Statistica Sinica Preprint No: SS-2022-0008	
Title	Propensity Score Regression for Causal Inference With Treatment Heterogeneity
Manuscript ID	SS-2022-0008
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202022.0008
Complete List of Authors	Peng Wu, Shasha Han, Xingwei Tong and Runze Li
Corresponding Authors	Peng Wu
E-mails	pengwu@btbu.edu.cn
Notice: Accepted version subject to English editing.	

Statistica Sinica

Propensity score regression for causal inference with treatment heterogeneity

Peng Wu[a,b]†, Shasha Han[c,d]†, Xingwei Tong[b] and Runze Li[e]∗

[a] Beijing Technology and Business University, [b] Beijing Normal University,
[c] Chinese Academy of Medical Sciences and Peking Union Medical College,
[d] Peking University, and [e] The Pennsylvania State University

Abstract: Understanding how treatment effects vary on several key characteristics is critical in the practice of personalized medicine. To estimate these conditional average treatment effects, non-parametric estimation is often desirable, but few methods are available due to the computational difficulty. Existing non-parametric methods such as the inverse probability weighting methods have limitations that hinder their use in many practical settings where the values of propensity scores are close to 0 or 1. We propose the propensity score regression (PSR) that allows the non-parametric estimation of such conditional average treatment effects in a wide context. PSR includes two non-parametric regressions in turn, where it first regresses on the propensity scores together with the characteristics of interest, to obtain an intermediate estimate; and then, regresses the intermediate estimate on the characteristics of interest only. By including propensity scores

†contributed equally.
∗correspond to: rzli@psu.edu

as regressors in a non-parametric manner, PSR is capable of substantially easing the computational difficulty while remaining less sensitive to values of propensity scores. We present its several appealing properties, including the consistency and asymptotical normality, and in particular, the existence of an explicit variance estimator, from which the analytical behavior of PSR and its precision can be assessed. Simulation studies indicate that PSR outperforms existing methods in varying settings with extreme values of propensity scores. We apply our method to the national 2009 flu survey (NHFS) data to investigate the effects of seasonal influenza vaccination and having paid sick leave across different age groups.

Key words and phrases: Heterogeneous treatment effect, nonparametric estimation, propensity score, high-dimensional covariates.

1. Introduction

The heterogeneous treatment effect describes the effect variability due to varying characteristics and is widely utilized in the contexts of personalized medicine, policy design, and customized marketing recommendation (Kent et al., 2018; Yin, 2018; Imai and Strauss, 2011; Sato et al., 2019). In many of the application settings, the characteristics of treatment relevance are only a subset of baseline covariates ($X = (X^l, X^{-l})$). Understanding how treatment works for individuals that differ on these few core characteristics (X^l) is particularly critical for developing tailored treatment decisions. For example, in clinical practice, patients at old ages tend to suffer more from side effects or drug-drug interactions. The age-dependent drug efficacy has been recommended for trade-off on benefits and risks of therapies (Velentgas et al., 2013). In health policy, the age-dependent vaccine effectiveness has been used to guide targeted vaccination programs (Soiza et al., 2021). Compared to the conditional treatment effects given the full covariates (Nie and Wager, 2021; Wager and Athey, 2018), these conditional treatment effects given the key characteristics are more easily interpretable and have been more widely used in clinical settings.

However, estimating such covariates conditional treatment effect is challenging: methods should be able to flexibly distinguish the heterogeneous effect (which is conditional on X^l) from the effect due to the remaining confounding covariates (X^{-l}). Since X^{-l} (could be high dimensional), may still confound with the effects of treatments on outcomes, conditioning on X^l is not sufficient. Moreover, the degree of confounding may vary with X^l, making modelling the conditional outcome particularly challenging. Non-parametric estimation method allows a fully flexible model and is therefore desirable. Yet, few non-parametric methods for estimating these heterogeneous treatment effects are available and the weighting-based methods have limitations that hinder their use in a wide context. For example, the inverse probability weighting proposed by Abrevaya et al. (2015) uses the inverse of propensity scores as weights to adjust outcomes. The method can result in unstable estimates when the values of propensity scores are close to 0 or 1— i.e., when the weights are very large, as typically observed in weighting methods to population average treatment effects (See e.g., Hahn, 1998; Rubin, 2001; Kang and Schafer, 2007). The augmented inverse probability weighting methods (Lee et al., 2017) also use the inverse of propensity scores as weights, and they require correctly parametric modelling outcomes to achieve efficiency (Seaman and Vansteelandt, 2018). Although the requirement can be relaxed by leveraging the machine learning methods (e.g., Fan et al., 2022; Zimmert and Lechner, 2019; Semenova and Chernozhukov, 2021), these methods can rely heavily on extrapolation, which could be a critical concern in the context with extreme values of propensity scores (e.g. Kang and Schafer, 2007; Tan, 2007; Wu et al., 2022).

Alternative methods generally rely on a two-step estimation: estimating the conditional treatments effect defined on the full covariate first, and then integrating out the obtained estimates into the desired level of granularity. However, it is often difficult to non-parametrically estimate the conditional treatment effect for a high dimensional covariate (Abrevaya et al., 2015; Lechner, 2019; Zimmert and Lechner, 2019; Wu et al., 2022). For example, in our real example with the national 2009 flu survey (NHFS) data (See §5 for the detail), the dimension of full covariates is as high as 65 (See Supplementary Table S2). With such a high dimension, typical non-parametric estimation methods, e.g., the local linear regression, would suffer from the so-called curse-of-dimensionality (Fan and Gijbels, 1996).

In line with the two-step estimation, we propose the non-parametric propensity score regression(PSR). It consists of two non-parametric regressions in turn: It first regresses on the propensity scores together with the covariates of interest; Then, it integrates out the scores by regressing the estimates from the first regression on the covariates of interest only. Propensity score has the crucial balancing property, namely, the distributions of full covariates between the treatment groups are identical at each level of the propensity scores (including the one-to-one functions of propensity scores). In our context, the balancing property is useful for controlling the confounding due to the remaining covariates X^{-l}, and for easing the computational difficulty. PSR utilizes a continuous and bounded function of the score, and therefore is less sensitive to the extreme values of propensity scores. Further, by utilizing the propensity scores in a non-parametric manner in the first step, PSR reduces the influences of errors in propensity scores on the estimates of the second step (Mammen et al., 2012), and thus enjoys increased robustness to the estimation errors on the propensity scores. On the other hand, the weighting-based estimation methods achieve covariate balance for a hypothetical super-population that is constructed from reweighing units in the studying population, where small changes in propensity scores could lead to large discrepancies in weights, and even non-parametric estimation can result in highly unstable estimates.

The idea of including propensity scores in regression is not new in the context of parametric estimation for estimating average treatment effects (See e.g., Little and An, 2004; Zhang and Little, 2009; Zhou et al., 2019; Wu et al., 2021). However, these approaches still rely on technical modeling assumptions for the outcome. When using propensity score in parametric regression, the key is to correctly specify the elusive relationship between the propensity score and the outcome, which is intrinsically connected to difficulties in specifying outcome models. Unlike these methods, we propose to use propensity scores as regressors in a non-parametric manner for estimating heterogeneous treatment effects.

We theoretically validate the approach and show its appealing advantages. To obtain the theoretical results, we assume a parametric estimation of propensity scores, but non-parametric estimation of propensity scores is allowed. Of note, even under the parametric assumption, unlike the weighting-based methods, PSR allows one-to-one transformation of propensity scores and the functional form of propensity score is less important. We present the theoretical properties, including the consistency and asymptotical normality, and the explicit variance estimator, from which the analytical behavior of propensity score regression and its precision can be assessed.

PSR is not only valuable in exploring the treatment heterogeneity for practical guidance, but also useful in understanding treatment heterogeneity with the high-dimensional full covariate. For example, we could decompose the full covariate into many subsets, each with only a few covariates, and estimate the heterogeneous treatment effects on different subsets; With the ensemble of such heterogeneous treatment effects, we may approach the full picture of the treatment heterogeneity with the high dimensional full covariate, the direct estimation of which is acknowledged to be computational difficult (Abrevaya et al., 2015; Lechner, 2019; Zimmert and Lechner, 2019; Wu et al., 2022; Semenova and Chernozhukov, 2021).

The paper proceeds as follows. In §2, we introduce the basic framework and the motivation of the analysis. §3 presents the propensity score regression, where §3.1 outlines the method and provides the theoretical validation, and §3.2 provides the non-parametric estimator. In §3.3, we show the theoretical properties. We conduct several simulation studies in §4. In

§5, we apply our method to the national 2009 flu survey (NHFS) data to investigate the effects of seasonal influenza vaccination and having paid sick leave across different age groups. We conclude with a discussion in §6.

2. Motivations

2.1 Notations and assumptions

We adopt the framework of Rubin Causal Model (Rubin, 1974) — often also called the potential outcome approach to causal inference (Imbens and Rubin, 2015). Consider a study with N units. Each unit $i = 1, \dots, N$ is associated with a vector-valued covariate X_i $X_i = (X_{i1}, \dots, X_{il}, \dots, X_{ip}) \in \mathcal{X} \subseteq \Re^p$, measured before being exposed to the treatment D_i. The low dimensional covariate of interest is denoted by X_i^l, $X_i^l = (X_{i1}, \dots, X_{il}) \in \mathcal{X}^l \subset \Re^p$. We write $X_i = (X_i^l, X_i^{-l})$. The outcome variable Y is measured on each unit after its treatment exposure. Associated with treatment d, $d = 0, 1$, is the potential outcome $Y_i(d)$, the value of Y when the unit i is exposed to treatment t, which implicitly assumes the stable unit treatment value assumption (SUTVA, Rubin, 1980). When referring to a generic unit, we drop the subscript and write $X, D, Y(0), Y(1), Y, X^l, X^{-l}$, etc.

As in literature (e.g. Abrevaya et al., 2015), we assume that the unconfounded assumption, namely, $D \perp\!\!\!\perp (Y(1), Y(0))\ |X$. We denote the propensity score $e(X) := \mathbb{P}(D = 1|X)$, assuming that $0 < e(x) < 1$, for any $x \in \mathcal{X}$. The heterogeneous treatment effect of interest, $\tau(x^l)$, is defined on the subspace of the covariates, $\mathcal{X}^l$, as

$$\tau(x^l) := \mathbb{E}\left(Y(1) - Y(0)|X^l = x^l\right), \tag{2.1}$$

$x^l \in \mathcal{X}^l$. Often, the dimension of $\mathcal{X}^l$ is much smaller than that of the full covariate space $\mathcal{X}$. Therefore, $\tau(x^l)$ is at a higher level of granularity than the treatment effects conditional on the full covariate.

2.2 Two-step estimation

To estimate an estimand at a higher level of granularity, an intuitive way is to estimate treatment effects at a lower level of granularity first; and then, integrate out the obtained estimates into the subspace of interest. Our idea broadly belongs to the two-step estimation. The key intuition is to explore an estimand at the lower level of granularity that can be estimated unbiasedly and non-parametrically. Below, we briefly describe the idea.

We write $\tau(x^l)$ using the tower property of conditional expectation as

$$\tau(x^l) = \mathbb{E}\left[\mathbb{E}\left(Y(1) - Y(0)|X^l = x^l, X^{-l}\right)|X^l = x^l\right].$$

Let $\eta(x) := \mathbb{E}\left(Y(1) - Y(0)|X^l = x^l, X^{-l} = x^{-l}\right)$, where $x = (x^l, x^{-l})$. In principle, we could estimate the insider expectation $\eta(X)$ first; and then integrate X^{-l} out with respect to the conditional distribution of X^{-l} given $X^l = x^l$. In this case, our task would become estimating the finest estimand $\eta(x)$ for each $x \in \mathcal{X}$, which can be identified as

$$\begin{aligned}\eta(x) &= \mathbb{E}\left(Y(1)|D = 1, X^l = x^l, X^{-l} = x^{-l}\right) - \mathbb{E}\left(Y(0)|D = 0, X^l = x^l, X^{-l} = x^{-l}\right) \\ &= \mathbb{E}\left(Y|D = 1, X^l = x^l, X^{-l} = x^{-l}\right) - \mathbb{E}\left(Y|D = 0, X^l = x^l, X^{-l} = x^{-l}\right).\end{aligned} \tag{2.2}$$

However, when the dimension p of the full covariate space is large, the non-parametric estimation of $\eta(X)$ could be difficult (Abrevaya et al., 2015; Lechner, 2019; Zimmert and Lechner, 2019; Wu et al., 2022).

This motivates us to explore an alternative estimand for the first-step estimation. Specifically, we aim to find an estimand that lies in a much higher level of granularity than $\eta(X)$ while still in a lower level than $\tau(X^l)$, so that in practice it is possible to estimate the new estimand non-parametrically. Notably, the lower auxiliary variable needs to be able to replace the important role that X^{-l} plays in the identification. As illustrated in Equation (2.2), conditioning on X^{-l} as well as X^l, i.e., the full covariate X, facilitates the identification of $\eta(X)$ using the observed data, due to the natural balancing property of the full covariate X.

Together, the new auxiliary variable needs to rest on the subspace with a dimension much smaller than p; and it is also desirable to have the balancing property as we mentioned above. The propensity score, defined as the probability of assignment to the treatment given the full covariate (Rosenbaum and Rubin, 1983), is one candidate. As a summary of covariates, the propensity score is able to reduce the p dimensional covariates into a scaler while having the desired balancing property. Clearly, any one-to-one functions of the propensity score are candidates as well.

2.3 Compared to existing methods using propensity scores

Existing methods have leveraged propensity scores to achieve covariate balance through reweighing the units. For example, Abrevaya et al. (2015) proposed the inverse probability weighting (IPW) estimator,

$$\tau^{IPW}(x^l) = \mathbb{E}\Big(DY/e(X) - (1-D)Y/(1-e(X))\Big| X^l = x^l\Big). \tag{2.3}$$

As in the weighting methods (See e.g., Hahn, 1998; Rubin, 2001; Kang and Schafer, 2007), the IPW estimator in (2.3) is sensitive to the values of estimated propensity scores and their estimates are highly unstable if the values of propensity scores are close to 0 or 1. When the parametric function of the outcome model is knowable, Lee et al. (2017) propose the augmented inverse probability weighting (AIPW) estimator

$$\tau^{AIPW}(x^l) = \mathbb{E}\left(\frac{D\,(Y-\mu_1(X))}{e(X)} - \frac{(1-D)\,(Y-\mu_0(X))}{1-e(X)} + (\mu_1(X)-\mu_0(X))\,\middle|\, X^l = x^l\right), \tag{2.4}$$

where $\mu_d(\cdot)$, $d = 0, 1$ are specified outcome functions. AIPW allows the misspecification of the propensity score model, if the parametric outcome functions $\mu_d(\cdot)$ are correctly specified. However, in the setting with a high dimensional covariate, the correct specification of the outcome functions $\mu_d(\cdot)$ is not easy. Literatures have instead explored techniques to estimate $\mu_d(\cdot)$ using machine learning methods (e.g. Fan et al., 2022; Semenova and Chernozhukov, 2021), but the estimation methods heavily rely on the extrapolation (e.g. Kang and Schafer, 2007; Tan, 2007).

Like IPW, AIPW is also sensitive to the extreme values of propensity scores, even when the propensity score models are specified correctly (Rotnitzky and Vansteelandt, 2014). Instead of using propensity scores as weights, we include the propensity scores as one regressor in non-parametric regression. We call the generic method propensity score regression (PSR).

PSR is conceptually different from the propensity score weighting to achieve the balance. The weighting methods use the inverse of propensity scores as weights to construct a hypothetical super-population for which the distributions of the covariates between the treated units and control units could be balanced. Clearly, these weights are unbounded around 0 or 1, with small changes in propensity scores potentially leading to large discrepancies in weights. This is likely to be a concern in the context of extreme values of propensity scores. The current setting is different. $\beta(X^l, e)$ defined in equation (3.5) is a bounded function of the propensity score, and therefore is less sensitive to the extreme values of propensity scores.

As such, PSR is analogous to propensity score matching and subclassification, and they all are based on the balancing property of propensity score in the studying population. But unlike matching on propensity scores, which implicitly involves model specifications (e.g., we need to specify the matching criteria and different criteria generally lead to different matched sets), PSR utilizes propensity scores in a non-parametric manner, i.e., using existing non-parametric methodologies. Hence, whereas with parametric modeling on propensity scores, difficulty in matching with extreme values of propensity scores is likely to lead to substantial bias, here the estimation is non-parametric and $\beta(x^l, e)$ is estimated smoothly. Thus, the results should be less sensitive to minor differences between the inexact match. We show this using a variant of PSR where the regression procedure is replaced by matching on propensity scores (§4).

Acknowledgedly, with large differences in propensity scores between treated units and control units — for the treated units near a propensity of 1.0 only and for the control units near a propensity of 0.0 only — PSR may not be suitable as well. We note that in a mild situation, where both the treated units and control near a propensity of 1.0 and 0.0, IPW and AIPW estimators could generate highly sensitive estimates.

3. Propensity score regression

Propensity score regression (PSR) includes two non-parametric regressions as follows. For notional simplicity, we refer hereafter "e" the values of propensity scores $e(X)$ or any one-to-one function of $e(X)$. As discussed in §2, the key idea is to explore an intermediate estimand at the lower level of granularity that can be estimated unbiasedly and non-parametrically. We denote $\beta(X^l, e)$ the intermediate estimand, which is conditional on the values of propensity scores as well as covariates of interest, as

$$\beta(x^l, e) = \mathbb{E}\left(Y(1) - Y(0) | X^l = x^l, e(X) = e\right). \tag{3.5}$$

We would like to note that the estimand $\beta(x^l, e)$ is conditional on the $l+1$ dimensional "covariates", where $l+1$ is substantially smaller than the dimension of the full covariate p, and therefore can mitigate the problem of high-dimensional covariates. Estimating $\beta(x^l, e)$ is our central task. Note that the definition of $\beta(x^l, e)$ includes both two potential outcomes $Y(1)$ and $Y(0)$, but only one of them is observed in the real world. By conditioning on propensity scores, we could replace the potential outcomes with the observed outcome Y. Below, we prove this non-parametrically.

Proposition 1. Suppose that $\mathbb{E}\left(Y|D, X^l, e\right)$ is a non-parametric regression function. Then, $\beta(X^l, e)$ and $\mathbb{E}\left(Y(0)|X^l, e\right)$ are the functional coefficients correspond to the treatment indicator D and the intercept respectively,

$$\mathbb{E}\left(Y|D, X^l, e\right) = \beta(X^l, e) \cdot D + \mathbb{E}\left(Y(0)|X^l, e\right), \tag{3.6}$$

where $\beta(x^l, e) = \mathbb{E}\left(Y \mid D = 1, X^l = x^l, e(X) = e\right) - \mathbb{E}\left(Y \mid D = 0, X^l = x^l, e(X) = e\right)$.

Model (3.6) is a varying coefficient model (Hastie and Tibshirani, 1993) and can be estimated through standard local linear regression techniques (Fan and Zhang, 1999). Note that here we do not make any parametric assumptions. The model is generally enough to capture any model specification. The counterpart working model is given by

$$Y = \beta(X^l, e) \cdot D + \mathbb{E}\left(Y(0)|X^l, e\right) + \xi, \tag{3.7}$$

where $\mathbb{E}\left(\xi|D, X^l, e\right) = 0$.

Remarkably, here $\beta(x^l, e)$ is estimated using the full sample of data under both two treatment conditions, $d = 0, 1$. Unlike the existing methods such as AIPW, our approach does not extrapolate the unobserved outcomes by the predicted values from the other treatment group. Consequently, it is more likely to have good finite sampling performance. After having the estimates of $\beta(x^l, e)$, we simply integrate out the propensity scores e to get the estimates for $\tau(x^l)$. To do so, we conduct a second non-parematric estimation based on the projection relationship of $\tau(x^l)$ and $\beta(x^l, e)$.

Proposition 2. $\tau(x^l)$ is, geometrically, the projection of $\beta(X^l, e)$ into the subspace spanned by X^l, specifically, $\tau^{PSR}(x^l) = \mathbb{E}\left(\beta(X^l, e)|X^l = x^l\right)$.

Proposition 2 suggests that $\tau(x^l)$ can be estimated non-parametrically, e.g., through local linear regression of $\beta(X^l, e)$ on X^l.

3.1 Description of the approach

When $e(X)$ is known, PSR is implemented with two non-parametric regressions, built on Propositions 1 and 2 respectively. When $e(X)$ also estimated from the data, PSR is implemented in total three steps, as below:

Step 0: Estimate $e(X)$ in either parametric or non-parametric manner.

Step 1: Estimate $\beta(X^l, e)$ through non-parametrically regressing the outcome Y on the covariate X^l and the propensity scores $e(X)$.

Step 2: Estimate $\tau(X^l)$ through non-parametrically regressing the estimated values of $\beta(X^l, e)$ on the covariate X^l only.

For practical use, we are interested in the large sample properties of $\hat{\tau}(x^l)$. An important question is: how do the errors of the estimated propensity scores in Step 0 affect the estimates for $\tau(x^l)$? Briefly answering, PSR is robust to the estimation errors of the propensity scores, provided that the influence of the errors of estimated propensity scores on the second step estimation is negligible (Mammen et al., 2012). In our context, this is because the propensity scores are used in the non-parametric manner (Step 1), where the true propensity scores and their estimated counterparts are asymptotically indistinguishable.

Below, we will focus on continuous X^l, but our results hold in the general setting including discrete X^l. Readers who are interested in the details for the discrete X^l, please refer to §7 of the Supplementary Material.

3.2 Non-parametric estimator

We adopt the mostly used non-parametric methods, the standard local linear regressions, to estimate $\beta(x^l, e)$ and $\tau(x^l)$ in Steps 1 and 2 respectively.

For notational simplicity, we focus on the case with $l = 1$ and consider a more general case in Supplementary Material.

We consider the case where the propensity scores are known, and denote the response vector $\boldsymbol{Y} = (Y_1, \cdots, Y_N)^\intercal$, the regressor vector $\boldsymbol{\Gamma} = (\Gamma_1, \cdots, \Gamma_N)^\intercal$, with the regressors $\Gamma_i = (D_i, 1, D_i(X_i^l - x^l)/h_1, (X_i^l - x^l)/h_1, D_i(e_i - e)/h_2, (e_i - e)/h_2)^\intercal$, $i = 1, \ldots, N$, and the kernel vector $\boldsymbol{W} = \mathrm{diag}\{K_{h_1}(X_1^l - x^l)K_{h_2}(e_1 - e), \ldots, K_{h_1}(X_N^l - x^l)K_{h_2}(e_N - e)\}$, with $K_{h_j}(u) = K(u/h_j)/h_j$, $j = 1, 2$. The standard local linear estimator $\tilde{\beta}(x^l, e)$ of $\beta(x^l, e)$ is

$$\tilde{\beta}(x^l, e) = (1, 0, 0, 0, 0, 0)(\boldsymbol{\Gamma}^\intercal \boldsymbol{W} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^\intercal \boldsymbol{W} \boldsymbol{Y}, \tag{3.8}$$

which is the first component of $(\boldsymbol{\Gamma}^\intercal \boldsymbol{W} \boldsymbol{\Gamma})^{-1} \boldsymbol{\Gamma}^\intercal \boldsymbol{W} \boldsymbol{Y}$. Further, we denote that the response vector $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}(X_1^l, e_1), \ldots, \tilde{\beta}(X_N^l, e_N))^\intercal$, the regressor vector $\boldsymbol{G} = (G_1, \ldots, G_n)^\intercal$ with $G_i = (1, (X_i^l - x^l)/h_3)^\intercal$, and $\boldsymbol{\Lambda} = \mathrm{diag}\{K_{h_3}(X_1^l - x^l), \ldots, K_{h_3}(X_N^l - x^l)\}$. The local linear estimator of $\tau(x^l)$ is

$$\tilde{\tau}(x^l) = (1, 0)(\boldsymbol{G}^\intercal \boldsymbol{\Lambda} \boldsymbol{G})^{-1} \boldsymbol{G}^\intercal \boldsymbol{\Lambda} \tilde{\boldsymbol{\beta}}. \tag{3.9}$$

Next, we consider the case where the propensity scores are estimated from the data. We replace $\boldsymbol{\Gamma}$,$\boldsymbol{W}$ and $\tilde{\boldsymbol{\beta}}$ in (3.8) and (3.9) by $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{\Gamma}}$, $\hat{\boldsymbol{\beta}}$ respectively. Note that we replace the true propensity scores e by the estimated $\hat{e}$. Using the estimated propensity scores, the local linear estimator of $\beta(X^l, e)$ and $\tau(x^l)$ is then given by

$$\hat{\beta}(x^l, e) = (1,0,0,0,0,0)(\hat{\mathbf{\Gamma}}^\intercal \hat{\boldsymbol{W}} \hat{\mathbf{\Gamma}})^{-1} \hat{\mathbf{\Gamma}}^\intercal \hat{\boldsymbol{W}} \boldsymbol{Y}, \tag{3.10}$$

$$\hat{\tau}(x^l) = (1,0)(\boldsymbol{G}^\intercal \mathbf{\Lambda} \boldsymbol{G})^{-1} \boldsymbol{G}^\intercal \mathbf{\Lambda} \hat{\boldsymbol{\beta}}. \tag{3.11}$$

3.3 Theoretical properties

To answer the question proposed in §3.1, we first present the following key assumption.

Assumption 1. The propensity score model can be written as $e(X) = g(X^\intercal \alpha)$, where α is the true unknown parameter and $g(\cdot)$ is a known function (e.g., generalized linear model).

(i) The estimates of α, $\hat{\alpha}$, satisfies $\hat{\alpha} - \alpha = O_p(N^{-1/2})$;

(ii) The second-order derivative of g is uniformly bounded, i.e., $\sup_t |g''(t)|$ is bounded.

Assumption 1(i) is critical to our main results and is made for simplifying the asymptomatic analysis on $\hat{\tau}(X^l)$. We note that the parametric modeling assumption on propensity scores can be relaxed. Details are provided in the Discussion.

Theorem 1. Under Assumption 1 and regularity Assumption 2 in Supplementary Material, $\hat{\beta}(x^l, e)$ in (3.10) and $\tilde{\beta}(x^l, e)$ in (3.8) are asymptotically indistinguishable, i.e., $\sup_{\left(x^l, e\right) \in \mathcal{X}^l \times (0,1)} \left|\hat{\beta}(x^l, e) - \tilde{\beta}(x^l, e)\right| = O_p(N^{-1/2})$.

Theorem 1 demonstrates that the impact of the estimation error of propensity scores on the estimator of $\beta(x^l, e)$ is negligible. Intuitively, this is because the propensity score is used in a non-parametrical manner in Step 1, where locally indistinguishable estimates of the propensity scores from the true values are needed only. The finding is consistent with the earlier results by (Mammen et al., 2012) in a different setting on studying the coverage rates of the final estimates, where they show that the influence of the first-step estimation error on second-step estimation is restricted in a smoothed way through the estimation bias in the first step. Using the results from Theorem 1, it is easy to show that the impact on the estimator of $\tau(x^l)$ is small as well (e.g., Gu and Yang, 2015). We next establish the asymptotical normality of $\hat{\tau}(x^l)$. Let $\bar{K}(x) = \int K(t)K(x + h_1 t/h_3)dt$, $\nu = \int K^2(t)dt$, $f(x^l)$ is the density function for X^l.

Theorem 2. Under Assumption 1 and regularity Assumptions 2–3 in Supplementary Material, $\hat{\tau}(x^l)$ in (3.11) is a consistent estimator of $\tau(x^l)$, and

$$\mathbb{V}\left(x^l\right)^{-1/2}\left(\hat{\tau}(x^l) - \tau(x^l)\right) \xrightarrow{d} N(0,1),$$

where $\mathbb{V}\left(x^l\right)$ represents the asymptotic variance and

$$\mathbb{V}\left(x^l\right)=\frac{1}{Nh_3 f(x^l)}\left(\nu\cdot\mathbb{V}\left(\beta(X^l,e)|X^l=x^l\right)+\int\bar{K}^2(t)dt\cdot\mathbb{E}\left(\frac{(D-e)^2}{e^2(1-e)^2}\xi^2|X^l=x^l\right)\right).$$

We would like to note that Theorem 2 relies on the Assumption 3(iv) $\sqrt{Nh_3}(h_1^2+h_2^2+h_3^2)\to 0$ as $N\to\infty$. The common bias term, i.e., $O(h_1^2+h_2^2+h_3^2)$, is vanished under the condition. However, our main conclusion still holds if the assumption is relaxed. Importantly, the asymptotic variance $\mathbb{V}\left(x^l\right)$ can be estimated by the plug-in method. Specifically, the estimated asymptotic variance for $\hat{\tau}(x^l)$ is

$$\hat{\mathbb{V}}(x^l):=\frac{1}{Nh_3\hat{f}(x^l)}\left(\nu\cdot\hat{\mathbb{V}}\left(\beta(X^l,e)|X^l=x^l\right)+\int\bar{K}^2(t)dt\cdot\hat{\mathbb{E}}\left(\frac{(D-e)^2}{e^2(1-e)^2}\xi^2|X^l=x^l\right)\right), \tag{3.12}$$

where $\hat{f}(x^l)$ is the kernel density estimation, $\hat{\mathbb{V}}(\beta_1(X^l,e)|X^l=x^l)$ can be estimated via conducting nonparametric regressing $(\hat{\beta}(X^l,e)-\hat{\tau}(X^l))^2$ on X^l (Fan and Yao, 1998). Since ξ is estimated using the residual in model (3.7), we obtain $\hat{E}\left[\xi^2(D-e)^2/e^2(1-e)^2|X^l=x^l\right]$ by regressing $(D-\hat{e})^2\hat{\xi}^2/\hat{e}^2(1-\hat{e})^2$ on X^l. In addition, ν and $\int\bar{K}^2(x)dx$ can be calculated directly. Finally, we have the following conclusion.

Corollary 1. *The estimated asymptotic variance is a consistent estimator, namely,* $\hat{\mathbb{V}}(x^l)-\mathbb{V}(x^l)=o_p(1)$.

4. Simulation studies

Extensive simulation studies are conducted to assess the finite sample performance of PSR, compared to the existing IPW method of Abrevaya et al. (2015), and AIPW method of Lee et al. (2017). Also, we consider a matching variant of PSR, where we use the matching method to estimate $\beta(X^l, e)$. Specifically, we replace Step 1 in the PSR by first creating matched pairs through matching on $(X^l, \hat{e}(X))$ and then using the matched pairs to impute the missing potential outcomes. Finally, we calculate $\beta(X^l, e)$ as using the imputed potential outcomes. In the matching, we use one-to-one matching and the Mahalanobis metric. In the following, we focus on the settings with extreme values of propensity scores. We also considered an alternative scenario where propensity scores are distributed far from 0 and 1.

4.1 Simulation Setup

We set $l = 1$ and the covariate dimension $p = 5$, 20 and 50. $X^l \sim \text{Uniform}(-0.5, 0.5)$ and $X^{-l} = (X_1^{-l}, ..., X_{p-1}^{-l}) \sim \text{Norm}(0, \Sigma)$ with $\Sigma_{j,k} = 2^{-|j-k|}$ for $1 \leq j, k \leq p-1$. The assignment of treatment D follows the logistic model, with $\mathbb{P}(D = 1|X) = \exp(X^\intercal\alpha)/(1 + \exp(X^\intercal\alpha))$.

In the setting with extreme values of propensity scores, we consider two assignment mechanisms: Mechanism A, $\alpha = (1, -1, -1, 1, -1, 0, \ldots, 0)^\intercal$, with five non-zero entries; Mechanism B, $\alpha = (1, 1, 1, 1, 1, 0, \ldots, 0)^{\intercal}$, with five non-zero entries. We conduct four simulation settings, with the heterogeneous treatment effects $\tau(x^l)$ being linear, quadratic, polynomial and complex functions of x^l respectively. The potential outcomes under the four contexts are modelled as follows:

I: $Y(1) = X^l(1 + 2X^l)^2(X^l - 1)^2 + f(X) + \epsilon(1)$, $Y(0) = f(X) + \epsilon(0)$, $f(X) = (X^l)^2 X_1^{-l} X_2^{-l} X_3^{-l} X_4^{-l}$,

II: $Y(1) = X^l(1 - X^l)\cos(X^l)\log(X^l + 2)\exp(X^l) + f(X) + \epsilon(1)$, $Y(0) = f(X) + \epsilon(0)$, $f(X) = (X^l)^2 X_1^{-l} X_2^{-l} X_3^{-l} X_4^{-l}$.

III: $Y(1) = X^l + f(X) + \epsilon(1)$, $Y(0) = f(X) + \epsilon(0)$, $f(X) = \{X^l X_1^{-l} + \exp(X_2^{-l} - 3)(\sin(X_3^{-l}) + \cos(X_4^{-l}))\}/2$,

IV: $Y(1) = 5(X^l)^2 + X^l + f(X) + \epsilon(1)$, $Y(0) = f(X) + \epsilon(0)$, $f(X) = (X^l)^2\{\sum_{j=1}^{4} X_j^{-l}/2^{j+1}\}$.

Error terms $\epsilon(1)$ and $\epsilon(0)$ are independently and identically distributed with $\text{Norm}(0, 1)$. The true $\tau(x^l)$ in the four simulation settings are $x^l(2x^l + 1)^2(x^l - 1)^2$, $x^l(1 - x^l)\cos(x^l)\log(x^l + 2)\exp(x^l)$, x^l and $5(x^l)^2 + x^l$ respectively. Note that the potential outcome models have a complex function forms and hence it is very difficult to correctly specify the outcome models.

Simulation I and II use the assignment Mechanism A and Simulation III and IV use the assignment Mechanism B. The distributions of propensity scores under the two mechanisms are plotted in Figure 1. We observe that many values of propensity score close to 0 or 1. It is expected that the weighting-based propensity score methods will be highly unstable in the context.

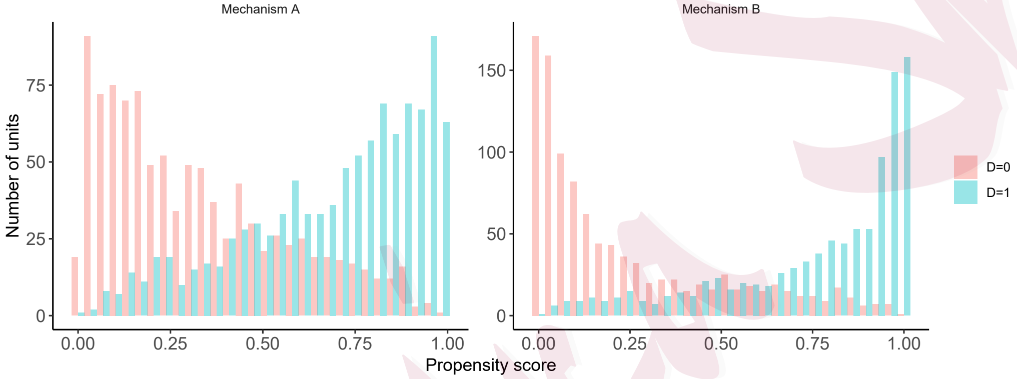

Figure 1: Distribution of true propensity score under the two treatment assignment mechanisms $(N = 2000, p = 5)$.

We estimate the propensity scores using the logistic regression. The bandwidths are chosen using the existing methods (Li and Racine, 2007; Ruppert et al., 1995) and implemented with R function `npscoefbw` from package **np** (Racine and Hayfield, 2021) and `dpill` from package **KernSmooth** (Wand et al., 2021). For competing AIPW method, the outcome regression functions are estimated with linear models. Each simulation study is based on 1000 replicates with sample sizes 500, 1000 and 2000. We

Table 1: The performance of PSR for cases I–IV.

	Simulation I				Simulation II				Simulation III				Simulation IV			
(N, p)	Bias (SD)	MAE	MSE	CP95	Bias (SD)	MAE	MSE	CP95	Bias (SD)	MAE	MSE	CP95	Bias (SD)	MAE	MSE	CP95
	$\times 10^{-2}$	$\times 10^{-2}$	$\times 10^{-2}$	%	$\times 10^{-2}$	$\times 10^{-2}$	$\times 10^{-2}$	%	$\times 10^{-2}$	$\times 10^{-2}$	$\times 10^{-2}$	%	$\times 10^{-2}$	$\times 10^{-2}$	$\times 10^{-2}$	%
$(500, 5)$	0.0 (16.9)	13.2	2.8	87.4	0.1 (15.5)	12.3	2.4	86.5	3.3 (17.4)	13.7	3.1	88.1	2.0 (22.1)	17.2	4.9	88.9
$(1000, 5)$	-0.4 (12.7)	10.0	1.6	91.8	0.1 (11.8)	9.3	1.4	90.7	2.7 (13.2)	10.6	1.8	91.4	1.4 (16.8)	13.2	2.8	92.2
$(2000, 5)$	-0.5 (9.4)	7.5	0.9	94.4	0.2 (8.4)	6.6	0.7	93.6	3.2 (10.5)	8.6	1.2	93.9	1.6 (13.0)	10.2	1.7	95.2
$(500, 20)$	0.0 (17.3)	13.7	3.0	87.5	0.4 (15.4)	12.4	2.6	86.5	2.8 (17.8)	14.2	3.4	86.9	2.1 (22.3)	17.5	5.1	88.6
$(1000, 20)$	-0.0 (12.7)	10.3	1.7	89.8	-0.0 (11.6)	9.2	1.4	90.6	2.8 (13.9)	10.8	1.9	91.0	1.6 (16.9)	13.2	2.9	92.7
$(2000, 20)$	0.0 (9.6)	7.5	0.9	94.3	-0.0 (9.0)	6.8	0.7	93.8	2.5 (10.7)	8.3	1.1	93.4	1.5 (12.8)	10.1	1.7	95.1
$(500, 50)$	-0.4 (17.3)	13.7	3.1	87.3	-0.0 (16.5)	12.7	2.6	86.9	2.9 (19.4)	14.6	3.5	85.8	1.8 (22.1)	17.4	4.9	88.3
$(1000, 50)$	-0.2 (13.0)	10.2	1.7	91.5	0.0 (11.8)	9.6	1.5	90.4	3.1 (14.4)	11.4	2.1	90.2	1.2 (16.7)	13.0	2.8	92.2
$(2000, 50)$	-0.3 (9.4)	7.5	0.9	94.2	0.0 (9.1)	6.8	0.8	93.8	2.4 (10.6)	8.6	1.2	93.7	1.3 (13.2)	10.4	1.8	94.6

evaluate the performance of PSR on sample average bias (Bias) and sample average standard deviation (SD), mean-absolute-error (MAE), mean-square-error (MSE), and average 95% confidence interval coverage proportion (CP95). For PSR, CP95 is estimated using the asymptotic variance formula (3.12). For the IPW and AIPW methods, CP95 is estimated as in the associated literature of Abrevaya et al. (2015) and Lee et al. (2017). Finally, for the matching variant of PSR and Random Forest methods, CP95 is estimated via 100 bootstraps.

Table 1 summarizes the results of PSR for cases I–IV. We observe that as the sample size increases, MAE and MSE decrease, and CP95 becomes closer to the nominal value of 0.95. Moreover, the results were very similar for different values of p, suggesting that PSR is insensitive to the dimension of covariates.

In addition, we consider an alternative scenario where propensity scores are distributed far from 0 and 1. We replace the data generation mechanisms in Simulation I-IV with new mechanisms. Specifically, we replace Mechanism A in simulations I and II with Mechanism C, where α is set as $0.25(1, -1, -1, 1, -1, 0, ..., 0)^\intercal$ with five non-zero entries, and denoted them as Simulation V and VI respectively, we replace Mechanism B in simulations III and IV with Mechanism D, $\alpha = 0.125(1, -1, -1, 1, -1, 0, \dots, 0)^\intercal$ with five non-zero entries, and denoted them as Simulation VII and VIII respectively. Figure 2 shows the propensity score distributions of the treatment assignment mechanisms C and D.

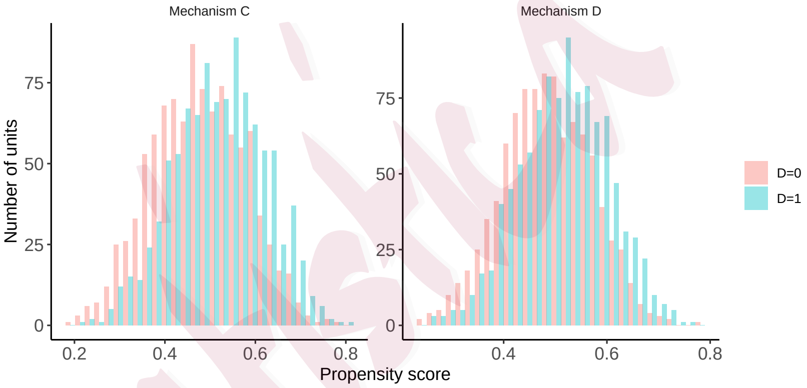

Figure 2: Distribution of true propensity score under the treatment assignment mechanisms C and D ($N = 2000, p = 5$).

We present the results for Simulations V–VIII and contrast them with Simulations I–IV in Table 2 ($N = 2000, p = 5$). Each row illustrates the results for a pair of simulations that differ in the mechanisms for propensity scores only. We found there are no substantial differences in Bias (SD),

Table 2: The performance of PSR, IPW, AIPW, and Matching variant of PSR under different simulation settings $(N = 2000, p = 5)$

.

Pair	Case	Bias (SD)	MAE	MSE	CP95	Case	Bias (SD)	MAE	MSE	CP95
		$\times 10^{-2}$	$\times 10^{-2}$	$\times 10^{-2}$	%		$\times 10^{-2}$	$\times 10^{-2}$	$\times 10^{-2}$	%
PSR method										
(1)	I	-0.5 (9.4)	7.5	0.9	94.4	V	-0.2 (8.9)	6.9	0.8	93.7
(2)	II	0.2 (8.4)	6.6	0.7	93.6	VI	0.0 (7.7)	6.1	0.6	94.0
(3)	III	3.2 (10.5)	8.6	1.2	93.9	VII	1.1 (8.3)	6.7	0.7	93.2
(4)	IV	1.6 (13.0)	10.2	1.7	95.2	VIII	1.1 (12.1)	9.6	1.5	94.1
IPW method										
(1)	I	0.6 (33.9)	14.7	11.5	92.7	V	0.6 (11.5)	8.6	1.3	91.1
(2)	II	0.3 (31.1)	14.2	9.7	93.7	VI	0.0 (10.9)	8.1	1.2	91.8
(3)	III	-0.1 (53.8)	20.7	28.9	94.2	VII	-0.2 (10.7)	8.1	1.1	92.1
(4)	IV	2.2 (49.0)	22.3	24.1	93.9	VIII	2.4 (11.7)	9.1	1.4	91.2
AIPW method										
(1)	I	0.4 (29.8)	17.0	8.9	94.9	V	-0.2 (13.7)	10.4	1.9	94.0
(2)	II	-0.2 (26.9)	17.0	7.2	94.6	VI	-0.0 (13.9)	10.4	1.9	94.3
(3)	III	0.8 (66.7)	25.6	44.6	94.0	VII	-0.1 (13.3)	10.1	1.8	93.8
(4)	IV	0.5 (61.8)	25.9	38.2	95.0	VIII	0.6 (14.0)	10.8	2.0	94.1
Matching variant of PSR										
(1)	I	0.4 (22.7)	17.7	5.2	93.4	V	0.0 (18.0)	13.9	3.2	96.1
(2)	II	0.5 (22.8)	17.8	5.2	93.2	VI	0.3 (17.9)	13.8	3.2	96.2
(3)	III	0.7 (29.3)	23.1	8.6	90.4	VII	0.2 (18.0)	14.0	3.2	95.8
(4)	IV	0.7 (29.2)	23.0	8.5	90.5	VIII	0.7 (17.8)	13.9	3.2	95.8

MAE, and MSE when PSR is used. However, when IPW, AIPW and the matching variant of PSR are used, SD, MAE, and MSE in settings with extreme values of propensity scores have substantial differences from those in settings with general propensity scores. This indicates that PSR is indeed robust to extreme values of propensity scores. In addition, we found that the matching variant of PSR performs close to our PSR in terms of bias, but has larger SD, MAE, and MSE. This is because both our proposed PSR and its matching variant leverage the idea of propensity score matching. However, unlike the matching method, PSR utilizes propensity scores in a non-parametric manner and smoothly estimate $\beta(x^l, e)$. Thus, it is less sensitive to minor differences between the inexact match. This explains why PSR has smaller SD, MAE, and MSE than its matching variant.

Table 3: The performance of PSR under different estimation errors of propensity scores ($N = 2000, p = 5$).

	Simulation I				Simulation II				Simulation III				Simulation IV			
Method	Bias (SD)	MAE	MSE	CP95	Bias (SD)	MAE	MSE	CP95	Bias (SD)	MAE	MSE	CP95	Bias (SD)	MAE	MSE	CP95
Method	$\times 10^{-2}$	$\times 10^{-2}$	$\times 10^{-2}$	%	$\times 10^{-2}$	$\times 10^{-2}$	$\times 10^{-2}$	%	$\times 10^{-2}$	$\times 10^{-2}$	$\times 10^{-2}$	%	$\times 10^{-2}$	$\times 10^{-2}$	$\times 10^{-2}$	%
Logistic	-0.5 (9.4)	7.5	0.9	94.4	0.2 (8.4)	6.6	0.7	93.6	3.2 (10.5)	8.6	1.2	93.9	1.6 (13.0)	10.2	1.7	95.2
Probit	-0.1 (9.6)	7.5	0.9	94.1	0.1 (8.9)	6.9	0.8	93.9	3.2 (10.6)	8.7	1.2	93.8	1.7 (13.1)	10.3	1.7	94.6
Random Forest	0.2 (9.6)	7.6	0.9	91.4	0.0 (8.7)	6.9	0.8	92.7	3.0 (10.6)	8.8	1.2	91.2	2.5 (12.2)	10.0	1.6	91.6

4.2 Alternative estimation methods on propensity scores

We consider two alternative scenarios in estimating propensity scores. In the first scenario, we estimate the propensity scores using a probit model; In the second scenario, we estimate the propensity scores non-parametrically, using the random forest method (R package **grf**). We compare the two with the baseline scenario where propensity scores are estimated through the true logit model. We found that the obtained Bias, SD, MAE, and MSE are all very close under the three scenarios.

5. Application

We illustrate our method in two studies using the data from the National 2009 H1N1 Flu Survey (NHFS). The NHFS was a large one-time telephone survey conducted in the USA from October 2009 through June 2010, by the Centers for Disease Control and Prevention (CDC). The survey asked questions on participants' seasonal influenza vaccination status, whether having been sick with the influenza-like-illness in the past month, the number of days off work due to influenza, whether having paid sick leave benefits, and the number of times seeing doctors, as well as other relevant information (e.g., the influenza-related behaviors, opinions about influenza vaccine safety and effectiveness, the number of household, and demographic characteristics) (Centers for Disease Control and Prevention, 2010b). The NHFS public dataset was released by the CDC, National Center for Immunization and Respiratory Diseases (NCRID), and National Center for Health Statistics (NCHS). The datasets were used to analyze the vaccination coverage, vaccination beliefs, and behaviors in literature (Centers for Disease Control and Prevention, 2010a; Ding et al., 2011; Burger et al., 2021). Using a subset of the data with adults (i.e., age ≥ 18) and English-speaking participants, we conduct two studies, as detailed in the following.

5.1 Effect of seasonal influenza vaccination on taking sick days

In the study, we estimate the effect of seasonal influenza vaccination on taking sick days due to the influenza-like illness. Our primary outcome is the number of days off work (after taking "log") when sick with the influenza-like illness, which was reported by the participants during the interview. For the study, we consider a subset of the data, including only the adult participants who had reported being infected with the influenza-like illness, and excluding participants with missing outcomes or treatments. We consider covariates that have non-missing values for at least 70% of participants. For each selected covariate, the missing value was treated as a new category. Our final sample comprised 2442 participants and 65 (i.e., $p = 65$) dimensional covariates, where 1145 individuals have seasonal flu vaccination. We include all the informative covariates in the analysis, because conditioning on any given covariates is generally better than not conditioning (Rosenbaum, 2002; Rubin, 2009; Ding and Miratrix, 2015b). Nevertheless, the potential bias introduced by adjustment needs further attention (Pearl, 2015; Ding and Miratrix, 2015a). Descriptive statistics of the sample and covariates are listed in Supplementary Table S2.

For our approach, we followed the three steps described in Section 3.1. We estimate the propensity scores using the logistic models and the heterogeneous effects using the standard local linear regression as described in Section 3.2. We display the results in Figure 3.

As shown in Figure 3a, the propensity score values vary from 0.001 to 0.994 for the group of people with seasonal vaccination and 0.001 to 0.978 for the group of people without seasonal vaccination. Therefore, methods that are sensitive to propensity scores are not applicable here. Using our PSR method, we show that the effects of seasonal vaccination on taking sick

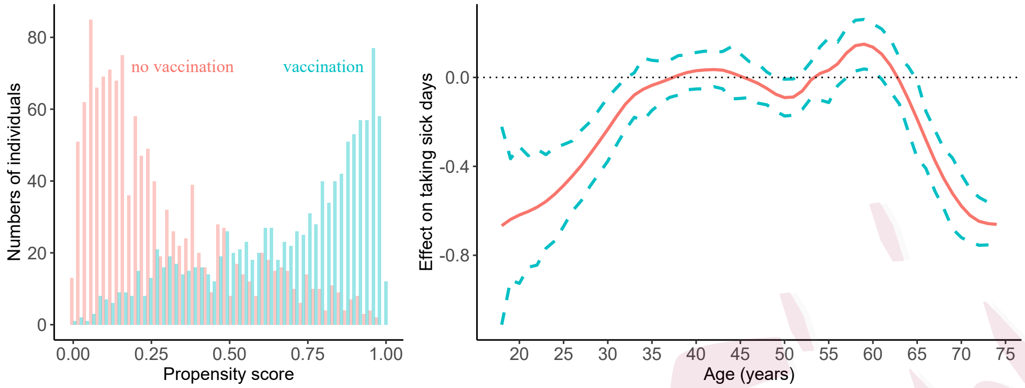

Figure 3: (a) Distribution of the propensity scores. (b) Effects of seasonal vaccination on taking sick days across age groups. Dashed lines refer to 95% confidence interval.

days indeed vary across age groups (Figure 3b). Vaccination with seasonal vaccines decreases the number of sick days off work for the population aged over 60 and under 35, while it has a negligible impact for the remaining age groups. Because seasonal vaccines are able to prevent the developing of severe symptoms of influenza (Deiss et al., 2015), in general, people with seasonal vaccination are more likely to develop light symptoms and therefore are less likely to leave off work when sick with influenza. However, our results suggest that people aged 36–59 are equally likely to leave off work to see doctors regardless of the severity of the symptoms.

5.2 Effect of having paid sick leave on visiting doctors

In the second study, we estimate the effect of having paid sick leave on the number of times seeing doctors regardless of the disease types. Our treatment is whether the adult earns paid sick time off from employment, with $D = 1$ representing having paid sick leave and $D = 0$ not having paid sick leave. The primary outcome Y is the self-reported number of times seeing a doctor. During the interview, the participants were asked to provide an estimated number of times seeing a doctor or other health professional about health since August 2009. For the study, we also consider a subset of the data, including only adults whose paid sick leave indicator D is known. Our final sample comprised 8425 participants and the 62 (i.e., $p = 62$) dimensional covariates, where 5502 have paid sick leave and 2923 do not. Descriptive statistics of the sample and covariates are listed in Supplementary Table S3. Similarly, we estimate the propensity scores using the logistic models and the heterogeneous effects using the standard local linear regression as described in Section 3.2. We display the results in Fig. 4.

As in the first study, the propensity scores are distributed with values varying from 0.033 to 0.984 for the group of people with paid sick leave and 0.001 to 0.956 for the group of people without paid sick leave. The effects

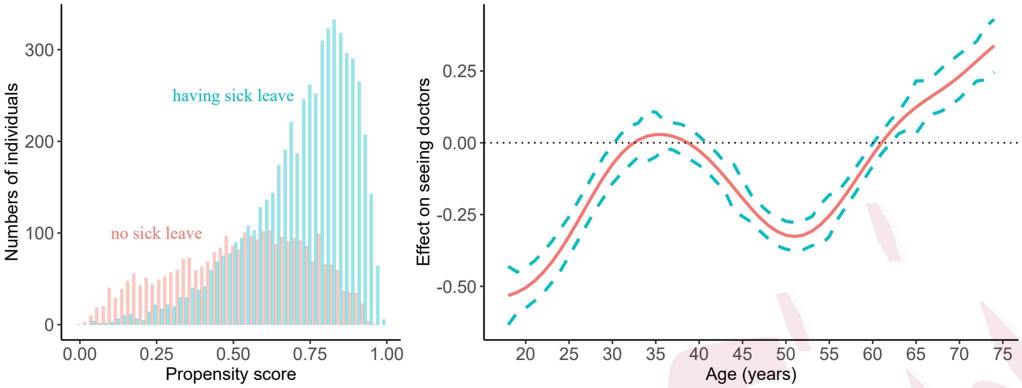

Figure 4: (a) Distribution of the propensity scores. (b) Effects of having paid sick leave on seeing doctor across age groups. Dashed lines refer to 95% confidence interval.

of having paid sick leave vary substantially across age groups (Fig. 4b). Specifically, having paid sick leave indeed increases the times of seeing doctors for people aged over 65, but has no significant effects for people aged 33-40. Interestingly, having paid sick leave motivates people aged under 33 or aged 41-64 to reduce the times of seeing doctors. One possible explanation is that these people may have a bundle of paid-time-off benefits that combines sick days, vacation days, and other types of leave while reducing the times of visiting doctors may increase their overall benefits (Zhai et al., 2018; Smith and Kim, 2010).

6. Discussion

In the paper, we propose non-parametric propensity score regression for estimating the heterogeneous treatment effects in a wide context, including the settings where the propensity scores are close to 0 or 1 and the number of full covariates is large. We have established its large sample properties. We show that it has better performance than the existing methods using simulation studies. Although in the main text we consider the continuous X^l, our methods hold generally regardless of the type of variable X^l. In §7 of the Supplementary Material, we present the theoretical results also for the discrete X^l, where we use the typical kernel smoothing method for estimation (Aitchison and Aitken, 1976; Li and Racine, 2010).

We emphasize that our theoretical results remain when the parametric model specification of propensity score is replaced by the semiparametric model, e.g., the single index model. In this case, we may simply use $X^\intercal\hat{\alpha}$ instead of $e(X)$. This is because PSR is built on the balancing property of the propensity score and any one-to-one function of propensity scores has the same balancing property. For a single index model, many available estimators can satisfy the assumptions in Assumption 1 with *unknown* link functions. See literatures, e.g., Horowitz and Härdle (1996); Ichimur (1993); Klein and Spady (1993); Härdle et al. (1997); Wang and Yang (2009), for detail. Also, methods that can improve the balance property (Huang and Chan, 2017; Wei et al., 2017; Tan, 2020; Imai and Ratkovic, 2014; Ning et al., 2020) could be useful for improving the performance of PSR.

For nonparametric estimation, we use kernel-based methods in the study. Acknowledgedly, for these methods, appropriate choice of the bandwidths, i.e., h_1, h_2 and h_3, is important to achieve good accuracy. In the simulation, we simply use the existing bandwidth-selection methods (Li and Racine, 2007) and the bandwidths are chosen independently in the corresponding varying coefficient models or local linear models. To improve accuracy, approaches that are capable of simultaneously accounting for the bandwidth choice in estimating $\beta(X^l, e)$ and $\tau(X^l)$ are probably helpful.

Finally, we have focused on continuous outcomes. But PSR is not restricted to continuous outcome. Note that $\tau(x^l) = \mathbb{E}[\beta(X^l, e)|X^l = x^l]$ can always be estimated via local linear regression of $\beta(X^l, e)$ on X^l, regardless of the type of outcome. However, when PSR is used for the discrete outcome, particular care is needed for the potential model extrapolation when estimating $\beta(X^l, e)$. Because in the context of discrete outcomes, the estimation of the model (3.7), $Y = \beta(X^l, e) \cdot D + \mathbb{E}\left(Y(0)|X^l, e\right) + \xi$, are often done separately for each treatment group, instead of two treatment groups together. We note that $\beta(X^l, e)$ can be rewritten with two components,

$$\beta(X^l, e) = \mathbb{E}\left(Y(1) - Y(0)|X^l, e\right) = \mathbb{E}\left(Y(1)|X^l, e\right) - \mathbb{E}\left(Y(0)|X^l, e\right).$$

As such, alternative methods are needed to estimate $\beta(X^l, e)$ without separating the two treatment groups. We leave that work for the future.

Supplementary Materials

Supplementary Material available online includes technical proofs, additional numerical results from the simulation study and empirical application, and extensions of the proposed method.

Acknowledgements

The authors thank AE and anonymous reviewers for their helpful comments and valuable suggestions on this article. This research was supported by the National Natural Science Foundation of China (No. 11971064, 12071015, and 12171374) and the Major Project of National Statistical Science Foundation of China (No. 2021LD01).

References

Abrevaya, J., Y. C. Hus, and R. P. Lieli (2015). Estimating conditional average treatment effect. *Journal of Business and Economic Statistics 33*, 485–505.

Aitchison, J. and C. Aitken (1976). Multivariate binary discrimination by the kernel method. *Biometrika 63*(3), 413–420.

Burger, A. E., E. N. Reither, S.-E. Mamelund, and S. Lim (2021). Black-white disparities in 2009 h1n1 vaccination among adults in the united states: A cautionary tale for the covid-19 pandemic. *Vaccine 39*, 943–951.

Centers for Disease Control and Prevention (2010a). *Interim results: state-specific influenza A (H1N1) 2009 monovalent vaccination coverage-United States, October 2009-January 2010*. Centers for Disease Control and Prevention.

Centers for Disease Control and Prevention (2010b). *National 2009 H1N1 Flu Survey Questionnaire [2010 Q1]*. Centers for Disease Control and Prevention.

Deiss, R. G., J. C. Arnold, W. J. Chen, S. Echols, M. P. Fairchok, C. Schofield, P. J. Danaher, E. McDonough, M. Ridoré, D. Mor, T. H. Burgess, and E. V. Millar (2015). Vaccine-associated reduction in symptom severity among patients with influenza a/h3n2 disease. *Vaccine 33*, 7160–7167.

Ding, H., T. A. Santibanez, D. J. Jamieson, C. M. Weinbaum, G. L. Euler, L. A. Grohskopf, P.-J. Lu, and J. A. Singleton (2011). Influenza vaccination coverage among pregnant women–national 2009 h1n1 flu survey (nhfs). *American Journal of Obstetrics and Gynecology 204*, 96–106.

Ding, P. and L. W. Miratrix (2015a). Reply to professor pearl's comment. *Journal of Causal Inference 3*, 251–252.

Ding, P. and L. W. Miratrix (2015b). To adjust or not to adjust? sensitivity analysis of m-bias and butterfly-bias. *Journal of Causal Inference 3*, 41–57.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. Chapman and Hall/CRC.

Fan, J. and Q. W. Yao (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika 85*, 645–660.

Fan, J. and W. Zhang (1999). Statistical estimation in varying coefficient models. *Annals of Statistics 27*, 1491–1518.

Fan, Q., Y. C. Hsu, R. P. Lieli, and Y. Zhang (2022). Estimation of conditional average treatment effects with high-dimensional data. *Journal of Business and Economic Statistics 40*, 313–327.

Gu, L. and L. Yang (2015). Oracally efficient estimation for single-index link function with simultaneous confidence band. *Electronic Journal of Statistics 9*, 1540–1561.

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

Härdle, W., V. Spokoiny, and S. Sperlich (1997). Semiparametric single index versus fixed link function modelling. *Annals of Statistics 25*, 212–243.

Hastie, T. and R. Tibshirani (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B 55*, 757–796.

Horowitz, J. L. and W. Härdle (1996). Direct semiparametric estimation of single-index models with discrete covariates. *Journal of the American Statistical Association 91*, 1632–1640.

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

Ichimur, H. (1993). Semiparametric least squares (sls) and weighted sls estimation of single-index models. *Journal of Econometrics 58*, 71–120.

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

Imai, K. and A. Strauss (2011). Estimation of heterogeneous treatment effects from randomized experiments, with application to the optimal planning of the get-out-the-vote campaign. *Political Analysis 19*, 1–19.

Imbens, G. W. and D. B. Rubin (2015). *Causal inference for statistics social and biomedical science*. Cambridge University Press.

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

Kent, D. M., E. Steyerberg, and D. van Klaveren (2018). Personalized evidence based medicine: predictive approaches to heterogeneous treatment effects. *The British Medical Journal 363*, k4245.

Klein, R. and R. H. Spady (1993). An efficient semiparametric estimator for binary response models. *Econometrica 61*, 387–421.

Lechner, M. (2019). Modified causal forests for estimating heterogeneous causal effects. *arxiv.org/abs/1812.09487v2*.

Lee, S., R. Okui, and Y. J. Whang (2017). Doubly robust uniform confidence band for the conditional average treatment effect function. *Journal of Applied Econometrics 32*, 1207–1225.

Li, Q. and J. S. Racine (2007). *Nonparametric econometrics*. Princeton University Press.

Li, Q. and J. S. Racine (2010). Smooth varying-coefficient estimation and inference for qualitative and quantitative data. *Ecomometric Theory 26*(6), 1607–1637.

Little, R. and H. An (2004). Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica 14*, 949–968.

Mammen, E., C. Rothe, and M. Schienle (2012). Nonparametric regression with nonparametrically generated covariates. *The Annals of Statistics 40*, 1132–1170.

Nie, X. and S. Wager (2021). Quasi-oracle estimation of heterogeneous treatment effects. *Biometrika 108*, 299–319.

Ning, Y., P. Sida, and K. Imai (2020). Robust estimation of causal effects via a high-dimensional covariate balancing propensity score. *Biometrika 107*, 533–554.

Pearl, J. (2015). Comment on ding and miratrix: “to adjust or not to adjust?”. *Journal of Causal Inference 3*, 59–60.

Racine, J. S. and T. Hayfield (2021). *np: Nonparametric Kernel Smoothing Methods for Mixed Data Types*. https://CRAN.R-project.org/package=np.

Rosenbaum (2002). *Observational Studies*. Springer.

Rosenbaum, P. R. and D. B. Rubin (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika 70*, 41–55.

Rotnitzky, A. and S. Vansteelandt (2014). *Doublerobust methods*, pp. 185–212. in ”Handbook of Missing Data Methodology”. CRC Press, Boca Raton, FL.

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

Rubin, D. B. (1980). Discussion of randomization analysis of experimental data in the fisher randomization test by basu. *Journal of the American Statistical Association 75*, 591–593.

Rubin, D. B. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation. *Health Services and Outcomes Research Methodology 2*, 169–188.

Rubin, D. B. (2009). Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine 28*, 1420–1423.

Ruppert, D., S. J. Sheather, and M. P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association 90*, 1257–1270.

Sato, M., J. Singh, S. Takemori, T. Sonoda, Q. Zhang, and T. Ohkuma (2019). Uplif-based evaluation and optimization of recommenders. In *Conference on Recommender Systems*.

Seaman, S. R. and S. Vansteelandt (2018). Introduction to double robust methods for incomplete data. *Statistical Science 33*, 184–197.

Semenova, V. and V. Chernozhukov (2021). Debiased machine learning of conditional average treatment effects and and other causal functions. *The Econometrics Journal 24*, 264–289.

Smith, T. W. and J. Kim (2010). *Paid Sick Days: Attitudes and Experiences*. Public Welfare Foundation.

Soiza, R. L., C. Scicluna, and E. C. Thomson (2021). Efficacy and safety of covid-19 vaccines in older people. *Age and Ageing 50*, 279–283.

Tan, Z. (2007). Comment: understanding or, ps and dr. *Statistical Science 22*, 560–568.

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

Velentgas, P., N. A. Dreyer, P. Nourjah, S. R. Smith, and M. M. Torchia (2013). *Developing a Protocol for Observational Comparative Effectiveness Research: A User's Guide*. Agency for Healthcare Research and Quality (AHRQ).

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*, 1228–1242.

Wand, M., C. Moler, and B. Ripley (2021). *KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones (1995)*. https://CRAN.R-project.org/package=KernSmooth.

Wang, L. and L. Yang (2009). Spline estimation of single index model. *Statistica Sinica 19*, 765–783.

Wei, L., Z. Yeying, and G. Debashis (2017). On estimating regression-based causal effects using sufficient dimension reduction. *Biometrika 104*, 51–65.

Wu, P., Z. Tan, W. Hu, and X.-H. Zhou (2022). Model-assisted inference for covariate-specific treatment effects with high-dimensional data. *Statistica Sinica (To Appear)*.

Wu, P., X. Tong, Y. Wang, J. Liang, and X.-H. Zhou (2022). Robust quasi-oracle estimation of average causal effects. *Biostatistics and Epidemiology 6*, 144–163.

Wu, P., X. Xu, X. Tong, Q. Jiang, and B. Lu (2021). Semiparametric estimation for average causal effects using propensity score-based spline. *Journal of Statistical Planning and Inference 212*, 153–168.

Yin, Y. (2018). Assessing the treatment effect heterogeneity with a latent variable. *Statistica Sinica 28*, 115–135.

Zhai, Y., T. A. Santibanez, K. E. Kahn, C. L. Black, and M. A. de Perio (2018). Paid sick leave benefits, influenza vaccination, and taking sick days due to influenza-like illness among u.s. workers. *Vaccine 36*, 7316–7323.

Zhang, G. and R. Little (2009). Extensions of the penalized spline of propensity prediction method of imputation. *Biometrics 65*, 911–918.

Zhou, T., M. R. Elliott, and R. J. A. Little (2019). Penalized spline of propensity methods for treatment comparison. *Journal of the American Statistical Association 114*, 1–19.

Zimmert, M. and M. Lechner (2019). Nonparametric estimation of causal heterogeneity under high-dimensional confounding. *https://arxiv.org/abs/1908.08779v1*.