Journal of Statistical Planning and Inference 212 (2021) 153–168

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Semiparametric estimation for average causal effects using propensity score-based spline

Peng Wu [a], Xinyi Xu [b], Xingwei Tong [a], Qing Jiang [c], Bo Lu [d,*]

[a] School of Statistics, Beijing Normal University, China
[b] Department of Statistics, The Ohio State University, United States of America
[c] School of Statistics, Southwestern University of Finance and Economics, China
[d] Division of Biostatistics, College of Public Health, The Ohio State University, United States of America

ARTICLE INFO

Article history:
Received 6 October 2019
Received in revised form 7 October 2020
Accepted 16 October 2020
Available online 12 November 2020

Keywords:
Average causal effect
Heterogeneous variance
Propensity score
Semiparametric estimation
Spline

ABSTRACT

When estimating the average causal effect in observational studies, researchers have to tackle both self-selection of treatment and outcome modeling. This is difficult because the parametric form of the outcome model is often unknown and there exists a large number of covariates. In this work, we present a semiparametric strategy for estimating the average causal effect by regressing on the propensity score. Furthermore, we show that regression error terms usually depend on the propensity score as well, which could cause heteroscedastic variances, and thus construct a refined estimator to improve the estimation efficiency. Both estimators are shown to be consistent and asymptotically normally distributed, with the latter one having a smaller asymptotic variance. The simulation studies indicate that our methods compare favorably with many competing estimators. Our methods are easy to implement and avoid hazardous impact due to extreme weights as often seen in weighting estimators. They can also be extended to handle subgroup effects with known structure. We apply the proposed methods to data from the Ohio Medicaid Assessment Survey 2012, estimating the effect of having health insurance on self-reported health status for a population with subsidized insurance plan choices under the Affordable Care Act.

1. Introduction

Causal inference with observational data is challenging as it has to handle two tasks at the same time, self-selection of treatment and outcome modeling. The self-selection bias refers to the bias in causal effect estimation associated with the difference in relevant pre-treatment covariates. Even under the assumption of ignorable treatment assignment (i.e. all relevant covariates are known), this is not easy since there are usually a large number of variables related to treatment selection, plus many other predictors for the outcome. Early strategies for estimating average causal effects involve mostly parametric models, i.e. linear regression for continuous outcomes. The validity of a fully parametric approach depends heavily on the correct model specification. Results from regression models are very sensitive to small discrepancies in the model setup (Imbens and Rubin, 2015). Especially, with a large number of covariates, it becomes extremely difficult to specify all functional forms correctly.

An improvement to conventional regression models is to use the generalized additive model (GAM) when the dimension of covariates is high (Hastie and Tibshirani, 1986). GAM is advantageous, as it not only allows to fit a

* Corresponding author.
E-mail address: lu.232@osu.edu (B. Lu).

https://doi.org/10.1016/j.jspi.2020.10.004
0378-3758/

non-linear function for each covariate by using B-spline approximation or other local smoothing methods, but also avoids the curse of dimensionality (Hastie et al., 2009). However, GAM is restricted to be additive. Non-additive functional forms cannot be used and important interaction terms of the covariates may be missed.

In causal inference, a primary parameter of interest is the average treatment effect, while coefficients for other covariates can be considered nuisance. Rosenbaum and Rubin (1983) introduced the concept of balancing score to address the high-dimensional covariates in causal effect estimation. A balancing score, $b(X)$, is a function of the covariates X such that the conditional distribution of X given $b(X)$ is the same between treated ($Z = 1$) and control ($Z = 0$) groups, that is,

$$X \perp Z \mid b(X).$$

They further developed a series of methods for estimating the average causal effect based on the propensity score. The propensity score is a scalar function of covariates defined as $e(X) = P(Z = 1 \mid X)$ and it is shown to be the coarsest balancing score (Rosenbaum and Rubin, 1983). The propensity score finds the lowest-dimensional function of the covariates that suffices for removing the selection bias due to pre-treatment variables.

The propensity score can be incorporated in constructing both nonparametric and semiparametric causal estimators. For nonparametric estimation, we may use the propensity score to form matched sets or strata, which mimics a block randomized experiment design (Rosenbaum, 2002). We may also use it as a weight, which leads to estimators similar to the Horvitz–Thompson estimator in survey sampling (Horvitz and Thompson, 1952). This approach is also known as inverse probability weighting (IPW) estimation. For semiparametric methods, Robins et al. (1994) considered a class of augmented IPW (AIPW) estimators by combining propensity score with outcome regression models. AIPW estimators enjoy a property of doubly robustness, which guarantees the consistent estimation of causal effects if either the outcome regression model or the propensity score model is correctly specified (Bang and Robins, 2005). Moreover, such estimators achieve the semiparametric efficiency bound when both the propensity score model and the outcome regression model are correctly specified. AIPW estimator is not always the optimal one, as Tan (2007) showed that the asymptotic variance of the outcome regression estimator is no larger than that of AIPW if the outcome model is correctly specified. Tan (2006, 2010) considered other regression estimators based on nonparametric likelihood. van der Laan (2010) also developed doubly robust estimator using machine learning tools based on targeted maximum likelihood estimation (TMLE), both of which further improved.

Doubly robust estimators rely on correct model specification of the propensity score or the outcome, which may not be an easy task in practice. When both models are specified somewhat wrong, the results could be seriously biased (Kang and Schafer, 2007). Even when model specification is not an issue, the estimator can be very volatile when there are propensity score values very close to zero or one (Hahn, 1998; Rubin, 2001; Tsiatis and Davidian, 2007; Lee et al., 2011). This is because the resulting weights are extremely large and they tend to inflate the variance estimation. Tan (2010) limited the effects of large wights and considered double robust estimation based on the restricted nonparametric likelihood. It requires either parametric or nonparametric modeling of the conditional mean $E(Y \mid X)$ (Y is outcome) or the conditional distribution (Lin et al., 2017), which has similar problems with AIPW.

A recent strand of literature tried to address the problems of extreme weights and the outcome model specification through specially designed weights or machine learning techniques. Li et al. (2018) unified the existing weighting methods and proposed an overlap weighting approach that avoids extreme weights by focusing on the most similar subpopulations between two treatment groups. Chernozhukov et al. (2018) proposed a double/debiased machine learning estimators in case of high-dimensional covariates. Taking a step further, Wager and Athey (2018) and Athey et al. (2019) discussed algorithms to construct causal forest for estimating the heterogeneous treatment effects using generalized random forests.

In this paper, we propose to combine the propensity score with nonparametric regression models to obtain high quality average causal effect estimators. Instead of using propensity scores as weights to remove the selection bias, we include propensity scores directly in the regression model. To improve the robustness, we adopt nonparametric modeling for the outcome, i.e. splines. Though spline methods are widely used in statistical modeling, it has received little discussion in causal effect estimation. We aim to fill the gap by establishing the asymptotic properties of spline estimators for average treatment effects. This approach can be implemented easily and requires minimum model specification. We further modify the estimator by incorporating adjustment for error heteroscedasticity to improve the efficiency. The primary goal of our method is to estimate a constant average causal effect. In practice, to test whether the causal effect is constant, our model can be extended by including interactions between the treatment and subgroups. The main advantage of our approach is to handle high dimensional covariates with arbitrary functional form in causal estimation. The dimension reduction is achieved via the propensity score and the use of spline models avoids the harmful impact of extreme propensity scores on variance results. The simulation studies show that our methods compare favorably with many competing estimators.

The paper is organized as follows. Section 2 introduces the notations and our estimation strategy. Section 3 shows the asymptotic results. In Section 4, we conduct extensive simulations to compare our methods with conventional linear regression model, GAM, subclassification on propensity score, AIPW estimator, Tan's calibrated likelihood estimator and regression estimator, TMLE and the overlap weight augmented estimation. In Section 5, we apply our methods to estimate the health insurance policy effect on self-reported health status. A brief discussion is presented in Section 6.

2. Notations and the estimation strategy

2.1. The setup

Let $Y = (Y_1, \ldots, Y_n)^T$ be a vector of continuous outcomes of interest, $Z = (Z_1, \ldots, Z_n)^T$ be the vector of treatment indicators, where $Z_i = 1$ represents being treated and 0 otherwise. (Y^1, Y^0) are the potential outcomes under corresponding treatment arms with $Y = ZY^1 + (1 - Z)Y^0$. Let $X = (X_1, \ldots, X_n)^T$, where X_i is a p-dimensional vector of covariates that potentially influences Z and Y. We assume that the observed outcome model has an additive structure

$$Y = Z\beta + f(X) + \epsilon, \tag{1}$$

where β captures the average treatment effect, f is a unknown smooth function of covariates, and ϵ is the error term with $E(\epsilon \mid X, Z) = 0$. To proceed with causal inference based on the propensity score, two common assumptions are used: the stable unit treatment value assumption (Rubin, 1980) and the ignorable treatment assignment assumption (Rosenbaum and Rubin, 1983).

We are mostly concerned with how to estimate β accurately and the specific functional form of f is not of our interest here. Let $e = e(X) = (e_1, \ldots, e_n)^T$ be the vector of propensity scores. The following proposition implies we can estimate β through regression on (Z, e), which significantly reduces the dimensionality in modeling.

Proposition 1. *Assume model (1) is true. Denote $g(e) = E(f(X) \mid e(X) = e)$ be the projection of $f(X)$ in the propensity score space. Then the treatment effect $\beta = E(Y^1 - Y^0)$ can be obtained by regressing Y on $(Z, e(X))$, that is,*

$$Y = Z\beta + g(e) + \xi, \quad E(\xi \mid Z, e) = 0.$$

Proof. Since $e(X)$ is a balancing score, we have $Z \perp X \mid e(X)$. When we regress Y on $(Z, e(X))$,

$$\begin{aligned} E(Y \mid Z, e(X)) &= E(Z\beta \mid Z, e(X)) + E(f(X) \mid Z, e(X)) + E(\epsilon \mid Z, e(X)) \\ &= Z\beta + E(f(X) \mid e(X)) + E\{E(\epsilon \mid Z, X) \mid Z, e(X)\} \\ &= Z\beta + E(f(X) \mid e(X)) = Z\beta + g(e). \end{aligned}$$

Following the strongly ignorable assumption,

$$\begin{aligned} E(Y^1 - Y^0) &= E\{E(Y^1 - Y^0 \mid e(X))\} = E\{E(Y^1 \mid Z, e(X)) - E(Y^0 \mid Z, e(X))\} \\ &= E\{E(Y \mid Z = 1, e(X)) - E(Y \mid Z = 0, e(X))\} \\ &= E\{\beta + g(e) - g(e)\} = \beta. \end{aligned}$$

which shows that β is the average treatment effect.

Since $\xi = f(X) - g(e) + \epsilon$, it is easy to see $E(\xi \mid Z, e) = E(f(X) - g(e) + \epsilon \mid Z, e) = E[f(X) \mid Z, e] - g(e) + E(\epsilon \mid Z, e) = 0$.

This proposition suggests that we can replace $f(X)$ in model (1) with the function $g(e)$. This would substantially reduce the model complexity, because the covariate X_i is usually a high dimensional vector while the propensity score e_i is a scalar, $i = 1, \ldots, n$. Based on observing $Y = (Y_1, \ldots, Y_n)^T$ and $Z = (Z_1, \ldots, Z_n)^T$, we rewrite the outcome model as

$$Y_i = Z_i\beta + g(e_i) + \xi_i, \qquad i = 1, \ldots, n, \tag{2}$$

where $\xi_i = f(X_i) - g(e_i) + \epsilon_i$ is the new random error. Following the proposition, we have $E(\xi_i \mid Z_i, e_i) = 0$. Moreover, ξ_i is uncorrelated with the regression mean $Z_i\beta + g(e_i)$, because

$$\begin{aligned} \operatorname{cov}(\xi_i, Z_i\beta + g(e_i)) &= E\,(\xi_i \cdot (Z_i\beta + g(e_i))) \\ &= E\,\{E\,(\xi_i \cdot (Z_i\beta + g(e_i)) \mid e_i)\} \\ &= E\,\{E\,((f(X_i) - g(e_i) + \epsilon_i)(Z_i\beta + g(e_i)) \mid e_i)\} \\ &= E\,\{E(f(X_i) - g(e_i) \mid e_i) \cdot E(Z_i\beta + g(e_i) \mid e_i)\} \\ &= 0, \end{aligned}$$

where the second last line follows from the fact that X_i and Z_i are conditionally independent given a balancing score.

For estimating the unknown function g, we adopt the B-spline method (Schumaker, 1981), as the propensity scores may be unevenly placed across the interval [0, 1]. The idea of estimating causal effects by modeling the expectation of exposure conditional on covariates has been discussed by other researchers, including Robins et al. (1992) and Lee (2018). But they used a weighting type of adjustment that may suffer from propensity scores close to 0 or 1. The spline based method is not affected by extreme propensity score values. Gutman and Rubin (2015) adopted the spline model as a flexible imputation tool for the missing potential outcomes. Let $B(\cdot) = \{b_1(\cdot), \ldots, b_{k_n}(\cdot)\}^T$ be a vector of order $m + 1$ normalized B-spline basis functions and set the knots to be $0 = u_0 < u_1 < \cdots < u_{k_n-(m+1)} < u_{k_n-m} = 1$. Then the

function $g(e)$ can be approximated by

$$g(e) \approx B(e)^T\gamma,$$

where γ is a k_n-dimensional vector.

When the propensity scores are known, it is natural to estimate the parameter (β, γ) using the least square method, that is, by minimizing $\sum_{i=1}^{n}(Y_i - Z_i\beta - B(e_i)^T\gamma)^2$. However, in practice, the propensity scores are usually unknown and need to be estimated. Following the common practice, we use the logistic regression for the estimation, that is, we assume

$$e(x_i) = P(Z_i = 1 \mid X_i = x_i) = \frac{\exp(x_i^T\alpha)}{1+\exp(x_i^T\alpha)}, \tag{3}$$

where α is the unknown parameter. We denote the maximum likelihood estimator of α by $\hat{\alpha}_n$, and denote the propensity score estimates of e_i by $\hat{e}_i = (1+\exp(-X_i^T\hat{\alpha}_n))^{-1}$. Plugging $\hat{e}_i$ into the likelihood yields the least squares estimator of the averaged treatment effect β as

$$\hat{\beta}_n^{homo} = \left[Z^T(I-\hat{H})Z\right]^{-1}Z^T(I-\hat{H})Y, \tag{4}$$

where $\hat{H} = B(\hat{e})(B(\hat{e})^TB(\hat{e}))^{-1}B(\hat{e})^T$ and $B(\hat{e}) = (B(\hat{e}_1), \ldots, B(\hat{e}_n))^T$. Technically, the propensity scores can be estimated more flexibly by nonparametric models, such as kernel methods. However, when the dimension of X is high, the nonparametric models suffer from "curse of dimensionality", which may cause big bias and finally influence the accuracy of causal effect estimation. In our simulation studies in Section 4, we will investigate the impact of misspecified propensity score models.

2.2. Treatment effect estimation with heterogeneous variances

For many practical situations, the covariates X_i's are independent, so the error terms $\xi_i = f(X_i) - g(e_i) + \epsilon_i$ in our reformulated model (2) are (conditionally) uncorrelated, that is, $\mathrm{cov}(\xi_i, \xi_j \mid e_i, e_j) = 0$. Therefore, the (conditional) error variance matrix

$$\mathrm{var}(\xi \mid e) = \Sigma = diag(w_1, \ldots, w_n),$$

where $\xi = (\xi_1, \ldots, \xi_n)^T$, $e = (e_1, \ldots, e_n)^T$ and $w_i = \mathrm{var}(\xi_i \mid e_i)$.

Note that the (conditional) variances w_i varies with e_i and so are heterogeneous. In such situations, the estimation efficiency can be improved by using weighted least squares, which incorporates consistent estimates of the variances in the weights. Let

$$w_i = \mathrm{var}(\xi_i \mid e_i) = \exp(h(e_i)),$$

where h is an unknown function. We use the B-spline method again to approximate $h(e)$, i.e., $h(e_i) \approx B(e_i)^T\eta$, where η is an k_n-dimensional vector. Under regular assumptions, the log likelihood function is given by

$$l_n(\beta, \gamma, \eta) = -\frac{1}{2}\left\{\sum_{i=1}^{n}\log(w_i(\eta)) + \sum_{i=1}^{n}1/w_i(\eta)\cdot(Y_i - Z_i\beta - B(e_i)^T\gamma)^2\right\},$$

where $w_i(\eta) = \exp(B(e_i)^T\eta)$. Therefore, the weighted least squares estimate of (β, γ, η) can be obtained by solving a set of equations

$$\begin{cases} \dfrac{\partial l_n(\beta,\gamma,\eta)}{\partial\beta} = 0 \\ \dfrac{\partial l_n(\beta,\gamma,\eta)}{\partial\gamma} = 0 \\ \dfrac{\partial l_n(\beta,\gamma,\eta)}{\partial\eta} = 0 \end{cases} \tag{5}$$

We estimate the propensity score e by $\hat{e}$ as in Section 2.1, and derive the treatment effect estimate using an iterative algorithm as follows:

step 1. Given $\hat{e}$, $\hat{\gamma}$ and $\hat{\eta}$, solve the first equation in (5) to obtain

$$\hat{\beta}_n^{hetero} = (Z'_{\hat{\Sigma}}(I-\hat{H}_{\hat{\Sigma}})Z_{\hat{\Sigma}})^{-1}Z'_{\hat{\Sigma}}(I-\hat{H}_{\hat{\Sigma}})\hat{\Sigma}^{-1/2}Y, \tag{6}$$

where $Z_{\hat{\Sigma}} = \hat{\Sigma}^{-1/2}Z$, $\hat{H}_{\hat{\Sigma}} = \hat{\Sigma}^{-1/2}B(\hat{e})\left[B(\hat{e})^T\hat{\Sigma}^{-1}B(\hat{e})\right]^{-1}B(\hat{e})'\hat{\Sigma}^{-1/2}$, and $\hat{\Sigma} = diag(w_1(\hat{\eta}), \ldots, w_n(\hat{\eta}))$.

step 2. Given $\hat{e}$, $\hat{\beta}_n^{hetero}$ and $\hat{\eta}$, solve the second equation in (5) to obtain $\hat{\gamma}$.

step 3. Given $\hat{e}$, $\hat{\beta}^{hetero}$ and $\hat{\gamma}$, solve the third equation in (5) to obtain $\hat{\eta}$.

step 4. Repeat Steps 1 through 3 for k iterations until

$$\|\hat{\beta}^{(k+1)} - \hat{\beta}^{(k)}\|_\infty \le c_0,$$

where $\|x\|_\infty = \max_i x_i$ and c_0 is a pre-specified small positive number.

3. Asymptotic properties

In this section, we establish the asymptotic properties of the treatment effect estimators $\hat{\beta}_n^{homo}$ and $\hat{\beta}_n^{hetero}$. In particular, we show that they both are consistent and have asymptotic normal distributions. The following regularity conditions are assumed throughout:

C1. The distribution of the propensity score e is absolutely continuous and its density is bounded on interval [0,1].
C2. There exist constants m, $0 < s \le 1$ and $C > 0$ and an m-times continuously differentiable function g, such that for any $0 \le e_1, e_2 \le 1$,

$$|g^{(m)}(e_1) - g^{(m)}(e_2)| \le C|e_1 - e_2|^s.$$

The parameter $r = m + s$ is often considered as a measure of the smoothness of the function g.
C3. $E\xi_1^2 < \infty$.
C4. $E[e_1(1-e_1)]$ and $E[e_1(1-e_1)/w_1(e)]$ are finite, where $w_1(e) = \mathrm{var}(\xi_1 \mid e_1)$.

These conditions are common for partial linear models (e.g., Green and Yandell, 1985; Heckman, 1986; Chen, 1988). We first show the asymptotic results of the homoscedastic treatment effect estimator $\hat{\beta}_n^{homo}$.

Theorem 1. *Assume that the regularity conditions C1–C4 hold and the dimension of the B-spline basis function vector $k_n = n^{1/(2r+1)}$. Denote the true values of β and α by β_0 and α_0, respectively. Then*
When the propensity scores e_i's are known, the homoscedastic treatment effect estimator $\hat{\beta}_n^{homo}$ is consistent and

$$n^{1/2}(\hat{\beta}_n^{homo} - \beta_0) \to N\left(0, \frac{var\{(Z_1 - e_1)\xi_1\}}{E^2\{e_1(1-e_1)\}}\right). \tag{7}$$

When the propensity scores e_i's are unknown but have consistent estimators $\hat{e}_i$'s, the homoscedastic treatment effect estimator $\hat{\beta}_n^{homo}$ is consistent and

$$n^{1/2}(\hat{\beta}_n^{homo} - \beta_0) \to N\left(0, \frac{var[(Z_1 - e_1)\xi_1] - A^T I^{-1}(\alpha_0)A}{E^2[e_1(1-e_1)]}\right), \tag{8}$$

where $A = E\{e_1(1-e_1)\xi_1 X_1\}$ and $I^{-1}(\alpha_0)$ is the Fisher information matrix of α at α_0.

Proof. See Appendix.

It is interesting to note that

$$\mathrm{var}\{(Z_1 - e_1)\xi_1\} = E[E\{(Z_1 - e_1)^2 \mid e_1\} \cdot E(\xi_1^2 \mid e_1)] = E\{\mathrm{var}(Z_1 \mid e_1) \cdot \mathrm{var}(\xi_1 \mid e_1)\},$$

that is, the asymptotic variance of (7) is the mean of the product of $\mathrm{var}(Z_1 \mid e_1)$ and $\mathrm{var}(\xi_1 \mid e_1)$. Furthermore, comparing (7) and (8) reveals that using estimated propensity scores reduces the asymptotic variance and thus can lead to more accurate estimation. This phenomenon has been noticed in literature, such as Joffe and Rosenbaum (1999) and Hirano et al. (2003), and we verify it in our simulation studies as well. A heuristic explanation is that the estimated propensity score cannot distinguish systematic bias from an imbalance in covariates that is due to chance or bad luck, and adjustment for the estimated propensity score tends to remove both types of imbalance. While adjustment for the true propensity score only removes systematic bias (Joffe and Rosenbaum, 1999). Moreover, another connection we observe is that the asymptotic variance in Theorem 1 is equal to that from the two-stage estimator in Robins et al. (1994). Next, we present the asymptotic results of the heteroscedastic treatment effect estimator (6).

Theorem 2. *Assume that the regularity conditions C1–C4 hold and the dimension of the B-spline basis function vector $k_n = n^{1/(2r+1)}$. Denote the true values of β and α by β_0 and α_0, respectively.*
When the propensity scores e_i's are known, the heteroscedastic treatment effect estimator $\hat{\beta}_n^{hetero}$ is consistent and

$$n^{1/2}(\hat{\beta}_n^{hetero} - \beta_0) \to N\left(0, \frac{var\{(Z_1 - e_1)\xi_1/w_1(e)\}}{E^2\{e_1(1-e_1)/w_1(e)\}}\right), \tag{9}$$

where $w_1(e) = var(\xi_1 \mid e_1)$ is the first diagonal element in the conditional error variance matrix $\Sigma = var(\xi \mid e)$.

When the propensity scores e_i's are unknown but have consistent estimators $\hat{e}_i$'s, the heteroscedastic treatment effect estimator $\hat{\beta}_n^{hetero}$ is consistent and

$$n^{1/2}(\hat{\beta}_n^{homo} - \beta_0) \to N\left(0, \frac{var\{(Z_1 - e_1)\xi_1/w_1(e)\} - A_\Sigma^T I^{-1}(\alpha_0)A_\Sigma}{E^2\{e_1(1-e_1)/w_1(e)\}}\right), \tag{10}$$

where $A_\Sigma = E[e_1(1-e_1)\xi_1 X_1/w_1(e)]$ and $I^{-1}(\alpha_0)$ is the Fisher information matrix of α at α_0.

Proof. See Appendix.

As is the case in Theorem 1, the propensity score plays a role in efficiency of the heteroscedastic estimator and its asymptotic variance is reduced by using the estimated propensity score. Furthermore, with respect to the efficiency between the heteroscedastic and homoscedastic estimator, we have the following proposition.

Proposition 2. *When the propensity scores are known, the asymptotic variance in (9) for $\hat{\beta}^{hetero}$ is less or equal to the asymptotic variance in (7) for $\hat{\beta}^{homo}$.*

Proof. It is straightforward to calculate that the asymptotic variance of $\hat{\beta}^{homo}$ is

$$\frac{\text{var}\{(Z_1 - e_1)\xi_1\}}{E^2\{e_1(1-e_1)\}} = \frac{E\{\text{var}(Z_1 \mid e_1)\cdot \text{var}(\xi_1 \mid e_1)\}}{E^2\{e_1(1-e_1)\}} = \frac{E\{w_1(e)e_1(1-e_1)\}}{E^2\{e_1(1-e_1)\}},$$

and the asymptotic variance of $\hat{\beta}^{hetero}$ is

$$\frac{\text{var}\{(Z_1 - e_1)\xi_1/w_1(e)\}}{E^2\{e_1(1-e_1)/w_1(e)\}} = \frac{1}{E\{e_1(1-e_1)/w_1(e)\}}.$$

By the Cauchy–Schwarz inequality,

$$E\{e_1(1-e_1)/w_1(e)\}\cdot E\{w_1(e)e_1(1-e_1)\} \geq E^2\{e_1(1-e_1)\},$$

this completes the proof.

The asymptotic variances of both $\hat{\beta}_n^{homo}$ and $\hat{\beta}_n^{hetero}$ can be estimated by plug-in method. Specifically, if the propensity scores are unknown, the estimated asymptotic variance for $\hat{\beta}_n^{homo}$ is

$$\frac{n^{-1}\sum_{i=1}^n (Z_i - \hat{e}_i)^2\hat{\xi}_i^2 - \hat{A}^T\hat{I}^{-1}(\alpha_0)\hat{A}}{n^{-1}\{\sum_{i=1}^n \hat{e}_i(1-\hat{e}_i)\}^2},$$

where $\hat{\xi}_i = Y_i - Z_i\hat{\beta} - \hat{g}(\hat{e}_i)$, $\hat{I}(\alpha_0) = n^{-1}\sum_{i=1}^n \hat{e}_i(1-\hat{e}_i)X_iX_i'$, $\hat{A} = n^{-1}\sum_{i=1}^n \hat{e}_i(\hat{e}_i - 1)\hat{\xi}_i X_i$. Similarly, the estimated asymptotic variance for $\hat{\beta}_n^{hetero}$ is given by

$$\frac{n^{-1}\sum_{i=1}^n (Z_i - \hat{e}_i)^2\hat{\xi}_i^2/\hat{w}_i^2 - \hat{A}_\Sigma^T\hat{I}^{-1}(\alpha_0)\hat{A}_\Sigma}{n^{-1}\{\sum_{i=1}^n \hat{e}_i(1-\hat{e}_i)/\hat{w}_i\}^2},$$

where $\hat{w}_i = \exp(B(\hat{e}_i)'\hat{\eta}_n)$. In next section, our simulation study show that the estimated asymptotic variance formulas perform well.

4. Simulation

We conduct an extensive simulation study to evaluate the finite sample performance of the proposed methods and compare with competing approaches. The true response model is set up as follows:

$$Y_i = Z_i\beta + f(X_i) + \epsilon_i,\ i = 1, \ldots, n,$$

where $\epsilon_i, i = 1, \ldots, n$ are independent identically distributed from $N(0, 1)$ and n is sample size. The treatment indicator Z_i is a binary variable with treatment reception probability $\text{pr}(Z_i = 1 \mid X_i) = \exp(X_i^T\alpha)/(1 + \exp(X_i^T\alpha))$ and $X_i = (X_{i1}, X_{i3}, X_{i3}, X_{i4}, X_{i5})^T$ is a five-dimensional covariate. We set the true treatment effect as $\beta_0 = 1$. We consider four different data generating scenarios of X_i and $f(\cdot)$:

(a) $f(X)$ is linear with $f(X_i) = X_{1i} + X_{2i} + X_{3i} + X_{4i} + X_{5i}$, $\alpha_0 = (1, 1, 1, 1, 1)^T$, $X_i \sim 0.5N(0, I_5) + 0.5N(1, 16I_5)$, where I_5 is an identity matrix;

(b) $f(X)$ is additive with $f(X_i) = 0.2\exp(X_{1i}) + 2X_{2i}$, $\alpha_0 = (1, -1, 1, -1, 1)^T$, $X_i \sim 0.5N(0, I_5) + 0.5N(0, 2I_5)$, where I_5 is an identity matrix;

(c) $f(X)$ is not additive with $f(X_i) = X_{1i}X_{2i} + \exp(X_{3i})(\sin(X_{4i}) + \cos(X_{5i}))$, $\alpha_0 = (1, -1, 1, -1, 1)^T$, $X_i \sim 0.5N(0, I_5) + 0.25N(1, 2I_5) + 0.25N(-1, 2I_5)$;

(d) $f(X)$ is not additive and X follows an asymmetric distribution, where $f(X_i) = X_{1i}X_{2i} + 0.2\exp((2X_{3i} - 3)/2)(\sin(X_{4i}) + \cos(X_{5i}))$, $\alpha_0 = (1, -1, 1, -1, 1)^T$, $X_{ij} \sim 0.5\exp(1) + 0.5\exp(2), j = 1, \ldots, 5$.

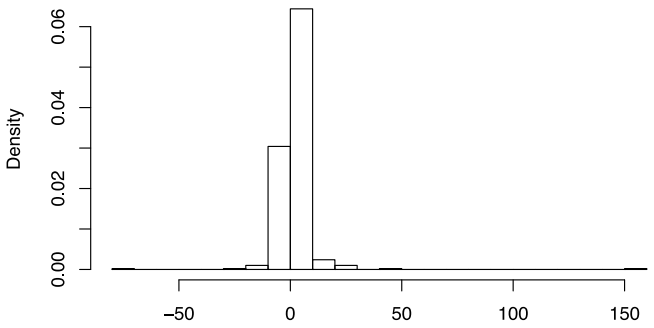

Fig. 1. Histogram of Y for Scenario (c) for one simulation.

Table 1
Comparison of the estimators with estimated or true propensity score for case (a)-(d).

		$n = 200$				$n = 500$				$n = 2000$			
Case	Approach	Bias	SSE	ESE	CP	Bias	SSE	ESE	CP	Bias	SSE	ESE	CP
		Estimated propensity score											
(a)	PSBS.homo	0.009	0.244	0.234	0.935	−0.004	0.151	0.151	0.951	0.001	0.075	0.076	0.949
	PSBS.hetero	0.008	0.249	0.232	0.934	−0.004	0.149	0.150	0.949	0.002	0.072	0.076	0.956
(b)	PSBS.homo	0.031	0.398	0.365	0.933	0.005	0.254	0.234	0.939	−0.001	0.129	0.122	0.949
	PSBS.hetero	0.015	0.406	0.359	0.910	0.009	0.246	0.226	0.930	−0.002	0.122	0.116	0.950
(c)	PSBS.homo	0.040	1.387	1.030	0.940	0.021	0.841	0.724	0.955	0.002	0.458	0.407	0.947
	PSBS.hetero	−0.025	0.783	0.658	0.908	−0.013	0.538	0.504	0.934	0.002	0.308	0.298	0.950
(d)	PSBS.homo	0.010	0.267	0.240	0.947	0.009	0.178	0.156	0.945	−0.002	0.082	0.080	0.962
	PSBS.hetero	0.005	0.226	0.202	0.925	0.003	0.137	0.137	0.948	0.000	0.069	0.072	0.966
		True propensity score											
(a)	PSBS.homo	−0.014	0.246	0.248	0.950	0.006	0.155	0.158	0.951	−0.002	0.078	0.080	0.949
	PSBS.hetero	0.013	0.248	0.230	0.928	0.006	0.153	0.150	0.946	−0.002	0.076	0.077	0.951
(b)	PSBS.homo	0.019	0.656	0.634	0.934	−0.008	0.421	0.411	0.943	0.000	0.208	0.210	0.955
	PSBS.hetero	0.014	0.652	0.589	0.917	0.002	0.411	0.392	0.934	0.001	0.201	0.203	0.951
(c)	PSBS.homo	−0.037	1.413	1.140	0.941	0.024	0.912	0.805	0.950	0.003	0.443	0.431	0.954
	PSBS.hetero	−0.040	0.810	0.703	0.913	−0.006	0.578	0.544	0.935	−0.002	0.332	0.322	0.943
(d)	PSBS.homo	−0.011	0.299	0.274	0.940	0.006	0.180	0.176	0.955	−0.003	0.094	0.090	0.962
	PSBS.hetero	−0.008	0.245	0.221	0.921	0.006	0.150	0.150	0.953	−0.000	0.078	0.080	0.955

Each simulation study is based on 1000 replicates with sample size 200, 500 and 2000. In addition to our proposed methods, we exploit ten commonly used approaches to estimate β including linear model (LM); generalized additive model (GAM, Hastie and Tibshirani, 1986); subclassification based on propensity score (SUBPS, Imbens and Rubin, 2015); augmented inverse probability weighting estimation (AIPW, Robins et al., 1994); the non-calibrated likelihood estimator (LIK, Tan, 2006); the non-calibrated regression estimator (REG, Tan, 2006); the calibrated likelihood estimator (CLIK, Tan, 2010); the calibrated regression estimator(CREG, Tan, 2010); the TMLE based on generalized random forests (TMLE-GRF, van der Laan, 2010; Wager and Athey, 2018); the overlap weight augmented estimator based on generalized random forests (OWAE-GRF, Li et al., 2018; Athey et al., 2019). Our proposed propensity score based spline methods are denoted as PSBS.homo and PSBS.hetero. AIPW, CLIK and CREG use a combination of logistic regression for propensity score model and GAM for outcome regression model. We run LIK, REG, CLIK, CREG with R package **iWeigReg** and implement TMLE-GRF and OWAE-GRF by using R package **grf** with default tuning parameters.

In the following tables, Bias and SSE are the sample average bias (i.e. $\hat{\beta} - \beta_0$) and the sample average standard error of the 1000 simulations respectively. ESE and CP are the average of estimated asymptotic standard error and estimated 95% confidence interval coverage by using the asymptotic variance formula, respectively.

In scenarios (c) and (d), we consider skewed distributions of the outcome, Y. Fig. 1 shows the histogram of Y in scenario (c) for one simulation with a sample size of 500, which indicates that the distribution of Y is skewed to the right with a high peak around zero. Such distribution scenario has important practical implication as many health outcomes have non-normal distributions with certain degree of skewness.

Table 1 summarizes the results of treatment effect estimation for our proposed methods, where the top panel presents results under the estimated propensity score and the bottom panel presents results under the true propensity score. As shown in Table 1, the proposed approaches are asymptotically unbiased, ESE is close to SSE and CP is close to 0.95. This justifies the asymptotic results in Theorems 1 and 2. As expected, PSBS.hetero has a better performance than PSBS.homo

Table 2
Comparison of various approaches for case (a)-(d).

	$n = 200$		$n = 500$		$n = 2000$		$n = 200$		$n = 500$		$n = 2000$	
Approach	Bias	SSE	Bias	SSE	Bias	SSE	Bias	SSE	Bias	SSE	Bias	SSE
	Case (a)						Case (b)					
PSBS.homo	0.009	0.244	−0.004	0.151	0.001	0.075	0.031	0.398	0.005	0.254	−0.001	0.129
PSBS.hetero	0.008	0.249	−0.004	0.149	0.002	0.072	0.015	0.406	0.009	0.246	−0.002	0.122
LM	−0.002	0.115	0.000	0.074	0.001	0.036	0.276	0.258	0.282	0.182	0.278	0.090
GAM	−0.007	0.191	0.003	0.121	−0.001	0.060	−0.040	0.210	−0.026	0.155	−0.025	0.076
SUBPS	0.350	0.316	0.503	0.245	0.461	0.215	−0.044	0.481	−0.075	0.340	−0.058	0.214
AIPW	0.003	0.507	0.029	0.544	−0.013	0.551	0.059	0.647	0.060	0.488	0.028	0.277
LIK	0.065	0.510	0.016	0.400	−0.010	0.449	0.081	0.325	0.081	0.226	0.049	0.146
CLIK	−0.044	1.012	−0.048	0.825	−0.016	0.769	0.086	0.315	0.084	0.221	0.050	0.144
REG	0.028	0.353	0.001	0.330	−0.007	0.396	0.076	0.320	0.079	0.223	0.053	0.143
CREG	−0.091	1.054	−0.052	0.882	−0.019	0.701	0.088	0.314	0.085	0.219	0.054	0.139
TMLE-GRF	5.028	0.426	4.155	0.240	2.753	0.106	0.104	0.286	0.077	0.166	0.035	0.076
OWAE-GRF	4.693	0.440	3.722	0.250	2.396	0.111	0.097	0.285	0.068	0.170	0.026	0.076
	Case (c)						Case (d)					
PSBS.homo	0.040	1.387	0.021	0.841	0.002	0.458	0.010	0.267	0.009	0.178	−0.002	0.082
PSBS.hetero	−0.025	0.783	−0.013	0.538	0.002	0.308	0.005	0.226	0.003	0.137	0.000	0.069
LM	2.570	0.978	2.588	0.574	2.635	0.307	−0.885	5.093	−1.437	11.801	−1.848	21.613
GAM	0.607	1.215	0.637	0.829	0.715	0.477	0.075	0.396	0.066	0.194	0.070	0.231
SUBPS	0.186	1.480	0.152	1.031	0.132	0.594	0.027	0.246	0.013	0.168	0.012	0.111
AIPW	0.207	6.187	−0.225	3.617	0.071	2.548	0.147	1.507	0.077	1.217	0.028	0.340
LIK	−0.114	1.494	−0.249	1.190	−0.143	0.808	0.123	1.094	0.176	2.083	0.151	1.318
CLIK	−0.076	1.482	−0.221	1.146	−0.135	0.802	0.139	1.548	0.158	1.843	0.182	2.019
REG	−0.118	1.370	−0.198	1.072	−0.113	0.767	0.084	0.762	0.167	1.953	0.100	0.653
CREG	−0.057	1.459	−0.171	1.111	−0.105	0.772	0.171	2.402	0.155	1.847	0.181	2.021
TMLE-GRF	0.078	1.516	0.017	0.703	0.084	0.333	0.337	4.250	0.237	1.856	0.130	0.648
OWAE-GRF	0.119	1.494	0.100	0.730	0.175	0.331	0.342	4.384	0.225	1.659	0.102	0.519

These competing approaches includes linear model (LM); generalized additive model (GAM); subclassification based; on propensity score (SUBPS); augmented inverse probability weighting estimation (AIPW); the non-calibrated likelihood estimator (LIK); the non-calibrated regression estimator (REG); the calibrated likelihood estimator (CLIK); the calibrated regression estimator(CREG); the targeted maximum likelihood estimation based on generalized random forests (TMLE-GRF); overlap weight augmented estimator based on generalized random forests (OWAE-GRF).

in terms of smaller SSEs and ESEs, under most scenarios. Moreover, SSEs and ESEs for the estimators with the estimated propensity score are smaller than that with the true propensity score.

Table 2 presents results for all twelve approaches being compared in terms of Bias and SSE. PSBS.homo and PSBS.hetero perform well under all scenarios, in comparison to competing methods. When $f(X)$ is linear, most of methods are asymptotically unbiased (except SUBPS, TMLE-GRF and OWAE-GRF). The poor performance of GRF based methods is likely due to the fact that, in case (a), one group has a much larger variance than the other one, so the overlap is as good as other cases. We generate the data this way to assess the sensitivity of the results to potentially extreme propensity score weights. As expected, LM and GAM have the best performance with smaller SSEs. When $f(X)$ is additive and nonlinear, GAM yield the smallest SSE but a larger bias. PSBS.homo and PSBS.hetero are asymptotically unbiased, and their SSEs are slightly larger than GAM and smaller than other methods. When $f(X)$ is not additive, only PSBS.homo and PSBS.hetero have both small bias and small variance, while other approaches tend to have large bias or large variance. The results are similar when X follows an asymmetric distribution. For case (b) to (d), TMLE and OWAE methods tend to produce estimates with similar bias to most of other methods but with smaller standard error.

We also conduct additional simulation to assess the performance of the proposed estimators under misspecified propensity score models. We set $\alpha_0 = (1, -1, 1, -1, 1)^T$, $X_i \sim 0.5N(0, I_5) + 0.25N(1, 4I_5) + 0.25N(-1, 4I_5)$, $f(X) = X_1X_2 + 0.25\exp(X_3)(\sin(X_4) + \cos(X_5))$. Two scenarios for misspecified propensity score models are as follows:

Scenario (e) (Slightly misspecified): The true propensity score model is a probit model, i.e. binary variable Z with success probability $\text{pr}(Z = 1 \mid X) = \Phi(X^T\alpha)$, where $\Phi(\cdot)$ is probability distribution function of the standard normal distribution;

Scenario (f) (Severely misspecified): The true propensity score model includes higher order terms of X, i.e. binary variable Z with success probability $\text{pr}(Z = 1 \mid X) = \exp((X^3)^T\alpha)/(1 + \exp((X^3)^T\alpha))$.

In this simulation, logistic regression with linear terms of X is used for estimating the propensity score, as commonly seen in practice. Results for treatment effect estimation are summarized in Table 3. For scenario (e), both proposed methods still perform well, while biases from other methods have increased substantially.

Unfortunately, for the severely misspecified scenario, where the success probability is not a linear function of X, no consistent estimation can be obtained using a linear propensity score model. All methods present large bias, but with large sample size, TMLE-GRF and OWAE-GRF estimators report bias much smaller than other methods. It seems that they might be more robust to highly misspecified propensity score models due to the use of random forest.

Table 3
Simulation results under misspecified propensity score model.

	$n = 200$		$n = 500$		$n = 2000$		$n = 200$		$n = 500$		$n = 2000$	
Approach	Bias	SSE	Bias	SSE	Bias	SSE	Bias	SSE	Bias	SSE	Bias	SSE
	Scenario (e)						Scenario (f)					
PSBS.homo	0.039	1.090	0.006	0.714	0.003	0.458	0.577	1.171	0.613	0.835	0.601	0.434
PSBS.hetero	0.018	0.748	−0.008	0.553	−0.004	0.336	0.397	0.779	0.481	0.588	0.539	0.323
LM	0.336	2.086	0.311	1.482	0.382	0.990	0.630	1.899	0.784	1.362	0.750	0.775
GAM	0.537	1.400	0.606	1.056	0.698	0.770	0.532	1.364	0.668	1.079	0.665	0.702
SUBPS	0.048	1.085	0.006	0.700	0.026	0.540	0.588	1.226	0.650	0.995	0.752	0.703
AIPW	0.077	1.765	0.031	1.327	0.168	0.802	0.429	3.120	0.452	3.328	0.590	2.939
LIK	−0.035	1.612	−0.113	1.426	−0.103	1.291	0.539	1.965	0.635	1.575	0.710	1.276
CLIK	0.009	1.811	−0.112	1.524	−0.102	1.313	0.512	2.016	0.615	1.590	0.700	1.249
REG	−0.016	1.324	−0.072	1.103	−0.074	1.009	0.540	1.802	0.603	1.548	0.647	1.396
CREG	−0.066	1.635	−0.106	1.343	−0.106	1.106	0.492	2.067	0.538	1.677	0.614	1.347
TMLE-GRF	0.263	2.601	0.189	1.754	0.214	0.600	0.441	1.863	0.233	1.022	0.161	0.434
OWAE-GRF	0.309	2.812	0.256	1.872	0.298	0.625	0.482	1.893	0.276	1.104	0.198	0.533

These competing approaches includes linear model (LM); generalized additive model (GAM); subclassification based; on propensity score (SUBPS); augmented inverse probability weighting estimation (AIPW); the non-calibrated likelihood estimator (LIK); the non-calibrated regression estimator (REG); the calibrated likelihood estimator (CLIK); the calibrated regression estimator(CREG); the targeted maximum likelihood estimation based on generalized random forests (TMLE-GRF); overlap weight augmented estimator based on generalized random forests (OWAE-GRF).

5. Application to health policy evaluation

5.1. Background

Since its passage in 2010, ACA has reduced the number of uninsured Americans substantially through newly created health insurance marketplace and an expansion of Medicaid program. Ohio is one of the 33 states adopting Medicaid expansion and it provides health insurance coverage to all non-senior adults (19–64) whose household income did not exceed 138% of the Federal Poverty Level (FPL). For those with FPL between 138%–400%, they could purchase subsidized health insurance plans from the marketplace. Changes in coverage are associated with improvements in self-reported measures of access to care and reduced out-of-pocket medical expenditures. But its impact on quality of medical care, including self-reported health status (SRHS), is less clear. The voluntary nature of insurance purchase for those aged 19–64 with FPL 138%–400% leaves an important causal question to be addressed: what would have happened to their SRHS for those who did not have health insurance had they purchased the insurance?

We use the data from Ohio Medicaid Assessment Survey (OMAS) 2012 to answer this question. OMAS 2012 is a representative survey of Ohio residents with a total sample size of 22,929. The outcome of interest is SRHS, which was collected using a 5-point likert scale, i.e. 1 for excellent, and 5 for poor.

The exposure variable is the health insurance status. We create a binary insurance status variable, Z, by grouping those with known insurance type as exposed ($Z = 1$) and those without insurance as unexposed ($Z = 0$). Based on discussion with health service experts, we include 13 important covariates in the propensity score model: race/ethnicity, age, gender, working status, education, disability status, income, county type, children in the household, marital status, smoking status, drink alcohol, and mental health distress. By subsetting the data to the target population with family income FPL between 138%–400% and age 19–64, and removing missing values, we get an analytical dataset with a total of 4880 samples, where 4195 have health insurance and 685 do not.

5.2. Estimating a constant average policy effect

Prior to the outcome analysis, we need to assess the overlap of the covariates distribution between the group with insurance ($Z = 1$) and the group without ($Z = 0$). This is to check the common support assumption to ensure the validity of the inference. If the support differ substantially between the two groups, the estimate of treatment effect depends largely on the model extrapolation. A common method to assess the difference between the two groups is using the propensity score. As discussed in Imbens and Rubin (2015), any imbalance in the covariate distributions leads to a difference in the propensity score distributions. In practice, we use the linear propensity score, which is the logit transformation of the propensity score, because it is more likely to be approximated well by a normal distribution.

The left panel of Fig. 2 presents the boxplot of the estimated linear propensity score for insured group and uninsured group. The right panel of Fig. 2 presents the histogram of the estimated linear propensity score with a kernel density estimation curve. It is clear that the linear propensity score distributions are different in two groups, which is expected for an observational study. The two distributions overlap well, which shows little violation to the common support assumption. The insured group seems to have a slightly wider range with a longer tail to the right. This is not surprising as the sample size in the insured group is six times as big as that in the uninsured group.

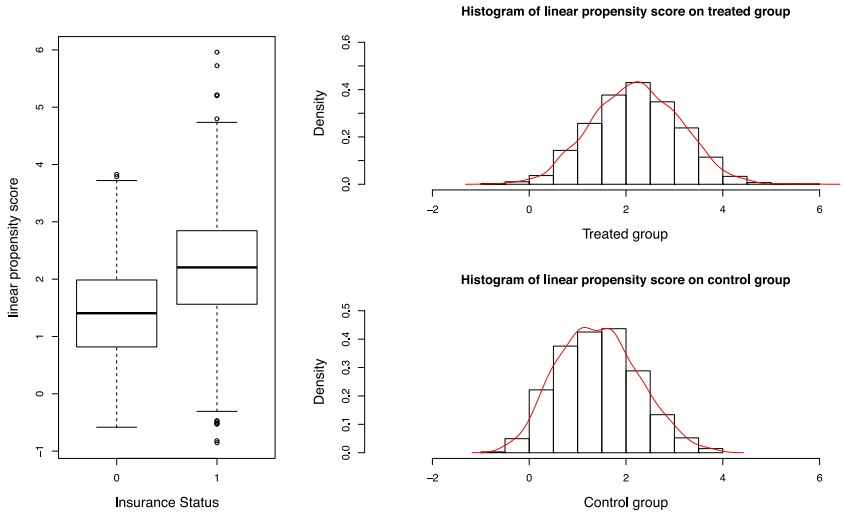

Fig. 2. Histogram-based estimate of the distribution of the linear propensity score for treated and control group.

Table 4
Counts of different values of outcome.

Value of outcome		1	2	3	4	5
Count	Insured	755	1451	1210	575	204
	Uninsured	104	188	236	115	42
	Total	859	1639	1446	690	246

Table 5
Estimated effects of insurance status with nine methods.

	Approach	$\hat{\beta}_n$	ESE	P-value[a]	Approach	$\hat{\beta}_n$	ESE	P-value[a]
AIC	PSBS.homo	−0.0798	0.0419	0.0283	AIPW	−0.0746	0.0527	0.0785
	PSBS.hetero	−0.0807	0.0419	0.0272	LIK	−0.0761	0.0521	0.0724
BIC	PSBS.homo	−0.0813	0.0419	0.0463	CLIK	−0.0756	0.0522	0.0740
	PSBS.hetero	−0.0817	0.0420	0.0258	REG	−0.0741	0.0527	0.0801
	LM	−0.0718	0.0427	0.0463	CREG	−0.0739	0.0529	0.0813
	GAM	−0.0697	0.0428	0.0518	TMLE-GRF	−0.107	0.0455	0.0094
	SUBPS	−0.1704	0.0465	0.0001	OWAE-GRF	−0.100	0.0415	0.0078

[a]The p-values are obtained by one-sided test, that is, $H_0 : \beta = 0$ against $H_1 : \beta < 0$.

We also notice that the distribution of SRHS is right skewed as shown in Table 4. It is not easy to specify a simple parametric model for such outcome. Therefore, we apply the propose method with using the cubic spline to approximate $g(\cdot)$ and find the optimal number of knots by using grid search with Akaike information criterion (AIC, Akaike, 1970, 1974) and Bayesian information criterion (BIC, Schwarz, 2005).

Table 5 presents the estimated effect, $\hat{\beta}_n$, of having insurance on SRHS. The corresponding estimated standard errors (ESE) for PSBS.homo and PSBS.hetero are based on the estimated variance formula, and ESEs for the other ten approaches are obtained via 200 bootstraps. The *p-values* are from one-sided test of no insurance effect, as we expect having insurance would improve SRHS. For comparison purpose, we report the estimated effects using all twelve methods as described in the simulation study section.

As shown in Table 5, all methods produce similar point estimates except SUBPS. The SUBPS result almost doubles the effect estimated from all other methods, which is between 0.07 to 0.1. Since subclassification does not fully utilize the information in covariates or the propensity score, it may yield more biased results (like case (a) in simulation) and we do not consider it further. Both LM and GAM have similar point estimates and relatively small ESEs, which are larger than our proposed methods, but smaller than other methods. This may suggest a slight violation of the linear and additive assumptions from the true outcome model.

AIPW, LIK, CLIK, REG and CREG all produce very similar results, but their variance estimates are much larger than the proposed methods. It implies that our approaches are advantageous for this data analysis. Moreover, the use of logistic regression to estimating the propensity score seems reasonable given that AIPW, CLIK and CREG yield close results. Otherwise, the variance estimates would be different in theory (Tan, 2006, 2010). TMLE-GRF and OWAE-GRF report treatment effects 20% larger than most of other methods. Due to the blackbox nature of random forest, it is not

Table 6
Estimated effects of insurance status for the proposed methods.

		PSBS.homo			PSBS.hetero		
		Estimate	ESE	p-value	Estimate	ESE	p-value
$\hat{\beta}_n$	Insurance status	0.019	0.081	0.816[a]	0.028	0.083	0.739[a]
$\hat{\theta}_n$	Gender (Female)	0.032	0.036	0.365[a]	0.032	0.036	0.376
	Race (Black)	0.104	0.086	0.228[a]	0.102	0.087	0.238[a]
	Race (White)	−0.170	0.073	0.019[a]	−0.178	0.072	0.014[a]
$\hat{T}_n$	Test statistic ($H_0 : \theta = 0$)	29.42		< 0.001[a]	34.08		< 0.001[a]
$\widehat{ATE}$	Average treatment effect	−0.084	0.042	0.024[b]	−0.083	0.042	0.025[b]

[a] p-value for two-sided test.
[b] p-value for one-sided test, $H_0 : ATE = 0$ vs. $H_1 : ATE < 0$.

clear if this departure represents a correction from the potentially misspecified propensity score model or just a reflection of more bias. Overall, having insurance seems to have a beneficial effect on SRHS. But the magnitude of the effect is rather small, as having insurance only improves the SRHS score by 0.08 unit on average.

5.3. Model checking for potential subgroup effects

Our model finds small beneficial effect of having insurance on SRHS, assuming a constant overall effect. It is robust since it is unrestrictive of the covariate functional forms. But it is restrictive in the sense that it only allows a constant causal effect. In practice, this may not hold exactly as people with different characteristics might respond differently to a given intervention. In health disparity research, one major concern is that a health policy may have heterogenous effects in subgroups defined by gender or racial groups. So we extend our constant effect model to the following setup:

$$Y = Z\beta + f(X) + Z \cdot W^T\theta + \epsilon, \quad E(\epsilon \mid X, Z) = 0 \tag{11}$$

where W is a set of effect modifiers, which could be a subset of X. We choose this model because of its clear causal interpretation, as the coefficient of the interaction term, θ, indicates potential subgroup effects as defined by W.

Denoting $g(e) = E[f(X)|e]$, we can rewrite model (11) as

$$Y = Z\beta + g(e) + Z \cdot W^T\theta + \xi, \tag{12}$$

where $\xi = f(X) - g(e) + \epsilon$ with $E[\xi|Z, e] = 0$. We can adopt B-spline basis to approximate $g(e)$, then obtain the estimator of θ by (weighted) least squares. Similar to the proof in Section 3, we can show that $\sqrt{n}(\hat{\theta}_n - \theta_0) \to N(0, \Sigma_\theta)$. So the test statistic is given by $T_n = n\hat{\theta}_n^T \hat{\Sigma}_\theta^{-1} \hat{\theta}_n$, where $\hat{\Sigma}_\theta$ is a consistent estimate of Σ_θ. Under the null hypothesis $\theta = 0$, $T_n \to \chi^2(q)$ in distribution, where q is the dimension of W.

To check the potential health disparity related subgroup policy effects, we include gender (Male vs. Female) and racial groups (White, Black, or Other) interactions with the insurance indicator. We rerun the data with model (12) and the results are reported in Table 6, where the corresponding ESEs are obtained based on 1000 bootstraps. $\hat{T}_n$ is the value of test statistic for the interactions. As shown in Table 6, its p-value is very significant indicating heterogeneous effects due to gender and racial subgroups. Then we recalculate the estimate of average causal effect using the sample mean of $\{\hat{\beta}_n + X_i^T\hat{\theta}_n, i = 1, \ldots, n\}$ (denoted as $\widehat{ATE}$ in the table). Both point estimates and standard errors are quite close to the results under constant effect model, which implies that the magnitude of the subgroup effects is kind of minor and does not shift the population average effect much.

6. Discussion

We develop a semiparametric strategy for estimating the population average treatment effect, using the popular spline method and establish its large sample properties. Both simulation studies and real data analysis show that our method outperforms the commonly used competing estimators. The main advantages of the proposed approach are: (1) no need to specify the outcome model with a large number of covariates; (2) overcome the unstable estimation when the propensity score is close to one or zero; (3) the estimators are consistent with superior variance results.

Our approach needs a consistently estimated propensity score and we use a logistic regression model to estimate it, following the common practice. In theory, the propensity score can be modeled generally as $\text{pr}(Z = 1 \mid X) = F(X^T\alpha)$ for a vector α and only the linear component $X^T\alpha$ matters. So our proposed estimator still works even if the function form F is unknown. But as shown in the simulation study, with a severely misspecified propensity score model, our methods do not work well, like many competing methods. Under such scenario, it is recommended to consider the targeted maximum likelihood estimation or the overlap weight augmented estimation, as they show smallest bias in our simulation.

One interesting extension is to consider heterogeneous subgroup effects as our approach mainly focuses on a constant causal effect. With a known structure of interaction terms, our model can be extended as shown in Section 5. This is

reasonable for many applications, as investigators usually have some prior knowledge of the subgroup effects based on literature or experience. When the subgroup structure is not known, as in some machine learning applications, more work is needed to develop a better model and corresponding estimation strategy. Another extension is to consider the case when the outcome is discrete, modeled by generalized linear models, which has broader applications in health and social sciences. Among the methods we have compared, AIPW, (C)REG, (C)LIK, TMLE and OWAE can be applicable to more complex outcome regression.

It is well known that observational studies may suffer from hidden bias, when important covariates are not collected. Sensitivity analysis strategies have been discussed to evaluate how much the estimated treatment effect may change in the presence of hidden bias at various hypothetical magnitudes (Rosenbaum, 2002; Hsu and Small, 2013). We will explore how to embed such strategy in our semiparametric estimator in future research.

CRediT authorship contribution statement

Peng Wu: Methodology, Programming, Data analysis, Writing - original draft. **Xinyi Xu:** Methodology, Writing - review & editing. **Xingwei Tong:** Conceptualization, Methodology, Writing - review & editing. **Qing Jiang:** Methodology. **Bo Lu:** Conceptualization, Methodology, Writing - original draft.

Acknowledgments

This work was partially supported by grant 1R01 HS024263-01 from the Agency of Healthcare Research and Quality of the U.S. Department of Health and Human Services and grant DMS-1613110 from National Science Foundation, United States. We thank Dr. Runze Li for insightful comments. We also thank the Associate Editor and two anonymous referees for their constructive reviews, which lead to substantial improvement of the paper.

Appendix

Lemma 1. *Let* $H = B(e)\left[B(e)^T B(e)\right]^{-1} B(e)^T$, *where* $B(e) = (B(e_1), \ldots, B(e_n))^T$. *Then under the conditions 1–4, as* $n \to \infty$,

$$\frac{1}{n}\left(Z^T(I-H)Z\right) \xrightarrow{P} E[e_1(1-e_1)]. \tag{13}$$

Proof. Note that

$$\begin{aligned}\frac{1}{n}\left(Z^T(I-H)Z\right) &= \frac{1}{n}\left[(Z-e+e)^T(I-H)(Z-e+e)\right]\\ &= \frac{1}{n}\left[(Z-e)^T(I-H)(Z-e)+(Z-e)^T(I-H)e+e^T(I-H)e\right].\end{aligned}$$

Since H is a non-negative definite matrix and is idempotent, following the proof of Lemma 5 in Chen (1988), we have that as $n \to \infty$,

$$\frac{1}{n}(Z-e)^T(I-H)(Z-e) \xrightarrow{P} var(Z_1-e_1) = E[e_1(1-e_1)]. \tag{14}$$

Moreover, as shown by Chen (1988), under the condition C3, there exists a positive constant C_n with $\lim_{n\to\infty} C_n = 0$ such that $|(I-H)e_i| \le C_n$, for $1 \le i \le n$. Therefore,

$$\frac{1}{n}e^T(I-H)e \xrightarrow{P} 0. \tag{15}$$

Finally, since $I-H$ is an idempotent matrix, it follows from (14) and (15), the Markov inequality and the Cauchy–Schwarz inequality that as $n \to \infty$,

$$\frac{1}{n}(Z-e)^T(I-H)e \xrightarrow{P} 0. \tag{16}$$

Combining (14), (15) and (16) yields the asymptotic behavior (13) in Lemma 1.

Proof of Theorem 1. When the propensity scores vector e is known, the homoscedastic treatment effect estimator $\hat{\beta}_n^{homo}$ can be represented by

$$\hat{\beta}_n^{homo} = \left[Z^T(I-H)Z\right]^{-1} Z^T(I-H)Y,$$

where $H = B(e)(B(e)^TB(e))^{-1}B(e)^T$ and $B(e) = (B(e_1), \ldots, B(e_n))^T$. To study the asymptotic properties of $\hat{\beta}_n^{homo}$, we decompose $\sqrt{n}(\hat{\beta}_n^{homo} - \beta_0)$ as

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n^{homo} - \beta_0) &= \sqrt{n}\left([Z^T(I-H)Z]^{-1}Z^T(I-H)(Z\beta_0 + g(e) + \xi) - \beta_0\right)\\ &= \sqrt{n}[Z^T(I-H)Z]^{-1}Z^T(I-H)(g(e)+\xi)\\ &= \left[\frac{1}{n}Z^T(I-H)Z\right]^{-1}\frac{1}{\sqrt{n}}Z^T(I-H)(g(e)+\xi).\end{aligned}$$

By Lemma 1, $Z^T(I-H)Z/n$ converges to a positive constant $E[e_1(1-e_1)]$. Also, Chen (1988) showed in his Eq. (8) that under the conditions (C1) and (C2),

$$\frac{1}{\sqrt{n}}Z^T(I-H)g(e) \xrightarrow{P} 0. \tag{17}$$

Therefore, we only need to show that $\sqrt{n}[Z^T(I-H)Z]^{-1}Z^T(I-H)\xi$ is asymptotically normal. Let r_{ii} be the ith diagonal element of the projection matrix $(I-H)Z[Z^T(I-H)Z]^{-1}Z^T(I-H)$, and then it follows from Lemma 3 of Wu (1981) that

$$\max_i r_{ii} \to 0, \qquad \text{as } n \to \infty.$$

According to Proposition 2.2 of Huber (1973), when $max_i r_{ii} \to 0$ as $n \to \infty$, the term $\sqrt{n}[Z^T(I-H)Z]^{-1}Z^T(I-H)\xi$ converges in distribution to a normal random variable with mean 0. Since H is a projection matrix, HZ is a consistent estimator of $E(Z \mid e) = e$. Thus, by the Central Limit Theorem,

$$\frac{1}{\sqrt{n}}Z^T(I-H)\xi \xrightarrow{d} N\left(0, var[(Z_1 - e_1)\xi_1]\right). \tag{18}$$

The asymptotic property (7) follows immediately from Lemma 1, (17) and (18).

When propensity score vector e is unknown but can be consistently estimated by $\hat{e}$, the homoscedastic treatment effect estimator $\hat{\beta}^{homo}$ is written as

$$\hat{\beta}_n^{homo} = (Z^T(I-\hat{H})Z)^{-1}Z^T(I-\hat{H})Y,$$

where $\hat{H} = B(\hat{e})(B(\hat{e})^TB(\hat{e}))^{-1}B(\hat{e})^T$ and $B(\hat{e}) = (B(\hat{e}_1), \ldots, B(\hat{e}_n))^T$. Similar as in part (a), we can represent $\sqrt{n}(\hat{\beta}_n^{homo} - \beta_0)$ by

$$\begin{aligned}\sqrt{n}(\hat{\beta}_n^{homo} - \beta_0) &= \sqrt{n}[Z^T(I-\hat{H})Z]^{-1}Z^T(I-\hat{H})(g(e)+\xi),\\ &= \left[\frac{1}{n}Z^T(I-\hat{H})Z\right]^{-1}\frac{1}{\sqrt{n}}Z^T(I-\hat{H})(g(e)+\xi).\end{aligned}$$

Since $\hat{e}$ is an consistent estimator of e, by the continuous mapping theorem, Lemma 1 and Eq. (17), we have

$$\frac{1}{n}\left(Z^T(I-\hat{H})Z\right) \xrightarrow{P} E[e_1(1-e_1)] \tag{19}$$

and

$$\frac{1}{\sqrt{n}}Z^T(I-\hat{H})g(e) \xrightarrow{P} 0. \tag{20}$$

Moreover, by a similar argument as in part (a), $\sqrt{n}[Z^T(I-\hat{H})Z]^{-1}Z^T(I-\hat{H})\xi$ also converges in distribution to a normal random variable with mean 0. To derive the asymptotic variance, we decompose $n^{-1/2}Z^T(I-H)\xi$ as

$$\begin{aligned}\frac{1}{\sqrt{n}}Z^T(I-H)\xi &= \frac{1}{\sqrt{n}}\left[(Z-e) + (e-\hat{e}) + (\hat{e}-\hat{H}Z)\right]^T\xi\\ &= \frac{1}{\sqrt{n}}(Z-e)^T\xi + \frac{1}{\sqrt{n}}(e-\hat{e})^T\xi + \frac{1}{\sqrt{n}}(\hat{e}-\hat{H}Z)^T\xi\end{aligned}$$

Because of the consistency of $\hat{e}$, it is easy to see that

$$\frac{1}{\sqrt{n}}(\hat{e}-\hat{H}Z)'\xi \xrightarrow{P} 0.$$

Using the logistic regression (3) and Taylor expansion, we obtain that for $i = 1, \ldots, n$,

$$e_i - \hat{e}_i = -e_i(1-e_i)X_i'(\hat{\alpha}_n - \alpha_0)(1 + o_n(1)).$$

The Fisher information matrix of α at α_0 can be calculated as

$$I(\alpha_0) = -E[l_1''(\alpha_0)] = E[e_1(1-e_1)X_1X_1'],$$

where $l_1(\alpha_0)$ is the likelihood of e_1 at α_0 based on the logistic regression model (3). Thus,

$$\begin{aligned}\frac{1}{\sqrt{n}}(e-\hat{e})'\xi &= -\frac{1}{\sqrt{n}}\sum_{i=1}^{n} e_i(1-e_i)\xi_i X_i'(\hat{\alpha}_n-\alpha_0)(1+o_p(1))\\ &= -\left[\frac{1}{n}\sum_{i=1}^{n} e_i(1-e_i)\xi_i X_i\right]'\cdot\sqrt{n}(\hat{\alpha}_n-\alpha_0)(1+o_p(1))\\ &= -\left[\frac{1}{n}\sum_{i=1}^{n} e_i(1-e_i)\xi_i X_i\right]'\cdot\left[\frac{1}{n}l_n''(\alpha_0)\right]^{-1}\left(-\frac{1}{\sqrt{n}}l_n'(\alpha_0)\right)(1+o_p(1))\\ &= -\frac{1}{\sqrt{n}}A'I^{-1}(\alpha_0)\sum_{i=1}^{n}(Z_i-e_i)X_i\,(1+o_p(1))\end{aligned}$$

where $A=E[e_1(1-e_1)\xi_1X_1]$ and $l_n(\alpha_0)$ is the likelihood of $e=(e_1,\ldots,e_n)$ at α_0 based on the logistic regression model (3). Therefore,

$$\begin{aligned}\frac{1}{\sqrt{n}}Z'(I-H)\xi &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[(Z_i-e_i)\xi_i - A'I^{-1}(\alpha_0)(Z_i-e_i)X_i\right](1+O_n(1))\\ &\xrightarrow{d} N\left(0,\, var[(Z_1-e_1)\xi_1 - A'I^{-1}(\alpha_0)(Z_1-e_1)X_1]\right).\end{aligned}\tag{21}$$

Also note that the asymptotic variance

$$\begin{aligned}&var[(Z_1-e_1)\xi_1 - A'I^{-1}(\alpha_0)(Z_1-e_1)X_1]\\ =\;& var[(Z_1-e_1)\xi_1] - 2cov[(Z_1-e_1)\xi_1, A'I^{-1}(\alpha_0)(Z_1-e_1)X_1]\\ &+var[A'I^{-1}(\alpha_0)(Z_1-e_1)X_1]\\ =\;& var[(Z_1-e_1)\xi_1] - 2A'I^{-1}(\alpha_0)E[(Z_1-e_1)^2\xi_1X_1]\\ &+A'I^{-1}(\alpha_0)E[(Z_1-e_1)^2X_1X_1']I^{-1}(\alpha_0)A\\ =\;& var[(Z_1-e_1)\xi_1] - 2A'I^{-1}(\alpha_0)A + A'I^{-1}(\alpha_0)I(\alpha_0)I^{-1}(\alpha_0)A\\ =\;& var[(Z_1-e_1)\xi_1] - A'I^{-1}(\alpha_0)A\end{aligned}\tag{22}$$

The asymptotic property (8) follows from (19), (20), (21) and (22). □

Lemma 2. *Let the conditional error covariance matrix $\Sigma = diag\{w_1(e),\ldots,w_n(e)\}$, where $w_i(e) = var(\xi_i \mid e_i)$. Also, let $H_\Sigma = \Sigma^{-1/2}B(e)\left[B(e)'\Sigma^{-1}B(e)\right]^{-1}B(e)'\Sigma^{-1/2}$ and $Z_\Sigma = \Sigma^{-1/2}Z$. Then under the conditions C1–C4, as $n\to\infty$,*

$$\frac{1}{n}\left(Z_\Sigma'(I-H_\Sigma)Z_\Sigma\right)\xrightarrow{P} E[e_1(1-e_1)/w_1(e)].\tag{23}$$

Proof. Similar as in the proof of Lemma 1, we rewrite $\left(Z_\Sigma'(I-H_\Sigma)Z_\Sigma\right)/\sqrt{n}$ as

$$\begin{aligned}\frac{1}{n}\left(Z_\Sigma'(I-H_\Sigma)Z_\Sigma\right) &= \frac{1}{n}\left[(Z_\Sigma-\Sigma^{-1/2}e+\Sigma^{-1/2}e)'(I-H_\Sigma)(Z_\Sigma-\Sigma^{-1/2}e+\Sigma^{-1/2}e)\right]\\ &= \frac{1}{n}\left[(Z_\Sigma-\Sigma^{-1/2}e)'(I-H_\Sigma)(Z_\Sigma-\Sigma^{-1/2}e)\right]\\ &\quad+\frac{1}{n}\left[(Z_\Sigma-\Sigma^{-1/2}e)'(I-H_\Sigma)\Sigma^{-1/2}e + e'\Sigma^{-1/2}(I-H_\Sigma)\Sigma^{-1/2}e\right]\end{aligned}$$

It is easy to verify that H_Σ is also non-negative definite matrix and is idempotent. Thus, following the proof of Lemma 5 in Chen (1988), we have that as $n\to\infty$,

$$\begin{aligned}\frac{1}{n}(Z_\Sigma-\Sigma^{-1/2}e)'(I-H_\Sigma)(Z_\Sigma-\Sigma^{-1/2}e) &\xrightarrow{P} var(w_1(e)^{-1/2}(Z_1-e_1))\\ &= E[e_1(1-e_1)/w_1(e)].\end{aligned}\tag{24}$$

Furthermore, following the argument in the proof of Lemma 1, we can similarly show that

$$\frac{1}{n}e'\Sigma^{-1/2}(I-H_\Sigma)\Sigma^{-1/2}e\xrightarrow{P}0\tag{25}$$

and

$$\frac{1}{n}(Z_\Sigma - \Sigma^{-1/2}e)'(I - H_\Sigma)\Sigma^{-1/2}e \xrightarrow{P} 0. \tag{26}$$

This completes the proof of Lemma 2.

Proof of Theorem 2. When the propensity score vector e is known, the heteroscedastics treatment effect estimator $\hat{\beta}_n^{hetero}$ can be represented by

$$\hat{\beta}_n^{hetero} = (Z_\Sigma'(I - \hat{H}_\Sigma)Z_\Sigma)^{-1}Z_\Sigma'(I - \hat{H}_\Sigma)\Sigma^{-1/2}Y,$$

where $Z_\Sigma = \Sigma^{-1/2}Z$ and $\hat{H}_\Sigma = \Sigma^{-1/2}B(e)\left[B(e)'\Sigma^{-1}B(e)\right]^{-1}B(e)'\Sigma^{-1/2}$. Therefore,

$$\begin{aligned}
&\sqrt{n}(\hat{\beta}_n^{hetero} - \beta_0)\\
&= \sqrt{n}\left([Z_\Sigma'(I - H_\Sigma)Z_\Sigma]^{-1}Z_\Sigma'(I - H_\Sigma)\Sigma^{-1/2}(Z\beta_0 + g(e) + \xi) - \beta_0\right)\\
&= \sqrt{n}[Z_\Sigma'(I - H_\Sigma)Z_\Sigma]^{-1}Z_\Sigma'(I - H_\Sigma)\Sigma^{-1/2}(g(e) + \xi)\\
&= \left[\frac{1}{n}Z_\Sigma'(I - H_\Sigma)Z_\Sigma\right]^{-1}\frac{1}{\sqrt{n}}Z_\Sigma'(I - H_\Sigma)\Sigma^{-1/2}(g(e) + \xi).
\end{aligned}$$

By Lemma 2, $Z_\Sigma'(I - H_\Sigma)Z_\Sigma/n$ converges to a positive constant $E[e_1(1 - e_1)/w_1(e)]$. Following the argument in the proof of Theorem 2(a), we can show

$$\frac{1}{\sqrt{n}}Z_\Sigma'(I - H_\Sigma)\Sigma^{-1/2}g(e) \xrightarrow{P} 0. \tag{27}$$

and $\sqrt{n}[Z_\Sigma'(I - H_\Sigma)Z_\Sigma]^{-1}Z_\Sigma'(I - H_\Sigma)\xi$ converges in distribution to a normal random variable, that is,

$$\frac{1}{\sqrt{n}}Z_\Sigma'(I - H_\Sigma)\Sigma^{-1/2}\xi \xrightarrow{d} N\left(0, var[(Z_1 - e_1)\xi_1/w_1(e)]\right). \tag{28}$$

The asymptotic property (9) follows immediately from Lemma 2, (27) and (28).

When the propensity score vector e is unknown but can be consistently estimated by $\hat{e}$, the heteroscedastic treatment effect estimator $\hat{\beta}_n^{hetero}$ is

$$\hat{\beta}_n^{hetero} = (Z_{\hat{\Sigma}}'(I - \hat{H}_{\hat{\Sigma}})Z_{\hat{\Sigma}})^{-1}Z_{\hat{\Sigma}}'(I - \hat{H}_{\hat{\Sigma}})\hat{\Sigma}^{-1/2}Y,$$

where $Z_{\hat{\Sigma}} = \hat{\Sigma}^{-1/2}Z$, $\hat{H}_{\hat{\Sigma}} = \hat{\Sigma}^{-1/2}B(\hat{e})\left[B(\hat{e})'\hat{\Sigma}^{-1}B(\hat{e})\right]^{-1}B(\hat{e})'\hat{\Sigma}^{-1/2}$, and $\hat{\Sigma} = diag(w_1(\hat{\eta}), \ldots, w_n(\hat{\eta}))$. We decompose $\sqrt{n}(\hat{\beta}_n^{hetero} - \beta_0)$ as

$$\begin{aligned}
\sqrt{n}(\hat{\beta}_n^{hetero} - \beta_0) &= \sqrt{n}[Z_{\hat{\Sigma}}'(I - \hat{H}_{\hat{\Sigma}})Z_{\hat{\Sigma}}]^{-1}Z_{\hat{\Sigma}}'(I - \hat{H})\hat{\Sigma}^{-1/2}(g(e) + \xi),\\
&= \left[\frac{1}{n}Z_{\hat{\Sigma}}'(I - \hat{H}_{\hat{\Sigma}})Z_{\hat{\Sigma}}\right]^{-1}\frac{1}{\sqrt{n}}Z_{\hat{\Sigma}}'(I - \hat{H}_{\hat{\Sigma}})\hat{\Sigma}^{-1/2}(g(e) + \xi).
\end{aligned}$$

Following the argument in the proof of Theorem 1(b), we have

$$\frac{1}{n}\left(Z_{\hat{\Sigma}}'(I - \hat{H}_{\hat{\Sigma}})Z_{\hat{\Sigma}}\right) \xrightarrow{P} E[e_1(1 - e_1)/w_1(e)] \tag{29}$$

and

$$\frac{1}{\sqrt{n}}Z_{\hat{\Sigma}}'(I - \hat{H}_{\hat{\Sigma}})\hat{\Sigma}^{-1/2}g(e) \xrightarrow{P} 0. \tag{30}$$

Furthermore,

$$\begin{aligned}
&\frac{1}{\sqrt{n}}Z_{\hat{\Sigma}}'(I - H_{\hat{\Sigma}})\hat{\Sigma}^{-1/2}\xi\\
&= \frac{1}{\sqrt{n}}\left[(Z_{\hat{\Sigma}} - \hat{\Sigma}^{-1/2}e) + (\hat{\Sigma}^{-1/2}e - \hat{\Sigma}^{-1/2}\hat{e}) + (\hat{\Sigma}^{-1/2}\hat{e} - \hat{H}_{\hat{\Sigma}}Z_{\hat{\Sigma}})\right]'\hat{\Sigma}^{-1/2}\xi\\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[w_i(e)^{-1/2}(Z_i - e_i)\xi_i - A_\Sigma'I^{-1}(\alpha_0)(Z_i - e_i)X_i\right](1 + O_n(1))\\
&\xrightarrow{d} N\left(0, var[(Z_1 - e_1)\xi_1/w_1(e)] - A_\Sigma'I^{-1}(\alpha_0)A_\Sigma\right).
\end{aligned} \tag{31}$$

where $A_\Sigma = E[e_1(1 - e_1)\xi_1X_1/w_1(e)]$ and $I(\alpha_0)$ is the Fisher information matrix of α at α_0. The asymptotic property (10) follows from (29), (30) and (31). □

References

Akaike, H., 1970. Statistical predictor identification. Ann. Inst. Statist. Math. 22 (1), 203–217.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Automat. Control 19 (6), 716–723.

Athey, S., Tibshirani, J., Wager, S., 2019. Generalized random forests. Ann. Statist. 47 (2), 1148–1178.

Bang, H., Robins, J.M., 2005. Doubly robust estimation in missing data and causal inference models. Biometrics 61, 962–972.

Chen, H., 1988. Covergence rates for parametric components in a partly linear model. Ann. Statist. 16 (1), 136–146.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., Robins, J., 2018. Double/debiased machine learning for treatment and structural parameters. Econometrics 21, 1–68.

Green, P., Yandell, B., 1985. Semi-parametric generalized linear models. In: Proceedings 2nd International GLIM Conference, Vol. 32. pp. 44–45.

Gutman, R., Rubin, D., 2015. Estimation of causal effects of binary treatments in unconfounded studies. Stat. Med. 34 (26), 3381–3398.

Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrica 66 (2), 315–331.

Hastie, T., Tibshirani, R., 1986. Generalized additive models. Statist. Sci. 1 (3), 297–318.

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, second ed. Springer.

Heckman, N.E., 1986. Spline smoothing in a partly linear model. J. R. Stat. Soc. Ser. B Stat. Methodol. 48 (2), 244–248.

Hirano, K., Imbens, G., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrica 71 (4), 1161–1189.

Horvitz, D.G., Thompson, D.J., 1952. A generalization of sampling without replacement from a finite universe. J. Amer. Statist. Assoc. 47 (260), 663–685.

Hsu, J., Small, D., 2013. Calibrating sensitivity analyses to observed covariates in observational studies. Biometric 69 (4), 803–811.

Huber, P., 1973. Robust regression: Asymptotics, conjectures and Monte Carlo. Ann. Statist. 1 (5), 799–821.

Imbens, G.W., Rubin, D.B., 2015. Causal Inference For Statistics Social and Biomedical Science. Cambridge University Press.

Joffe, M.M., Rosenbaum, P.R., 1999. Invited commentary: Propensity scores. Am. J. Epidemiol. 150 (4), 327–333.

Kang, J.D., Schafer, J.L., 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statist. Sci. 22 (4), 523–539.

van der Laan, M.J., 2010. Targeted maximum likelihood based causal inference. Int. J. Biostat. 6, 2.

Lee, M., 2018. Simple least squares estimator for treatment effects using propensity score residuals. Biometrika 105 (1), 149–164.

Lee, B.K., Lessler, J., Stuart, E.A., 2011. Weight trimming and propensity score weighting. Plos One 6 (3).

Li, F., Morgan, K.L., Zaslavsky, A.M., 2018. Balancing covariates via propensity score weighting. J. Amer. Statist. Assoc. 113 (521), 390–400.

Lin, H., Fu, B., Qin, G., Zhu, Z., 2017. Doubly robust estimation of generalized partial linear models for longitudinal data with dropouts. Biometrics 52 (1), 84–98.

Robins, J., Mark, S., Newey, W., 1992. Estimating exposure effects by modelling the expectation of exposure conditional onconfounders. Biometrics 48 (2), 479–495.

Robins, J., Rotnitzky, A., Zhao, L., 1994. Estimation of regression coefficients when some regressors are not always observed. J. Amer. Statist. Assoc. 89 (427), 846–866.

Rosenbaum, 2002. Observational Studies. Springer.

Rosenbaum, P., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal. Biometric 70 (1), 41–55.

Rubin, D.B., 1980. Discussion of randomization analysis of experimental data in the fisher randomization test by basu. J. Amer. Statist. Assoc. 75 (371), 591–593.

Rubin, D.B., 2001. Using propensity scores to help design observational studies: application to the tobacco litigation. Health Serv. Outcomes Res. Methodol. 2, 169–188.

Schumaker, L.L., 1981. Spline Functions: Basic Theory. Cambridge University Press.

Schwarz, G., 2005. Estimating the dimension of a model. Ann. Statist. 6 (2), 15–18.

Tan, Z., 2006. A distributional approach for causal inference using propensity scores. J. Amer. Statist. Assoc. 101 (476), 1619–1637.

Tan, Z., 2007. Comment: Understanding OR, PS and DR. Statist. Sci. 22 (4), 560–568.

Tan, Z., 2010. Bounded, efficient, and doubly robust estimation with inverse weighting. Biometric 92 (2), 1–22.

Tsiatis, A., Davidian, M., 2007. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statist. Sci. 22 (4), 569–573.

Wager, S., Athey, S., 2018. Estimation and inference of heterogeneous treatment effects using random forests. J. Amer. Statist. Assoc. 113 (523), 1228–1242.

Wu, C.-F., 1981. Asymptotic theory of nonlinear least squares estimation. Ann. Statist. 9 (3), 501–513.