{\alpha}_j = \arg\min_{\alpha}\left[\sum_{i:D_i=j}\left(Y_i - X_i^\top\alpha\right)^2 + \lambda\left\{(1-\rho)\|\alpha\|_2^2 + \rho\|\alpha\|_1\right\}\right],$$

and

$$\gamma_j = \arg\min_{b}\left\{(1-\zeta)\|b\|_2^2 + \zeta\left\|\bar{X} - \bar{X}_j^\top b\right\|_\infty^2 \text{ subject to } \sum_{i:D_i=j} b_i = 1 \text{ and } 0 \leqslant b_i \leqslant n_j^{-2/3}\right\}$$

with $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $\bar{X}_j = \frac{1}{n_j}\sum_{i:D_i=j} X_i$, $j = 0, 1$. Here, $0 < \zeta < 1$, $0 < \rho < 1$, and $\lambda > 0$ are tuning parameters.

As argued by Athey, Imbens and Wager (2018), the penalized linear regression step can be effective at capturing the main effects while the weighting regression residuals step can be effective at capturing additional small effects and makes compensations to the previous step, leading to an effective procedure for estimating the ATE in high-dimensional cases. Different from the methods discussed in Section 3, the approximate residual balancing method does not rely on modeling the relationship between the treatment variable and the covariates, so that the sparsity assumption for the propensity score is not required. The ATE estimator considered in Athey, Imbens and Wager (2018) is closely related to the DR estimators if one views the inverse propensity score in the DR estimator of (4) as a covariate balance weight. Also, as argued by Athey, Imbens and Wager (2018), the approximate residual balancing method is closely related to the de-biased Lasso studies for the regression model of high dimensions in the statistical literature.

4.3 High-Dimensional CBPS

Ning, Peng and Imai (2020) generalize the CBPS method by Imai and Ratkovic (2014) to introduce model selection methods into covariate balancing procedures, termed as high-dimensional CBPS (HD-CBPS). The CBPS method does not rely on modeling the outcome regression, while the HD-CBPS method requires the sparsity assumptions for both the outcome regression and the propensity score models to allow for model selections. To adjust for the bias caused by regularized estimation, Ning, Peng and Imai (2020) apply the covariate balancing procedure to the covariates selected from the penalized outcome regression model to obtain calibrated propensity score estimators. For the estimation of $\mu_1 = E[Y(1)]$, Ning, Peng and Imai (2020) first estimate the propensity score function via penalized logistic regression model

$$\tilde{\beta} = \underset{\beta}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} \left[D_i \cdot \left(X_i^\top \beta\right) - \log\left(1 + \exp\left(X_i^\top \beta\right)\right)\right] + \lambda\|\beta\|_1,$$

where $\lambda > 0$ is a tuning parameter, and estimate the outcome regression function by penalized regression model

$$\tilde{\alpha} = \underset{\alpha}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n} D_i(Y_i - X_i^\top \alpha)^2 + \lambda'\|\alpha\|_1,$$

where $\lambda' > 0$ is a tuning parameter, and then, they calibrate the estimated propensity score by balancing the covariates selected in the outcome regression model as follows,

$$\tilde{\gamma} = \underset{\gamma \in \mathbb{R}^{|\mathcal{S}|}}{\operatorname{argmin}} \left\| \frac{1}{n}\sum_{i=1}^{n} \left(\frac{D_i}{\pi\left(X_{i\mathcal{S}}^\top \gamma + X_{i\mathcal{S}^c}^\top \tilde{\beta}_{\mathcal{S}^c}\right)} - 1 \right) X_{i\mathcal{S}} \right\|_2^2,$$

where $\mathcal{S} = \{1 \le k \le p : |\tilde{\alpha}_k| > 0\}$ and $\mathcal{S}^c = \{1 \le k \le p : |\tilde{\alpha}_k| = 0\}$. Finally, μ_1 is estimated by the IPW estimator

$$\hat{\mu}_1^{HD-CBPS} = \frac{1}{n}\sum_{i=1}^{n} \frac{D_i Y_i}{\pi(X_{i\mathcal{S}}^\top \tilde{\gamma} + X_{i\mathcal{S}^c}^\top \tilde{\beta}_{\mathcal{S}^c})}.$$

Although the strong covariate balance property as in Imai and Ratkovic (2014) may not be held with high dimensions, Ning, Peng and Imai (2020) argue that once the outcome regression model is properly estimated, the HD-CBPS method would still approximately satisfy weak covariate balancing property in the sense that the covariate balancing property holds for the linear combination of the covariates. Ning, Peng and Imai (2020) also show that the HD-CBPS procedure is robust to the misspecification of either the outcome regression or the propensity score model and extend the linear outcome regression model to generalized linear cases.

Finally, Fan et al. (2021) propose a new method to combine the GMM-Lasso type estimation with the covariate balancing procedure under the framework of the CBPS method. The proposed method can not only balance the covariates but also select the relevant ones. Under sparsity conditions, Fan et al. (2021) explore the effectiveness of the proposed method via simulation studies.

4.4 Model-Assisted Inference for Treatment Effects with Regularized Calibrated Estimation

Tan (2020a) proposes a calibrated estimation procedure for the propensity score from the perspective of the loss function. Similar to the CBW method based on the propensity score model, a logistic regression model is assumed to the propensity score. Tan (2020a) argues that solving the covariate balance equation between the treated (or control) and combined groups is equivalent to minimizing the calibrated loss function, which is a key component in Tan (2020a, 2020b). To derive an ATE estimator in the high-dimensional settings, Tan (2020a) adds a Lasso penalty to the calibrated loss function. By plugging the minimal solution of the regularized calibrated loss function into the logistic regression model, one obtains the calibrated propensity score estimator and leads to a regularized calibrated IPW estimator for the ATE. Finally, Tan (2020a) illustrates the advantages of the proposed methods over the conventional maximum likelihood method via a simulation study.

Tan (2020b) proposes a DR type regularized calibrated estimation procedure for the ATE with high-dimensional data. The regularized calibrated ATE estimator provides not only DR point estimators but also model-assisted confidence intervals. The term "model-assisted" means that the confidence intervals are valid when the propensity score function is correctly specified or when a linear outcome regression model is assumed and correctly specified. The model-assisted confidence intervals property makes the proposed method in Tan (2020b) one step further when compared with the methods introduced in Section 3. Under the unconfoundedness and overlap assumptions, the DR ATE estimators are used in Tan (2020b) and the calibrated loss function with a Lasso penalty as in Tan (2020a) is used for the propensity score model while a carefully chosen weighted least-square loss function with Lasso penalty is used for the outcome regression model.

Clearly, causal inference with covariate balance can either circumvent the modeling of propensity score function or make inference robust to the misspecification of the propensity score model. This is certainly an advantage over the methods based on the DR estimators in

Section 3.

§5 Causal Inference Based on SDR

A third line of making causal inference with high dimensions is the SDR method, which captures full information of covariates through a linear combination of the whole covariates. Indeed, Huang and Chan (2017) propose the joint SDR method for the ATE, where the "joint" means that the proposed method aims at finding a linear combination $\mathbf{B}^\top X$ to simultaneously satisfy

$$D\perp X\,|\,\mathbf{B}^\top X,\quad Y(0)\perp X\,|\,\mathbf{B}^\top X\quad\text{and}\quad Y(1)\perp X\,|\,\mathbf{B}^\top X,$$

where $\mathbf{B}$ is $p\times r$ unknown parameter matrix with $r\leq p$. Once the parameter matrix $\mathbf{B}$ is estimated, denoted as $\hat{\mathbf{B}}$, the ATE is then estimated by the outcome regression estimator (1), where the outcome regression model is estimated by sieve approach with respect to $\hat{\mathbf{B}}X_i$. Furthermore, Luo, Zhu and Ghosh (2017) propose regression-based SDR to estimate the ATE, where the focus is to find $\mathbf{B}_0^\top X$ and $\mathbf{B}_1^\top X$ such that

$$E(Y(0)|X)\perp X\,|\,\mathbf{B}_0^\top X\quad\text{and}\quad E(Y(1)|X)\perp X\,|\,\mathbf{B}_1^\top X,$$

where $\mathbf{B}_j$ is $p\times r(j)$ unknown parameter matrix with $r(j)\leq p$, $j=0,1$. The minimum average variance estimation method as in Xia et al. (2002) is adopted to estimate the parameter matrix $\mathbf{B}_j$, $j=0,1$ and to recover the outcome regression function simultaneously.

However, the theoretical results in Huang and Chan (2017) and Luo, Zhu and Ghosh (2017) are derived only for fixed dimensions. Recently, Ma et al. (2019) propose the sparse SDR method for causal inference of ATE with high dimensions. Notice that both the outcome regression and propensity score models can be viewed as conditional expectations. Let W be a random variable that can be taken as the potential outcomes $Y(j)$, $j=0,1$ or the treatment variable D. The goal of the sparse SDR method is to find a $p\times r$ sparse parameter matrix $\mathbf{B}$ such that

$$E(W|X)=E(W|\mathbf{B}^\top X),$$

where $\mathbf{B}^\top X=(\mathbf{B}_{\mathcal{R}}^\top,\mathbf{0}_{(p-s_{\mathcal{R}})\times r}^\top)^\top X=\mathbf{B}_{\mathcal{R}}^\top X_{\mathcal{R}}$, $\mathcal{R}$ denotes as the set of indices of relevant covariates, and $s_{\mathcal{R}}$ is the number of the relevant covariates. To estimate the sparse parameter matrix $\mathbf{B}$, Ma et al. (2019) adopt the sparse reduced-rank regression method proposed by Chen and Huang (2012). Then, the conditional expectation $E(W|X)$ is estimated by multivariate kernel method and the ATE is estimated by the DR estimator (4). Finally, Ma et al. (2019) show that the proposed ATE estimator is $\sqrt{n}$ consistent, asymptotic normal and semi-parametrically efficient.

Although the inference of the ATE is based on the DR estimator, the sparse SDR method does not rely on the restrictive parametric modeling assumptions on the outcome regression and propensity score functions. In this regard, the modeling strategy is similar to the generalized random forest method introduced in the next section.

§6 Machine Learning Methods

The fourth line for making valid inference of treatment effects lies in the machine learning, especially the random forest methods. Random forest, firstly introduced by Breiman (2001), is known as one of the most attractive machine learning methods. Similar to the bagging method, developed by Breiman (1996), random forest is an ensemble algorithm consisting of many decision or regression trees and making an average of the results produced by different trees. In the bagging method, all covariates are used to generate a single tree while in the random forest method, only a fraction of covariates are used to generate a single tree. This small modification makes the trees in the random forest more diverse than those in the bagging algorithm and improves the generalization property of the resulting model.

It is well known that the traditional machine learning methods are good at prediction rather than inference. However, Wager and Athey (2018) propose the causal forest method to introduce the random forest method into the causal inference and derive the consistency and asymptotical normality results as well as asymptotic confidence intervals for the treatment effect estimators. The treatment effects are estimated via the causal forest consisting of many causal trees which can be viewed as the nearest-neighbor methods with the leaves in the causal tree as the neighborhood metrics. An important advantage of the causal forest over the nearest-neighbor methods is that the weights in the causal forest are data-driven allowing for more flexibility. Thinking of the leaves in a causal tree as small enough, the individuals in the same leaf can be treated as randomly assigned so that the simple calculation between the treated and control individuals as in the randomization experiment can be applied. The causal forest then averages the results produced by the causal trees to obtain the treatment effect estimate.

Recently, Athey, Tibshirani and Wager (2019) propose the generalized random forest method that adopts the random forest algorithm to estimate various types of conditional expectations. Obviously, the causal forest method by Wager and Athey (2018) directly aims at the treatment effect estimation while the generalized random forest method is indirectly used to estimate the treatment effect through intermediate parameters. Estimating conditional expectations via the random forest algorithm can be viewed as the nonparametric method. In classical nonparametric kernel methods, a series of weights are determined by the kernel function whose values vary with the distance to the given conditional variable. In practice, these classical methods suffer from the "curse of dimensionality" problem and may lead to poor performance as the dimensions of the covariates increase. However, the weights assigned to the response variable in Athey, Tibshirani and Wager (2019) are determined by the random forest method, making it possible to deal with the high-dimensional settings. The basic idea in Athey, Tibshirani and Wager (2019) is similar to that in Wager and Athey (2018) by viewing the leaves in the generating trees as the closeness metrics. Furthermore, the large sample theories of consistency and asymptotic normality are derived in Athey, Tibshirani and Wager (2019). Although the theoretical results in both Wager and Athey (2018) and Athey, Tibshirani and Wager (2019) are derived for low-dimensional settings, as long as the sparsity assumptions are made to the true model, the algorithms would still work for the high dimensions since the irrelevant covariates

would be ignored by the forests, which is implicitly pointed out in Section 3 of Wager and Athey (2018).

The advantage of the random forest methods is that they are flexible and do not rely on postulating specific models for the outcome regression and potential outcome functions. Specifically, the random forest-based methods can be viewed as nonparametric estimations that can deal with the curse of dimensionality issue.

Remark 2: *In addition to the random forest or the generalized random forest as mentioned above, it should be very attractive and demanding to develop other machine learning methods to study causal inferences, which is clearly warranted to investigate.*

§7 Conclusion

In the big data environment, on the one hand, it can make the unconfoundedness assumption more credible, on the other hand, one needs to determine which covariate (or its functional form) should be included in the true model. The inference of treatment effects based on the DR estimators is valid in high-dimensional settings, indicating that the DR estimator has the self de-biased property. Another line of making valid inference of treatment effects with high-dimensional data is to exploit the covariate balance property. The utilization of the covariate balance property can either circumvent modeling the propensity score or relax the model specification assumptions. Furthermore, one can base on the SDR method to make causal inference in high dimensions and one advantage of this method is that it does not require consistency for variable selection. Finally, one can directly or indirectly make inference of treatment effect via the machine learning algorithms such as random forest. The approximate residual balancing method and the inference of treatment effects based on the DR estimators are closely related to the de-biased Lasso method for the regression model with high dimensions in the statistical literature.

The DR estimators for causal inference need to consider model selection in both the outcome regression and propensity score functions. However, the estimators based on the covariate balance methods can ease the dependence on the PS models by using the auxiliary information. The estimators based on the SDR method do not rely on restrictive assumptions on OR and PS models. The methods to estimate the treatment effects via the random forest algorithm are flexible and do not rely on modeling the OR and PS functions. For empirical applications, the first two methods are recommended when one believes that it is reasonable to model the OR and PS functions in parametric forms, while if one has no idea of what functional forms of the OR and PS models should be, it is better to try the last two methods.

Currently, the causal inference in high dimensions mainly focuses on the ATE estimate. How to make valid inference for quantile treatment effects besides the Neyman orthogonality condition? Can one generalize the existing methods to the other setups, such as, difference-in-differences, local average treatment effects and regression discontinuity design? Do other machine learning methods such as support vector machines work for causal inferences in high-

dimensional settings? These questions are left as future research topics.

Acknowledgement

The authors thank the editor and two anonymous referees for their helpful and constructive comments, which have improved the presentation of this article.

References

[1] S Athey, G W Imbens, S Wager. *Approximate residual balancing: debiased inference of average treatment effects in high dimensions*, Journal of the Royal Statistical Society, Series B, 2018, 80(4): 597-623.

[2] S Athey, J Tibshirani, S Wager. *Generalized random forests*, Annals of Statistics, 2019, 47(2): 1148-1178.

[3] A Belloni, V Chernozhukov. *Least squares after model selection in high-dimensional sparse models*, Bernoulli, 2013, 19(2): 521-547.

[4] A Belloni, V Chernozhukov, C Hansen. *Inference on treatment effects after selection among high-dimensional controls*, Review of Economic Studies, 2014, 81(2): 608-650.

[5] A Belloni, V Chernozhukov, I Fernández-Val, C Hansen. *Program evaluation and causal inference with high-dimensional data*, Econometrica, 2017, 85(1): 233-298.

[6] L Breiman. *Bagging predictors*, Machine Learning, 1996, 24(2): 123-140.

[7] L Breiman. *Random forests*, Machine Learning, 2001, 45(1): 5-32.

[8] Z Cai. *Recent developments in estimating treatment effects for panel data*, China Journal of Econometrics, 2021, 1(2): 233-249.

[9] C Carvalho, R Masini, M C Medeiros. *ArCo: An artificial counterfactual approach for high-dimensional panel time-series data*, Journal of Econometrics, 2018, 207(2): 352-380.

[10] G Cerulli. *Econometric Evaluation of Socio-Economic Programs*. Advanced Studies in Theoretical and Applied Econometrics, Berlin Heidelber: Springer, 2015, 49.

[11] K C G Chan, S C P Yam, Z Zhang. *Globally efficient nonparametric inference of average treatment effects by empirical balancing calibration weighting*, Journal of the Royal Statistical Society, Series B, 2016, 78(3): 673-700.

[12] L Chen, J Z Huang. *Sparse reduced-rank regression for simultaneous dimension reduction and variable selection*, Journal of the American Statistical Association, 2012, 107(500): 1533-1545.

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

[14] J Fan, M Zhan, Z Cai, Y Fang, M Lin. *Covariate balancing in propensity score estimation with variable selection: Based on GMM-LASSO approach*, Systems Engineering - Theory & Practice, 2021, 41(10): 2631-2639.

[15] C Hsiao, S H Ching, K S Wan. *A panel data approach for program evaluation: measuring the benefits of political and economic integration of Hong kong with mainland China*, Journal of Applied Econometrics, 2012, 27(5): 705-740.

[16] M Y Huang, K C G Chan. *Joint sufficient dimension reduction and estimation of conditional and average treatment effects*, Biometrika, 2017, 104(3): 583-596.

[17] K Imai, M Ratkovic. *Covariate balancing propensity score*, Journal of the Royal Statistical Society, Series B, 2014, 76(1): 243-263.

[18] G W Imbens, J M Wooldridge. *Recent developments in the econometrics of program evaluation*, Journal of Economic Literature, 2009, 47(1): 5-86.

[19] A Javanmard, A Montanari. *Confidence intervals and hypothesis testing for high-dimensional regression*, Journal of Machine Learning Research, 2014, 15(1): 2869-2909.

[20] Z Liu, Z Cai, Y Fang, M Lin. *Statistical analysis and evaluation of macroeconomic policies: a selective review*, Applied Mathematics - A Journal of Chinese Universities, 2020, 35(1): 57-83.

[21] W Luo, Y Zhu, D Ghosh. *On estimating regression-based causal effects using sufficient dimension reduction*, Biometrika, 2017, 104(1): 51-65.

[22] S Ma, L Zhu, Z Zhang, C L Tsai, R J Carroll. *A robust and efficient approach to causal inference based on sparse sufficient dimension reduction*, Annals of Statistics, 2019, 47(3): 1505-1535.

[23] Y Ning, S Peng, K Imai. *Robust estimation of causal effects via a high-dimensional covariate balancing propensity score*, Biometrika, 2020, 107(3): 533-554.

[24] P R Rosenbaum, D B Rubin. *The central role of the propensity score in observational studies for causal effects*, Biometrika, 1983, 70(1): 41-55.

[25] D B Rubin. *Matching to remove bias in observational studies*, Biometrics, 1973, 29(1), 159-183.

[26] D B Rubin. *Estimating causal effects of treatments in randomized and nonrandomized studies*, Journal of Educational Psychology, 1974, 66(5): 688-701

[27] D B Rubin. *Assignment to treatment group on the basis of a covariate*, Journal of Educational Statistics, 1977, 2(1): 1-26.

[28] P H Sant'Anna, X Song, Q Xu. *Covariate distribution balance via propensity scores*, 2020, arXiv preprint arXiv:1810.01370v4.

[29] Z Shi, J Huang. *Forward-selected panel data approach for program evaluation*, Journal of Econometrics, 2021, https://doi.org/10.1016/j.jeconom.2021.04.009.

[30] Z Tan. *Regularized calibrated estimation of propensity scores with model misspecification and high-dimensional data*, Biometrika, 2020, 107(1): 137-158.

[31] Z Tan. *Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data*, Annals of Statistics, 2020b, 48(2): 811-837.

[32] R Tibshirani. *Regression shrinkage and selection via the Lasso*, Journal of the Royal Statistical Society, Series B, 1996, 58(1): 267-288.

[33] S Van de Geer, P Bühlmann, Y Ritov, R Dezeure. *On asymptotically optimal confidence regions and tests for high-dimensional models*, Annals of Statistics, 2014, 42(3): 1166-1202.

[34] S Wager, S Athey. *Estimation and inference of heterogeneous treatment effects using random forests*, Journal of the American Statistical Association, 2018, 113(523): 1228-1242.

[35] Y Xia, H Tong, W K Li, L X Zhu. *An adaptive estimation of dimension reduction space*, Journal of the Royal Statistical Society, Series B, 2002, 64(3): 363-410.

[36] M Yuan, Y Lin. *Model selection and estimation in regression with grouped variables*, Journal of the Royal Statistical Society, Series B, 2006, 68(1): 49-67.

[37] C H Zhang, S S Zhang. *Confidence intervals for low dimensional parameters in high dimensional linear models*, Journal of the Royal Statistical Society: Series B, 2014, 76(1): 217-242.

[38] H Zou, T Hastie. *Regularization and variable selection via the elastic net*, Journal of the Royal Statistical Society, Series B, 2005, 67(2): 301-320.

[39] J R Zubizarreta. *Stable weights that balance covariates for estimation with incomplete outcome data*, Journal of the American Statistical Association, 2015, 110(511): 910-922.

[1]Wang Yanan Institute for Studies in Economics and Fujian Key Laboratory of Statistical Sciences, Xiamen University, Xiamen 361005, China.

[2]Department of Economics, University of Kansas, Lawrence, KS 66045, USA.

[3]Department of Statistics and Data Science, Xiamen University, Xiamen 361005, China.

Email: linming50@xmu.edu.cn