Wei Luo, Wenbo Wu, and Yeying Zhu*

Learning Heterogeneity in Causal Inference Using Sufficient Dimension Reduction

https://doi.org/10.1515/jci-2018-0015
Received June 4, 2018; revised September 26, 2018; accepted October 2, 2018

Abstract: Often the research interest in causal inference is on the regression causal effect, which is the mean difference in the potential outcomes conditional on the covariates. In this paper, we use sufficient dimension reduction to estimate a lower dimensional linear combination of the covariates that is sufficient to model the regression causal effect. Compared with the existing applications of sufficient dimension reduction in causal inference, our approaches are more efficient in reducing the dimensionality of covariates, and avoid estimating the individual outcome regressions. The proposed approaches can be used in three ways to assist modeling the regression causal effect: to conduct variable selection, to improve the estimation accuracy, and to detect the heterogeneity. Their usefulness are illustrated by both simulation studies and a real data example.

Keywords: Central causal effect subspace, Conditional causal effect, Heterogeneity, Variable selection

1 Introduction

Causal inference has been widely applied for decades to draw cause-and-effect conclusions based on observational studies, in which treatments are assigned to observations in a non-random fashion. In many cases, theories in causal inference are developed under the potential outcome framework [1]. That is, when the treatment assignment is binary, i. e., either treated or untreated, the outcome variable in the hypothetical complete data set has two components (Y_0, Y_1), in which Y_0 is the outcome if the subject is untreated and Y_1 is the outcome if treated. Let X be a set of covariates recorded in the study that collects subject's personal information. The regression causal effect, defined by $E(Y_1 - Y_0 \mid X)$, the mean difference in the potential outcomes given the subject's personal information, has received increasing attention in the recent years. As the average causal effect, the conventional parameter of interest in causal inference, does not consider subjects' personal information, the regression causal effect uniquely serves as the parameter of interest in applications such as individualized treatment assignment in precision medicine.

Let T be the treatment assignment with support $\{0, 1\}$. Since for each subject, only Y_T is observable, the missing value mechanism must be regulated in order to estimate the regression causal effect. In the literature, it is commonly assumed that X includes all the potential confounders; that is,

$$T \perp\!\!\!\perp Y_t \mid X, \quad \text{for } t = 0, 1, \tag{1}$$

under which the distribution of Y_t given X is identical whether or not restricted to a specific treatment group. As (1) applies to the individual outcomes Y_0 and Y_1, most existing methods implicitly or explicitly adopt a two-step strategy: first, to perform regression analysis within each treatment group; that is, to estimate the individual outcome regression functions $E(Y_0 \mid X)$ and $E(Y_1 \mid X)$ separately; second, to estimate the regression causal effect by the difference of the two functions. Examples include G-computation [2, 3] and difference lasso [4], etc. In particular, Luo, Zhu, and Ghosh [5] used sufficient dimension reduction to propose

***Corresponding author: Yeying Zhu,** Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Canada, e-mail: yeying.zhu@uwaterloo.ca
Wei Luo, Center for Data Science, Zhejiang University, Hangzhou, China, e-mail: weiluo@zju.edu.cn
Wenbo Wu, Department of Management Science and Statistics, The University of Texas at San Antonio, San Antonio, United States, e-mail: wenbo.wu@utsa.edu

a semiparametric estimator that is model-free and effective in finite samples. Exceptions that do not adopt this strategy include Tian, Alizadeh, Gentles, and Tibshirani [6], Abrevaya, Hsu, and Lieli [7], and some recent literature on the optimal treatment study [8, 9] that focused on doubly-robust approaches.

Because the regression causal effect is only a part of the two-dimensional function $\{E(Y_0 \mid X), E(Y_1 \mid X)\}$, the two-step strategy introduces nuisance functional parameters. To see this point, let $\eta(\cdot)$ be a function of X, and

$$E(Y_t \mid X) = \eta(X) + t, \quad \text{for } t = 0, 1. \tag{2}$$

Then η must be estimated in the first step. However, this estimation would not be needed had we known that the regression causal effect is a constant, in which case we could estimate $E(Y_t)$ as the first step instead. The gain is substantial when the nuisance function $\eta(X)$ has a complex form and hard to fit. The issue of estimating nuisance functions has also been observed in Tian et al. [6], in the case of balanced completely randomized experimental studies.

In addition to fitting the regression function, another common interest about the regression causal effect is to find a pre-assumed low-dimensional transformation of the covariates, denoted by $\tau(X)$, such that the regression causal effect is a function of $\tau(X)$; that is, $\tau(X)$ satisfies

$$Y_1 - Y_0 = f(\tau(X)) + \epsilon, \tag{3}$$

where f is an free and unknown function and the error term ϵ satisfies $E(\epsilon \mid X) = 0$. In particular, $\tau(X)$ is a constant in Model (2). In general, estimation of $\tau(X)$ is useful in three ways. First, it tells which part of the covariates is informative to the regression causal effect, so that researchers, such as policy makers, can develop simpler mechanism by controlling this part only. Second, by using $\tau(X)$ in place of X to estimate the regression causal effect, data visualization is possible and the number of parameters is reduced in the subsequent modeling, both of which enhance the accuracy and interpretability of the estimation. Third, hypothesis testing on whether $\tau(X)$ is a constant can be of interest to some researchers, as the acceptance implies a constant regression causal effect, which would be useful in different areas such as subgroup identification [10]. For consistency with the literature in causal inference [11], hereafter we call the regression causal effect homogeneous when it is a constant, and call it heterogeneous otherwise.

In the literature, estimation of $\tau(X)$ has been studied under different scenarios. When $\tau(X)$ is restricted to a lower-dimensional linear combination of covariates, Luo et al. [5] employed sufficient dimension reduction on each of the individual outcome regressions. However, by doing so, their method requires the existence of low dimensional linear combinations of covariates that are sufficient for regressing each of Y_0 and Y_1, which is more restrictive than (3). In other words, Luo et al.'s method may introduce redundant covariates in the resulting estimate of $\tau(X)$. This adversely affects the effectiveness of the method, for example, in Model (2) if $\eta(\cdot)$ does not have a low dimensional structure.

When $\tau(X)$ is restricted to a subset of covariates, variable selection methods have been conducted, see the lasso approach [12, 6], virtual twins method [10], the hypothesis testing procedure [13], and cross-validation [13]. For testing the heterogeneity of $\tau(X)$, Crump et al. [11] proposed parametric and nonparametric tests. The former can detect $n^{1/2}$-order fluctuation from the homogeneity, and the latter is more conservative. Similar to Luo et al. [5], these methods are based on the individual outcome regressions, so they are ineffective when the "main effect" of X, e. g. $\eta(X)$ in Model (2), is complex.

In this paper, we use sufficient dimension reduction to propose a new and model-free estimator of $\tau(X)$, assuming that a consistent estimation of the propensity score, the probability that a subject is treated conditional on X, is given a priori. Compared with Luo et al. [5], the estimator is directly built on the regression causal effect rather than the individual outcome regressions, so it is more efficient in reducing dimensionality. In practice, it is applicable when the propensity score is easy to tackle, and particularly useful when the individual outcome regressions are complex and need to be fitted nonparametrically. In addition, it can be slightly modified to perform variable selection, and detect the heterogeneity of the regression causal effect. For simplicity, we assume X to be continuous and have zero mean throughout the theoretical development.

2 Sufficient dimension reduction

Sufficient dimension reduction is a family of methods that aims to reduce the dimension of covariates prior to subsequent modeling. When the regression of a response variable W on the p-dimensional covariates X is of interest, it assumes the existence of $\beta \in \mathbb{R}^{p\times d}$, where $d < p$, such that $E(W \mid X)$ is a measurable function of $\beta^{\mathsf{T}}X$, or equivalently,

$$W = f(\beta^{\mathsf{T}}X) + \epsilon, \tag{4}$$

where f is a free and unknown function and ϵ satisfies $E(\epsilon \mid X) = 0$. Because for any β that satisfies (4) and any matrix A of full row-rank, βA still satisfies (4) if f is adjusted accordingly, one needs to consider the linear space spanned by the columns of β with minimal dimension for identifiable parametrization. Cook and Li [14] showed that under fairly general conditions on X, which we assume throughout the article, such space with minimal dimension is unique. This space, commonly called the central mean subspace and denoted by $\mathcal{S}_{E(W|X)}$, is then the parameter of interest in sufficient dimension reduction.

Existing methods for estimating the central mean subspace include ordinary least square [15], principal Hessian directions [16], iterative Hessian transformations [14], minimal average variance estimation [17], and other semiparametric methods [18, 19], etc. A method is called Fisher-consistent if it recovers a subspace of $\mathcal{S}_{E(W|X)}$, and is called exhaustive if this subspace further coincides with $\mathcal{S}_{E(W|X)}$. For example, the ordinary least square estimate is spanned by $\Sigma_X^{-1/2}\nu_{OLS}$, where Σ_X is the covariance matrix of X and $\nu_{OLS} = \Sigma_X^{-1/2}E[X\{W - E(W)\}]$. Its Fisher-consistency requires the linearity condition on X:

$$E(X|\beta^{\mathsf{T}}X) = \Sigma_X\beta(\beta^{\mathsf{T}}\Sigma_X\beta)^{-1}\beta^{\mathsf{T}}X, \tag{5}$$

where β spans $\mathcal{S}_{E(W|X)}$. If one strengthens (5) to that the equation is satisfied for an arbitrary β, then the condition is equivalent to an elliptical distribution of X. Principal Hessian directions first estimates $M_{pHd} = \Sigma_X^{-1/2}E[XX^{\mathsf{T}}\{W - E(W)\}]\Sigma_X^{-1/2}$, and then multiplies the column space of the matrix with $\Sigma_X^{-1/2}$ to estimate $\mathcal{S}_{E(W|X)}$. In addition to the linearity condition, its Fisher-consistency requires the constant variance condition on X:

$$\mathrm{var}(X|\beta^{\mathsf{T}}X) \equiv \Sigma_X - \Sigma_X\beta(\beta^{\mathsf{T}}\Sigma_X\beta)^{-1}\beta^{\mathsf{T}}\Sigma_X, \tag{6}$$

where β spans $\mathcal{S}_{E(W|X)}$. If one strengthens both (5) and (6) to that they are satisfied for an arbitrary β, then these conditions together require X to have a joint normal distribution. These conditions also hold approximately when the dimension of X is relatively large [20], for which they are not considered restrictive in applications. Using the sample moments to estimate the corresponding population moments, both ordinary least square and principal Hessian directions are easy to implement.

Because ordinary least square is built upon $E(XW)$ and principal Hessian directions is built upon $E(XX^{\mathsf{T}}W)$, the former is effective when the regression function $E(W|X)$ is relatively asymmetric, and the latter is effective when the regression is relatively symmetric. In practice, these methods are often used complementarily to form an ensemble. Throughout the article, we assume that the ensemble method is exhaustive.

A separate issue in sufficient dimension reduction is to estimate the dimension d of $\mathcal{S}_{E(W|X)}$, also known as order determination. Existing methods include the sequential tests [16, 14, 21], the information criterion [22], and the ladle estimator [23], etc. All these methods can detect the case $d = 0$, which corresponds to a homogeneous regression function $E(W|X)$.

3 The central causal effect subspace

Before digging into more details, we introduce the notations and some regularity conditions. Let $\Delta Y = Y_1 - Y_0$. The regression causal effect is $E(\Delta Y|X)$. We assume that $\mathrm{var}(Y_t|X)$ exists and is integrable for $t = 0, 1$, which

implies the existence and integrability of $\mathrm{var}(\Delta Y|X)$. For any random element R, let $\Omega(R)$ be the support of R. We denote the propensity score by $\pi(X)$, and assume the common support condition. That is, there exists $c > 0$ such that

$$\Omega(\pi(X)) \subseteq (c, 1-c). \tag{7}$$

For any real vector v, denote its Euclidean norm by $\|v\|_2$. We treat v as a matrix with one column whenever needed. For any matrix β, let $\mathrm{vec}(\beta)$ be the vectorization of β, $\beta^{\otimes 2}$ be $\beta\beta^{\mathsf{T}}$, $\mathcal{S}(\beta)$ be the column space of β, and for any index set A, let β_A consist of rows of β indexed by A and β_{-A} be the rest of β. When β is a square matrix, let $tr(\beta)$ be the trace of β. We use $\mathbf{0}$ to denote the origin of a real space of arbitrary dimension, if no ambiguity is caused. For any two linear subspaces $\mathcal{S}_1$ and $\mathcal{S}_2$ in $\mathbb{R}^p$, let $\mathcal{S}(\mathcal{S}_1, \mathcal{S}_2)$ be the space spanned by their union and $\Pi(\mathcal{S}_1)$ be the projection matrix of $\mathcal{S}_1$ under $\|\cdot\|_2$. The deviation between $\mathcal{S}_1$ and $\mathcal{S}_2$ is measured by the maximum eigenvalue of $\{\Pi(\mathcal{S}_1) - \Pi(\mathcal{S}_2)\}^{\otimes 2}$, denoted by $D(\mathcal{S}_1, \mathcal{S}_2)$.

When $\tau(X)$ in (3) is forced to be a linear combination of covariates, (3) coincides with the sufficient dimension reduction assumption (4) with ΔY as the response variable. Naturally, to estimate $\tau(X)$ in this case, our parameter of interest is the central mean subspace $\mathcal{S}_{E(\Delta Y|X)}$.

In the literature, multiple papers have studied the application of sufficient dimension reduction in causal inference. Ghosh [24] applied it to the propensity score and introduced $\mathcal{S}_{E(T|X)}$, which can be irrelevant to $\mathcal{S}_{E(\Delta Y|X)}$. As mentioned in § 1, Luo et al. [5] applied it to the individual outcome regressions and introduced $\mathcal{S}(\mathcal{S}_{E(Y_0|X)}, \mathcal{S}_{E(Y_1|X)})$. Hu, Follmann, and Wang [25] and Huang and Chan [26] combined the two and introduced $\mathcal{S}(\mathcal{S}_{E(Y_0|X)}, \mathcal{S}_{E(Y_1|X)}, \mathcal{S}_{E(T|X)})$. While $\mathcal{S}(\mathcal{S}_{E(Y_0|X)}, \mathcal{S}_{E(Y_1|X)})$ can be easily shown to include $\mathcal{S}_{E(\Delta Y|X)}$, their difference can be substantial. An example is Model (2) with η measurable of $\|X\|_2$, where both $\mathcal{S}_{E(Y_0|X)}$ and $\mathcal{S}_{E(Y_1|X)}$ are p-dimensional but $\mathcal{S}_{E(\Delta Y|X)}$ vanishes; a similar example can be seen later in § 8.

Because it is the regression causal effect, rather than the outcome regressions or the propensity score, that serves as the primary interest in causal inference, it must be $\mathcal{S}_{E(\Delta Y|X)}$ that serves as the parameter of interest when applying sufficient dimension reduction. For this reason, we call $\mathcal{S}_{E(\Delta Y|X)}$ the central causal effect subspace. As we are aware of, there has been no application of sufficient dimension reduction in causal inference that target exactly at this space.

Because ΔY is unobserved, the aforementioned sufficient dimension reduction methods are not directly applicable to estimate $\mathcal{S}_{E(\Delta Y|X)}$. However, had the propensity score been known a priori, we can use the inverse probability weighting to construct

$$Y_\Delta = \frac{TY_1}{\pi(X)} - \frac{(1-T)Y_0}{1-\pi(X)} \tag{8}$$

which is observable from the data and can substitute ΔY for our purpose. That is,

$$E(\Delta Y \mid X) = E(Y_\Delta \mid X). \tag{9}$$

This observation is the key to our theoretical development. The proof of (9) is straightforward and is omitted. Its weaker version $E(\Delta Y) = E(Y_\Delta)$ has been commonly used in the literature of causal inference when the average causal effect is the parameter of interest. When the propensity score is known to be degenerate at 0.5, which happens in a balanced completely randomized experimental study, Tian et al. [6] proposed a similar result to (9) by assuming a parametric model for the regression causal effect. (9) gives a model-free result and is applicable for general observational studies. In particular, it readily implies the following theorem.

Theorem 1. *For Y_Δ defined in (8), we have $\mathcal{S}_{E(\Delta Y|X)} = \mathcal{S}_{E(Y_\Delta|X)}$.*

The proof of this theorem can be found in Appendix A.1. In practice, the propensity score is unknown and needs to be estimated. Throughout the article, we assume that a consistent estimator $\hat{\pi}(X)$ is given a priori. Considering the rich literature of propensity score estimation, this assumption is realistic. A variety of methods, including the conventional logistic regression, the more recent covariate balancing propensity score [27], and in particular the super learner [28] which allows more flexible model structure, have been

shown effective in applications. A semiparametric estimation can also be constructed using $\mathcal{S}_{E(T|X)}$ [24], which is $\sqrt{n}$-consistent when $\mathcal{S}_{E(T|X)}$ is low dimensional, details omitted.

For ease of asymptotic study, we additionally assume $\hat{\pi}(X)$ is constructed based on an appropriate parametric model, in which the estimation of parameters is asymptotic linear; that is,

(C1) Assume $\pi(X) = \phi(\alpha_0^{\mathsf{T}} h(X))$, where $\alpha_0 \in \mathbb{R}^r$ and ϕ and h are known functions such that $E\{h^4(X)\}$ exists and ϕ is continuously twice differentiable. $\hat{\pi}(X)$ is given by $\phi(\hat{\alpha}^{\mathsf{T}} h(X))$, where $\hat{\alpha}$ is asymptotic linear, i. e., there exists $g : \Omega(X, T) \to \mathbb{R}^r$ with $E\{g(X, T)\} = 0$ and $E\{g^{\otimes 2}(X, T)\} < \infty$, such that $\hat{\alpha} = \alpha_0 + E_n\{g(X, T)\} + O_P(n^{-1})$.

This asymptotic linearity assumption holds for the logistic regression, the covariate balancing propensity score, and the super learner, etc., so it is fairly general. The case where (C1) is violated will be discussed in § 8.

Using $\hat{\pi}(X)$, we can estimate Y_Δ accordingly. By Theorem 1, we can estimate the central causal effect subspace $\mathcal{S}_{E(\Delta Y|X)}$ by equivalently estimating $\mathcal{S}_{E(Y_\Delta|X)}$, for which all the sufficient dimension reduction methods in § 2 can be applied. As an illustration, we use the ensemble of ordinary least square and principal Hessian directions, for its exhaustiveness and ease of implementation. The implementation procedure is listed in the following. Because the proposed estimator consists of moments, we call it the ensemble moment-based estimator.

Step 0. Let $\hat{Y}_\Delta = TY_1/\hat{\pi}(X) - (1-T)Y_0/\{1-\hat{\pi}(X)\}$ and $Z = \Sigma_X^{-1/2}\{X - E(X)\}$. Estimate Σ_X by $\hat{\Sigma}_X = E_n(X^{\otimes 2}) - E^{\otimes 2}(X)$, and let $\hat{Z} = \hat{\Sigma}_X^{-1/2}\{X - E_n(X)\}$.

Step 1. Estimate $v_{OLS} = E(Z\Delta Y)$ by $\hat{v}_{OLS} = E_n(\hat{Z}\hat{Y}_\Delta)$.

Step 2. Estimate $M_{pHd} = E[Z^{\otimes 2}\{\Delta Y - E(\Delta Y)\}]$ by $\hat{M}_{pHd} = E_n[\hat{Z}^{\otimes 2}\{\hat{Y}_\Delta - E_n(\hat{Y}_\Delta)\}]$.

Step 3. Let $M_{ens} = v_{OLS}^{\otimes 2} + M_{pHd}^{\otimes 2}$ and $\hat{M}_{ens} = \hat{v}_{OLS}^{\otimes 2} + \hat{M}_{pHd}^{\otimes 2}$. Let $\hat{v}_{ens}$ be the eigenvectors of $\hat{M}_{ens}$ corresponding to positive (nonzero) eigenvalues in M_{ens}, and $\hat{\beta} = \hat{\Sigma}_X^{-1/2}\hat{v}_{ens}$. The ensemble moment-based estimator of $\mathcal{S}_{E(\Delta Y|X)}$ is $\mathcal{S}(\hat{\beta})$.

In Step 3, one needs to determine the rank of M_{ens}, or equivalently the dimension of the central causal effect subspace, for which all the order-determination methods mentioned in § 2 can be used. To avoid distraction from the main message of the paper, we assume that the rank of M_{ens} is known a priori, but we leave the hypothesis of zero rank to be tested in § 6, which is of separate interest in practice.

We next develop the asymptotic normality of the matrix estimator $(\hat{v}_{OLS}, \hat{M}_{pHd})$, which induces the $n^{1/2}$-consistency of the ensemble moment-based estimator as a natural consequence.

Theorem 2. *Suppose the unconfoundedness assumption (1), the common support condition (7), and (C1) hold, and the sample observations are independent. As $n \to \infty$, we have*

$$n^{1/2}\{vec(\hat{v}_{OLS}, \hat{M}_{pHd}) - vec(v_{OLS}, M_{pHd})\} \to N(0, \Gamma\Lambda\Gamma), \tag{10}$$

where Γ is a block diagonal matrix with diagonal blocks $\Sigma_X^{-1/2}$ and $\Sigma_X^{-1/2} \otimes \Sigma_X^{-1/2}$, and

$$\begin{aligned}\Lambda = {} & E[[v(X) - E\{v(X)\}]\{Y_\Delta - E(Y_\Delta)\} - \mathrm{cov}\{v(X), H\}g(X, T)]^{\otimes 2} \\ & - E^{\otimes 2}[v(X)\{\Delta Y - E(\Delta Y)\}].\end{aligned} \tag{11}$$

Here, $v(X) = vec(X, X^{\otimes 2})$ and $H = \phi'(\alpha_0^{\mathsf{T}} h(X))h^{\mathsf{T}}(X)[Y_1/\pi(X) + Y_0/\{1 - \pi(X)\}]$.

The proof of this theorem can be found in Appendix A.2. As mentioned in § 1, an estimator of the central causal effect subspace can be used in three ways: to perform variable selection for the regression causal effect, to improve the estimation accuracy of the regression causal effect, and, to detect the heterogeneity of the regression causal effect. We next discuss these in details.

4 A sparse estimation

To enhance the interpretability of the central causal effect subspace and its estimator, we now conduct sparse sufficient dimension reduction. That is, in addition to (4) with $W = \Delta Y$, we assume the existence of a min-

imal set of covariates X_A called the active set, where $A \subset \{1, \ldots, p\}$, such that $E(\Delta Y|X)$ is a function of X_A. Equivalently, we have

$$E(\Delta Y|X) = E(\Delta Y|X_A). \tag{12}$$

This assumption is commonly adopted in variable selection. It means that all the components of X_A are informative to the regression causal effect, while all the others are redundant. Under (12), it is easy to see that X_A will exactly be used to form the central causal effect subspace, i. e., for any basis matrix β of $\mathcal{S}_{E(\Delta Y|X)}$, each row of β_A is nonzero and each row of β_{-A} is zero. Thus, the active set X_A can be selected if we can identify all the nonzero rows of β. Compared with variable selection, sparse sufficient dimension reduction further tells by which linear combination the active set affects the regression causal effect. Thus, it provides researchers with additional information, and is more efficient in reducing dimensionality.

To incorporate the sparsity structure into the ensemble moment-based estimator, we follow Chen, Zou, and Cook [29] to convert the eigen-decomposition of $\hat{M}_{ens}$ into a least square problem, and impose a group-Lasso type penalty function. Accordingly, continued from Step 3 in § 3, we have:

Step 4. Let $\hat{\beta}^S$, where S stands for sparsity, be a minimizer of

$$-tr(\beta^{\mathsf{T}}\hat{\Sigma}_X^{1/2}\hat{M}_{ens}\hat{\Sigma}_X^{1/2}\beta) + \theta \sum_{i=1}^{p} \|\hat{\beta}_i\|_2^{-1/2}\|\beta_i\|_2 \tag{13}$$

subject to $\beta^{\mathsf{T}}\hat{\Sigma}_X\beta = I_d$, where $\hat{\beta}$ is derived in Step 3. Estimate $\mathcal{S}_{E(\Delta Y|X)}$ by $\mathcal{S}(\hat{\beta}^S)$.

In (13), θ is the tuning parameter that determines how sparse the resulting estimate is. Chen et al. [29] suggested selecting θ by minimizing a Bayesian information criterion, which we slightly modify to be

$$-p^{-1}tr\{(\hat{\beta}_\theta^S)^{\mathsf{T}}\hat{\Sigma}_X^{1/2}\hat{M}_{ens}\hat{\Sigma}_X^{1/2}\hat{\beta}_\theta^S\} + n^{-1}\log n\{(p_\theta - d)d\}, \tag{14}$$

where $\hat{\beta}_\theta^S$ is the sparse estimator $\hat{\beta}^S$ given θ and p_θ is the number of nonzero rows in $\hat{\beta}_\theta^S$.

To minimize (13), Chen et al. [29] suggested an algorithm that alternates between the local quadratic approximation and spectral decomposition until convergence. Chen et al. [29] also showed the selection consistency and the oracle property of the resulting sparse estimator. These properties can be directly parallelized for our case, proof omitted here.

Theorem 3. *Let $\hat{\beta}^O$ be the oracle estimator with $\hat{\beta}_{-A}^O$ fixed at zero and $\hat{\beta}_A^O$ estimated by Steps 1 – 3 using X_A as the covariates. If Assumptions (1), (7), (12), and (C1) hold, and θ in (13) satisfies that $n^{1/2}\theta \to 0$ and $n^{3/4}\theta \to \infty$ as $n \to \infty$, then we have $P(\hat{\beta}_{-A}^S = \mathbf{0}) \to 1$, $D\{\mathcal{S}(\hat{\beta}^S), \mathcal{S}_{E(\Delta Y|X)}\} = O_P(n^{-1/2})$, and $D\{\mathcal{S}(\hat{\beta}^S), \mathcal{S}(\hat{\beta}^O)\} = o_P(n^{-1/2})$.*

By Theorem 3, the sparse ensemble moment-based estimator consistently selects the active set for the regression causal effect, and is asymptotically equally accurate as the oracle estimator. A simulation study (see simulation study 1 of Supplementary Material) showed that in the finite-sample level, it consistently outperforms the ordinary ensemble moment-based estimator when the sparsity assumption (12) holds. This differs from the commonly observed phenomenon in variable selection, where sparse estimators are always suboptimal to their ordinary counterparts in terms of larger finite-sample bias. However, it is reasonable in the sufficient dimension reduction scenario, as the estimation accuracy is not measured for the coefficients of individual covariates in the active set, but rather for the entire central mean subspace, which would be improved if the estimation error associated with the irrelative covariates is wiped out.

The sparse ensemble moment-based estimator inherits the advantage that it avoids fitting the individual outcome regressions. This is shared by the variable selection procedure in Tian et al. [6], not by the others mentioned in § 1.

5 Estimation of the regression causal effect

Equation (9) suggests estimating the regression causal effect by equivalently estimating $E(Y_\Delta|X)$, for which we can use $\hat{Y}_\Delta$ as the response and the reduced covariates from the central causal effect subspace in place

of X. When the central causal effect subspace is zero-dimensional, this amounts to averaging $\hat{Y}_\Delta$, which coincides with the conventional inverse probability weighting strategy to estimate the average causal effect. In general, the use of the reduced covariates and $\hat{Y}_\Delta$ makes data visualization possible, based on which researchers can adopt suitable parametric or nonparametric models. In either case, the reduced dimensionality of the covariates also enhances the accuracy in model fitting. Thus, our procedure is advantageous to that in Abrevaya et al. [7] mentioned in § 1, which models the regression causal effect nonparametrically using $\hat{Y}_\Delta$ as the response and suffers from the curse of dimensionality.

As an illustration, we next use local linear regression to estimate the regression causal effect. Based on the scatter plot of the reduced covariates and this estimate, one may also adopt appropriate parametric models to further improve the estimation. Following Step 4 in § 4, we have

Step 5. For any $x \in \Omega(X)$, estimate $E(\Delta Y|X = x)$ by a_x, which, together with $b_x \in \mathbb{R}^d$, minimizes

$$s(x) = E_n[\{\hat{Y}_\Delta - a_x - b_x^{\mathrm{T}}(\hat{\beta}^{S\mathrm{T}}X - \hat{\beta}^{S\mathrm{T}}x)\}^2 K_\ell(\hat{\beta}^{S\mathrm{T}}X - \hat{\beta}^{S\mathrm{T}}x)]. \tag{15}$$

Here, $K_\ell(\cdot)$ is a kernel density function with bandwidth ℓ. The sparse ensemble moment-based estimator $\mathcal{S}(\hat{\beta}^S)$ is used instead of the ordinary $\mathcal{S}(\hat{\beta})$, for its superiority to the latter. The minimization of (15) can be easily implemented by a weighted least-squares algorithm; see Xia et al. [17] for more detail.

In the literature, a common strategy to estimate the regression causal effect is to treat it as the difference between the individual outcome regression functions, and estimate the latter within each treatment group. Unfortunately, this strategy can not be parallelized if we use the reduced covariates from the central causal effect subspace in place of X. The reason is that these reduced covariates may not be sufficient to predict the individual outcomes, so the un-confoundedness assumption (1) would fail and the individual outcome regression would not be estimable by the observed data. For example, in Model (2) where $\mathcal{S}_{E(\Delta Y|X)}$ is trivial, the un-confoundedness assumption would reduce to the marginal independence between Y_t and T, which is fully unspecified.

6 Detecting heterogeneous causal effect

Based on the asymptotic normality result in Theorem 2, all the aforementioned order-determination methods in § 2 can be applied to detect whether the central casual effect subspace is zero dimensional, which corresponds to a homogeneous regression causal effect. As an example, we employ the hypothesis testing procedure proposed in Bura and Yang [21].

The test is based on the observation that when $\mathcal{S}_{E(\Delta Y|X)}$ is zero dimensional, the matrix parameter (ν_{OLS}, M_{pHd}) is also zero, so that the singular values of its estimate $(\hat{\nu}_{OLS}, \hat{M}_{pHd})$, denoted by $\hat{\lambda}_1, \ldots, \hat{\lambda}_p$, are asymptotically negligible. A direct application of Bura and Yang [21] on Theorem 2 implies that, under this null hypothesis,

$$n \sum_{i=1}^{p} \hat{\lambda}_i^2 \to \sum_{i=1}^{r(p)} \omega_i \chi_i^2 \tag{16}$$

in distribution, where $r(p) = p(p+3)/2$ is the rank of the matrix $\Gamma\Lambda\Gamma$, $\omega_1, \ldots, \omega_{r(p)}$ are the corresponding positive eigenvalues, and χ_i^2's denote the independent random variables that follow the chi-squared distribution with one degree of freedom. Under the alternative hypothesis that d is nonzero, the test statistic in (16) is stochastically larger and diverging to infinity in probability. Thus, for a pre-specified significance level α, we reject the null hypothesis, i. e. a homogeneous regression causal effect, if $n \sum_{i=1}^{p} \hat{\lambda}_i^2$ exceeds the correspondingly critical value, and the power of the test converges to one.

To estimate the ω_i's in (16), we estimate Γ using the sample covariance matrix of X, and estimate Λ by bootstrap re-sampling and using the bootstrap sample covariance matrix of $(X, XX^{\mathrm{T}})\hat{Y}_\Delta$. To comply with the fact that the true propensity score is unknown, we re-estimate this score in each bootstrap sample. Alternative estimators of Λ can be constructed by estimating the moments in (11). However, such an estimator would depend on the specific form of the working propensity score estimator, and omitted simulation studies showed that it is not equally consistent as the bootstrap estimator in finite samples.

Given the estimates $\hat{\omega}_1, \ldots, \hat{\omega}_{r(p)}$, we follow Bura and Yang [21] to approximate $\sum_{i=1}^{r(p)} \omega_i \chi_i^2$ by $\{\sum_{i=1}^{r(p)} \omega_i / r(p)\}\chi^2_{r(p)}$. The latter can be easily simulated in statistical software like R.

Compared with the parametric and nonparametric tests in Crump et al. [11], our test enjoys the advantages of both: it can detect a $n^{1/2}$-order fluctuation from a homogeneous regression causal effect (see Theorem 3.4 in Crump et al. [11], details omitted), and avoids the risk of model misspecification on the individual outcome regressions.

7 Simulation studies

We use the simulated models to illustrate the effectiveness of the sparse ensemble moment-based estimator in estimating the regression causal effect and in variable selection, and that of the proposed χ^2 test for the homogeneity of the regression causal effect. For simplicity, we assume the dimension of the central causal effect subspace to be known a priori, except when testing the homogeneity of the regression causal effect. Throughout the section, we set $n = 600$ and $p = 10$. Results for higher-dimensional cases can be found in the simulation study 2 of Supplementary Material.

7.1 Estimating the regression causal effect

We consider the following four models. In each model, the treatment assignment T is generated independently of the outcomes conditional on the propensity score, so the un-confoundedness assumption (1) holds.

Model 1. $Y_t = t + 2(1-t)X_1 + \varepsilon_t$, $\text{logit}(\pi(X)) = X_p$

Model 2. $Y_t = |X_1| + tX_1 + (1-t)X_2 + \varepsilon_t$, $\text{logit}(\pi(X)) = X_1$

Model 3. $Y_t = \cos(2X_1) + tX_2^2 + \varepsilon_t$, $\text{logit}(\pi(X)) = (X_1 + X_2 + X_3)/3$

Model 4. $Y_t = 2(t-0.5)\{\cos(2X_1) + \sin(X_2)\} + \sum_{i=3}^{5} \sin(3X_i)/6 + \varepsilon_t$, $\text{logit}(\pi(X)) = 0$

For $t = 0, 1$, ε_t is a random error distributed under $N(0, 0.5^2)$. To assess the robustness of the proposed methods to the linearity condition (5) and the constant variance condition (6), in Model 1, we generate (X_1, X_p), the first and last components of X, under independent Bernoulli distribution with mean equal to 0.5, and generate the other components under independent standard normal distributions; in Model 2, we generate the components of X from independent uniform distribution on $(-2, 2)$; in Models 3 and 4, we generate X under $N(\mathbf{0}, \Sigma_X)$, where the (i, j)th entry of Σ_X is $0.25^{|i-j|}$. The two conditions are satisfied only in the latter two models.

In Models 1 and 2, the regression causal effect is linear, although the individual outcome regression functions are more complex in Model 2 with non-monotone structure. Theoretically, the proposed methods will perform consistently in both models, while all the existing methods that rely on individual outcome regressions are expected to be competent for Model 1. In Models 3 and 4, both the regression causal effect and the outcome regression functions are nonlinear. In conjunction with the various propensity scores, these models represent a variety of cases in practice.

From the dimension reduction point of view, the ensemble space $\mathcal{S}(\mathcal{S}_{E(Y_0|X)}, \mathcal{S}_{E(Y_1|X)})$ used in [5] is one dimensional in Model 1, two dimensional in Models 2 and 3, and five dimensional in Model 4. By contrast, the central causal effect subspace $\mathcal{S}_{E(\Delta Y|X)}$ is one dimensional in Models 1–3 and two-dimensional in Model 4. Thus, the latter is more efficient in reducing dimensionality in all the models.

As mentioned in § 4, we use the sparse ensemble moment-based estimator in estimating $\mathcal{S}_{E(\Delta Y|X)}$ and the regression causal effect. We next evaluate its effectiveness for the latter. For reference, we include G-computations with the outcome regressions fitted by linear model, quadratic model, semiparametric model [5] and random forest, respectively, which vary in model complexity. We also include the G-estimation [8] and dynamic weighted least squares approach [9], both of which allow mis-specification of one of the propensity score and $E(Y_0|X)$, subject to a truly specified parametric model of the regression causal effect. For these two methods, we fit the regression causal effect by both linear model and quadratic model, which are true

in Models 1–2 and Models 1–3, respectively. In addition, to study the cost of sufficient dimension reduction estimation, we include an oracle estimator by using the true central casual effect subspace in Step 5 in § 4.

To measure the overall deviation of an estimator $\hat{E}(\Delta Y \mid X)$ from the true regression causal effect, we use the sample median integrated absolute error,

$$\hat{\mathcal{E}}\{|\hat{E}(\Delta Y \mid X) - E(\Delta Y \mid X)|\}, \tag{17}$$

where $\hat{\mathcal{E}}(\cdot)$ denotes the sample median of a random variable. The performance of all the estimators is summarized in Table 1.

Table 1: Accuracy of regression causal effect estimation. The number in the top (bottom) of each cell is the sample average (standard deviation) of the deviation between the regression causal effect and its estimate over 200 replicates, multiplied by 100. "Oracle" stands for the oracle estimator, "S-ENS" for the proposed method based on the sparse ensemble moment-based estimator, "LZG", "GCL", "GCQ" and "RF" for the semiparametric, the linear, the quadratic, and the random forest G-computations, "GEL","GEQ", "WML", and "WMQ" for G-estimation and Wallace and Moodie's approach with the regression causal effect fitted by linear model and quadratic model, respectively.

Model	Oracle	S-ENS	LZG	GCL	GCQ	RF	GEL	GEQ	WML	WMQ
1	6.6	6.6	8.9	8.8	11.4	29.0	9.9	12.4	9.9	12.4
	(4.4)	(4.4)	(3.8)	(3.8)	(3.1)	(3.3)	(2.3)	(2.2)	(2.3)	(2.3)
2	8.9	16.8	15.5	52.6	14.1	55.7	16.6	15.9	16.7	15.9
	(2.7)	(4.1)	(2.3)	(5.8)	(0.8)	(2.5)	(3.7)	(2.8)	(3.6)	(2.8)
3	18.7	20.9	21.2	57.3	19.5	28.2	71.6	19.6	71.5	19.6
	(4.9)	(6.9)	(5.3)	(6.5)	(3.4)	(2.2)	(6.8)	(3.1)	(6.8)	(3.2)
4	19.5	19.9	53.8	125.6	87.7	96.4	125	87.6	125	87.6
	(3.2)	(2.9)	(3.5)	(5)	(6.6)	(12.7)	(5.2)	(6.4)	(5.2)	(6.3)

From Table 1, the linear G-computation is consistent only in Model 1. It is slightly improved by the linear G-estimation, as the latter truly specifies both the regression causal effect and the propensity score in Models 1 and 2. The quadratic G-computation is consistent in Models 1–3, as well as the quadratic G-estimation. The dynamic weighted least square approaches perform almost identically to the G-estimations, which conforms to the results in Wallace and Moodie [9]. All these methods fail in Model 4, where they mis-specify parametric models on the regression causal effect. On the other hand, the random forest estimator, which is non-parametric, is lack of effectiveness in most models due to the limited sample size.

By contrast, both the proposed estimator and the semiparametric G-computation are consistent in all the models. In particular, the proposed estimator outperforms the linear G-computation in Model 1, for which the latter adopts parsimonious and appropriate parametric models. This is not surprising, as the proposed estimator uses additional sparsity structure in the model. Compared with the semiparametric G-computation, the proposed estimator is substantially superior in Model 4, where the outcome regression functions are complex. Referring to the discussion in § 5, this conforms to our theoretical expectation. Compared with the oracle estimator, the proposed estimator is less effective in Model 2, indicating that the cost of estimating the central causal effect subspace can be non-negligible.

7.2 Variable selection

We now examine the variable selection consistency of the sparse ensemble moment-based estimator, by evaluating its true positive rate and false positive rate of selecting the active set of covariates, when applied to Models 1–4.

Persson et al. [13] proposed a variable selection procedure that estimates two active sets, each for an individual outcome regression. Naturally, the union of the two sets can be treated as an estimate of the active

set for the regression causal effect, although their estimate can contain redundant covariates, for example, in Models 3 and 4. The difference Lasso approach [4] and the virtual twins method [10] first impute the missing outcomes using nonparametric techniques and then use the imputed ΔY to conduct variable selection directly for the regression causal effect, for which the former employs the Lasso method and the latter employs the regression tree. These three methods are included for comparison. The results are shown in Table 2.

Table 2: Accuracy of variable selection methods. Each cell in the column TPR (FPR) is the average true (false) positive rate of the variable selection method, over 200 simulation samples. "S-ENS" stands for the sparse moment-based estimator, "VT" for the virtual twins method, "dLasso" for the difference lasso approach, and "PHWD" for the method based on Persson et al. [13].

Model	S-ENS		VT		dLasso		PHWD	
	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
1	1.000	0.000	1.000	0.007	1.000	0.388	1.000	0.433
2	1.000	0.008	1.000	0.000	1.000	0.485	1.000	0.434
3	1.000	0.099	1.000	0.185	0.530	0.290	1.000	0.443
4	1.000	0.000	1.000	0.000	0.702	0.306	1.000	0.361

Due to the use of the Lasso method, the difference Lasso approach favors a linear regression causal effect. This is supported by the results in Table 2, which show that the method is incompetent in Models 3 and 4. The approach based on Persson et al. [13] has a desired sensitivity in all the models, but with a worrisome specificity by its nature. By contrast, both the virtual twins method and the sparse ensemble moment-based estimator constantly select the exact active set in all the models, with the former slightly outperformed by the latter in Model 3.

7.3 Testing the heterogeneity

We now evaluate the proposed χ^2-test for detecting the heterogeneity of the regression causal effect. We use Models 1, 3, and 4 to evaluate the power of the test, but with X in Model 1 changed to be normally distributed as in the other two models. In addition, to examine the actual significance level of the test, we simulate the following three models that have homogeneous regression causal effect.

Model 5. $Y_t = X_1 + X_2 + X_3 + X_4 + \varepsilon_t$

Model 6. $Y_t = e^{X_1} + e^{X_2} + \varepsilon_t$

Model 7. $Y_t = \cos(X_1 + X_2 + X_3) + \varepsilon_t$

The distribution of (X, T, ϵ_t) in these models follows exactly as in Model 3. These models vary in the complexity of the outcome regression function, which is linear in Model 5, monotone but nonlinear in Model 6, and a composition of linear and trigonometric functions in Model 7.

We perform the proposed χ^2 test over 1000 samples for each model with $\alpha = 0.05$, and record the percentage of correct decision, i. e. acceptance if the regression causal effect is homogeneous and rejection otherwise, in Table 3. Theoretically, for Models 1, 3, and 4, a test with large power should have small p-values, which also induces high percentage of rejection; for Models 5–7, a test that achieves its nominal significance level should give p-values that are approximately uniformly distributed on $(0, 1)$, which makes the percentage of acceptance around 95 %.

As mentioned in § 1, Crump et al. [11] proposed both a normal test and a χ^2 test, along with a robust version for each that allows more flexibility in covariates' distribution. These four tests are included in the comparison for reference.

From Table 3, the proposed χ^2 test approximately reaches its nominal level in Models 1, 3, and 4, and its power is close to one in Models 5–7. Thus, it consistently makes correct decision in all the cases. The tests in Crump et al. [11] perform well in Models 1, 4, and 5, but fail to give consistent results otherwise.

Table 3: Percentage of correct decision made by each test. "SDR χ^2" stands for the proposed χ^2 test, "r-Normal", "r-χ^2", "Normal", and "χ^2" for the robust normal test, the robust χ^2 test, the ordinary normal test, and the ordinary χ^2 test in Crump et al. [11], respectively.

	1	3	4	5	6	7
SDR χ^2	93.3	93.1	99.9	98.7	96.7	99.7
r-Normal	100	58.2	100	89.1	23.9	8.8
r-χ^2	100	53.8	100	92.5	27.2	10.0
Normal	100	79.4	100	91.3	4.7	4.0
χ^2	100	76.5	100	94.5	5.2	5.4

To give a closer look at the performance of the tests, in Figure 1, we draw the box-plot of p-values for each model and each test, except for the normal tests as they perform similarly to Crump et al.'s χ^2 tests. These box-plots reconfirm the consistency of the proposed χ^2 test, and suggest that the p-value of the test is approximately uniformly distributed on $(0,1)$ when the regression causal effect is homogeneous. By contrast, the moderate p-values in Models 3, 6, and 7 for Crump et al.'s χ^2 tests indicate that these tests can be insensitive sometimes and over optimistic otherwise.

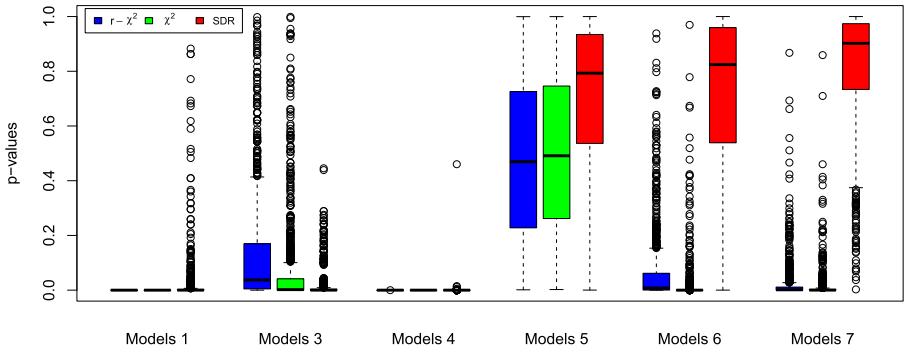

Figure 1: Boxplot of p-values among 1000 replications. The left, middle, and right boxes represent p-values of the robust and the ordinary χ^2 tests in Crump et al. [11], and the proposed χ^2 test, respectively.

7.4 Data analysis

We analyzed the data from the health evaluation and linkage to primary care study, publicly available with the approval of the Institutional Review Board of Boston University Medical Center and the Department of Health and Human Services. The data set contains 453 patients recruited from a detoxification unit, who possibly spent at least one night on the street or shelter within six months before entering the study, in which case the patient is marked as homeless. Our interest is to estimate the causal effect of the homeless experience on patients' physical health condition, measured when entering the study and by the SF-36 physical component score, with higher scores indicating better functioning.

To make the un-confoundedness assumption plausible, we included all the covariates collected in the data who do not have many missing values, some of which were transformed to favor the linearity condition (5) and the constant variance condition (6). They are: age at baseline, a scale indicating depressive symptoms, the square root of the number of friends, the square of a total score of inventory of drug use consequences, the square root of a sex risk score, gender, the average and the maximum number of drinks consumed per day in the past month, and the number of times hospitalized for medical problems. All the nine covariates were standardized to have zero mean and unit variance.

By applying the proposed test, we detected that the regression causal effect is heterogeneous with p-value 0.04. The sequential tests [21] further suggested that $\mathcal{S}_{E(\Delta Y|X)}$ is one dimensional. We then applied the sparse ensemble moment-based estimator, which gave

$$\hat{\beta} = (-0.005, -0.012, -0.081, 0, -0.079, 0.993, 0.002, 0, -0.004)^{\mathsf{T}}.$$

Thus, gender is the dominating factor for the causal effect of homeless experience, and the number of friends and the sex risk are also affective. Cross-validation [17, 5] showed that both $\mathcal{S}_{E(Y_0|X)}$ and $\mathcal{S}_{E(Y_1|X)}$ are nine-dimensional, which means sufficient dimension reduction is not useful for the individual outcome regressions in this data set.

To illustrate the sufficiency and effectiveness of the univariate reduced covariate, we generated a pseudo ΔY by imputing the missing outcome using random forest [10], and tentatively set the dimension of $\mathcal{S}_{E(\Delta Y|X)}$ at two. We drew the scatter plots of the pseudo response against each of the two reduced covariates, along with the fitted loess curve and the 95 % confidence band. From the right panel of Figure 2, the second reduced covariate is irrelevant to the response, as the loess confidence band includes a horizontal line. By contrast, the left panel shows that the first reduced covariate clearly affects the response. In particular, the causal effect of the homeless experience is nearly homogeneous and negligible for the females, and is evident for the males. The effect also increases on those males who have less number of friends or lower sex risk.

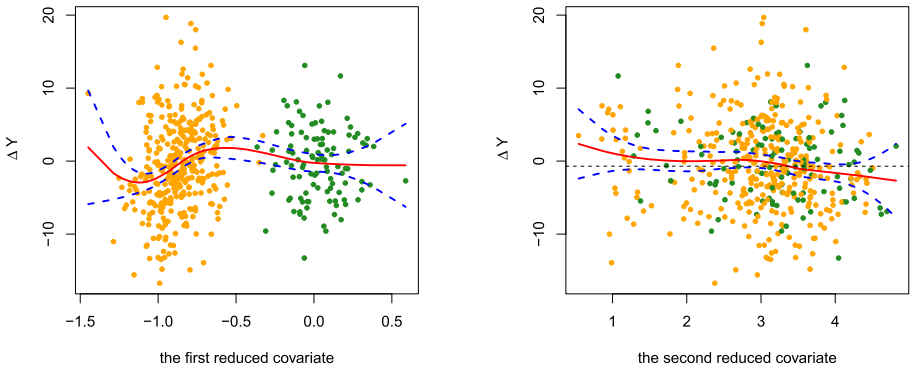

Figure 2: Homeless data with reduced covariates. The left (right) panel is the scatter plot of the imputed ΔY against the first (second) reduced covariate from the estimated central causal effect subspace. In each plot, the male (female) subjects are marked in orange(green).

8 Discussion

When a set of appropriate parametric models is not available for the propensity score $\pi(X)$, the quantity is commonly estimated either semiparametrically under a weak low-dimensional structure assumption [24], or estimated fully non-parametrically. The resulting estimator is generally consistent, but the desired asymptotic property, in particular the asymptotic linearity in (C1), is not satisfied.

In these cases, it is easily seen that the ensemble moment-based estimator is still consistent but its asymptotically normality (10) fails, for which the inferential results must be adjusted. For the variable selection consistency, the order of the tuning parameter θ in Theorem 3 needs to be adjusted according to the convergence order of $\hat{M}_{ens}$. For detecting the heterogeneity of the regression causal effect, we suggest using the permutation test instead. That is, randomly permute Y_0 and Y_1 within each treatment group, and use the permuted data to implement $\hat{M}_{ens}$ and the corresponding $\sum_{i=1}^{p} \hat{\lambda}_i^2$. A large repetition of such procedure will simulate the null distribution of $\sum_{i=1}^{p} \hat{\lambda}_i^2$ for the original data. Because these adjustments are straightforward, we omit the details.

As mentioned earlier, an alternative that avoids estimating the propensity score is to estimate the regression causal effect as the difference between the outcome regression functions, the latter being estimated semiparametrically [5]. Following the literature, one may think of constructing a doubly-robust estimator that combines the two approaches. However, in contrast to the case where the parameter of interest is the average causal effect and both the propensity score and the outcome regression functions are estimated parametrically, such an estimator will always inherit the drawback of outcome regression-based estimator, i. e. the estimation of the nuisance functional parameters mentioned in § 1 and the redundant directions in $\mathcal{S}(\mathcal{S}_{E(Y_0|X)}, \mathcal{S}_{E(Y_1|X)})$ mentioned in § 3. For this reason, in practice, we do not recommend constructing and using a doubly-robust estimator, and instead suggest using the proposed ensemble moment-based estimator whenever there is evidence that support the consistency of propensity score estimation, and using the semiparametric estimator based on the outcome regressions proposed by Luo et al. [5] otherwise.

Funding: Dr. Zhu's research was supported by Award Number 430-2016-00163 from the Social Sciences and Humanities Research Council and by Grant Number RGPIN-2017-04064 from the Natural Sciences and Engineering Research Council of Canada.

Appendix A

A.1 Proof of Theorem 1

Proof. By (1) and (7), we have

$$E\{TY_1/\pi(X) \mid X\} = E\{\pi(X)E(Y_1 \mid X)/\pi(X)\} = E(Y_1 \mid X),$$

and likewise, $E[(1-T)Y_0/\{1-\pi(X)\} \mid X] = E(Y_0 \mid X)$. Thus $E(Y_\Delta \mid X) = E(\Delta Y \mid X)$. □

A.2 Proof of Theorem 2

Proof. We first show the asymptotic normality of

$$\{\hat{\Sigma}_X^{1/2}\hat{v}_{OLS}, \mathrm{vec}(\hat{\Sigma}_X^{1/2}\hat{M}_{pHd}\hat{\Sigma}_X^{1/2})\} - \{\Sigma_X^{1/2}v_{OLS}, \mathrm{vec}(\Sigma_X^{1/2}M_{pHd}\Sigma_X^{1/2})\},$$

which we denote by V_n. Let $S = \phi'(\alpha_0^{\mathsf{T}}h(X))h^{\mathsf{T}}(X)$. By (C1) and Taylor's expansion, we have $\hat{\pi}(X) - \pi(X) = SE_n g(X,T) + O_P(n^{-1})$, which implies that

$$\hat{Y}_\Delta = Y_\Delta - [TY_1/\pi^2(X) + (1-T)Y_0/\{1-\pi(X)\}^2]SE_n g(X,T) + O_P(n^{-1}).$$

Because $V_n = E_n[v(X)\{\hat{Y}_\Delta - E_n(\hat{Y}_\Delta)\}] - E[v(X)\{\Delta Y - E(\Delta Y)\}]$, it is easy to see that

$$\begin{aligned} V_n =& E_n[[v(X) - E\{v(X)\}]\{Y_\Delta - E(Y_\Delta)\} - E[v(X)\{\Delta Y - E(\Delta Y)\}] \\ &+ \mathrm{cov}\{v(X), H\}g(X,T)] + o_P(n^{-1/2}) \\ =& U_n + o_P(n^{-1/2}), \end{aligned}$$

and that by the central limit theorem, $\sqrt{n}U_n \to N(0,\Lambda)$. The desired asymptotic normality (10) follows by that $\hat{\Sigma}_X \to \Sigma_X$ in probability and Slutsky's theorem. □

References

1. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66:688–701.

2. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Math Model. 1986;7:1393–512.
3. Snowden JM, Rose S, Mortimer KM. Implementation of G-computation on a simulated data set: demonstration of a causal inference technique. Am J Epidemiol. 2011;173:731–8.
4. Ghosh D, Zhu Y, Coffman DL. Penalized regression procedures for variable selection in the potential outcomes framework. Stat Med. 2015;34:1645–58.
5. Luo W, Zhu Y, Ghosh D. On estimating regression-based causal effects using sufficient dimension reduction. Biometrika. 2017;104:51–65.
6. Tian L, Alizadeh AA, Gentles AJ, Tibshirani R. A simple method for estimating interactions between a treatment and a large number of covariates. J Am Stat Assoc. 2014;109:1517–32.
7. Abrevaya J, Hsu Y-C, Lieli RP. Estimating conditional average treatment effects. J Bus Econ Stat. 2015;33:485–505.
8. Robins JM. Optimal structural nested models for optimal sequential decisions. In: Proceedings of the second seattle Symposium in Biostatistics. Springer; 2004. p. 189–326.
9. Wallace MP, Moodie EE. Doubly-robust dynamic treatment regimen estimation via weighted least squares. Biometrics. 2015;71:636–44.
10. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. Stat Med. 2011;30:2867–80.
11. Crump RK, Hotz VJ, Imbens GW, Mitnik OA. Nonparametric tests for treatment effect heterogeneity. Rev Econ Stat. 2008;90:389–405.
12. Imai K, Ratkovic M. Estimating treatment effect heterogeneity in randomized program evaluation. Ann Appl Stat. 2013;7:443–70.
13. Persson E, Häggström J, Waernbaum I, de Luna X. Data-driven algorithms for dimension reduction in causal inference. Comput Stat Data Anal. 2017;105:280–92.
14. Cook RD, Li B. Dimension reduction for conditional mean in regression. Ann Stat. 2002;30:455–74.
15. Li K-C, Duan N. Regression analysis under link violation. The Annals of Statistics. 1989. 1009–1052.
16. Li K-C. On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma. J Am Stat Assoc. 1992;87:1025–39.
17. Xia Y, Tong H, Li WK, Zhu L-X. An adaptive estimation of dimension reduction space. J R Stat Soc, Ser B, Stat Methodol. 2002;64:363–410.
18. Luo W, Li B, Yin X. On efficient dimension reduction with respect to a statistical functional of interest. Ann Stat. 2014;42:382–412.
19. Ma Y, Zhu L. On estimation efficiency of the central mean subspace. J R Stat Soc, Ser B, Stat Methodol. 2014;76:885–901.
20. Hall P, Li K-C. On almost linearity of low dimensional projections from high dimensional data. Ann Stat. 1993;47:867–89.
21. Bura E, Yang J. Dimension estimation in sufficient dimension reduction: a unifying approach. J Multivar Anal. 2011;102:130–42.
22. Zhu L, Miao B, Peng H. On sliced inverse regression with high-dimensional covariates. J Am Stat Assoc. 2006;101:630–42.
23. Luo W, Li B. Combining eigenvalues and variation of eigenvectors for order determination. Biometrika. 2016;103:875–87.
24. Ghosh D. Propensity score modelling in observational studies using dimension reduction methods. Stat Probab Lett. 2011;81:813–20.
25. Hu Z, Follmann DA, Wang N. Estimation of mean response via the effective balancing score. Biometrika. 2014;101:613–24.
26. Huang M-Y, Chan KCG. Joint sufficient dimension reduction and estimation of conditional and average treatment effects. Biometrika. 2017;104:583–96.
27. Imai K, Ratkovic M. Covariate balancing propensity score. J R Stat Soc B. 2014;76:243–63.
28. van der Laan MJ, Polley EC, Hubbard AE. Super learner. Stat Appl Genet Mol Biol. 2007;6:1–21.
29. Chen X, Zou C, Cook R. Coordinate-independent sparse suffcient dimension reduction and variable selection. Ann Stat. 2010;6:3696–723.

Supplemental Material: The online version of this article offers supplementary material (https://doi.org/10.1515/jci-2018-0015).