econometrics

MDPI

Article

Efficiency of Average Treatment Effect Estimation When the True Propensity Is Parametric

Kyoo il Kim

Department of Economics, Michigan State University, 486 W. Circle Dr., East Lansing, MI 48824, USA; kyookim@msu.edu; Tel.: +1-517-353-9008

Received: 8 March 2019; Accepted: 28 May 2019; Published: 31 May 2019

Abstract: It is well known that efficient estimation of average treatment effects can be obtained by the method of inverse propensity score weighting, using the estimated propensity score, even when the true one is known. When the true propensity score is unknown but parametric, it is conjectured from the literature that we still need nonparametric propensity score estimation to achieve the efficiency. We formalize this argument and further identify the source of the efficiency loss arising from parametric estimation of the propensity score. We also provide an intuition of why this overfitting is necessary. Our finding suggests that, even when we know that the true propensity score belongs to a parametric class, we still need to estimate the propensity score by a nonparametric method in applications.

Keywords: average treatment effect; efficiency bound; propensity score; sieve MLE

JEL Classification: C14; C18; C21

1. Introduction

Estimating treatment effects of a binary treatment or a policy has been one of the most important topics in evaluation studies. In estimating treatment effects, a subject's selection into a treatment may contaminate the estimate, and two approaches are popularly used in the literature to remove the bias due to this sample selection. One is regression-based *control function* method (see, e.g., Rubin (1973); Hahn (1998); and Imbens (2004)) and the other is *matching* method (see, e.g., Rubin and Thomas (1996); Heckman et al. (1998); and Abadie and Imbens (2002, 2006)). When there are many covariates or pre-treatment variables that govern this selection, the matching method may be less practical. In this case, due to Rosenbaum and Rubin (1983, 1984), we can control for the sample selection bias using the propensity score to reduce the dimensionality problem.

Although adjusting for sub-population differences in the propensity score removes the bias, the resulting treatment effect estimators may not be all efficient. Hahn (1998) shows that, using a nonparametric series estimation of the propensity score, we can achieve the efficiency bound. Hirano et al. (2003) also develop an efficient estimation of average treatment effects using the logit series estimation of the propensity score overcoming some practical limitations of Hahn (1998)'s series estimator (see also Li et al. (2009)).

Based on these studies, empirical researchers are encouraged to estimate treatment effects using the imputation method of the inverse weighting of the estimated propensity score. However, a nonparametric method of estimating the propensity score may require a large data set, especially when covariates or pre-treatment variables are high dimensional. For this reason, many empirical researchers estimate the propensity score parametrically using the probit or logit specification, given the idea that these parametric models are still good approximations to the true propensity score. Also in the statistics literature such as Rosenbaum (1987); Rubin and Thomas (1996); and Robins et al. (1995),

they show that using parametric estimates of the propensity score can improve the efficiency of the treatment effect estimation.

However, from the existing literature (Hahn (1998); Hirano et al. (2003); Kang and Schafer (2007); and Tan (2007)), we can infer that, even when the true propensity score is parametric and the parametric estimator is consistent, we still need to estimate the propensity score nonparametrically to achieve the full efficiency. The first contribution of this paper is to formalize this efficiency argument and confirm that parametric estimation of the propensity score yields an inefficient estimator of the average treatment effect if some or all of covariates are continuous.[1] The second contribution of this paper, which is more interesting, is to identify the source of this inefficiency, and formally characterize the efficiency loss due to parametric estimation of the propensity score.

For our results, we find that a nonparametric sieve estimation of the propensity score has two roles in the efficient estimation of average treatment effects. First, it approximates the true propensity score, and second it approximates the conditional expectation of the derivative of the moment function for the treatment effect with respect to the propensity score. We show that parametric estimation of propensity score accomplishes the first role when the true propensity score is indeed parametric, but cannot achieve the second role, if some of covariates are continuous. In other words, consistent estimation of the propensity score alone is not enough to obtain the efficient estimation of average treatment effects.

This finding also suggests that the performance of the treatment effect estimator in finite samples may depend not only on how precisely the propensity score is estimated, but also how well the conditional expectation of the derivative of the moment condition is approximated by the same sieve basis functions or regressors used to estimate the propensity score. We note that the literature has focused on the former, but the latter has been somewhat ignored. Moreover, because these two objects are quite different in nature, a sieve approximation solely targeted for the propensity score does not necessarily well approximate the conditional expectation of the derivative of the moment function in finite samples.

The rest of the paper is organized as follows. Section 2 outlines the average treatment effect estimation using the inverse propensity score weighting. Section 3 examines the role of the nonparametric propensity score estimation when the true one is parametric. We also provide an illustrative example. We conclude in Section 4. Some technical details are provided in Appendix A.

2. Estimation of Average Treatment Effect

In this section, we review estimation of average treatment effects using the inverse propensity score weighting in a standard setting. For this purpose, suppose we have a random sample of size n individuals where some of them received a treatment and others did not. Let T_i denote the treatment status with $T_i = 1$ if individual i receives the treatment and $T_i = 0$ otherwise. Using the same notation with Rubin (1973), denote $Y_i(0)$ as the potential outcome for each individual i under control and $Y_i(1)$ as the outcome under treatment. We observe T_i, X_i, and $Y_i = T_iY_i(1) + (1 - T_i)Y_i(0)$ where X_i is a vector of observable covariates of the individual. Here, we have a fundamental missing data problem since we observe only either $Y_i(1)$ or $Y_i(0)$ but not both for each individual depending on the treatment status.

The parameter of interest is the population average treatment effect defined as

$$\tau^* = E[Y(1) - Y(0)].$$

[1] As it is pointed out by one referee, in the literature, there have been several studies, related to our findings, that show using an estimated nuisance parameter rather than the true value improves the efficiency of the parameter estimate of interest (see, e.g., Prokhorov and Schmidt (2009); Hitomi et al. (2008); and Hristache and Patilea (2017)). Our work provides a new insight to this problem by illustrating parametric estimation of the nuisance parameter may not achieve the full efficiency.

If we had both $Y_i(1)$ and $Y_i(0)$ for all individuals, we can simply estimate the average treatment effect using its sample analogue, but it is not feasible due to the missing data problem. One important way to circumvent this missing data problem in the literature is using the imputation method based on the propensity score, motivated by Rosenbaum and Rubin (1983, 1984). The propensity score of an individual whose observable characteristics X_i equals x is defined by

$$p^*(x) = \Pr(T_i = 1|X_i = x) \text{ or } E[T_i|X_i = x].$$

According to Rosenbaum and Rubin, if (i) there exist covariates X_i such that the treatment status T_i is ignorable given X_i and (ii) $0 < p^*(x) < 1$ for all $x \in \mathcal{X} \equiv \text{Supp}(X)$, then T_i and $(Y_i(0), Y_i(1))$ are independent of each other given the propensity score. This implies that

$$\tau^* = E\left[E[Y_i|T_i = 1, p^*(X_i)] - E[Y_i|T_i = 0, p^*(X_i)]\right]. \tag{1}$$

This allows us to estimate the treatment effect using a sample analogue of Equation (1). To be precise, define

$$\widehat{\beta}_1(x) = \frac{\widehat{E}[T_iY_i|X_i = x]}{\widehat{E}[T_i|X_i = x]} \text{ and } \widehat{\beta}_0(x) = \frac{\widehat{E}[(1-T_i)Y_i|X_i = x]}{1 - \widehat{E}[T_i|X_i = x]},$$

where $\widehat{E}[\cdot|\cdot]$'s denote suitable conditional mean function estimators. Then, we can construct *complete* data using the imputation such that $\widehat{Y}_i(1) \equiv T_iY_i + (1-T_i)\widehat{\beta}_1(X_i)$ and $\widehat{Y}_i(0) \equiv T_i\widehat{\beta}_0(X_i) + (1-T_i)Y_i(0)$, and we can estimate the average treatment effect as $\widehat{\tau}_1 = \frac{1}{n}\sum_{i=1}^n(\widehat{Y}_i(1) - \widehat{Y}_i(0))$ or alternatively as $\widehat{\tau}_2 = \frac{1}{n}\sum_{i=1}^n(\widehat{\beta}_1(X_i) - \widehat{\beta}_0(X_i))$. These nonparametric imputation methods were proposed by Hahn (1998), and he further shows that these treatment effect estimators achieve the semiparametric efficiency bound.[2]

Hirano et al. (2003) propose an alternative estimator for which the propensity score is estimated using a logit series estimation, and the propensity score is given by $p^*(x) = \frac{\exp(h_0(x))}{1+\exp(h_0(x))}$ for some unknown function $h_0(x)$. In the logit series estimation, we approximate $h_0(x)$ using linear sieves and the estimated propensity score is given by $\widehat{p}_L(x) = \frac{\exp(\widehat{h}_n(x))}{1+\exp(\widehat{h}_n(x))}$, where $\widehat{h}_n(x)$ denotes the sieve Maximum Likelihood (ML) estimator.[3] The proposed treatment effect estimator is given by $\widehat{\tau}_3 = \frac{1}{n}\sum_{i=1}^n\left(\frac{T_iY_i}{\widehat{p}_L(X_i)} - \frac{(1-T_i)Y_i}{1-\widehat{p}_L(X_i)}\right)$. This estimator also achieves the semiparametric efficiency bound, and improves over Hahn (1998)'s estimator in two practical ways. First, we do not need to estimate the conditional mean functions of $\widehat{E}[T_iY_i|X_i]$ and $\widehat{E}[(1-T_i)Y_i|X_i]$. Second, the estimated propensity score lies between zero and one by construction.

Estimation of average treatment effects using the estimated propensity score with a general link function that includes the logit or probit specification was proposed by Kim (2013). We will use this general setting to argue that the inefficiency of the treatment effect estimate with the estimated parametric propensity score is not specific to a particular functional form assumption like logit or probit. To obtain a sieve ML estimator for the propensity score with a general link function, we assume

---

[2] Hahn (1998) proposes to estimate $\widehat{E}[T_iY_i|X_i]$, $\widehat{E}[(1-T_i)Y_i|X_i]$, and $\widehat{E}[T_i|X_i]$ using series estimations (e.g., Newey (1997)). The resulting treatment effect estimators, however, are subject to some practical issues, e.g., the propensity score estimate $\widehat{E}[T_i|X_i]$ may lie outside the zero and one interval.

[3] See, e.g., Shen and Wong (1994) and Chen and Shen (1998) for further details on the sieve extremum estimations that include the sieve ML estimation.

the true function h_0 belongs to a class of bounded and smooth functions such as a Hölder ball, and let $p^*(x) = F(h_0(x))$ for a known link function $F(\cdot)$.[4]

Then, based on a triangular sequence of orthonormal basis functions such as polynomials or splines, we construct a tensor-product sieve space $\mathcal{H}_n$ as

$$\mathcal{H}_n = \{h(X) | h(X) = R^{K(n)}(X)'\pi \text{ for all } \pi \text{ satisfying } \|h\|_{\Lambda^{\gamma_1}} \leq c_1\},$$

where $\|\cdot\|_{\Lambda^{\gamma_1}}$ denotes a Hölder norm, and we let $K(n) \to \infty$ as $n \to \infty$. The sieve ML estimator is obtained by solving

$$\widehat{h}_n = \text{argmax}_{h \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \log \left\{ F(h(X_i))^{T_i} \left(1 - F(h(X_i))\right)^{1-T_i} \right\} \tag{2}$$

or equivalently $\widehat{\pi}_K = \text{argmax}_{\pi, R^K(X)'\pi \in \mathcal{H}_n} \frac{1}{n} \sum_{i=1}^{n} \log \left\{ F(R^K(X_i)'\pi)^{T_i} \left(1 - F(R^K(X_i)'\pi)\right)^{1-T_i} \right\}$ such that $\widehat{h}_n(x) = R^K(x)'\widehat{\pi}_K$, and the resulting propensity score estimator becomes $\widehat{p}(x) = F(\widehat{h}_n(x))$.

Finally, using the estimated propensity score, we estimate the average treatment effect as

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{Y_i T_i}{\hat{p}(X_i)} - \frac{Y_i(1 - T_i)}{1 - \hat{p}(X_i)} \right).$$

Define $\mu_t(x) \equiv E[Y(t)|X = x]$ and $\sigma_t^2(x) \equiv \text{Var}[Y(t)|X = x]$. For the general class of $F(\cdot)$, as long as the function is continuous and monotonic in h, Kim (2013) shows that this treatment effect estimator achieves the semiparametric efficiency bound such that

$$\sqrt{n}(\hat{\tau} - \tau^*) \xrightarrow{d} \mathcal{N}(0, V),$$

where $\tau(X) = E[Y(1) - Y(0)|X]$ and

$$\begin{aligned} V &= E\left[\left(\left(\frac{YT}{p^*(X)} - \frac{Y(1-T)}{1-p^*(X)} - \tau^* \right) - \left(\frac{\mu_1(X)}{p^*(X)} + \frac{\mu_0(X)}{1-p^*(X)} \right) (T - p^*(X)) \right)^2 \right] \\ &= E\left[(\tau(X) - \tau^*)^2 + \frac{\sigma_1^2(X)}{p^*(X)} + \frac{\sigma_0^2(X)}{1-p^*(X)} \right], \end{aligned} \tag{3}$$

---

[4] The Hölder space is a space of functions $g \in \Lambda^{\gamma}(\mathcal{X})$, $g : \mathcal{X} \longrightarrow \mathcal{R}$ such that the first $\underline{\gamma}$ derivatives are bounded, and the γ-th derivatives are Hölder continuous with the exponent $\gamma - \underline{\gamma} \in (0, 1]$, where $\underline{\gamma}$ is the largest integer smaller than γ. The Hölder space becomes a Banach space when endowed with the Hölder norm:

$$||g||_{\Lambda^{\gamma}} = \sup_x |g(x)| + \max_{a_1 + a_2 + \ldots + a_{d_x} = \underline{\gamma}} \sup_{x \neq x'} \frac{|\nabla^a g(x) - \nabla^a g(x')|}{(||x - x'||_E)^{\gamma - \underline{\gamma}}} < \infty,$$

where $\nabla^a g(x) \equiv \frac{\partial^{a_1 + a_2 + \ldots + a_{d_x}}}{\partial x_1^{a_1} \ldots \partial x_{d_x}^{a_{d_x}}} g(x)$ and $||\cdot||_E$ denotes the Euclidean norm. The Hölder ball $\Lambda_c^{\gamma}(\mathcal{X})$ is defined as $\Lambda_c^{\gamma}(\mathcal{X}) \equiv \{g \in \Lambda^{\gamma}(\mathcal{X}) : ||g||_{\Lambda^{\gamma}} \leq c < \infty\}$.

which is identical to the efficiency bound derived by Hahn (1998). This efficiency result with the general link function is obtained, similarly as in Hirano et al. (2003), following the influence function approach by Newey (1994). To see this, define

$$\begin{aligned} \psi(Z_i, \tau, p(X_i)) &= \left(\frac{Y_i T_i}{p(X_i)} - \frac{Y_i(1-T_i)}{1-p(X_i)} - \tau \right) \\ \psi_p(Z_i, \tau, p(X_i)) &= -\left(\frac{Y_i T_i}{p(X_i)^2} + \frac{Y_i(1-T_i)}{(1-p(X_i))^2} \right) \\ s_p(X_i) &= E[\psi_p(Z_i, \tau^*, p^*(X_i))|X_i], \end{aligned} \tag{4}$$

where $\psi_p(\cdot)$ denotes the derivative of the moment function for the treatment effect, $\psi(\cdot)$, with respect to the propensity score $p(\cdot)$, and $s_p(\cdot)$ denotes its conditional expectation at the true parameter values. The asymptotic variance result of Equation (3) is obtained by showing that the estimator is asymptotically linear with influence function decomposed into two terms:

$$\left| \sqrt{n}(\hat{\tau} - \tau^*) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\psi(Z_i, \tau^*, p^*(X_i)) + s_p(X_i)(T - p^*(X_i)) \right) \right| = o_p(1). \tag{5}$$

The first term in Equation (5) is the influence function when we know the true propensity score $p^*(\cdot)$, and the second term represents the contribution of the estimated propensity score on the asymptotic distribution of $\hat{\tau}$. It follows that the asymptotic variance V in Equation (3) equal to

$$V = \text{Var}\left[\psi(Z_i, \tau^*, p^*(X_i)) + s_p(X_i)(T - p^*(X_i)) \right],$$

which derives the result.

3. Efficient Estimation When the True Propensity Score Is Parametric

As we discuss in the previous section, the efficiency of the treatment effect estimator depends on whether the estimator has the asymptotically linear representation as Equation (5). When the propensity score is estimated using a nonparametric sieve ML, we achieve this representation and hence the efficiency bound. Here, we pose the question of whether we can achieve this asymptotic linear representation if the true propensity is parametric, and is estimated under the correct parametric specification. We confirm that, in this case, the semiparametric efficiency bound is not achieved as can be inferred from the existing literature. This suggests that, even though we know the true propensity score belongs to a parametric class, we still need to estimate the propensity score by a nonparametric method.

Our intuition behind this result is that the nonparametric sieve estimation of the propensity score plays two roles in the estimation of the treatment effect. First, it approximates the true propensity score, and second it approximates the conditional expectation of the derivative of the moment condition for the treatment effect with respect to the propensity score. For the purpose of illustration, without loss of generality, suppose $p^*(x) = \Phi(x'\pi_0)$, where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function (CDF), so the true propensity is a probit model. We then can estimate π_0 with MLE, denoted by $\widehat{\pi}$, and obtain the parametric convergence rate such that $\sqrt{n}\,(\widehat{\pi} - \pi_0) = O_p(1)$ and hence $\sup_{x \in \mathcal{X}} |\hat{p}(x) - p^*(x)| = O_p(n^{-1/2})$ with $\hat{p}(x) = \Phi(x'\widehat{\pi})$.

For ease of notation without losing the main idea, we consider the special case that $Y(0) = 0$ with probability one. Define $\beta_0 = E[Y(1)]$ as the average outcome of interest, where $Y(1)$ is missing at random conditional on the covariates X. We estimate the average outcome as $\hat{\beta} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i T_i}{\hat{p}(X_i)}$. For this estimator, following the Equation (5), if we can obtain the asymptotic linear representation as

$$\left| \sqrt{n}(\hat{\beta} - \beta_0) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{ \left(\frac{Y_i T_i}{p^*(X_i)} - \beta_0 \right) - \frac{\mu_1(X_i)}{p^*(X_i)}(T_i - p^*(X_i)) \right\} \right| = o_p(1), \tag{6}$$

then we will achieve the efficiency bound. To see whether this asymptotic linear representation is attainable with parametric estimation of the propensity score, we decompose $\sqrt{n}(\hat{\beta}-\beta_0)$ as

$$\sqrt{n}(\hat{\beta}-\beta_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\frac{T_iY_i}{\hat{p}(X_i)}-\frac{T_iY_i}{p^*(X_i)}+\frac{T_iY_i}{p^*(X_i)^2}(\hat{p}(X_i)-p^*(X_i))\right) \quad (7)$$

$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(-\frac{T_iY_i}{p^*(X_i)^2}(\hat{p}(X_i)-p^*(X_i))+\int_{\mathcal{X}}\frac{\mu_1(x)}{p^*(x)}(\hat{p}(x)-p^*(x))dF_0(x)\right) \quad (8)$$

$$-\sqrt{n}\int_{\mathcal{X}}\frac{\mu_1(x)}{p^*(x)}(\hat{p}(x)-p^*(x))dF_0(x)-\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\delta^*(X_i)\frac{T_i-p^*(X_i)}{\sqrt{p^*(X_i)(1-p^*(X_i))}} \quad (9)$$

$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\delta^*(X_i)-\delta_0(X_i))\frac{T_i-p^*(X_i)}{\sqrt{p^*(X_i)(1-p^*(X_i))}} \quad (10)$$

$$+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\left(\frac{T_iY_i}{p^*(X_i)}-\beta_0\right)+\delta_0(X_i)\frac{(T_i-p^*(X_i))}{\sqrt{p^*(X_i)(1-p^*(X_i))}}\right), \quad (11)$$

where $p^*(x)=\Phi(x'\pi_0)$, $\hat{p}(x)=\Phi(x'\hat{\pi})$, $F_0(\cdot)$ denotes the distribution function of X, $W=E[\frac{\phi(X_i'\pi_0)^2}{p^*(X_i)(1-p^*(X_i))}X_iX_i']$ with $\phi(\cdot)$ being the standard normal density function, and

$$\delta^*(x) = -\int_{\mathcal{X}}\frac{\mu_1(z)}{p^*(z)}\phi(z'\pi_0)z'dF_0(z)\ W^{-1}\frac{\phi(x'\pi_0)x}{\sqrt{p^*(x)(1-p^*(x))}},$$

$$\delta_0(x) = -\frac{\mu_1(x)}{p^*(x)}\sqrt{p^*(x)(1-p^*(x))}.$$

If we can show that all terms (7)–(10) are $o_p(1)$, we then obtain the desirable result of Equation (6). Following the steps in Hirano et al. (2003) or Kim (2013), it is straightforward to bound the terms (7)–(9) as $o_p(1)$. We focus on the term (10), from which we derive our main finding.

By inspecting $\delta^*(x)$ and $\delta_0(x)$, we see that $\delta^*(x)$ is the linear projection of $\delta_0(x)$ on $\frac{\phi(x'\pi_0)x}{\sqrt{p^*(x)(1-p^*(x))}}$. In other words,

$$\delta^*(x)-\delta_0(x)=\theta_0'\frac{\phi(x'\pi_0)x}{\sqrt{p^*(x)(1-p^*(x))}}-\delta_0(x), \quad (12)$$

where the projection coefficient is given by $\theta_0' \equiv -\int_{\mathcal{X}}\frac{\mu_1(z)}{p^*(z)}\phi(z'\pi_0)z'dF_0(z)W^{-1}$. Therefore, unless $\delta_0(x)$ is indeed linear in $\frac{\phi(x'\pi_0)x}{\sqrt{p^*(x)(1-p^*(x))}}$,[5] we will have $\inf_{x\in\mathcal{X}}|\delta^*(x)-\delta_0(x)|>C>0$ for some positive constant C. It follows that

$$\mathrm{Var}\left[(\delta^*(X_i)-\delta_0(X_i))\frac{T_i-p^*(X_i)}{\sqrt{p^*(X_i)(1-p^*(X_i))}}\right]=E[(\delta^*(X_i)-\delta_0(X_i))^2]>C. \quad (13)$$

Therefore, the term (10) remains as $O_p(1)$ and contributes to the asymptotic distribution of the treatment effect estimator. In other words, the asymptotic linear representation of Equation (6) is not obtained with parametric estimation of the propensity score in general, even when the true propensity score is parametric. In the Appendix, we derive the asymptotic variance of the treatment effect estimator with the estimated parametric propensity score, and characterize the efficiency loss due to parametric estimation of the propensity score. In particular, we show that this efficiency loss is exactly given by Equation (13).

---

[5] When the true propensity score is the logit model instead, this term is replaced by $\sqrt{p^*(x)(1-p^*(x))}\cdot x$ where $p^*(x)=\exp(x'\pi_0)/(1+\exp(x'\pi_0))$.

As the key difference, in the treatment effect estimation using the nonparametric sieve estimation of the propensity score like Equation (2), it can be shown that when $\delta_0(x)$ is t-times continuously differentiable, we have

$$\sup_{x \in \mathcal{X}} |\delta^*(x) - \delta_0(x)| = O(K^{-t/d_x}), \tag{14}$$

where K denotes the number of approximating sieve terms used in $\delta^*(x)$ (see Hirano et al. (2003) or Kim (2013)). Therefore, we can bound the term (10) as $o_p(1)$ for some large enough K. This is because Equation (12) becomes

$$\delta^*(x) - \delta_0(x) = \theta_0' \frac{\phi(R^K(x)'\pi_0)R^K(x)}{\sqrt{p^*(x)(1-p^*(x))}} - \delta_0(x)$$

when the sieve estimation like Equation (2) is used to estimate the propensity score, where $R^K(x)$ denotes a vector of approximating basis functions, and hence the bound (14) is obtained due to some approximation theories of sieves for a class of smooth functions such as a Hölder class (see, e.g., Chen (2007)). We, however, note that, because $p^*(x)$ and $\delta_0(x)$ are quite different in nature, the sieve approximation used to estimate the propensity score does not necessarily well approximate the latter in finite samples, which may contribute to the inefficiency of the treatment effect estimation.

Finally, by inspecting Equations (4) and (5) for the case $Y(0) = 0$ along with Equation (10), note that the term $\delta_0(x)$ is related to the conditional expectation of the derivative of the moment function with respect to the propensity score. This implies that the nonparametric sieve estimation of the propensity score plays two roles in the estimation of the treatment effect. It first approximates the true propensity score, and second approximates the conditional expectation of the derivative of the moment condition with respect to the propensity score. The parametric propensity score estimation can accomplish the first role, if the true one is parametric, but cannot achieve the second when some of covariates are continuous.

The asymptotic variance of a treatment effect estimator using parametric estimation of the propensity score can also depend on which parametric estimator is being used in practice. In this regard, given a parametric model of the propensity score, one can directly derive the asymptotic variance of the treatment effect estimator using the estimated parametric propensity score by combining two moments as a sequential estimation problem (see, e.g., Newey (1984)). The first moment is given by, e.g., the first order condition of the population ML objective function of the propensity score estimation such as the logit or probit ML, and the second moment is given by the moment condition to estimate the treatment effect $E[\psi(Z_i, \tau, p(X_i))] = 0$ defined in Equation (4). We can then directly compare the asymptotic variance of the treatment effect estimator resulting from using a specific parametric estimator of the propensity score to the semiparametric efficiency bound, instead of deriving the inefficiency term from the Equation (13). This joint moments approach for parametric estimation of the propensity score also allows us to explicitly derive the efficiency loss due to a specific parametric estimator of the propensity score, and hence compare different parametric models of the propensity score in terms of efficiency.[6]

3.1. Reconsidering the Simple Example in Hirano et al. (2003)

Hirano et al. (2003) present a simple example with a binary covariate, illustrating that, weighting by the inverse of the propensity score estimate, rather than the true one, we can improve the efficiency and indeed achieve the efficiency bound. Here, we reproduce the example and provide an intuition why in this case the efficiency bound is achieved in view of the results from the previous section. Consider a simple problem of estimating the population average of a variable Y, $\beta_0 = E[Y]$, given a

[6] We thank the referee for this useful suggestion.

random sample of size n of the triple $(T_i, X_i, T_i \cdot Y_i)$. Therefore, T_i and X_i are observed for all units in the sample, but Y_i is only observed if $T_i = 1$. Denote $\mu(x) = E[Y|X = x]$ and $\sigma^2(x) = \text{Var}(Y|X = x)$.

Now let N_{tx} denote the number of observations with $T_i = t$ and $X_i = x$, for $t, x \in \{0, 1\}$. Further assume that the true selection probability is $p(x) = \pi_0 + x(\pi_1 - \pi_0)$.[7] The estimated selection probability is then

$$\widehat{p}(x) = \left\{ \begin{array}{l} N_{10}/(N_{00} + N_{10}) \text{ if } x = 0 \\ N_{11}/(N_{01} + N_{11}) \text{ if } x = 1 \end{array} \right\}. \tag{15}$$

The *true weights* estimator is given by $\widehat{\beta}_{tw} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i T_i}{p(X_i)}$ while the *estimated weights* estimator is then $\widehat{\beta}_{ew} = \frac{1}{n}\sum_{i=1}^{n} \frac{Y_i T_i}{\widehat{p}(X_i)}$. Hirano et al. (2003) show that $\widehat{\beta}_{ew}$ is more efficient than $\widehat{\beta}_{tw}$, and $\widehat{\beta}_{ew}$ achieves the efficiency bound. Interestingly, one can easily see that $\widehat{p}(x)$ in Equation (15) is a nonparametric estimator of $p(x)$, and is also a parametric MLE of $p(x)$ since we can write $\widehat{p}(x) = \widehat{\pi}_0 + x(\widehat{\pi}_1 - \widehat{\pi}_0)$ with $\widehat{\pi}_0 = N_{10}/(N_{00} + N_{10})$ and $\widehat{\pi}_1 = N_{11}/(N_{01} + N_{11})$.

In this example, for the corresponding terms of $\delta^*(x)$ and $\delta_0(x)$ in Equation (10), we show below that indeed

$$\delta^*(x) - \delta_0(x) = 0 \tag{16}$$

for all x, and hence the efficiency bound is achieved for the estimator $\widehat{\beta}_{ew}$ because the asymptotic linear representation like Equation (6) is obtained (i.e., the term (14) is simply equal to zero in this case). To derive the result, consider the following terms corresponding to $\delta^*(x)$ and $\delta_0(x)$ in Equation (10) for the stochastic expansion of $\widehat{\beta}_{ew}$. Let

$$\begin{aligned} \widehat{W} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{X}_i \mathbf{X}_i'}{p(X_i)\,(1 - p(X_i))}, W = E\left[\frac{\mathbf{X}_i \mathbf{X}_i'}{p(X_i)\,(1 - p(X_i))}\right], \mathbf{X}_i \equiv \left(\begin{array}{c} 1 - X_i \\ X_i \end{array} \right), \\ \delta^*(x) &= -\left\{ \sum_x \frac{\mu(x)}{p(x)}\,(1 - x, x)\,q(x) \right\} \cdot W^{-1} \frac{(1 - x, x)'}{\sqrt{p(x)\,(1 - p(x))}}, \delta_0(x) = -\frac{\mu(x)}{p(x)}\sqrt{p(x)(1 - p(x))}, \end{aligned}$$

where $q(\cdot)$ denotes the probability mass of X.

By investigating $\delta^*(x)$ and $\delta_0(x)$, we can see that $\delta^*(x)$ is the linear projection of $\delta_0(x)$ on $\frac{(1-x,x)}{\sqrt{p(x)(1-p(x))}}$. In other words, $\delta^*(x) = \frac{1-x}{\sqrt{p(x)(1-p(x))}}\theta_0 + \frac{x}{\sqrt{p(x)(1-p(x))}}\theta_1$ for some constants θ_0 and θ_1 that are determined by the linear projection. Note that we have

$$\delta^*(x) - \delta_0(x) = \frac{(1 - x)\theta_0 + x\theta_1}{\sqrt{p(x)(1 - p(x))}} + \frac{\mu(x)}{p(x)}\sqrt{p(x)(1 - p(x))} = 0$$

if $\theta_x = -\mu(x)(1 - p(x))$ for $x \in \{0, 1\}$. Indeed, from the definition of $\delta^*(x)$, we find

$$\begin{aligned} (\theta_0, \theta_1) &= -\left\{ \sum_x \frac{\mu(x)}{p(x)}(1 - x, x)q(x) \right\} \cdot W^{-1} \\ &= -\left(\frac{\mu(0)}{p(0)}q(0), \frac{\mu(1)}{p(1)}q(1) \right) \left(\begin{array}{cc} \frac{q(0)}{p(0)(1-p(0))} & 0 \\ 0 & \frac{q(1)}{p(1)(1-p(1))} \end{array} \right)^{-1} \\ &= -\left(\frac{\mu(0)}{p(0)}q(0), \frac{\mu(1)}{p(1)}q(1) \right) \left(\begin{array}{cc} \frac{p(0)(1-p(0))}{q(0)} & 0 \\ 0 & \frac{p(1)(1-p(1))}{q(1)} \end{array} \right) \\ &= -\left(\mu(0)(1 - p(0)), \mu(1)(1 - p(1))\right) \end{aligned}$$

[7] In the orignial example, we have $\pi_0 = \pi_1 = 1/2$.

and therefore the efficiency result follows.

This example clearly illustrates why the condition like (16) is crucial to achieve the efficiency bound. This suggests that, when the covariates are multinomial, we can always achieve the condition like (16) since the parametric ML estimation becomes equivalent to the nonparametric ML estimation. Therefore, we can achieve the efficiency bound. However, when the covariates or a subset of covariates are continuous, using the parametric propensity score estimation cannot achieve the efficiency bound even though the true one is parametric. This also suggests that the efficiency loss due to using the parametric propensity score estimator is attributed to the fact that some covariates are continuous.

3.2. Generalization to Estimating the Weighted Average Treatment Effect

We generalize the efficiency comparison between treatment effect estimators using nonparametric or parametric estimation of the propensity score to the weighted average treatment effect, τ^*_{wate}, defined as

$$\tau^*_{wate} \equiv \frac{\int E[Y(1) - Y(0)|X = x]g(x)dF_0(x)}{\int g(x)dF_0(x)}$$

for a known weight function $g(x)$. We estimate τ^*_{wate} using the moment condition

$$\psi(Z_i, \tau_{wate}, p(X_i), g(X_i)) = g(X_i)\left(\frac{Y_iT_i}{p(X_i)} - \frac{Y_i(1-T_i)}{1-p(X_i)} - \tau_{wate}\right) \tag{17}$$

that yields the estimator as

$$\hat{\tau}_{wate} = \sum_{i=1}^{n} g(X_i)[\frac{Y_iT_i}{\hat{p}(X_i)} - \frac{Y_i(1-T_i)}{1-\hat{p}(X_i)}] / \sum_{i=1}^{n} g(X_i) \tag{18}$$

given an estimator of the propensity score $\hat{p}(x)$.

Because the function $g(x)$ is known and only appears as a weight in the moment function (17), following the same line of argument for the average treatment effect, one can obtain the asymptotic linear representation of $\hat{\tau}_{wate}$ using the nonparametric propensity score estimator (2) in Equation (18) as

$$\left|\sqrt{n}(\hat{\tau}_{wate} - \tau^*_{wate}) - \frac{1}{E[g(X)]}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\psi(Z_i, \tau^*_{wate}, p^*(X_i), g(X_i)) + s_p(X_i)(T - p^*(X_i))\right)\right| = o_p(1)$$

where $\psi_p(Z_i, \tau, p(X_i), g(X_i)) = -g(X_i)\left(\frac{Y_iT_i}{p(X_i)^2} + \frac{Y_i(1-T_i)}{(1-p(X_i))^2}\right)$ and $s_p(X_i) = E[\psi_p(\cdot)|X_i]$. Therefore, the semiparametric efficiency bound is achieved for the weighted average treatment effect estimator $\hat{\tau}_{wate}$ using the nonparametric propensity score estimator (see Hirano et al. (2003)). On the other hand, for the parametric propensity score estimator, we can derive the inefficiency term similar to Equation (13), the inefficiency term derived for the average treatment effect, as

$$\frac{1}{(E[g(X_i)])^2}E[g(X_i)^2(\delta^*(X_i) - \delta_0(X_i))^2]$$

and therefore a similar inefficiency result holds for $\hat{\tau}_{wate}$ using the parametric propensity score estimator.

Note that, under the unconfoundedness assumption (Rosenbaum and Rubin (1983, 1984)), with the weight function $g(x)$ being equal to the true propensity score $p^*(x)$, the weighted average treatment effect becomes the average treatment effect for the treated,

$$\tau^*_{treated} \equiv E[Y(1) - Y(0)|T = 1].$$

Based on this equivalence, $\tau^*_{treated}$ can be estimated using the moment condition

$$\psi(Z_i, \tau_{treated}, p(X_i), p(X_i)) = p(X_i)\left(\frac{Y_i T_i}{p(X_i)} - \frac{Y_i(1-T_i)}{1-p(X_i)} - \tau_{treated}\right)$$

by replacing $g(x)$ with $p(x)$. However, an efficiency comparison between treatment effect estimators for $\tau^*_{treated}$ using the nonparametric or parametric propensity score estimator is more complicated because, in this case, the propensity score has two roles in the moment function. One is the inverse weighting to control for the self-selection and the other is the weighting function in place of $g(x)$. To see this, let $\hat{p}(x)$ and $\hat{p}^*(x)$ denote the nonparametric and the correctly specified parametric estimator of the propensity score, respectively. Then, we can consider three alternative estimators for the average treatment effect for the treated. One is using the parametric propensity score $\hat{p}^*(x)$ everywhere and solving

$$0 = \sum_{i=1}^{n} \hat{p}^*(X_i) \cdot \left(\frac{Y_i T_i}{\hat{p}^*(X_i)} - \frac{Y_i(1-T_i)}{1-\hat{p}^*(X_i)} - \tau_{treated}\right), \tag{19}$$

the second one is using the nonparametric propensity score $\hat{p}(x)$ everywhere and solving

$$0 = \sum_{i=1}^{n} \hat{p}(X_i) \cdot \left(\frac{Y_i T_i}{\hat{p}(X_i)} - \frac{Y_i(1-T_i)}{1-\hat{p}(X_i)} - \tau_{treated}\right), \tag{20}$$

and the last one is using the parametric propensity score $\hat{p}^*(x)$ in place of $g(x)$ while using the nonparametric propensity score $\hat{p}(x)$ for the inverse weighting and solving

$$0 = \sum_{i=1}^{n} \hat{p}^*(X_i) \cdot \left(\frac{Y_i T_i}{\hat{p}(X_i)} - \frac{Y_i(1-T_i)}{1-\hat{p}(X_i)} - \tau_{treated}\right). \tag{21}$$

From the efficiency argument of Hahn (1998) and Hirano et al. (2003) when the true propensity score is known, one can conjecture that the treatment effect estimator that solves Equation (21) will be more efficient than other estimators that solve Equations (19) and (20), respectively. However, in terms of efficiency, the two estimators solving Equations (19) and (20) (or other variations) cannot be uniformly ranked in general, and studying these alternative estimators is beyond the scope of this paper.

4. Conclusions

One can obtain efficient estimation of average treatment effects by the method of inverse propensity score weighting based on the estimated propensity score, rather than the true one, even when the true one is known. From the literature, we can infer that, even when the true propensity score is a parametric function, we still need to estimate the propensity score nonparametrically to achieve the efficiency. We formalize this argument and further identify the source of the efficiency loss due to parametric estimation of the propensity score. We also provide an intuition as to why this overfitting is necessary. The idea is that the nonparametric estimation of the propensity score plays two roles in the treatment effect estimation. It first replaces the true propensity score, and second it approximates the conditional expectation of the derivative of the moment condition for the treatment effect with respect to the propensity score. The parametric propensity score estimation can achieve the first but cannot achieve the second when some of covariates are continuous. This also suggests that the finite sample performance of the treatment effect estimator, using the imputation method based on the estimated propensity score, may depend not only on how precisely the propensity score is estimated, but also how well the conditional expectation of the derivative of the moment condition is approximated by the same approximating sieves or regressors that are used to estimate the propensity score.

Funding: This research received no external funding.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Efficiency Loss Due to Parametric Estimation of the Propensity Score

For ease of notation, we assume $Y(0) = 0$ with probability one, and define $\beta_0 = E[Y(1)]$ as the average outcome of interest. In the main text, we have established the following:[8]

$$\begin{aligned}
&\sqrt{n}(\hat{\beta} - \beta_0) \\
= \; &\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \begin{pmatrix} \left(\frac{T_i Y_i}{p^*(X_i)} - \beta_0\right) + \delta_0(X_i) \frac{(T_i - p^*(X_i))}{\sqrt{p^*(X_i)(1-p^*(X_i))}} \\ +(\delta^*(X_i) - \delta_0(X_i)) \frac{T_i - p^*(X_i)}{\sqrt{p^*(X_i)(1-p^*(X_i))}} \end{pmatrix} + o_p(1),
\end{aligned}$$

where $\delta^*(x) = -\int_{\mathcal{X}} \frac{\mu_1(z)}{p^*(z)} \phi(z'\pi_0) z' dF_0(z) \; W^{-1} \frac{\phi(x'\pi_0)x}{\sqrt{p^*(x)(1-p^*(x))}}$ and $\delta_0(x) = -\frac{\mu_1(x)}{p^*(x)} \sqrt{p^*(x)(1-p^*(x))}$. Define

$$\begin{aligned}
\psi(y,t,x,\beta,p(\cdot)) &= \frac{t \cdot y}{p(x)} - \beta, \alpha_0(t,x) = -\frac{\mu_1(x)}{p^*(x)}(t - p^*(x)), \text{and} \\
c^*(t,x) &= (\delta^*(x) - \delta_0(x)) \frac{t - p^*(x)}{\sqrt{p^*(x)(1-p^*(x))}}.
\end{aligned}$$

The first term $\psi(\cdot)$ is the moment function that would be obtained when we do not estimate the propensity score $p^*(\cdot)$. The second and the third term, $\alpha_0(t,x)$ and $c^*(t,x)$, are the contribution of estimating $p^*(\cdot)$ using the parametric ML estimator to the asymptotic distribution of $\widehat{\beta}$. If we estimated the propensity score using a nonparametric ML estimation, even when the true propensity score is parametric, we would not have the third term since we can replace $\delta^*(x)$ with $\delta_0(x)$ without affecting the asymptotic distribution.

The asymptotic variance of $\widehat{\beta}$ is equal to the variance of the sum of $\psi(Y,T,X,\beta_0,p^*(\cdot))$, $\alpha_0(T,X)$, and $c^*(T,X)$. We obtain for each component that potentially determines the asymptotic variance:

$$\begin{aligned}
E[\psi(Y,T,X,\beta_0,p^*(\cdot))^2] &= E\left[\frac{\mu_1(X)^2}{p^*(X)}\right] + E\left[\frac{\sigma_1^2(X)}{p^*(X)}\right] - \beta_0^2 \\
E\left[\alpha_0(T,X)^2\right] &= E\left[\frac{\mu_1(X)^2}{p^*(X)}\right] - E\left[\mu_1(X)^2\right] \\
E\left[c^*(T,X)^2\right] &= E\left[(\delta^*(X) - \delta_0(X))^2\right] \\
E[\psi(Y,T,X,\beta_0,p^*(\cdot))\alpha_0(T,X)] &= -E\left[\frac{\mu_1(X)^2}{p^*(X)}\right] + E\left[\mu_1(X)^2\right] \\
E[\{\psi(Y,T,X,\beta_0,p^*(\cdot)) + \alpha_0(T,X)\}c^*(T,X)] &= 0.
\end{aligned}$$

Combining these results, we obtain

$$\sqrt{n}(\hat{\beta} - \beta_0) \xrightarrow{d} N(0, E\left[(\mu_1(X) - \beta_0)^2 + \sigma_1^2(X)/p^*(X)\right] + E\left[(\delta^*(X) - \delta_0(X))^2\right]),$$

where the first term in the asymptotic variance is identical to the efficiency bound derived by Hirano et al. (2003). Therefore, the efficiency loss due to parametric estimation of the propensity score is given by $E\left[(\delta^*(X) - \delta_0(X))^2\right]$.

---

[8] This is because the terms (7)–(9) are $o_p(1)$ and only the terms (10) and (11) remain in the stochastic expansion.

References

Abadie, Alberto, and Guido W. Imbens. 2002. *Simple and Bias-Corrected Matching Estimators for Average Treatment Effects*. NBER Working Paper. Cambridge: National Bureau of Economic Research, vol. 283.

Abadie, Alberto, and Guido W. Imbens. 2006. Large Sample Properties of Matching Estimators for Average Treatment Effects. *Econometrica* 74: 235–67. [CrossRef]

Chen, Xiaohong. 2007. Large Sample Sieve Estimation of Semi-Nonparametric Models. *Handbook of Econometrics* 6: 5549–632.

Chen, Xiaohong, and Xiaotong Shen. 1998. Sieve Extremum Estimates for Weakly Dependent Data. *Econometrica* 66: 289–314. [CrossRef]

Hahn, Jinyong. 1998. On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects. *Econometrica* 66: 315–31. [CrossRef]

Heckman, James J., Hidehiko Ichimura, and Petra Todd. 1998. Matching as an Econometric Evaluations Estimator. *Review of Economic Studies* 65: 605–54. [CrossRef]

Hirano, Keisuke, Guido W. Imbens, and Geert Ridder. 2003. Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score. *Econometrica* 71: 1161–89. [CrossRef]

Hitomi, Kohtaro, Yoshihiko Nishiyama, and Ryo Okui. 2008. A Puzzling Phenomenon in Semiparametric Estimation Problems with Infinite-Dimensional Nuisance Parameters. *Econometric Theory* 24: 1717–28. [CrossRef]

Hristache, Marian, and Valentin Patilea. 2017. Conditional Moment Models with Data Missing at Random. *Biometrika* 104: 735–42. [CrossRef]

Imbens, Guido W. 2004. Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review. *The Review of Economics and Statistics* 86: 4–29. [CrossRef]

Kang, Joseph D. Y., and Joseph L. Schafer. 2007. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science* 22: 523–39. [CrossRef]

Kim, Kyoo Il. 2013. An Alternative Efficient Estimation of Average Treatment Effects. *Journal of Market Economy* 42: 1–41.

Li, Qi, Jeffrey S. Racine, and Jeffrey M. Wooldridge. 2009. Efficient Estimation of Average Treatment Effects with Mixed Categorical and Continuous Data. *Journal of Business and Economics Statistics* 27: 206–23. [CrossRef]

Newey, Whitney K. 1984. A Method of Moments Interpretation of Sequential Estimators. *Economics Letters* 14: 201–6. [CrossRef]

Newey, Whitney K. 1994. The Asymptotic Variance of Semiparametric Estimators. *Econometrica* 62: 1349–82. [CrossRef]

Newey, Whitney K. 1997. Convergence Rates and Asymptotic Normality for Series Estimators. *Journal of Econometrics* 79: 147–68. [CrossRef]

Prokhorov, Artem, and Peter Schmidt. 2009. GMM Redundancy Results for General Missing Data Problems. *Journal of Econometrics* 151: 47–55. [CrossRef]

Robins, James M., Andrea Rotnitzky, and Lue Ping Zhao. 1995. Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of the American Statistical Association* 90: 106–21. [CrossRef]

Rosenbaum, Paul R. 1987. Model-Based Direct Adjustment. *Journal of the American Statistical Association* 82: 387–94. [CrossRef]

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The Central Role of the Propensity Score in Observational Studies for Casual Effects. *Biometrika* 70: 41–55. [CrossRef]

Rosenbaum, Paul R., and Donald B. Rubin. 1984. Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association* 79: 516–24. [CrossRef]

Rubin, Donald B. 1973. The Use of Matched Sampling and Regression Adjustments to Remove Bias in Observational Studies. *Biometrics* 29: 185–203. [CrossRef]

Rubin, Donald B., and Neal Thomas. 1996. Matching Using Estimated Propensity Scores: Relating Theory to Practice. *Biometrics* 52: 249–64. [CrossRef] [PubMed]

Shen, Xiaotong, and Wing Hung Wong. 1994. Convergence Rate of Sieve Estimates. *The Annals of Statistics* 22: 580–615. [CrossRef]

Tan, Zhiqiang. 2007. Comment: Understanding OR, PS and DR. *Statistical Science* 22: 560–68. [CrossRef]