Journal of Econometrics 244 (2024) 105829

Contents lists available at ScienceDirect

Journal of Econometrics

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

Tuning-parameter-free propensity score matching approach for causal inference under shape restriction

Yukun Liu [a,*], Jing Qin [b]

[a] KLATASDS - MOE, School of Statistics, East China Normal University, Shanghai 200062, China
[b] National Institute of Allergy and Infectious Diseases, National Institutes of Health, MD 20892, USA

ARTICLE INFO

JEL classification:
C13
C14
C18

Keywords:
Average treatment effect on the treated
Pool adjacent violated algorithm
Propensity score matching estimators
Shape-restricted inference
Semiparametric efficiency

ABSTRACT

Propensity score matching (PSM) is a pseudo-experimental method that uses statistical techniques to construct an artificial control group by matching each treated unit with one or more untreated units of similar characteristics. To date, the problem of determining the optimal number of matches per unit, which plays an important role in PSM, has not been adequately addressed. We propose a tuning-parameter-free PSM approach to causal inference based on the nonparametric maximum-likelihood estimation of the propensity score under the monotonicity constraint. The estimated propensity score is piecewise constant, and therefore automatically groups data. Hence, our proposal is free of tuning parameters. The proposed causal effect estimator is asymptotically semiparametric efficient when the covariate is univariate or the outcome and the propensity score depend on the covariate through the same index $\mathbf{X}^\top\boldsymbol{\beta}$. We conclude that matching methods based on the propensity score alone cannot, in general, be efficient.

1. Introduction

To assess the treatment effect in medical studies, randomized and controlled clinical trials are the gold standard because the baseline covariates are balanced in the treatment and control arms by the randomization. To evaluate the effectiveness of an economic program or policy in econometrics or political science, however, randomization is difficult or impossible to implement for various reasons. In observational studies, the available covariate information from people who participated in the program or not may be unbalanced. The simple two-sample t-test is likely to produce biased results. It is desirable to replicate a randomized experiment as closely as possible by obtaining treated and control groups with similar covariate distributions. Due to the simplicity and intuitiveness of adjusting the distribution of covariates among samples from different populations, matching methods are widely used in applied statistics, econometrics, and epidemiology. Many examples can be found in a comprehensive review paper by Stuart (2010).

A common feature of matching methods is the outcome-independence of matching, i.e., outcome values are not used in the matching process even if they are available at the time of matching. Commonly used matching methods are all based on covariate values. They use the "distance" between treated and untreated individuals, which is a measure of the similarity between two individuals. A well-known example is the Mahalanobis distance. Another popular matching method is propensity score matching, which was proposed by Rosenbaum and Rubin (1983). The propensity score is the conditional probability of assignment to a treatment given a vector of covariates. Suppose that adjusting for a set of covariates is sufficient to eliminate confounding. A key observation made by Rosenbaum and Rubin (1983) is that adjusting for the propensity score is also sufficient to eliminate

* Corresponding author.
E-mail address: ykliu@sfs.ecnu.edu.cn (Y. Liu).

https://doi.org/10.1016/j.jeconom.2024.105829
Received 21 February 2023; Received in revised form 16 January 2024; Accepted 29 July 2024
Available online 7 August 2024

confounding. In contrast to covariate-based distance matching methods, propensity score matching (PSM) has the advantage of reducing the dimensionality of matching to a single dimension, making the matching process much easier.

Although various matching methods have been used, the theoretical results have only recently been studied by Abadie and Imbens (2006, 2012, 2016). This series of papers showed that matching estimators based on the covariate distance involve a biased term that is only negligible under certain regularity conditions. Moreover, the matching estimators are not necessarily root-n-consistent, and some strong conditions are needed to guarantee the root-n-consistency. Furthermore, they demonstrated that, even in settings where matching estimators are root-n-consistent, simple matching estimators with a fixed number of matches do not attain the semiparametric efficiency bound. Furthermore, they found that matching estimators based on the estimated propensity score have smaller variances than those based on the true PSM. To make the matching methods accessible to practitioners, Imbens (2015) used three examples to demonstrate practical implementations from the theoretical literature, and provided detailed recommendations on how the procedures should be performed. In her review paper in Statistical Science, Stuart (2010) lists some of the major software packages that implement matching procedures. A regularly updated version is available at https://www.elizabethstuart.org/psoftware/.

In the PSM approach, there are a number of matching methods that can be employed. The most commonly used include: (1) Nearest neighbor matching (Rubin, 1973) matches for a given treated subject with K untreated subjects ($K \geq 1$), whose propensity scores are closest to that of the treated subject. (2) Caliper matching (Rosenbaum, 1985) establishes a caliper, and matches for each treated subjects all untreated subject whose propensity scores are within the given caliper of the treated subject. (3) Kernel matching (Heckman et al., 1998) compares the outcome of each treated subject with a kernel-based weighted average of the outcomes of all untreated subjects, where the weights are based on the distance of the propensity score of the untreated subjects to that of the treated subject's. The performance of the PSM approach is always influenced by the choice of the involved tuning parameter, which is generally artificial. A too small tuning parameter may lead to inflated variances, whereas a too large tuning parameter may lead to biased results (Cochran, 1968). A natural question is "Are there any optimal methods for choosing the involved tuning parameter?" To the best of our knowledge, no theoretical research has yet been conducted to address this issue.

In this paper, we present a tuning-parameter-free propensity score matching method under a monotone single-index model for the propensity score. Our model assumption on propensity score is a semi-parametric extension of the commonly-used parametric models including the popular linear logistic and probit models; it not only sufficiently alleviate the risk of model mis-specification but also circumvents the curse of dimensionality problem in the multivariate covariate case. In contrast, the method of Hirano et al. (2003) based on a nonparametric propensity score estimate is difficult to implement due to the curse of dimensionality problem. Our first contribution in this paper is to establish the semiparametric efficiency lower bounds (SELBs) for parameters of interest under a single-index model on the propensity score (See Theorem 1), which are new in the literature. We find that the SLEBs are the same whether the link function in the single index model is completely nonparametric or monotone.

Our second contribution is to develop an estimation method based on the semiparametric maximum-likelihood estimation of the propensity score function. This method can be efficiently implemented by the well-known pool adjacent violated algorithm (Ayer et al., 1955, PAVA). As the semiparametric maximum-likelihood estimator (MLE) of the propensity score is a piecewise step function, individuals in the treatment arm can be exactly matched by individuals in the control arm based on their estimated propensity scores. This matching method is purely data-driven and involves no artificial interference. We also find a surprising result: the proposed tuning-free matching estimator is numerically equivalent to an inverse probability weighting (IPW) estimator, which is not the case in general. Theoretically the proposed estimator is not only asymptotically unbiased, but also achieves the SELB if (1) the covariate is univariate, or (2) the covariate is multivariate and the outcome and propensity score depend on the covariate through the same index $\mathbf{X}^\top\boldsymbol{\beta}$. This finding discloses that semiparametric efficient estimation of causal effects depends on the coincidence of the directions through which the outcome and the propensity score depend on the covariate, or equivalently the so-called index bias is zero (Lee, 2018). The latter implies the zero covariance condition in Abadie and Imbens (2016). In the meanwhile, the matching estimator using the true propensity score cannot achieve the SELB, and therefore is less efficient than our estimator, which uses the PAVA estimator of the propensity score. Otherwise, the proposed estimator remains consistent, but is not efficient. In this situation, other PSM methods are relatively inefficient. Our theoretical results depend critically on shape-restricted inference and empirical process theory. Our numerical simulation results show that the proposed method outperforms existing commonly used PSM methods in terms of mean square error even when the parametric form of the propensity score is known.

The rest of the paper is organized as follows. Section 2 introduces our model assumptions, establishes two SELBs, presents the proposed PSM estimation method and investigates its large-sample properties. Simulation results are provided in Section 3. Section 4 applies our methods to a real econometric dataset. Section 5 contains concluding remarks. All technical proofs are given in the supplementary material for clarity.

2. Efficient estimation under shape constraints

This section establishes the SELBs for parameters of interest, and presents our PSM estimation procedure after introducing the basic setup, namely the potential outcome framework.

2.1. The potential outcome framework

We adopt the potential outcome framework (Neyman, 1923–1990; Rubin, 1974) with a binary treatment. Let $Y(1)$ and $Y(0)$ be the univariate potential outcomes of a treatment and a control, respectively, which cannot be observed simultaneously. Let $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^p$

be the baseline covariate, and D be the treatment indicator with $D = 1$ denoting a treatment and $D = 0$ denoting a control. The outcome is $Y(1)$ if $D = 1$ and $Y(0)$ otherwise, which can be written as $Y = Y(D) = DY(1)+(1-D)Y(0)$. Let $(Y_i, \mathbf{X}_i, D_i), i = 1, 2, \ldots, n$, be n independent and identically distributed (i.i.d.) observations from $(Y, \mathbf{X}, D)$. We focus on the estimations of the average treatment effect on the treated (ATT; Wang and Han, 2024) $\tau = \mathbb{E}\{Y(1) - Y(0)|D = 1\}$ and the treatment response mean $\mu_1 = \mathbb{E}\{Y(1)\}$; the average treatment effect $\mathbb{E}\{Y(1)-Y(0)\}$ can be estimated similarly. For the identifiability of treatment effects, we make the commonly used unconfoundedness assumption (Rubin, 1978; Rosenbaum and Rubin, 1983), i.e., conditional on the observed covariates, the treatment indicator is independent of the potential outcomes.

Assumption 1 (*Unconfounded Treatment Assignment*). $D \perp (Y(0), Y(1))|\mathbf{X}$.

Denote the propensity score as $e(\mathbf{x}) = \mathrm{pr}(D = 1|\mathbf{X} = \mathbf{x})$. Under Assumption 1, Rosenbaum and Rubin (1983) showed that D and $(Y(0), Y(1))$ are conditionally independent when $\mathbf{X}$ is replaced by $e(\mathbf{X})$ in the condition, namely $D \perp (Y(0), Y(1))|e(\mathbf{X})$.

The most popular parametric models for propensity score are the linear logistic and probit models, although they always suffer from model mis-specification, which may lead to inconsistent or misleading treatment effect estimators. A common feature of them is that they are monotonic increasing functions of a linear combination of covariates. This motivates us to consider a semiparametric single-index propensity score model with a nonparametric and monotone nondecreasing link function, namely

$$e(\mathbf{x}) = \mathrm{pr}(D = 1|\mathbf{X} = \mathbf{x}) = \pi(\mathbf{x}^\top\boldsymbol{\beta}), \tag{1}$$

where π is a monotone nondecreasing function and $\boldsymbol{\beta}$ is an unknown true p-variate parameter. We assume that $\|\boldsymbol{\beta}\| = 1$ for identifiability with its first component being positive. This model reduces to a completely nonparametric and monotone nondecreasing function for the propensity score if the covariate is univariate. Commonly-used parametric propensity score models such as the linear logistic, probit and complementary log–log models and, more generally, linear latent variable models all satisfy model (1). Thus, model (1) is a semiparametric extension of the commonly-used parametric probability models. It circumvents the choice of different link functions and hence alleviates of the risk of model misspecification. Moreover, our estimation method applies to more general propensity scores $e(\mathbf{x}) = \pi(h^\top(\mathbf{x})\boldsymbol{\beta})$, where $h(\mathbf{x})$ is a known function of $\mathbf{x}$. To our surprise, the nonparametric monotonicity of $\pi(\cdot)$ in model (1) together with the maximum likelihood estimation method gives natural partitions of propensity scores and data, and hence leads to a natural PSM estimation method with exact matching of propensity scores.

2.2. Efficiency bounds

In this subsection, we establish the SELBs for μ_1 and τ. Assumption 2, which requires the covariate to be nondegenerate and the outcome and covariate variables to have finite variances, is trivial.

Assumption 2. The variance matrix $\mathbb{V}\mathrm{ar}(\mathbf{X})$ is positive positive-definite. For $k = 0, 1$, the functions $\mu_k(\mathbf{X}) = \mathbb{E}\{Y(k)|\mathbf{X}\}$ and $\sigma_k^2(X) = \mathbb{V}\mathrm{ar}(Y(k)|\mathbf{X})$ are all well-defined, and the quantities $\mathbb{E}\{Y^2(k)\}$ and $\mathbb{E}\{\mu_k^2(\mathbf{X})\}$ are all finite.

Under a partially linear regression model, Tripathi (2000) disclosed that monotonicity of the nonparametric function does not improve the semiparametric efficiency lower bounds of the coefficient in the linear part. We show that this is also the case in the estimation of μ_1 and τ. See also Severini and Tripathi (2013). Throughout the paper, we use $\boldsymbol{\beta}$ to denote the true value of $\boldsymbol{\beta}$, and denote $\eta = \mathbb{E}(D)$, $\Delta(\mathbf{X}) = \mu_1(\mathbf{X}) - \mu_0(\mathbf{X})$, $Z = \mathbf{X}^\top\boldsymbol{\beta}$ and $\tilde{\mathbf{X}} = \mathbf{X} - \mathbb{E}(\mathbf{X}|Z)$.

Theorem 1. *Suppose that Assumptions 1–2 are valid and that the propensity score $e(\mathbf{X})$ satisfies model (1) with $\pi(\cdot)$ monotone or not. The SELBs for μ_1 and τ are,*

$$\sigma^2_{\mu_1,\mathrm{eff,sim}} = \mathbb{E}\left[\frac{\sigma_1^2(\mathbf{X})}{\pi(Z)} + \mathbb{V}\mathrm{ar}\{\mu_1(\mathbf{X})\}\right], \tag{2}$$

$$\begin{aligned}\sigma^2_{\tau,\mathrm{eff,sim}} = \frac{1}{\eta^2}\mathbb{E}\Big[\ & \pi(Z)(1-\pi(Z))\{\mathbb{E}(\Delta(\mathbf{X})|Z) - \tau\}^2 \\ & + \pi(Z)\{1-\pi(Z)\}\mathbb{E}\{\Delta(\mathbf{X})\tilde{\mathbf{X}}|Z\}^\top\{\mathbb{E}(\tilde{\mathbf{X}}\tilde{\mathbf{X}}^\top|Z)\}^{-1}\mathbb{E}\{\Delta(\mathbf{X})\tilde{\mathbf{X}}|Z\} \\ & + \pi^2(Z)\{\Delta(\mathbf{X}) - \tau\}^2 + \pi(Z)\sigma_1^2(\mathbf{X}) + \frac{\pi^2(Z)}{1-\pi(Z)}\sigma_0^2(\mathbf{X})\Big],\end{aligned} \tag{3}$$

respectively, provided the involved expectations are well defined.

If there is no model assumption on the propensity score, Hahn (1998) showed that the SELBs for μ_1 and τ are $\sigma^2_{\mu_1,\mathrm{eff,np}} = \mathbb{E}\left[\sigma_1^2(\mathbf{X})/e(\mathbf{X}) + \mathbb{V}\mathrm{ar}\{\mu_1(\mathbf{X})\}\right]$ and

$$\sigma^2_{\tau,\mathrm{eff,np}} = \frac{1}{\eta^2}\mathbb{E}\Big[e(\mathbf{X})\{\Delta(\mathbf{X}) - \tau\}^2 + e(\mathbf{X})\sigma_1^2(\mathbf{x}) + \frac{e^2(\mathbf{X})}{1-e(\mathbf{X})}\sigma_0^2(\mathbf{X})\Big],$$

respectively. Under model (1), $e(\mathbf{X}) = \pi(Z)$, therefore $\sigma^2_{\mu_1,\mathrm{eff,sim}} = \sigma^2_{\mu_1,\mathrm{eff,np}}$ but $\sigma^2_{\tau,\mathrm{eff}} \neq \sigma^2_{\tau,\mathrm{eff,np}}$. These results coincides with Hahn (1998)'s finding that the propensity score is ancillary for estimation of the average treatment effect (including μ_1) but is not ancillary for estimation of τ. Note that $\sigma^2_{\tau,\mathrm{eff,np}} - \sigma^2_{\tau,\mathrm{eff,sim}} = \eta^{-2}\mathbb{E}[\pi(Z)\{1-\pi(Z)\}g(Z)]$ with $g(Z) = \inf_{\boldsymbol{\gamma}}\mathbb{E}[\{\Delta(\mathbf{X}) - \mathbb{E}(\Delta(\mathbf{X})|Z) - \tilde{\mathbf{X}}^\top\boldsymbol{\gamma}\}^2|Z]$. Because

$g(Z) \geq 0$, the difference $\sigma^2_{\tau,\text{eff,np}} - \sigma^2_{\tau,\text{eff,sim}}$ is nonnegative and indicates the marginal value (in the estimation of τ) of the knowledge of the single-index structure in the propensity score.

If $e(\mathbf{X})$ is completely known, Hahn (1998) showed the SELB for τ is

$$\sigma^2_{\tau,\text{eff,kn}} = \frac{1}{\eta^2}\mathbb{E}\Big[e^2(\mathbf{X})\{\Delta(\mathbf{X}) - \tau\}^2 + e(\mathbf{X})\sigma_1^2(\mathbf{X}) + \frac{e(\mathbf{X})}{1-e(\mathbf{X})}\sigma_0^2(\mathbf{X})\Big].$$

Under model (1), $\sigma^2_{\tau,\text{eff,kn}} \leq \sigma^2_{\tau,\text{eff,sim}} \leq \sigma^2_{\tau,\text{eff,np}}$ as model (1) lies between a completely known $e(\mathbf{X})$ and a completely nonparametric model. When $e(\mathbf{X})$ satisfies a general parametric model, Chen et al. (2008) establishes the SELBs for parameters defined through general moment restrictions with missing data. Their results are applicable to μ_1, but not directly applicable to τ. We have derived the SELB, say $\sigma^2_{\tau,\text{eff,para}}$, for τ under a parametric propensity score model and shown that $\sigma^2_{\tau,\text{eff,para}} \leq \sigma^2_{\tau,\text{eff,sim}}$; see Theorem 2.1 and the followed remarks in the supplementary material. The difference $\sigma^2_{\tau,\text{eff,sim}} - \sigma^2_{\tau,\text{eff,para}}$ indicates the value of the knowledge of the link function in the (monotone) single-index model for the propensity score. In summary, the aforementioned SELBs for τ satisfy $\sigma^2_{\tau,\text{eff,kn}} \leq \sigma^2_{\tau,\text{eff,para}} \leq \sigma^2_{\tau,\text{eff,sim}} \leq \sigma^2_{\tau,\text{eff,np}}$, namely SELB increases as model assumption becomes weaker and weaker.

2.3. Proposed estimation procedure

For the time being, we suppose that a consistent estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ is available. We shall consider the estimation for $\boldsymbol{\beta}$ in Section 2.4. Let $\hat{Z}_i = \mathbf{X}_i^\top \hat{\boldsymbol{\beta}}$ for $i = 1, 2, \ldots, n$. Without loss of generality, we assume that $\hat{Z}_1 \leq \hat{Z}_2 \leq \cdots \leq \hat{Z}_n$. Based on $\{(\hat{Z}_i, D_i) : 1 \leq i \leq n\}$, the log-likelihood of π is

$$\sum_{i=1}^n \Big[D_i \log \pi(\hat{Z}_i) + (1-D_i)\log\{1-\pi(\hat{Z}_i)\}\Big], \text{ s.t. } \pi(\hat{Z}_1) \leq \cdots \leq \pi(\hat{Z}_n).$$

By Theorem 2.12 of Barlow et al. (1972), maximizing this likelihood with respect to π is equivalent to minimizing $\sum_{i=1}^n \{D_i - \pi(\hat{Z}_i)\}^2$ under the same monotonicity constraint, which can be efficiently solved by the well-known PAVA algorithm (Ayer et al., 1955). The solution or the MLE $\hat{\pi}$ always exists uniquely, and it is the left derivative of the greatest convex minorant of the cumulative sum diagram (Barlow et al., 1972, Theorem 1.1). Specifically, $\hat{\pi}(\cdot)$ is a step function determined by $\pi(\hat{Z}_1), \pi(\hat{Z}_2), \ldots \pi(\hat{Z}_n)$, which always satisfies the monotonicity restriction $\pi(\hat{Z}_1) \leq \pi(\hat{Z}_2) \cdots \leq \pi(\hat{Z}_n)$. If the steps of $\hat{\pi}(\cdot)$ are known, then the value of $\hat{\pi}(\cdot)$ in each step is simply the least square constant regression estimate based on $(D_i, \hat{Z}_i)$'s in the step.

Write $\hat{\pi}_i = \hat{\pi}(\hat{Z}_i)$. Suppose that there are k distinct values in $\hat{\pi}_i$, $1 \leq i \leq n$, and let $0 = m_0 < m_1 < \cdots < m_k = n$ be the locations of the inflection points of the greatest convex minorant of the cumulative sum diagram. Then,

$$\hat{\pi}_i = \hat{\pi}(\hat{Z}_i) = \hat{\pi}_{m_j}, \quad m_{j-1} < i \leq m_j, \quad j = 1, \ldots, k. \tag{4}$$

According to the lemma on page 34 of Barlow et al. (1972),

$$\hat{\pi}_{m_j} = \frac{\sum_{l=m_{j-1}+1}^{m_j} D_l}{m_j - m_{j-1}}, \quad 1 \leq j \leq k. \tag{5}$$

We propose to estimate $\mu_1 = \mathbb{E}\{Y(1)\}$ by

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^n \frac{D_i Y_i}{\hat{\pi}(\hat{Z}_i)} = \frac{1}{n}\sum_{j=1}^k \frac{1}{\hat{\pi}_{m_j}} \sum_{l=m_{j-1}+1}^{m_j} D_l Y_l = \sum_{j=1}^k \rho_j \hat{\mu}_{1j}, \tag{6}$$

where $\rho_j = (m_j - m_{j-1})/n$ is the proportion of observations $\mathbf{X}_s$'s satisfying $\hat{Z}_{m_{j-1}} = \mathbf{X}_{m_j-1}^\top \hat{\boldsymbol{\beta}} < \mathbf{X}_s^\top \hat{\boldsymbol{\beta}} \leq \mathbf{X}_{m_j}^\top \hat{\boldsymbol{\beta}} = \hat{Z}_{m_j}$, and $\hat{\mu}_{1j} = \sum_{l=m_{j-1}+1}^{m_j} D_l Y_l / \sum_{l=m_{j-1}+1}^{m_j} D_l$ is the group mean. Essentially, $\hat{\mu}_1$ is a weighted average of subgroup means, where the subgroups are formed by the steps of the shape-restricted nonparametric MLE $\hat{\pi}$. Note that this grouping method is *automatically data-driven* and is *free from any tuning parameter*.

Remark 1. It seems that $\hat{\mu}_1$ is not well defined if $\hat{\pi}(\hat{Z}_i) = 0$ for some i. We claim that this does not matter. Note that $\hat{\mu}_1 = (1/n)\sum_{i:D_i=1} Y_i/\hat{\pi}(\hat{Z}_i)$. For each i, according to result (4), there must be $1 \leq j \leq k$ such that $m_{j-1} < i \leq m_j$ and $\hat{\pi}(\hat{Z}_i) = \sum_{r=m_{j-1}+1}^{m_j} D_r/(m_j - m_{j-1})$ (by combining (4) and (5)). Thus $\hat{\pi}(\hat{Z}_i) > 1/(m_j - m_{j-1}) > 0$ for each i with $D_i = 1$, therefore $\hat{\mu}_1$ is well defined.

Let $n_1 = \sum_{i=1}^n D_i$. By PSM, we propose to estimate the ATT τ by

$$\hat{\tau} = \frac{1}{n_1}\sum_{i=1}^n D_i \left\{ Y_i - \frac{\sum_{j=1}^n (1-D_j) Y_j I(\hat{\pi}(\hat{Z}_j) = \hat{\pi}(\hat{Z}_i))}{\sum_{r=1}^n (1-D_r) I(\hat{\pi}(\hat{Z}_r) = \hat{\pi}(\hat{Z}_i))} \right\}. \tag{7}$$

Lemma 1. *The proposed PSM estimator $\hat{\tau}$ can be equivalently expressed as*

$$\hat{\tau} = \frac{1}{n_1}\sum_{j=1}^n \left\{ D_j Y_j - (1-D_j) Y_j \frac{\hat{\pi}_j}{1-\hat{\pi}_j} \right\}, \tag{8}$$

which is an IPW estimator.

Interestingly, Lemma 1 discloses that the PSM estimator $\hat{\tau}$ for the ATT τ is numerically equivalent to an IPW estimator based on the same shape-restricted propensity score estimator. Usually matching and weighting estimators have different expressions and perform quite differently, but Lemma 1 seems to unify them through estimating the propensity score using the shape-restricted semiparametric MLE. The IPW expression of $\hat{\tau}$ in Lemma 1 also makes it much more convenient to study its asymptotic properties. With the expression (8), by similar reasoning to that in Remark 1, we can show that $\hat{\pi}(\widehat{Z}_i) \le 1 - 1/(m_{j-1} - m_j) < 1$ if $D_i = 0$, therefore $\hat{\tau}$ is well defined.

With the estimated propensity scores $\hat{\pi}_i$, the estimator of τ developed by Hirano et al. (2003) is

$$\tilde{\tau} = \frac{1}{\sum_{j=1}^n \hat{\pi}_j} \sum_{i=1}^n \left\{ D_i Y_i - (1 - D_i) Y_i \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} \right\}.$$

By Theorem 1.7 of Barlow et al. (1972), the shape-restricted MLEs $\hat{\pi}_i$ satisfy $\sum_{i=1}^n (\hat{\pi}_i - D_i) = 0$ or, equivalently, $\sum_{i=1}^n \hat{\pi}_i = \sum_{i=1}^n D_i = n_1$. We find that $\hat{\tau} = \tilde{\tau}$, i.e., the proposed PSM estimator is equal to that of Hirano et al. (2003) in the form for the ATT.

Remark 2. Like $\hat{\mu}_1$, $\hat{\tau}$ may not be well defined if $\hat{\pi}(\widehat{Z}_i) = 1$ for some i. Suppose that i and j satisfy $m_{j-1} < i \le m_j$. We immediately have $\hat{\pi}(\widehat{Z}_i) < 1 - 1/(m_{j-1} - m_j) < 1$ if $D_i = 0$, therefore $\hat{\tau}$ is well defined.

Remark 3. One referee highlighted that our PSM estimator for the ATT can be interpreted as a regression estimator, employing the PAVA estimator $\hat{\pi}(x)$ as a generated regressor. Also it is equivalent to the kernel regression estimator in Lee (2018) when a rectangular/uniform kernel is applied. Since the PAVA estimator $\hat{\pi}(x)$ is a piecewise step function, this kernel regression estimator does not require the tuning parameter (i.e. the bandwidth) for the uniform kernel. Consequently, these three estimators — matching, inverse probability weighting, and regression estimator (Lee, 2018) — are considered equivalent.

Assumption 3. The ranges of $\mathbf{X}$ and $\boldsymbol{\beta}$, $\mathcal{X}$ and $\mathcal{B}$, are compact. Let $t_{\text{low}} = \inf\{\mathbf{X}^\top \boldsymbol{\gamma} : \mathbf{X} \in \mathcal{X}, \boldsymbol{\gamma} \in \mathcal{B}\} - \varepsilon_0$ and $t_{\text{up}} = \sup\{\mathbf{X}^\top \boldsymbol{\gamma} : \mathbf{X} \in \mathcal{X}, \boldsymbol{\gamma} \in \mathcal{B}\} + \varepsilon_0$ for some $\varepsilon_0 > 0$.

Assumption 4. There exists $c_0 \in (0, 1)$ such that $c_0 \le \pi(t) \le 1 - c_0$ and π has a continuous second derivative on $[t_{\text{low}},\ t_{\text{up}}]$, where t_{low} and t_{up} are defined in Assumption 3.

Under Assumption 4, $\pi'(t)$ is also continuous on the closed interval $[t_{\text{low}},\ t_{\text{up}}]$. Therefore, it must be Lipschitz-continuous, i.e., there exists $c_1 > 0$ such that $|\pi'(t) - \pi'(s)| \le c_1 |s - t|$ for any $t_{\text{low}} \le s, t \le t_{\text{up}}$.

Assumption 5. There exists a constant $M > 0$ such that the density function $f_{\mathbf{X}^\top \boldsymbol{\gamma}}(u)$ of $\mathbf{X}^\top \boldsymbol{\gamma}$ satisfies $f_{\mathbf{X}^\top \boldsymbol{\gamma}}(u) \le M$ for all $\mathbf{x} \in \mathcal{X}$ and $\boldsymbol{\gamma} \in \mathcal{B}$.

Define $\mu_1^*(u; \boldsymbol{\gamma}) = \mathbb{E}\{Y(1) | \mathbf{X}^\top \boldsymbol{\gamma} = u\} = \mathbb{E}\{\mu_1(\mathbf{X}) | \mathbf{X}^\top \boldsymbol{\gamma} = u\}$, and $\mu_0^*(u; \boldsymbol{\gamma}) = \mathbb{E}\{Y(0) | \mathbf{X}^\top \boldsymbol{\gamma} = u\} = \mathbb{E}\{\mu_0(\mathbf{X}) | \mathbf{X}^\top \boldsymbol{\gamma} = u\}$.

Assumption 6. The function $\mu_1^*(u; \boldsymbol{\gamma})$ is continuous in both u and $\boldsymbol{\gamma}$.

Because $\mathcal{X}$ and $\mathcal{B}$ are both compact, Assumptions 5 and 6 imply that the function $\mu_1^*(\mathbf{X}^\top \boldsymbol{\gamma}_1; \boldsymbol{\gamma}_2)$ is Lipschitz-continuous with respect to $(\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2)$, i.e., there exists a constant L such that

$$|\mu_1^*(\mathbf{X}^\top \boldsymbol{\gamma}_1; \boldsymbol{\gamma}_2) - \mu_1^*(\mathbf{X}^\top \boldsymbol{\gamma}_3; \boldsymbol{\gamma}_4)| \le L(\|\boldsymbol{\gamma}_1 - \boldsymbol{\gamma}_3\| + \|\boldsymbol{\gamma}_2 - \boldsymbol{\gamma}_4\|),\ \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_4 \in \mathcal{B}.$$

The function μ_0^* has the same property under Assumptions 5 and 7. In general, if $\hat{\boldsymbol{\beta}}$ is $\sqrt{n}$-consistent and asymptotically normal, the proposed estimators for μ_1 and τ both follow asymptotically normal distributions, and both are asymptotically semiparametric efficient under certain additional conditions. Let $\mathbb{P}_n$ denote the empirical measure based on data $\{(Y_i, \mathbf{X}_i, D_i) : 1 \le i \le n\}$.

Theorem 2. *Suppose that model (1) is true, and that Assumptions 1–6 are satisfied. Define* $\mathbf{B}_1 = \mathbb{E}[\{\mu_1(\mathbf{X}) - \mu_1^*(Z; \boldsymbol{\beta})\}\mathbf{X}^\top \pi'(Z)/\pi(Z)]$, *where* $Z = \mathbf{X}^\top \boldsymbol{\beta}$ *is defined previously. If* $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_p(n^{-1/2})$, *then the following results hold as* $n \to \infty$.

(1) A linear approximation for $\hat{\mu}_1$ *is*

$$\hat{\mu}_1 = \mu_1 + \mathbb{P}_n \left[\frac{D - \pi(Z)}{\pi(Z)} \{\mu_1(\mathbf{X}) - \mu_1^*(Z; \boldsymbol{\beta})\} + \frac{D(Y - \mu_1(\mathbf{X}))}{\pi(Z)} + \mu_1(\mathbf{X}) - \mu_1 \right] + \mathbf{B}_1(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + o_p(n^{-1/2}). \tag{9}$$

(2) If $\mu_1(\mathbf{X}) = \tilde{\mu}_1(Z)$ *for some* $\tilde{\mu}_1(\cdot)$, *then* $\mu_1^*(Z; \boldsymbol{\beta}) = \tilde{\mu}_1(Z)$, $\mathbf{B}_1 = \mathbf{0}$ *and*

$$\sqrt{n}(\hat{\mu}_1 - \mu_1) = \sqrt{n}\mathbb{P}_n \left\{ \frac{D(Y - \tilde{\mu}_1(Z))}{\pi(Z)} + \tilde{\mu}_1(Z) - \mu_1 \right\} + o_p(1) \xrightarrow{d} N(0, \sigma_{\mu,\text{m}}^2),$$

where $\xrightarrow{d}$ *means "converge in distribution to" and* $\sigma_{\mu,\text{m}}^2 = \mathbb{V}\text{ar}(\tilde{\mu}_1(Z)) + \mathbb{E}\{\sigma_1^2(\mathbf{X})/\pi(Z)\}$.

By Theorem 2, if $\mu_1(\mathbf{X}) = \tilde{\mu}_1(Z)$ for some function $\tilde{\mu}_1$ and model (1) is true, then $\sigma^2_{\mu,\mathrm{m}} = \sigma^2_{\mu,\mathrm{eff}}$, which implies that in this situation $\widehat{\mu}_1$ achieves the SELB in Theorem 1. Namely, if both the propensity score $e(\mathbf{X})$ and the regression function $\mu_1(\mathbf{X})$ depend on the covariate $\mathbf{X}$ through the same index $\mathbf{X}^\top\boldsymbol{\beta}$, or equivalently there exist functions $\pi(\cdot)$ and $\tilde{\mu}_1(\cdot)$ such that $e(\mathbf{X}) = \pi(\mathbf{X}^\top\boldsymbol{\beta})$ and $\mu_1(\mathbf{X}) = \tilde{\mu}_1(\mathbf{X}^\top\boldsymbol{\beta})$ for the same vector $\boldsymbol{\beta}$, then $\widehat{\mu}_1$ is asymptotically semiparametric efficient for any $n^{1/2}$-consistent estimator $\widehat{\boldsymbol{\beta}}$.

Assumption 7. The function $\mu_0^*(u;\boldsymbol{\gamma})$ is continuous in both u and $\boldsymbol{\gamma}$.

Theorem 3. *Suppose that model (1) is true, and that Assumptions 1–5 and 7 are satisfied. Define* $\mathbf{B}_2 = \mathbb{E}[\{\mu_0(\mathbf{X}) - \mu_0^*(Z;\boldsymbol{\beta})\}\mathbf{X}^\top\pi'(Z)/1-\pi(Z)]$. *If* $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = O_p(n^{-1/2})$, *then the following results hold as* $n \to \infty$.

(1) A linear approximation for $\widehat{\tau}$ *is*

$$\begin{aligned}\widehat{\tau} = \tau + \frac{1}{\eta}\mathbb{P}_n\Big[&\{D(\Delta(\mathbf{X}) - \tau) + D(Y(1) - \mu_1(\mathbf{X}))\} \\ &-(1-D)\frac{\{Y(0) - \mu_0(X)\}\pi(Z)}{1-\pi(Z)} \\ &+\frac{D-\pi(Z)}{1-\pi(Z)}\{\mu_0(\mathbf{X}) - \mu_0^*(Z;\boldsymbol{\beta})\}\Big] - \frac{1}{\eta}\mathbf{B}_2(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\ &+ o_p(n^{-1/2}).\end{aligned} \tag{10}$$

(2) If $\mu_0(\mathbf{X}) = \tilde{\mu}_0(Z)$ *for some function* $\tilde{\mu}_0(\cdot)$, *then* $\mu_0(\mathbf{X}) = \mu_0^*(Z;\boldsymbol{\beta}) = \tilde{\mu}_0(Z)$, $\mathbf{B}_2 = \mathbf{0}$ *and*

$$\begin{aligned}\sqrt{n}(\widehat{\tau} - \tau) = \frac{1}{\eta}\sqrt{n}\mathbb{P}_n\Big[&D(\Delta(\mathbf{X}) - \tau) + D(Y(1) - \mu_1(\mathbf{X})) \\ &-(1-D)\{Y(0) - \mu_0(\mathbf{X})\}\frac{\pi(Z)}{1-\pi(Z)}\Big] + o_p(1) \\ \xrightarrow{d}\ &N(0, \sigma^2_{\tau,\mathrm{m}}),\end{aligned}$$

where

$$\sigma^2_{\tau,\mathrm{m}} = \frac{1}{\eta^2}\mathbb{E}\Big[\pi(Z)(\Delta(\mathbf{X}) - \tau)^2 + \pi(Z)\sigma_1^2(\mathbf{X}) + \sigma_0^2(\mathbf{X})\frac{\{\pi(Z)\}^2}{1-\pi(Z)}\Big].$$

Under model (1), if $\mu_0(\mathbf{X}) = \tilde{\mu}_0(Z)$ for a function $\tilde{\mu}_0(\cdot)$, then $\sigma^2_{\tau,\mathrm{m}}$ is equal to the $\sigma^2_{\tau,\mathrm{eff,np}}$, i.e., $\widehat{\tau}$ achieves the SELB $\sigma^2_{\tau,\mathrm{eff,np}}$, although it is not the genuine SELB under model (1). If it also holds that $\mu_1(\mathbf{X}) = \tilde{\mu}_1(Z)$ for some function $\tilde{\mu}_1(\cdot)$, then $\Delta(\mathbf{X})$ can be written as $\tilde{\Delta}(Z) \equiv \tilde{\mu}_1(Z) - \tilde{\mu}_0(Z)$. Therefore $\mathbb{E}\{\Delta(\mathbf{X})|Z\} = \Delta(\mathbf{X})$ and $\mathbb{E}\{\Delta(\mathbf{X})\tilde{\mathbf{X}}|Z\} = 0$. In this situation, the SELB $\sigma^2_{\tau,\mathrm{eff,sim}}$ for τ derived in Theorem 1 is exactly $\sigma^2_{\tau,\mathrm{m}}$. In other words, if the propensity score and both the regression functions, $\mu_0(\mathbf{X})$ and $\mu_1(\mathbf{X})$, depend on covariate $\mathbf{X}$ through the same index $\mathbf{X}^\top\boldsymbol{\beta}$, then $\widehat{\tau}$ achieves the genuine SELB $\sigma^2_{\tau,\mathrm{eff,sim}}$ for any $n^{1/2}$-consistent estimator $\widehat{\boldsymbol{\beta}}$.

Besides the asymptotic normality and efficiency results, Theorems 2 and 3 also indicate that if the propensity score and regression functions depend on the covariate X in different directions, or the regression functions do not obey single-index models, then neither $\widehat{\mu}_1$ nor $\widehat{\tau}$ is asymptotically semiparametric efficient.

Remark 4. Imai and Ratkovic (2014) introduced a covariate balancing propensity score methodology that models treatment assignment while optimizing the covariate balance. Suppose that $\pi(\mathbf{X}^\top\boldsymbol{\beta})$ is a correctly specified model for the propensity score. Observing the fact that, for any function $\mathbf{h}(\mathbf{X})$,

$$\mathbb{E}\left\{\frac{D}{\pi(Z)}\mathbf{h}(\mathbf{X}) - \frac{1-D}{1-\pi(Z)}\mathbf{h}(\mathbf{X})\right\} = 0,$$

instead of estimating $\boldsymbol{\beta}$ by the maximum-likelihood method, Imai and Ratkovic (2014) proposed to estimate $\boldsymbol{\beta}$ by solving

$$\mathbb{P}_n\left[\frac{D-\pi(Z)}{\pi(Z)\{1-\pi(Z)\}}\mathbf{h}(\mathbf{X})\right] = 0$$

for some function $\mathbf{h}(\mathbf{x})$. For example, $\mathbf{h}(\mathbf{X}) = \mathbf{X}$, $\pi'(Z)\mathbf{X}$, or the vector consisting of all the linear and quadratic terms of $\mathbf{X}$. They argue that if the propensity score model is misspecified, the MLE of the propensity score might not balance the covariates, while their proposed approach can balance the first and second moments between the two arms. If the dimension of $\mathbf{h}(\mathbf{X})$ is greater than that of $\boldsymbol{\beta}$, this is a well-known over-identified estimation problem. They estimate $\boldsymbol{\beta}$ using the generalized method of moments and the empirical likelihood method. In general, a higher dimension of $\mathbf{h}$ produces more efficient estimators, but creates a heavier computational burden. In practical applications, one has to make a trade-off between computational cost and estimation efficiency.

Let $\widehat{\pi}(\cdot)$ be the MLE of $\pi(\cdot)$ under the monotonicity constraint based on observations $\{\mathbf{X}_i^\top\widehat{\boldsymbol{\beta}} : 1 \le i \le n\}$ for any given $\widehat{\boldsymbol{\beta}}$. By the characterization of such a shape-restricted MLE (Barlow et al., 1972), we have

$$\mathbb{P}_n[\{D - \widehat{\pi}(\mathbf{X}^\top\widehat{\boldsymbol{\beta}})\}\mathbf{h}(\widehat{\pi}(\mathbf{X}^\top\widehat{\boldsymbol{\beta}}))] = 0$$

for any function h. Moreover, we can show that

$$\mathbb{P}_n[\{D - \widehat{\pi}(\mathbf{X}^\top\widehat{\boldsymbol{\beta}})\}\mathbf{h}(\mathbf{X}^\top\widehat{\boldsymbol{\beta}})] = o_p(n^{-1/2}),$$

if $\mathbf{h}$ and π are continuously differentiable (see the proof of Lemma 4.4 in the supplementary material). In other words, our proposal can balance any covariate function of the form $\mathbf{h}(\mathbf{X}^\top\widehat{\boldsymbol{\beta}})$ up to $o_p(n^{-1/2})$, higher than root-n. As far as we know, the proposed PSM method is the first in the literature that can approximately balance so many functions.

Remark 5. When the covariate is univariate (Univariate variables are written in non-bold font), namely $p = 1$, the index coefficient β is exactly equal to 1, the index term Z equals X and the propensity score satisfying model (1) becomes a nonparametric monotone increasing function $\pi(X)$. Since $\tilde{X} = X - \mathbb{E}(X|Z) = 0$, the SELBs in (2) and (3) for μ_1 and τ are simplified to be $\sigma^2_{\mu_1,\mathrm{eff,sim}} = \mathbb{E}[\mathbb{V}\mathrm{ar}\{\mu_1(X)\} + \sigma_1^2(X)/\pi(X)]$ and

$$\sigma^2_{\tau,\mathrm{eff,sim}} = \frac{1}{\eta^2}\mathbb{E}\Big[\pi(X)\{\Delta(X)-\tau\}^2 + \pi(X)\sigma_1^2(X) + \frac{\pi^2(X)}{1-\pi(X)}\sigma_0^2(\mathbf{X})\Big].$$

Because $Z = X$ and $B_1 = B_2 = 0$ in this situation, the asymptotic variances of $\widehat{\mu}_1$ and $\widehat{\tau}$ are exactly $\sigma^2_{\mu,\mathrm{m}} = \sigma^2_{\mu_1,\mathrm{eff,sim}}$ and $\sigma^2_{\tau,\mathrm{m}} = \sigma^2_{\tau,\mathrm{eff,sim}}$, respectively. In other words, the proposed estimators $\widehat{\mu}_1$ and $\widehat{\tau}$ both automatically achieves their SELBs when the covariate is univariate.

2.4. Estimation of β

The proposed estimation procedure for μ_1 and τ both requires a $n^{1/2}$-consistent estimator of $\boldsymbol{\beta}$ in model (1). There are many well-developed methods that can produce $n^{1/2}$-consistent estimator for the index coefficient under a general monotone single-index model. Well-known examples include the maximum rank correlation estimator of Han (1987), the monotone rank estimators of Cavanagh and Sherman (1998), the partial rank estimator (Khan and Tamer, 2007), and the simple score estimator (SSE) of Balabdaoui et al. (2019), etc. We choose to estimate $\boldsymbol{\beta}$ by the SSE $\widehat{\boldsymbol{\beta}}$ as its calculation is relatively simple and convenient.

We briefly review the SSE of Balabdaoui et al. (2019). Given $\boldsymbol{\gamma}$, let $Z_i(\boldsymbol{\gamma}) = \mathbf{X}_i^\top\boldsymbol{\gamma}$ and assume that $Z_1(\boldsymbol{\gamma}) \le Z_2(\boldsymbol{\gamma}) \le \cdots \le Z_n(\boldsymbol{\gamma})$. Let $\widehat{\pi}_{\boldsymbol{\gamma}}$ denote the PAVA estimator of $\pi(\cdot)$ that minimizes $\sum_{i=1}^n\{D_i - \pi(Z_i(\boldsymbol{\gamma}))\}^2$. Define a $d-1$-dimensional sphere as $S_{d-1} = \{\boldsymbol{\gamma} : \boldsymbol{\gamma} \in \mathbb{R}^d, \|\boldsymbol{\gamma}\| = 1\}$, a one-to-one map $\mathbb{S} : [0,\pi]^{(d-2)} \times [0,2\pi) \mapsto S_{d-1}$ as

$$\begin{aligned}\boldsymbol{\zeta} \equiv (\zeta_{(1)},\zeta_{(2)},\ldots,\zeta_{(d-1)}) \mapsto (&\cos(\zeta_{(1)}),\ \sin(\zeta_{(1)})\cos(\zeta_{(2)}),\ldots,\\ &\sin(\zeta_{(1)})\ldots\sin(\zeta_{(d-2)})\cos(\zeta_{(d-1)}),\sin(\zeta_{(1)})\ldots\sin(\zeta_{(d-2)})\sin(\zeta_{(d-1)})),\end{aligned} \tag{11}$$

and a $d\times(d-1)$ matrix as $\mathbf{J}(\boldsymbol{\zeta}) = \partial\mathbb{S}^\top(\boldsymbol{\zeta})/\partial\boldsymbol{\zeta}$. Let $\boldsymbol{\zeta}_0$ satisfy $\boldsymbol{\beta} = \mathbb{S}(\boldsymbol{\zeta}_0)$ and $\widehat{\boldsymbol{\zeta}}$ be a *zero-crossing* of the function

$$\boldsymbol{\phi}_n(\boldsymbol{\zeta}) = \mathbb{P}_n[\mathbf{J}^\top(\boldsymbol{\zeta})\mathbf{X}\{D - \widehat{\pi}_{\mathbb{S}(\boldsymbol{\zeta})}(\mathbf{X}^\top\mathbb{S}(\boldsymbol{\zeta}))\}] \tag{12}$$

(see page 521 of Balabdaoui et al. (2019) for the definition of *zero-crossing*). Accordingly, we estimate $\boldsymbol{\beta}$ by $\widehat{\boldsymbol{\beta}} = \mathbb{S}(\widehat{\boldsymbol{\zeta}})$, and estimate the propensity score function by $\widehat{\pi}_{\widehat{\boldsymbol{\beta}}}(\cdot\widehat{\boldsymbol{\beta}})$. The resulting PSM estimators for μ_1 and τ are

$$\widehat{\mu}_1 = \frac{1}{n}\sum_{i=1}^n \frac{D_iY_i}{\widehat{\pi}_{\widehat{\boldsymbol{\beta}}}(\mathbf{X}_i^\top\widehat{\boldsymbol{\beta}})} \quad\text{and}\quad \widehat{\tau} = \frac{1}{n_1}\sum_{j=1}^n\left\{D_jY_j - (1-D_j)Y_j\frac{\widehat{\pi}_{\widehat{\boldsymbol{\beta}}}(\mathbf{X}_j^\top\widehat{\boldsymbol{\beta}})}{1-\widehat{\pi}_{\widehat{\boldsymbol{\beta}}}(\mathbf{X}_j^\top\widehat{\boldsymbol{\beta}})}\right\},$$

respectively. To study the large-sample properties of the two estimators, we assume the following conditions, which correspond to Assumptions A3, A5, A7, and A9, respectively, of Balabdaoui et al. (2019).

Assumption 8. There exists $\delta_0 > 0$ such that the function $\pi_{\boldsymbol{\gamma}}(u) = \mathbb{E}\{\pi(Z)|\mathbf{X}^\top\boldsymbol{\gamma} = u\}$ is monotone increasing on $I_{\boldsymbol{\gamma}} = \{\mathbf{X}^\top\boldsymbol{\gamma} : \mathbf{X}\in\mathcal{X}\}$ for all $\boldsymbol{\gamma}\in\mathcal{B}(\boldsymbol{\beta},\delta_0) = \{\boldsymbol{\gamma} : \|\boldsymbol{\gamma}-\boldsymbol{\beta}\| \le \delta_0\}$.

Assumption 9. The distribution of $\mathbf{X}$ admits a density $f_{\mathbf{X}}(\mathbf{x})$ that is differentiable on $\mathcal{X}$. In addition, there exist positive constants $c_1,c_2,c_3,c_4 > 0$ such that $c_1 \le f_{\mathbf{X}}(\mathbf{x}) \le c_2$ and $c_3 \le \partial f_{\mathbf{X}}(\mathbf{x})/\partial x_j \le c_4$ on $\mathcal{X}$ for all $1\le j\le d$.

Assumption 10. For all $\boldsymbol{\zeta}\neq\boldsymbol{\zeta}_0$ such that $\mathbb{S}(\boldsymbol{\zeta})\in\mathcal{B}(\boldsymbol{\beta},\delta_0)$, the random variable $\mathbb{C}\mathrm{ov}[(\boldsymbol{\zeta}_0-\boldsymbol{\zeta})^\top\mathbf{J}^\top(\boldsymbol{\zeta}_0)\mathbf{X},\ \pi(\mathbf{X}^\top\mathbb{S}(\boldsymbol{\zeta}_0))|(\mathbf{X}^\top\mathbb{S}(\boldsymbol{\zeta}_0))] \neq 0$ almost surely.

Assumption 11. $\mathbf{J}^\top(\boldsymbol{\zeta}_0)\mathbb{E}\{\pi'(Z)\mathbb{V}\mathrm{ar}(\mathbf{X}|Z)\}\mathbf{J}(\boldsymbol{\zeta}_0)$ is nonsingular.

If Assumptions 3 and 8–11 are satisfied, then Theorem 3 of Balabdaoui et al. (2019) implies that the estimator $\widehat{\boldsymbol{\beta}} = \mathbb{S}(\widehat{\boldsymbol{\zeta}})$ is consistent and asymptotically normal (see Lemma 12 in the supplementary material).

Theorem 4. *Suppose that model (1) is true, and that Assumptions 1–11 are satisfied. Define* $\mathbf{B}_3 = \mathbf{J}(\boldsymbol{\zeta}_0)[\mathbf{J}^\top(\boldsymbol{\zeta}_0)\mathbb{E}\{\pi'(Z)\mathbb{V}\mathrm{ar}(\mathbf{X}|Z)\}\mathbf{J}(\boldsymbol{\zeta}_0)]^{-1}\mathbf{J}^\top(\boldsymbol{\zeta}_0)$. *Then, the following results hold as* $n\to\infty$.

(1) $\sqrt{n}(\widehat{\mu}_1-\mu_1) \xrightarrow{d} N(0,\sigma^2_{\mu,\mathrm{sse}})$, *where*

$$\sigma^2_{\mu,\mathrm{sse}} = \mathbb{E}\left[\frac{1-\pi(Z)}{\pi(Z)}\Big\{\mu_1(\mathbf{X}) - \mu_1^*(Z;\boldsymbol{\beta}) + \mathbf{B}_1\mathbf{B}_3(\mathbf{X}-\mathbb{E}(\mathbf{X}|Z))\pi(Z)\Big\}^2\right]$$

$$+\mathbb{E}\left\{\frac{\sigma_1^2(\mathbf{X})}{\pi(Z)}\right\}+\mathbb{V}\mathrm{ar}\{\mu_1(\mathbf{X})\}.$$

(2) $\sqrt{n}(\widehat{\tau}-\tau)\xrightarrow{d}N(0,\sigma^2_{\tau,\mathrm{sse}})$, *where*

$$\begin{aligned}\sigma^2_{\tau,\mathrm{sse}} =\ & \frac{1}{\eta^2}\mathbb{E}\Big[\pi(Z)\sigma_1^2(\mathbf{X})\Big]+\frac{1}{\eta^2}\mathbb{E}\Big[\sigma_0^2(\mathbf{X})\frac{\{\pi(Z)\}^2}{1-\pi(Z)}\Big]+\frac{1}{\eta^2}\mathbb{E}\{\pi(Z)(\Delta(\mathbf{X})-\tau)^2\}\\ &+\frac{1}{\eta^2}\mathbb{E}\Big[\pi(Z)\{1-\pi(Z)\}\Big\{\frac{\mu_0(\mathbf{X})-\mu_0^*(Z;\boldsymbol{\beta})}{1-\pi(Z)}-\mathbf{B}_2\mathbf{B}_3(\mathbf{X}-\mathbb{E}(\mathbf{X}|Z))\Big\}^2\Big]\\ &+\frac{2}{\eta^2}\mathbb{E}\Big[(\Delta(\mathbf{X})-\tau)\pi(Z)\{1-\pi(Z)\}\Big\{\frac{\mu_0(\mathbf{X})-\mu_0^*(Z;\boldsymbol{\beta})}{1-\pi(Z)}\\ &-\mathbf{B}_2\mathbf{B}_3(\mathbf{X}-\mathbb{E}(\mathbf{X}|Z))\Big\}\Big].\end{aligned}$$

Theorem 4 indicates that, if $\boldsymbol{\beta}$ is estimated by Balabdaoui et al. (2019)'s SSE, the proposed estimators $\widehat{\mu}_1$ and $\widehat{\tau}$ are still consistent and asymptotically normal. However, by Theorems 2 and 3, they are not asymptotically semiparametric efficient if $Y(1)$ or $Y(0)$ does not depend on the covariate through the same index as the treatment indicator does.

2.5. Variance estimation

When constructing confidence intervals for τ based on the asymptotic normality results in Theorems 3–4, we need to construct consistent estimators for the asymptotic variances, which involve quantities such as $\sigma_0^2(\mathbf{X})$ or $\sigma_1^2(\mathbf{X})$; we have to resort to nonparametric techniques such as kernel methods, in which tuning parameters have to be chosen. To keep tuning-parameter-free, we propose to construct confidence intervals by the nonparametric bootstrap (Efron, 1979). Specifically, we use the nonparametric bootstrap procedure in Algorithm 1 to estimate the asymptotic variance of $\widehat{\tau}$.

Algorithm 1: *Nonparametric bootstrap variance estimation*

Data: $\mathcal{W}=\{W_i\}_{i=1}^n$ with $W_i=(D_i,\mathbf{X}_i,Y_i)$.
Input: B: number of bootstrap samples; $\mathcal{A}$: estimation method of $\boldsymbol{\beta}$
Output: Variance estimate of the ATT τ

1 **Prepare**
- Calculate an estimate $\widehat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$ by method $\mathcal{A}$ based on $\mathcal{W}$.
- Obtain the MLE $\widehat{\pi}$ by maximizing $\ell(\pi)$ over $\mathcal{G}=\{q(\cdot):0\le q(t)\le 1,\ q(t)\text{ is nondecreasing}\}$.
- Obtain the proposed ATT estimate $\widehat{\tau}$ based on $\mathcal{W}$.

2 **for** $b=1:B$ **do**
- Let $\mathcal{W}_b^*=\{W_{b1}^*,\dots,W_{bn}^*\}$ be a sample from $\mathcal{W}$ by the simple random sampling with replacement.
- Calculate $\widehat{\boldsymbol{\beta}}_b^*$, $\widehat{\pi}_b^*$, and $\widehat{\tau}_b^*$, which are the bootstrap version of $\widehat{\boldsymbol{\beta}}$, $\widehat{\pi}$, and $\widehat{\tau}$, respectively, based on the bootstrap sample $\mathcal{W}_b^*$

Result: An estimator for the asymptotic variance of $\widehat{\tau}$ is the sample variance, say $\widehat{\sigma}^2_{\mathrm{boot}}$, of $\{\widehat{\tau}_1^*,\dots,\widehat{\tau}_B^*\}$

In general, the nonparametric bootstrap is not valid for the nonparametric maximum likelihood estimator (NPMLE) under the monotonicity constraint, including the well-known Grenander estimator (Sen et al., 2010). Groeneboom and Hendrickx (2017) shows that it is possible to obtain the bootstrap validity for smooth functionals of the NPMLE. We show that the nonparametric bootstrap method is asymptotically valid for the proposed ATT estimator when the method $\mathcal{A}$ in Algorithm 1 is chosen to be the SSE of Balabdaoui et al. (2019).

Theorem 5. *Let* $\mathcal{W}=\{(D_i,\mathbf{X}_i,Y_i)\}_{i=1}^n$ *and* $\widehat{\tau}^*$ *be the proposed ATT estimator based on a nonparametric bootstrap sample* $\mathcal{W}^*$ *from* $\mathcal{W}$. *Under the Assumptions of Theorem 4,* $\sup_{t\in\mathbb{R}}|\mathbb{P}\{\sqrt{n}(\widehat{\tau}^*-\widehat{\tau})\le t|\mathcal{W}\}-\mathbb{P}\{\sqrt{n}(\widehat{\tau}-\tau)\le t\}|=o_p(1)$.

3. Simulations

In this section, we conduct simulations to evaluate the finite-sample performance of the proposed estimators. In particular, we compare the following estimation methods: (I) PAVA1: the proposed PSM method with $\boldsymbol{\beta}$ estimated by the MLE under the logistic propensity score model and π estimated by PAVA; (II) PAVA2: the proposed PSM method with $\boldsymbol{\beta}$ estimated by SSE and π estimated by PAVA; (III) PARA: the proposed PSM estimator (7) with $\widehat{\pi}$ replaced by the logistic function and $\widehat{\boldsymbol{\beta}}$ being the MLE under the logistic propensity score model; (IV) PSM or PSMk: the PSM method with the propensity score estimated by the logistic regression model and each case matched with k controls. Four choices of k are considered: 3, 5, 10, and 15. (V) KN or KNc: the IPW method

Table 1
Simulated biases and RMSEs based on 1000 samples of size $n = 500$ when $\pi(\cdot)$ is set to the standard logistic distribution function in the data generating process.

a	b	PAVA1	PAVA2	PARA	PSM3	PSM5	PSM10	PSM15	KN0.5	KN0.6	KN1.0	KN1.5
Model = 1: $h = \cos(X_1 + bX_6)$; RMSE												
1	1	0.477	0.527	1.704	0.657	0.771	0.990	1.141	13.005	0.301	0.605	0.727
1	0	0.395	0.398	0.902	0.492	0.475	0.493	0.520	11.049	0.958	1.350	1.477
1	−1	0.480	0.531	0.710	0.526	0.491	0.440	0.416	1.629	0.729	0.903	0.960
2	1	0.983	0.943	1.828	1.095	1.234	1.482	1.611	24.813	0.694	0.531	0.509
2	0	0.545	0.542	1.199	0.629	0.600	0.592	0.588	3.469	0.552	0.769	0.836
2	−1	0.820	0.818	1.112	1.004	0.946	0.856	0.813	4.024	0.681	0.814	0.864
Model = 2: $h = X_1$; RMSE												
1	1	0.421	0.483	0.953	0.497	0.451	0.415	0.416	2.199	0.278	0.218	0.201
1	0	0.393	0.498	0.671	0.403	0.412	0.456	0.505	4.123	0.523	0.417	0.381
1	−1	0.487	0.638	0.874	0.429	0.463	0.574	0.690	1.419	0.776	0.626	0.573
2	1	0.572	0.559	1.422	0.579	0.554	0.500	0.441	10.780	0.638	0.917	1.032
2	0	0.361	0.528	0.517	0.462	0.460	0.492	0.561	2.907	0.295	0.250	0.321
2	−1	0.649	0.768	0.936	0.850	0.839	0.839	0.909	62.139	0.659	1.075	1.232
Model = 1: $h = \cos(X_1 + bX_6)$; Bias												
1	1	−0.170	−0.296	0.033	−0.495	−0.657	−0.928	−1.098	−0.065	0.164	0.596	0.721
1	0	0.102	0.044	−0.023	−0.223	−0.265	−0.357	−0.416	0.886	0.926	1.346	1.474
1	−1	0.138	0.151	−0.047	−0.015	−0.011	0.009	0.005	0.399	0.705	0.898	0.956
2	1	−0.636	−0.548	−0.239	−0.908	−1.111	−1.413	−1.563	−0.632	−0.657	−0.510	−0.490
2	0	−0.029	0.000	−0.034	−0.301	−0.348	−0.404	−0.433	0.291	0.505	0.757	0.826
2	−1	−0.122	−0.097	−0.002	0.050	0.057	0.046	0.057	−0.303	−0.610	−0.795	−0.851
Model = 2: $h = X_1$; Bias												
1	1	0.078	0.162	−0.034	0.072	0.101	0.162	0.202	0.231	0.240	0.197	0.182
1	0	0.166	0.304	−0.003	0.157	0.208	0.323	0.407	0.505	0.501	0.404	0.369
1	−1	0.287	0.494	0.044	0.256	0.335	0.491	0.628	0.921	0.761	0.619	0.566
2	1	−0.302	−0.064	0.021	−0.290	−0.322	−0.304	−0.245	−0.118	−0.584	−0.903	−1.022
2	0	0.048	0.266	0.012	0.102	0.163	0.285	0.401	0.328	0.092	−0.189	−0.285
2	−1	−0.055	0.164	−0.012	0.248	0.345	0.514	0.646	−2.087	−0.546	−1.058	−1.222

with the propensity scores estimated by the kernel estimate, i.e. $(nh)^{-1}\sum_{j=1}^{n}\phi((\hat{Z}_j - \hat{Z}_i)/h)$. Here $\phi(\cdot)$ is the standard normal density function, $\hat{Z}_i = \mathbf{X}_i^\top\hat{\boldsymbol{\beta}}$ with $\hat{\boldsymbol{\beta}}$ being the SSE estimate of $\boldsymbol{\beta}$, $h = c \times n^{-1/5} \times s$ is the bandwidth and s is the standard deviation of $\hat{Z}_i$'s. Four choices of c are considered: 0.5, 0.6, 1, and 1.5.

To generate data, we consider a 10-dimensional $\mathbf{X} = (X_1, \ldots, X_{10})^\top$, where $X_1, \ldots, X_5$ are i.i.d. as $N(0,1)$, $X_6, \ldots, X_{10}$ are i.i.d. as centralized Binomial$(4, 0.5)$ and all of them are independent from each other. Given $\mathbf{X}$, we generate D from the propensity score model $\mathrm{pr}(D = 1|\mathbf{X}) = \pi(2 + X_1 + X_6)$. The potential outcomes satisfy the following regression models:

$$Y(1) = -(X_1 + X_6)^a + \epsilon, \quad Y(0) = 3h(\mathbf{X}) - (X_1 + bX_6)^a + \epsilon,$$

where $\epsilon \sim N(0,1)$. We choose $\pi(t) = e^t/(1+e^t)$ or the standard normal distribution function, $h(\mathbf{X}) = \cos(X_1 + bX_6)$ (Model 1) or X_1 (Model 2), $a = 1$ or 2, and $b = 1, 0$ or -1. From each case, we generate 1000 samples with a sample size of $n = 500$ and calculate the eleven estimators for the ATT τ.

We first examine the results in Table 1, which presents the Biases and root mean square errors (RMSEs) of the eleven estimators when $\pi(t) = e^t/(1+e^t)$. The overall rate of nonmissing data is about 81.6%. As $\pi(t)$ is the logistic function, the eleven estimators under comparison all have correctly specified propensity score models. The propensity score satisfies model (1) with $\boldsymbol{\beta} = (1,0,0,0,0,1,0,0,0,0)^\top/\sqrt{2}$. In all cases, although having negligible biases, the PARA estimator (which uses the true logistic propensity score function) always has the largest RMSE among the first seven estimators, meaning that it is always the most unreliable. This coincides with the finding of Hirano et al. (2003) that "weighting by the inverse of a nonparametric estimate of the propensity score, rather than the true propensity score, leads to an efficient estimate of the average treatment effect". The kernel-based IPW estimator, KNc, makes use of the single-index structure and does not suffer from the curse of dimensionality problem. However, it is still dramatically influenced by the bandwidth and often has very large Bias. The RMSE of KN0.5 is more than 10 times of that of KN0.6 in many cases although the coefficient c in the bandwidth changes only from 0.5 to 0.6, indicating that the KNc estimator is very sensitive to the choice of bandwidth.

Under Model 1, the regression function in the control group is a single-index model $\mu_0(\mathbf{X}) = \cos(\sqrt{1+b^2}\cdot\mathbf{X}^\top\boldsymbol{\theta}) - (\sqrt{1+b^2}\cdot\mathbf{X}^\top\boldsymbol{\theta})^a$ with $\boldsymbol{\theta} = (1,0,0,0,0,b,0,0,0,0)^\top/\sqrt{1+b^2}$. When $b = 1$, $\boldsymbol{\theta} = \boldsymbol{\beta}$. By Theorem 3, the proposed estimator $\hat{\tau}$ in (7) is asymptotically semiparametric efficient, regardless of whether $\boldsymbol{\beta}$ is estimated by the MLE or SSE. We see from Table 1 that PAVA1 and PAVA2 have almost the same nice performance, and both of them perform better than the PARA, PSMk and KNc estimators in most cases in terms of bias and RMSE. The performance of the PSM estimator is dramatically influenced by the number of matches, M, per unit; the PSM estimator has increasing Biases and RMSEs as M increases from 3 to 15, and PSM15 has twice the Biases and RMSEs as PSM3.

Table 2
Simulated biases and RMSEs based on 1000 samples of size $n = 500$ when $\pi(\cdot)$ is set to the standard normal distribution function in the data generating process.

a	b	PAVA1	PAVA2	PARA	PSM3	PSM5	PSM10	PSM15	KN0.5	KN0.6	KN1.0	KN1.5
Model = 1: $h = \cos(X_1 + bX_2)$; RMSE												
1	1	0.339	0.409	0.821	1.713	1.702	1.614	1.477	0.795	0.769	0.907	0.953
1	0	0.724	0.711	0.873	0.870	0.812	0.800	0.782	1.921	1.434	1.588	1.642
1	−1	0.734	0.779	1.000	1.089	0.848	0.629	0.524	0.976	0.921	0.981	1.002
2	1	0.832	0.751	1.414	1.957	1.801	1.378	0.959	0.405	0.242	0.244	0.259
2	0	0.546	0.602	0.726	0.950	0.855	0.679	0.566	0.977	0.886	0.952	0.972
2	−1	0.990	1.033	1.522	1.906	1.593	1.261	1.069	1.232	0.854	0.902	0.920
Model = 2: $h = X_1$; RMSE												
1	1	0.472	0.570	1.052	0.959	0.830	0.707	0.687	1.517	0.233	0.203	0.192
1	0	0.623	0.702	0.906	1.021	1.007	1.097	1.202	0.599	0.440	0.378	0.354
1	−1	0.856	0.955	1.097	1.263	1.352	1.586	1.785	0.785	0.645	0.552	0.515
2	1	0.562	0.571	0.741	1.010	0.909	1.058	1.364	0.833	0.787	0.954	1.027
2	0	0.466	0.627	0.820	1.357	1.348	1.499	1.709	1.863	0.230	0.319	0.368
2	−1	0.744	0.800	1.798	1.947	1.795	1.835	1.915	5.535	1.149	1.353	1.431
Model = 1: $h = \cos(X_1 + bX_2)$; Bias												
1	1	−0.102	−0.103	−0.477	−1.651	−1.660	−1.583	−1.447	0.614	0.761	0.904	0.950
1	0	0.628	0.583	0.241	−0.636	−0.650	−0.704	−0.712	1.204	1.429	1.586	1.640
1	−1	0.565	0.559	0.412	0.044	0.037	0.021	0.026	0.841	0.916	0.978	0.999
2	1	−0.748	−0.598	−1.035	−1.879	−1.724	−1.263	−0.761	−0.236	−0.197	−0.208	−0.227
2	0	0.360	0.386	0.097	−0.486	−0.478	−0.350	−0.266	0.783	0.875	0.944	0.964
2	−1	−0.519	−0.476	−0.396	−0.025	−0.025	−0.022	−0.036	−0.806	−0.834	−0.891	−0.910
Model = 2: $h = X_1$; Bias												
1	1	0.234	0.269	0.196	0.295	0.365	0.441	0.507	0.267	0.210	0.184	0.174
1	0	0.534	0.595	0.528	0.706	0.809	1.004	1.143	0.465	0.427	0.368	0.344
1	−1	0.815	0.904	0.846	1.110	1.251	1.533	1.750	0.714	0.635	0.545	0.509
2	1	−0.371	−0.199	−0.194	0.112	0.297	0.767	1.184	−0.629	−0.765	−0.944	−1.019
2	0	0.240	0.363	0.386	0.846	0.994	1.317	1.584	−0.040	−0.142	−0.286	−0.343
2	−1	−0.229	−0.138	0.213	1.155	1.272	1.549	1.728	−1.102	−1.128	−1.345	−1.424

When $b \neq 1$, we have $\boldsymbol{\theta} \neq \boldsymbol{\beta}$. The proposed estimator $\hat{\tau}$ loses its semiparametric optimality. Even so, when $b = 0$ (the angle between $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ is 45 degrees), the proposed estimators PAVA2 and PAVA1 still achieve better performances than the other estimators, but the relative advantage is smaller. When $b = -1$, $\boldsymbol{\theta}$ is perpendicular to $\boldsymbol{\beta}$, and the relative advantage of our estimators decreases further until PSMk and KNc becomes comparable or even better. Surprisingly, as M increases from 3 to 15, the RMSE and Bias of the PSM estimator both increase when $b = 1$ and 0, but its RMSE decreases and its Bias keeps negligible when $b = -1$. The KNc estimator has a similar property: any of the four choices of c can lead to smallest RMES or Bias. These findings indicate that the optimal choices of the tuning parameters for the PSMk and KNc methods critically depend on the true regression function $\mu_0(\mathbf{X})$, and without any information on $\mu_0(\mathbf{X})$, it is impossible to correctly specify the optimal M. Additionally, the performance of the proposed estimator may be improved by making use of information on $\mu_0(\mathbf{X})$.

Under model 2, $\mu_0(\mathbf{X})$ does not follow a single-index model, and it cannot be written as $\tilde{\mu}_0(\mathbf{X}^\top\boldsymbol{\theta})$ for some function $\tilde{\mu}_0$. By Theorem 3, the proposed estimator $\hat{\tau}$ is no longer semiparametric efficient, but is still asymptotically normal. The results in Table 1 corresponding to Model 2 suggest that, compared with the PSMk and KNc estimators, the proposed estimators are at least comparable and perform uniformly better in some cases.

Table 2 presents the results when $\pi(t)$ is chosen to be the standard normal distribution function. In this situation, the overall rate of nonmissing data is about 87.5% and only PAVA2 and KNc have correctly specified propensity score models. Compared with the four PSMk estimators, the proposed PAVA1 and PAVA2 estimators have uniformly smaller RMSEs and generally smaller biases. The KNc estimators still have quite unstable performance although they may outperform the proposed estimators in some situations. Again, the two proposed estimators exhibit rather similar performance, although the index coefficient in the PAVA1 method is estimated by the MLE under the misspecified logistic propensity score model. As M increases from 3 to 15, the PSM method may perform worse in some cases and better in other cases. When the true propensity score model changes from a logistic model to a probit model, the RMSEs of the PAVA1 and PAVA2 estimators increase by no more than 30% in eight out of the twelve cases. In contrast, the RMSE of PSM3 increases by at least 50% in all cases, and can be as large as 190% (e.g., the case with Model 2, $a = 1$, and $b = -1$). This suggests that the PSM estimators are more sensitive than the proposed estimator to the misspecification of the propensity score model.

To obtain further insights into the performance of the proposed estimators and the PSM estimators, Fig. 1 displays their boxplots in the case of $a = 1$ when $\pi(t)$ is correctly specified; those of the PARA and KN0.5 estimates are excluded as they spread too widely. As we can see from the boxplots, PAVA1 and PAVA2 have smaller biases and better overall performance in most cases compared with the PSMk and KNc estimators. As the number of matches M increases from 3 to 15, the PSM estimators exhibit decreasing variances, but increasing biases except in the case of Model 1, $a = 1$, and $b = -1$. The results are similar in the mis-specification case where $\pi(t)$ is set to the standard normal distribution function.

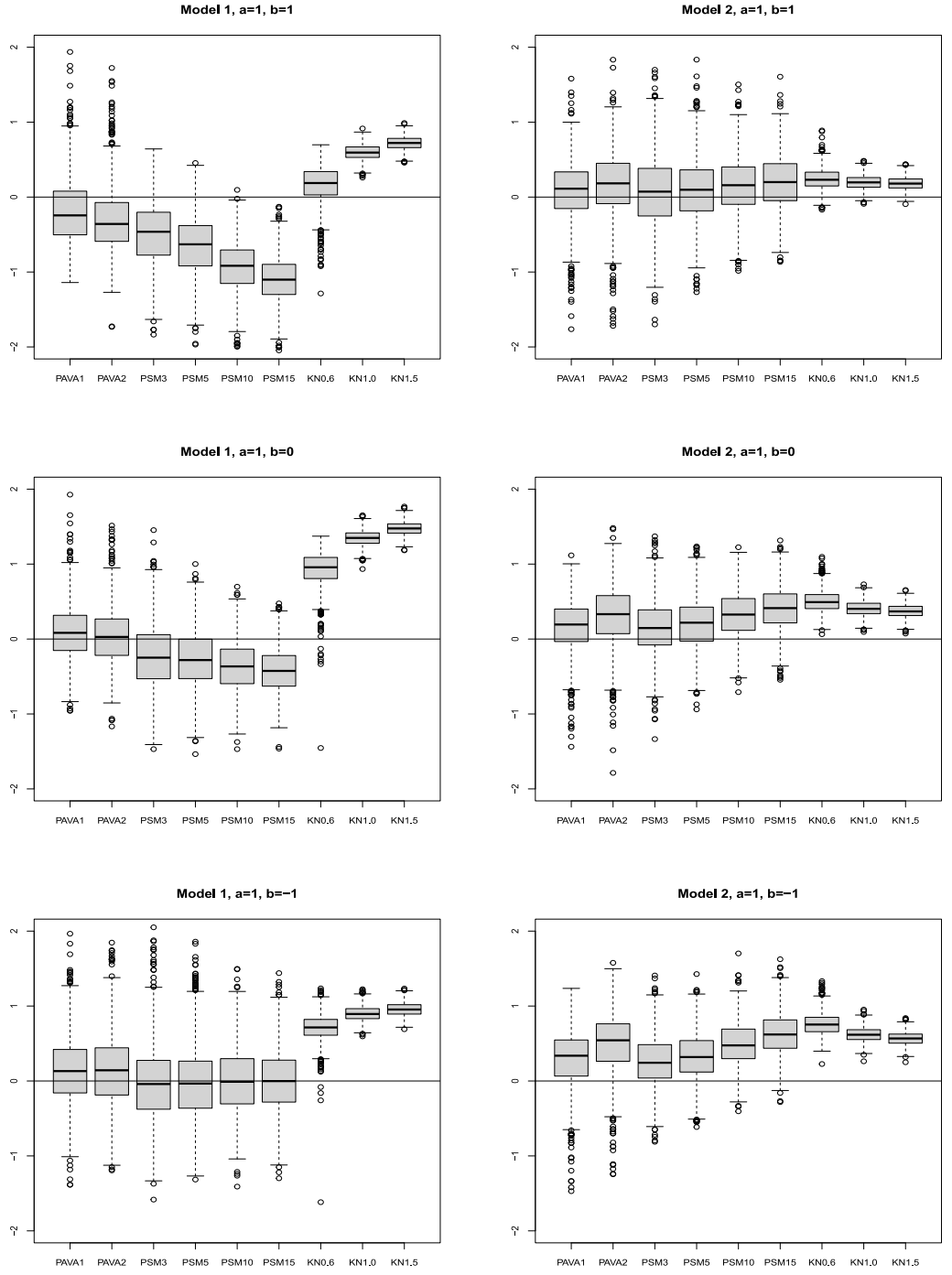

Fig. 1. Boxplots of the seven estimators when $\pi(\cdot)$ is correctly specified, $a = 1$, and $n = 500$. As b varies between 1, 0, and −1, the angle between $\boldsymbol{\beta}$ and $\boldsymbol{\beta}$ increases from 0 degrees to 90 degrees. Model 1: $h(\mathbf{X}) = \cos(X_1 + bX_2)$; Model 2: $h(\mathbf{X}) = X_1$.

Overall, the proposed PAVA-based estimation method is more reliable than the PSM and KN methods: it usually has smaller RMSEs and biases, and is not influenced by the tuning parameters. The performances of the PSM and KN methods are strongly influenced by tuning parameters, namely the number M of matches per unit for PSM and the bandwidth for KN; however, determining the optimal value of M is quite challenging and no research has been done on this issue. The direction in which the control response depends on the covariate does not influence the consistency, but does affect the estimation efficiency of the proposed method: if it coincides with the direction in which the treatment status depends on the covariate, the proposed method achieves optimal performance both numerically and theoretically.

Table 3
Estimation results for the ATT τ based on the Lalonde data. Point estimate: point estimate of τ based on the Lalonde data; Bootstrap mean and Standard Deviation: the sample mean and sample variance of the bootstrap estimates of τ based on 1000 bootstrap samples from the Lalonde data. Lower and Upper are the lower and upper bounds of the Wald confidence interval for τ with asymptotic variances estimated by bootstrap.

Methods	PAVA1	PAVA2	PARA	PSM-3	PSM-5	PSM-10	PSM-15
	Case (a): X_1 and X_2 are covariates						
Point estimate	917.33	911.47	875.37	514.42	714.39	802.88	898.99
Bootstrap mean	894.19	906.39	861.61	1071.51	1033.38	993.00	985.47
Standard deviation	496.56	495.75	487.74	603.06	558.29	523.40	510.66
Lower	−55.91	−60.18	−80.58	−667.55	−379.84	−222.97	−101.89
Upper	1890.57	1883.12	1831.32	1696.40	1808.62	1828.73	1899.87
	Case (b): X_1, X_2, X_1X_2, X_1^2 and X_2^2 are covariates						
Point estimate	913.10	918.46	809.43	88.21	337.75	947.25	903.67
Bootstrap mean	936.21	950.03	835.23	997.70	997.80	1019.41	1028.54
Standard deviation	507.33	508.74	503.50	645.31	592.54	532.30	515.47
Lower	−81.25	−78.65	−177.41	−1176.57	−823.61	−96.04	−106.63
Upper	1907.45	1915.57	1796.27	1352.99	1499.11	1990.54	1913.97

4. Application to the Lalonde data

In this section, we apply the proposed PAVA-based estimation method to data from the National Supported Work (NSW) Demonstration, which have previously been analyzed by LaLonde (1986) and Dehejia and Wahba (2002). The primary parameters of interest in these papers concern the average treatment effect of a job training program. We focus on the estimation of the ATT, τ. The data consist of 297 treated and 425 untreated observations. We take earnings in 1978 as the outcome variable of interest (Y) and take age and education as the basic covariates X_1 and X_2, respectively. To examine the sensitivity of the proposed method to the model specifications, we model the propensity score using a linear logistic model and a quadratic logistic model.

We calculate the point estimates of τ using the first seven estimation methods considered in the previous section. The KN method is too sensitive to the choice of bandwidth and hence is excluded. When constructing confidence intervals for τ, we use the nonparametric bootstrap procedure in Algorithm 1 to estimate the asymptotic variances of all the estimators under comparison. Table 3 presents point estimates of τ based on the Lalonde data, and bootstrap mean, bootstrap standard deviation, and the corresponding Wald confidence intervals with asymptotic variances estimated using 1000 bootstrap samples from the Lalonde data.

Under either the linear or quadratic logistic model, the PAVA1 and PAVA2 point estimates are all around 915, indicating that the proposed estimation methods are rather robust to different model specifications. As PAVA2 makes the weakest model assumption, we believe that the results of PAVA2 should be the most trustable among the seven methods considered here. The bootstrap means of all methods are around 950 and the point estimates of PSM-15 are about 900 in both cases, which seemingly provide evidence for the rationality of the PAVA1 and PAVA2 point estimates. Although PARA also has bootstrap standard deviations of around 500, it produces very different point estimates (875.37 and 809.43) in the linear and quadratic logistic propensity score models. The PSM method is rather sensitive to the number of matches per unit. Its point estimate changes from 514.42 to nearly 900 with the linear logistic model, and varies even more dramatically with the quadratic logistic model.

In Section 7 of the supplementary material, we report a small simulation study to investigate the performance of the bootstrap procedure for variance and interval estimations. Our general findings are that the bootstrap variance estimates are very close to the true asymptotic variances, and that the resulting Wald type confidence intervals usually have very accurate coverage probabilities. These observations provide evidence for the rationality of the confidence intervals in Table 3. As the PAVA1 and PAVA2 point estimates are more reliable than the other five estimates, we believe that the confidence intervals based on PAVA1 and PAVA2 are also the most reliable.

Fig. 2 displays the fitted propensity scores (versus the estimated index $X^\top\widehat{\boldsymbol{\beta}}$) using a parametrically logistic model and the estimations of the semiparametric PAVA method after $\boldsymbol{\beta}$ is replaced by its MLE under the logistic model. The parametric propensity score estimates for both the linear and quadratic logistic models apparently form straight lines; unlike the semiparametric PAVA-based propensity score estimates, they may not capture local changes in the propensity score. As the semiparametric method requires fewer model assumptions and is more flexible, we believe that the semiparametric PAVA-based propensity score estimates and the corresponding PAVA1 estimates are more reliable than those based on the parametric propensity score estimates, including PARA and the four PSM methods. This may explain why the proposed PAVA-based method is superior to PARA and the four PSM methods.

5. Concluding remarks

Motivated by the empirical likelihood method in the presence of auxiliary information and choice-based sampling, Hirano et al. (2003) proposed an efficient IPW method using the fully nonparametric estimated propensity score. Even though their theoretical results are elegant, the finite-sample performance of their method is unclear. Compared with other efficient estimates, Hirano et al. (2003) stated that "Which estimators have more attractive finite sample properties, and which have more attractive computational properties, remain open questions". The connection between their matching methods and the IPW method, however, is unclear. In contrast to Hirano et al. (2003), this paper has proposed an IPW method that uses the maximum shape-restricted semiparametric

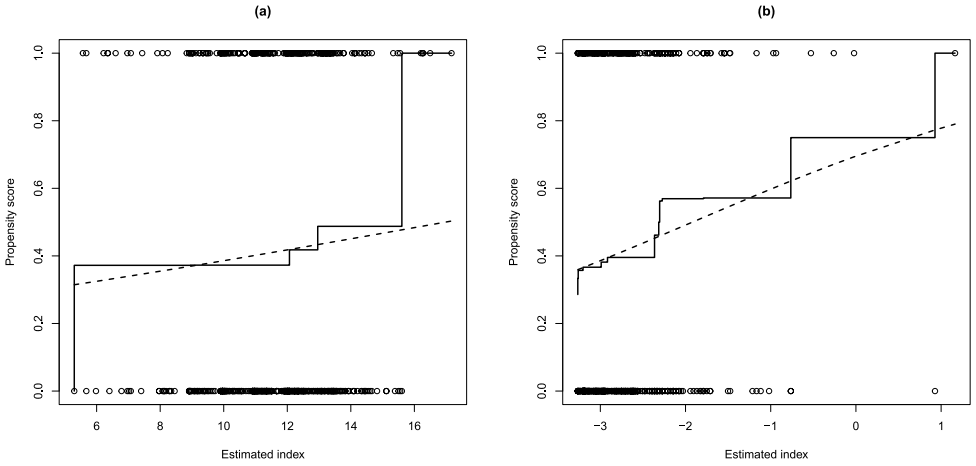

Fig. 2. Fitted propensity scores versus the estimated index $X^\top\widehat{\boldsymbol{\beta}}$ under a linear (left) and quadratic (right) logistic propensity score model based on the Lalonde data. Here, $\widehat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$ under the corresponding linear logistic model. Solid line: link function estimated by PAVA; dashed line: link function set to the logistic function.

likelihood estimation of the monotone index propensity score. Our method is very easy to implement using existing statistical software in R, such as the `Iso` and `Isotone` packages. Remarkably, our IPW method is seamlessly related to the tuning-parameter-free PSM method. Theoretical results show that our estimates can achieve the SELB for the average treatment effect and the ATT if the explanatory variable is univariate or the regression function and propensity score depend on the explanatory variables through the same index $\mathbf{X}^\top\boldsymbol{\beta}$. Our results underline the important role played by the propensity score and the regression function in estimating average causal effects. In general, the PSM method or the regression function matching method alone cannot be efficient. An efficient estimation method should take both of them into consideration (Hu et al., 2012).

The results in this paper are built on the unconfounded treatment assignment assumption, which is widely adopted in the causal inference literature. To guarantee this assumption to be satisfied, people are encouraged to collect as many covariates as possible in practical applications. However, the high dimensionality of the covariate vector may affect the performance of the proposed PAVA estimator, even though a single-index propensity score model is introduced for alleviating the impact of dimensionality. Motivated by Fan et al. (2020), it is of great theoretical value to investigate the asymptotic properties of the proposed PAVA estimator with a diverging number of covariates. To overcome the dimensionality challenge, Chen et al. (2024) proposed the use of feedforward artificial neural networks (ANN) to estimate the propensity score and developed an efficient estimation procedure for treatment effects. Taking shape restrictions into machine learning techniques such as ANN may produce more interpretable solutions to causal inference. We leave these topics for future research.

Acknowledgments

Liu's research is supported by the National Key R and D Program of China (2021YFA1000100 and 2021YFA1000101), the National Natural Science Foundation of China (12171157, 71931004, 32030063), Fundamental Research Funds for the Central Universities, China and the 111 Project, China (B14019).

Appendix A. Supplementary data

The supplementary materials contain proofs of all the technical results of this paper, and additional simulation results. Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jeconom.2024.105829.

References

Abadie, A., Imbens, G.W., 2006. Large sample properties of matching estimators for average treatment effects. Econometrika 74 (1), 237–267.
Abadie, A., Imbens, G.W., 2012. A martingale representation for matching estimators. J. Amer. Statist. Assoc. 107 (498), 833–843.
Abadie, A., Imbens, G.W., 2016. Matching on the estimated propensity score. Econometrika 84 (2), 781–807.
Ayer, M., Brunk, H.D., Ewing, G.M., Reid, W.T., Silverman, E., 1955. An empirical distribution function for sampling with incomplete information. Ann. Math. Stat. 26, 641–647.
Balabdaoui, F., Groeneboom, P., Hendrickx, K., 2019. Score estimation in the monotone single index model. Scand. J. Stat. 46, 517–544.
Barlow, R.E., Bartholomew, D.J., Bremner, J.M., Brunk, H.D., 1972. Statistical Inference under Order Restrictions: the Theory and Application of Isotonic Regression. In: Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, London–New York–Sydney.
Cavanagh, C., Sherman, R.P., 1998. Rank estimators for monotonic index models. J. Econometrics 84, 351–381.
Chen, X., Hong, H., Tarozzi, A., 2008. Semiparametric efficiency in GMM models with auxiliary data. Ann. Statist. 36 (2), 808–843.

Chen, X., Liu, Y., Ma, S., Zhang, Z., 2024. Efficient estimation of general treatment effects using neural networks with a diverging number of confounders. J. Econometrics 238.
Cochran, W.G., 1968. The effectiveness of adjustment by subclassification in removing bias in observational studies. Biometrics 24 (2), 295–313.
Dehejia, R.H., Wahba, S., 2002. Propensity score matching methods for non-experimental causal studies. Rev. Econ. Stat. 84, 151–161.
Efron, B., 1979. Bootstrap methods: another look at the jackknife. Ann. Statist. 7, 1–26.
Fan, Y., Han, F., Li, W., Zhou, X., 2020. On rank estimators in increasing dimensions. J. Econometrics 214 (2), 379–412.
Groeneboom, P., Hendrickx, K., 2017. The nonparametric bootstrap for the current status model. Electron. J. Stat. 11, 3446–3484.
Hahn, J., 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. Econometrika 66 (2), 315–331.
Han, A.K., 1987. Non-parametric analysis of a generalized regression model: the maximum rank correlation estimator. J. Econometrics 35, 303–316.
Heckman, J., Ichimura, H., Todd, P., 1998. Matching as an econometric evaluation estimator. Rev. Econ. Stud. 65, 261–294.
Hirano, K., Imbens, G.W., Ridder, G., 2003. Efficient estimation of average treatment effects using the estimated propensity score. Econometrika 71 (4), 1161–1189.
Hu, Z., Follmann, D.A., Qin, J., 2012. Semiparametric double balancing score estimation for incomplete data with ignorable missingness. J. Amer. Statist. Assoc. 107 (498), 247–257.
Imai, K., Ratkovic, M., 2014. Covariate balancing propensity score. J. R. Statist. Soc. Ser. B 76 (1), 243–263.
Imbens, G.W., 2015. Matching methods in practice: three examples. J. Hum. Resour. 50, 373–419.
Khan, S., Tamer, E., 2007. Partial rank estimation of duration models with general forms of censoring. J. Econometrics 136, 251–280.
LaLonde, R.J., 1986. Evaluating the econometric evaluations of training programs with experimental data. Amer. Econ. Rev. 76, 604–620.
Lee, Y.-Y., 2018. Efficient propensity score regression estimators of multivalued treatment effects for the treated. J. Econometrics 204 (2), 207–222.
Neyman, J., 1923–1990. On the application of probability theory to agricultural experiments. essay on principles. Section 9. (Translated and edited by D.M. Dabrowska and T.P. speed statistical science (1990), 5, 465–480). Ann. Agric. Sci. 10, 1–51.
Rosenbaum, P.R., 1985. The bias due to incomplete matching. Biometrics 41, 103–116.
Rosenbaum, P.R., Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.
Rubin, D.B., 1973. Matching to remove bias in observational studies. Biometrics 29, 159–184.
Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and non-randomized studies. J. Educ. Psychol. 66 (5), 688–701.
Rubin, D.B., 1978. Bayesian inference for causal effects: the role of randomization. Ann. Statist. 6, 34–58.
Sen, B., Banerjee, M., Woodroofe, M., 2010. Inconsistency of bootstrap: the grenander estimator. Ann. Statist. 38, 1953–1977.
Severini, T.A., Tripathi, G., 2013. Semiparametric efficiency bounds for microeconometric models: a survey. Found. Trends® Econometr. 6 (3–4), 163–397.
Stuart, E.A., 2010. Matching methods for causal inference: a review and a look forward. Statist. Sci. 25 (1), 1–21.
Tripathi, G., 2000. Local semiparametric efficiency bounds under shape restrictions. Econometric Theory 16, 729–739.
Wang, L., Han, P., 2024. Multiply robust estimation for average treatment effect among treated. Statist. Theory Relat. Fields 8 (1), 29–39.