The Value of Knowing the Propensity Score for Estimating Average Treatment Effects

Christoph Rothe

Abstract: In a treatment effect model with unconfoundedness, treatment assignments are not only independent of potential outcomes given the covariates, but also given the propensity score alone. Despite this powerful dimension reduction property, adjusting for the propensity score is known to lead to an estimator of the average treatment effect with lower asymptotic efficiency than one based on adjusting for all covariates. Moreover, knowledge of the propensity score does not change the efficiency bound for estimating average treatment effects, and many empirical strategies are more efficient when an estimate of the propensity score is used instead of its true value. Here, we resolve this "propensity score paradox" by demonstrating the value of knowledge of the propensity score. We show that by exploiting such knowledge properly, it is possible to construct an efficient treatment effect estimator that is not affected by the "curse of dimensionality", which yields desirable second order asymptotic properties and finite sample performance. The method combines knowledge of the propensity score with a nonparametric adjustment for covariates, building on ideas from the literature on double robust estimation. It is straightforward to implement, and performs well in simulations. We also show that confidence intervals based on our estimator and a simple variance estimate have remarkably robust coverage properties with respect to the implementation details of the nonparametric adjustment step.

JEL Classification: C13, C14, C21

Keywords: *Propensity score, treatment effects, semiparametric efficiency, randomized experiment, nonparametric estimation.*

First Version: April 25, 2015. *This Version:* June 13, 2016. I thank Joshua Angrist, Miikka Rokkanen and attendants of the 2016 Econometric Society Winter Meeting in San Francisco for helpful comments. Contact information: Department of Economics, Columbia University, 420 W 118th St., New York, NY 10027, Email: cr2690@columbia.edu, Web: http://www.christophrothe.net.

1. INTRODUCTION

Many economic studies estimate average treatment effects (ATEs) under the assumption that the treatment assignment is unconfounded, or independent of potential outcomes given a set of covariates; see Imbens (2004) or Imbens and Wooldridge (2009) for surveys. An important quantity in this context is the propensity score, which is defined the conditional probability of receiving the treatment given the covariates. In this paper, we are concerned with the role played by knowledge of the propensity score for estimating ATEs. While our main motivation for studying this question is theoretical, the issue is also relevant for certain empirical applications in which propensity scores can reasonably be modeled as known to the analyst. Examples include settings with data from randomized experiments, where the propensity score is defined by the study design, and settings with substantially more data on treatment assignments and covariates than on outcomes, where sampling uncertainty about the estimated value of the propensity score is negligible. The latter setting could arise for instance if treatment assignments and covariates are recorded in large and easily accessible administrative data sets, while information on outcomes is only available through expensive specialized, and thus rather small-scale, surveys.

The importance of the propensity score stems from the result that in models with unconfoundedness treatment assignments are not only independent of potential outcomes given the covariates, but also given the propensity score alone (Rosenbaum and Rubin, 1983). Working with the full set of covariates is thus not necessary to remove bias associated with differences in pre-treatment variables when the propensity score is known. Despite this powerful "dimension reduction" property of the propensity score, adjusting for covariates leads to an asymptotically efficient ATE estimator, while adjusting for a known propensity score does not (Hahn, 1998). Moreover, many empirical strategies for estimating ATEs have the puzzling feature that they are more efficient when an estimate of the propensity score is used rather than the true value. Results of this type have been obtained for example by Hirano, Imbens, and Ridder (2003) for inverse probability weighting, Imai and van Dyk (2004) for subclassification, Abadie and Imbens (2016) for nearest neighbor matching, and Hahn and Ridder (2013) and Mammen, Rothe, and Schienle (2016) for propensity score-based regression adjustment. Moreover, Hahn (1998) shows that the semiparametric efficiency bound for estimating the ATE is the same whether the propensity score is known or not, and many estimators achieve this bound when the propensity score is unknown (e.g. Hahn, 1998; Hirano, Imbens, and Ridder, 2003; Imbens, Newey, and Ridder, 2007; Chen, Hong, and Tarozzi, 2008; Rothe and Firpo, 2016). Existing results thus seem to suggest that there is no scope for exploiting knowledge of the propensity score for estimation purposes, and that any such attempt might even be harmful.

This paper makes progress towards resolving this "propensity score paradox" (Angrist and Pischke, 2008). We show, the above-mentioned results notwithstanding, that knowledge of the propensity score is indeed immensely useful, as it can be exploited for the construction of an ATE estimator that has superior theoretical and practical properties relative to procedures that ignore such knowledge. Building on ideas from the double robustness literature (e.g. Robins, Rotnitzky, and Zhao, 1994; Robins and Rotnitzky, 1995), the proposed estimator takes the form of an average of a sample analogue of the ATE's efficient influence function. This function depends on the data, the true propensity score, and the conditional expectation of the outcome given the treatment and the covariates. The latter is an unknown nuisance function that is estimated nonparametrically from the data.

We show that this construction leads to an ATE estimator that is fully efficient as long as the conditional expectation function is estimated consistently, and some weak regularity conditions hold. In contrast to other efficient ATE estimators, no restrictions on the bias of the nonparametric first-stage estimator are necessary for obtaining this result. Knowledge of the propensity score thus allows working with an "over-smoothed" nonparametric estimate of the conditional expectation function that converges very slowly, which alleviates concerns about the "curse of dimensionality" in settings with many covariates (cf. Robins and Ritov, 1997; Angrist and Hahn, 2004). This shows that knowing the propensity score not only serves as a dimension reduction device for the purpose of identification, but also for the purpose of efficient ATE estimation. To the best of our knowledge, this point seems to be new in the literature.

We also show that inference based on a simple estimate of the asymptotic variance of our estimator remains valid even if the conditional expectation function is estimated inconsistently. In this case our proposed ATE estimator becomes inefficient, but it remains consistent and asymptotically normal, and the variance estimate that we propose remains consistent as well. Inference also remains valid if the estimate of the conditional expectation is based on a parametric model, and the ATE estimator would be fully efficient if that parametric model was correctly specified. Knowledge of the propensity score thus allows for inference that is remarkably robust with respect to variation in the implementation details of the estimator of the conditional expectation function.

The remainder of the paper is structured as follows. In the new section, we introduce the model, review some existing estimation approaches, and introduce the new estimator. In Section 3, we derive its theoretical properties. Section 4 presents some simulation results, and Section 5 concludes. All proofs are collected in the Appendix.

2. ESTIMATION WITH KNOWLEDGE OF THE PROPENSITY SCORE

In this section we first introduce the model and review some existing results on ATE estimation under unconfoundedness, and then introduce the proposed estimator for settings where the propensity score is known, and describe the rationale behind it.

2.1. **The Model.**

We are interested in estimating the effect of a binary treatment on some economic outcome based on a random sample of n units from a large population. For each unit i, we observe the outcome $Y_i \in \mathcal{Y} \subset \mathbb{R}$, a treatment indicator $T_i \in \{0, 1\}$, with $T_i = 1$ if unit i is treated and $T_i = 0$ otherwise, and a vector of covariates $X_i \in \mathcal{X} \subset \mathbb{R}^d$. We define $Y_i(1)$ and $Y_i(0)$ as the potential outcomes of unit i with and without receiving the treatment, respectively, so that the observed outcome satisfies $Y_i = Y_i(T_i)$. The parameter of interest is the average treatment effect

$$\tau = \mathbb{E}(Y_i(1) - Y_i(0)).$$

Following Rosenbaum and Rubin (1983), the treatment assignment is assumed to be unconfounded, or independent of the potential outcomes conditional on the covariates. We also assume that the propensity score function $\pi(x) = P(T_i = 1|X_i = x)$ is bounded away from zero and one, and known to the analyst.

Assumption 1. *(i) $Y_i(1), Y_i(0) \perp T_i | X_i$ almost surely; (ii) $\epsilon \leq \pi(X_i) \leq 1 - \epsilon$ almost surely, for some $\epsilon > 0$; (iii) the propensity score function $\pi(x)$ is known.*

Assumption 1(i)–(ii) are standard in the treatment effects literature. The former condition implies that adjusting for X_i eliminates all biases due to confounding in comparisons between treated and untreated units, whereas the latter condition ensures that there are both treated and untreated units in every region of the support $\mathcal{X}$ of the covariates. Khan and Tamer (2010) point out that without Assumption 1(ii) no regular estimator of τ might exist. Assumption 1(iii) is natural in the context of randomized experiments, where the propensity score is specified by the study design. Assuming that the propensity score is known is generally difficult to justify with observational data, but it could be a reasonable approximation for example in settings where there is a large additional data set with information on treatment assignments and covariates, such that sampling uncertainty about the estimated value of the propensity score effectively becomes negligible.

2.2. **Previous Results.** Estimation of ATEs under unconfoundedness has been widely considered in the program evaluation literature with and without the assumption that the propensity score is known; and the problem also has close analogues in the literature on missing data. Let $\mu(t,x) = \mathbb{E}(Y_i|T_i = t, X_i = x)$ and $\sigma^2(t,x) = \mathbb{V}(Y_i|T_i = t, X_i = x)$ be the conditional expectation and variance, respectively, of Y_i given $T_i = t$ and $X_i = x$. Hahn (1998) shows that under Assumption 1(i)–(ii) the semiparametric efficiency bound for estimating τ is given by

$$V_{\text{eff}} = \mathbb{E}\left(\frac{\sigma^2(1,X_i)}{\pi(X_i)} + \frac{\sigma^2(0,X_i)}{1-\pi(X_i)}\right) + \mathbb{V}(\mu(1,X_i) - \mu(1,X_i)),$$

and that the corresponding efficient influence function is given by $\psi_i(\pi,\mu) - \tau$, where

$$\psi_i(\pi,\mu) = \frac{T_iY_i}{\pi(X_i)} - \frac{(1-T_i)Y_i}{1-\pi(X_i)} - (T_i - \pi(X_i))\left(\frac{\mu(1,X_i)}{\pi(X_i)} - \frac{\mu(0,X_i)}{1-\pi(X_i)}\right),$$

so that $V_{\text{eff}} = \mathbb{V}(\psi_i(\pi,\mu))$. This means that the asymptotic variance of any regular estimator $\hat{\tau}$ of τ is bounded from below by V_{eff}, and any regular estimator whose asymptotic variance achieves the bound is such that

$$\sqrt{n}(\hat{\tau} - \tau) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\psi_i(\pi,\mu) - \tau) + o_P(1) \xrightarrow{d} N(0, V_{\text{eff}}). \qquad (2.1)$$

Hahn (1998) also shows that additionally imposing Assumption 1(iii) does not change the semiparametric efficiency bound for estimating τ, and thus knowledge of the propensity score cannot be used to construct a regular estimator whose asymptotic variance is strictly smaller than V_{eff}.

A number of estimators that reach the semiparametric efficiency bound, and thus satisfy the representation (2.1), under certain regularity conditions have been proposed in the literature.[1] These estimators require nonparametric estimation of an unknown function, typically either the conditional expectation function or the propensity score. Hahn (1998),

[1] In addition, there are many empirical strategies for estimating ATEs that generally do not achieve the semiparametric efficiency bound, but that are nevertheless popular in practice for various reasons. Examples include different forms of matching, among many others; see Imbens (2004) or Imbens and Wooldridge (2009) for further details.

Imbens, Newey, and Ridder (2007) and Chen, Hong, and Tarozzi (2008) consider regression adjustment estimators (sometimes also called imputation estimators) of the form

$$\hat{\tau}_{\text{reg}} = \frac{1}{n}\sum_{i=1}^{n}\left(\hat{\mu}(1, X_i) - \hat{\mu}(0, X_i)\right),$$

with Imbens, Newey, and Ridder (2007) and Chen, Hong, and Tarozzi (2008) estimating $\mu(t, \cdot)$ directly through a nonparametric regression of Y_i on X_i in the subgroup with $T_i = t$, and Hahn (1998) using more indirect estimators of the form $\hat{\mu}(1, X_i) = \hat{\mathbb{E}}(Y_i T_i | X_i)/\hat{\pi}(X_i)$ and $\hat{\mu}(0, X_i) = \hat{\mathbb{E}}(Y_i(1 - T_i)|X_i)/(1 - \hat{\pi}(X_i))$. Hirano, Imbens, and Ridder (2003) propose an inverse probability weighting estimator, which weights outcomes by the inverse of the estimated propensity score:

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

Rothe and Firpo (2016) study a nonparametric version of the double robust estimator[2] of Robins, Rotnitzky, and Zhao (1994) or Robins and Rotnitzky (1995), which is of the form:

$$\hat{\tau}_{\text{dr}} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{T_i Y_i}{\hat{\pi}(X_i)} - \frac{(1 - T_i)Y_i}{1 - \hat{\pi}(X_i)} - (T_i - \hat{\pi}(X_i))\left(\frac{\hat{\mu}(1, X_i)}{\hat{\pi}(X_i)} - \frac{\hat{\mu}(0, X_i)}{1 - \hat{\pi}(X_i)}\right)\right);$$

This estimator can be written as $\hat{\tau}_{\text{dr}} = (1/n)\sum_{i=1}^{n}\psi_i(\hat{\pi}, \hat{\mu})$, and is thus a sample average of a sample analogue of the efficient influence function. Robins, Rotnitzky, and Zhao (1994) and Robins and Rotnitzky (1995) show that this estimator also reaches the semiparametric efficiency bound if estimates of both the propensity score and the conditional expectation function are based on correctly specified parametric models, a property to which they refer as *local* efficiency.[3] We remark that all the aforementioned estimators are affected by the "curse of dimensionality", in the sense that ever-stricter regularity conditions need to be imposed

[2] See Cattaneo (2010) for an earlier use of this construction in a settings with a multi-valued treatment.

[3] The eponymous property of doubly robust estimators is that they remain $\sqrt{n}$-consistent and asymptotically normal if only one of the parametric specifications is correct, although they are no longer efficient in this case.

in settings with multiple covariates in order for these estimators to reach the semiparametric efficiency bound, albeit to a different extent.

2.3. **A New Estimator.** Any of the estimators described in the previous subsection can also be used when the propensity score is known by simply ignoring such knowledge, and continues to reach the semiparametric efficiency bound in this case. Moreover, most ATE estimators that involve an estimate of the propensity score become less efficient when said estimate is replaced with the true propensity score. This is the case for the inverse probability weighting estimator, for example, but also for other estimators not explicitly mentioned in the previous subsection (e.g. Hahn, 1998; Imai and van Dyk, 2004; Abadie and Imbens, 2016).

It does not seem to be much appreciated in the economics literature, however, that this is not the case for *all* ATE estimators that use an estimate of the propensity score. In particular, the doubly robust estimator remains efficient if the true value of the propensity score is used instead of an estimate; see below for details. We denote this estimator by

$$\hat{\tau}_{\text{kps}} = \frac{1}{n}\sum_{i=1}^{n} \psi_i(\pi, \hat{\mu}),$$

where the subscript "kps" stands for "known propensity score", and we refer to it as the KPS estimator in the following.

We consider this estimator for the use in settings where the propensity score is known. Like the regression adjustment estimators, the KPS estimator also uses on an estimate $\hat{\mu}$ of the conditional expectation function μ, which we propose to obtain by running nonparametric regressions of Y_i on X_i separately in the subpopulations of treated and untreated units. We do not require the estimator $\hat{\mu}$ to be of a particular type otherwise. Instead, our main results below are derived under "high-level" conditions that hold for all commonly used nonparametric regression methods, such as local polynomial regression or series estimation, under weak regularity conditions. We also define

$$\hat{V}_{\text{kps}} = \frac{1}{n}\sum_{i=1}^{n}(\psi_i(\pi, \hat{\mu}) - \hat{\tau}_{\text{kps}})^2,$$

which is going to be a suitable estimate of the asymptotic variance of $\hat{\tau}_{\text{kps}}$ in great generality for reasons outlined below.

2.4. **Rationale Behind the KPS Estimator.** Since $\hat{\tau}_{\text{kps}}$ takes the form of the sample average of an estimate of the efficient influence function, following Newey (1994, Proposition 3) we expect that $\hat{\tau}_{\text{kps}}$ satisfies the representation (2.1), and thus achieves the semiparametric efficiency bound, under certain regularity conditions; just like the estimators described in Section 2.2. While using the KPS estimator does thus not yield an improvement in terms of first order asymptotic variance relative to existing methods, we argue that this estimator is nevertheless preferable in settings where the propensity score is known. This is due to its better *second order* asymptotic properties, which translate into substantial improvements in finite sample performance.

While we explain this point formally in the next section, it is nevertheless useful to consider an overview of the argument. Let $R_n(t) = t - (1/n)\sum_{i=1}^n \psi_i(\pi, \mu)$, and note that a necessary and sufficient condition for any regular estimator $\hat{\tau}$ of τ to reach the semiparametric efficiency bound is that $|R_n(\hat{\tau})| = o_P(n^{-1/2})$. For the estimators described in Section 2.2, this condition is achieved through assumptions that imply a high degree of accuracy of the nonparametric estimate of the respective nuisance function.[4] Typical requirements include a uniform rate of convergence that is faster than $n^{-1/4}$, which implies that both the bias and the variance of the nonparametric estimate are "sufficiently" small. As pointed out for example by Linton (1995) or Robins and Ritov (1997), asymptotic approximations based on such conditions can be fragile in practice, especially in settings with many covariates.

[4] Such assumptions are commonly made in the literature on the properties of "two-step" estimators that depend on a nonparametrically estimated function; see for example Newey (1994), Newey and McFadden (1994), Chen, Linton, and Van Keilegom (2003), or Ichimura and Lee (2010).

Correspondingly, the finite sample behavior of the estimators described in Section 2.2 can differ substantially from the predictions of first order asymptotic theory (e.g. Angrist and Hahn, 2004; Rothe and Firpo, 2016).

Knowledge of the propensity score solves this problem if it is exploited properly. Below, we show for the KPS estimator it holds that $|R_n(\hat{\tau}_{\text{kps}})| = o_P(n^{-1/2})$ *without* imposing restrictions on the rate at which the bias of $\hat{\mu}$ tends to zero. This turns out to be the case in great generality, and not just for a particular nonparametric estimation method. The KPS estimator can thus be fully efficient even if $\hat{\mu}$ converges arbitrarily slowly. Intuitively, this is because the KPS estimator is implicitly based on the moment condition $g(\mu, \tau) = 0$, where $g(m, t) = \mathbb{E}(\psi_i(\pi, m) - t)$, and this moment condition is very insensitive the variation in the conditional expectation function μ, in the sense that $\partial_m^k \left[g(m, \tau)\right]_{m=\mu} = 0$ for all $k \in \mathbb{N}$, where ∂_m^k is the kth order functional derivative operator with respect to m. That is, derivatives of the moment condition with respect to the nuisance function of *any order* are equal to zero. The KPS estimator then "inherits" the insensitivity of the moment condition, and is thus robust with respect to variation in the estimate of μ.

3. LARGE SAMPLE PROPERTIES

In this section, we first formally study the properties of $\hat{\tau}_{\text{kps}}$ under a set of general "high-level" conditions on $\hat{\mu}$. We then derive some more specific results for the special case that μ is estimated via local polynomial regression, and finally comment on the case where $\hat{\mu}$ is based on a parametric specification of μ.

3.1. **General Conditions.** Our first results concern the large sample properties of $\hat{\tau}_{\text{kps}}$ under general conditions on the properties of the estimator $\hat{\mu}$. The following notation is helpful for presenting them. For any class of functions $\mathcal{M}$ over $\{0, 1\} \times \mathcal{X}$, let $N_2(\epsilon, \mathcal{M})$ be the minimum number of ϵ-brackets with respect to the $L_2(P)$ norm needed to cover $\mathcal{M}$, where for two functions $u, l \in \mathcal{M}$ the set $\{f \in \mathcal{M} : l(t, x) \leq f(t, x) \leq u(t, x) \text{ for all } (t, x)\}$ is called an ϵ-bracket with respect to $L_2(P)$ if $\mathbb{E}((l(T_i, X_i) - u(T_i, X_i))^2) < \epsilon^2$. We also write $a(\eta) \lesssim b(\eta)$ for generic functions a, b if $a(\eta) \leq Cb(\eta)$ for some constant C not depending on η, and for $s = (s_1, \ldots, s_d)$ a vector of non-negative integers and $|s| = \sum_{j=1}^d s_j$ the notation $\partial_x^s m(t, x) = \partial_{x_1}^{s_1} \ldots \partial_{x_d}^{s_d} m(t, x)$ denotes the partial derivatives with respect to x of a generic function m. All limits are taken as $n \to \infty$. We impose two "high level" assumptions about the estimator $\hat{\mu}$.

Assumption 2. *There exists a sequence μ_n of non-random functions, a non-random function $\bar{\mu}$, and sequences $a_n = o(1)$ and $b_n = o(1)$ of constants such that $\|\hat{\mu} - \mu_n\|_\infty = O_P(a_n)$ and $\|\mu_n - \bar{\mu}\|_\infty = O(b_n)$.*

The idea behind this assumption is to put $\bar{\mu} = \mu$, and to define the function μ_n as the sum of the true conditional expectation function μ and the asymptotic bias of the respective nonparametric estimator. With such a choice of $\bar{\mu}$ and μ_n, Assumption 2 simply requires that the stochastic part and the bias of $\hat{\mu}$ converge to zero uniformly, and denotes the corresponding rates by a_n and b_n, respectively. Uniform convergence results for nonparametric regression estimators are widely available in the literature; see e.g. Newey (1997) for series estimators and Masry (1996) for local polynomial regression. For reasons that will become clear below, we also want to allow for the case that $\hat{\mu}$ is potentially inconsistent. This means that $\bar{\mu} \neq \mu$, and then the assumption only implies uniform convergence of $\hat{\mu}$ to a non-random probability limit, with a_n and b_n denoting the rates at which the stochastic part and the "pseudo-bias" tend to zero, respectively.

Assumption 3. *There exists a sequence $\mathcal{M}_n$ of function classes such that $P(\hat{\mu} - \mu_n \in \mathcal{M}_n) = 1 + o(1)$ and $N_2(\epsilon, \mathcal{M}_n^*) \lesssim \exp(\epsilon^{-\alpha} c_n)$ for $\alpha \in (0, 2)$, a sequence of constants $c_n = o(a_n^{\alpha-2})$, and all $\epsilon < a_n$, where $\mathcal{M}_n^* = \mathcal{M}_n \cap \{m \in \mathcal{M}_n : \|m - \mu_n\|_\infty \leq a_n\}$*

This assumption states that the estimator $\hat{\mu}$ takes values in a function class whose entropy with bracketing (which is defined as the natural logarithm of the covering number) does not grow too quickly as the sample size increases. Entropy restrictions of this type are commonly found in the literature on semiparametric two-step estimation. As explained below, this assumption does imply restrictions on the rate a_n of the stochastic component of $\hat{\mu}$, but it implies no restrictions on b_n. For most nonparametric estimators, it is natural to take $\mathcal{M}_n$ as a smoothness class, such as that of functions with bounded partial derivatives up to a particular order.[5] The role of Assumption 3 is to ensure that the estimator $\hat{\mu}$ is such that there is no overfitting of the data, by requiring that it takes values in a class whose elements cannot be "too complex"; see e.g. van der Vaart (1998) for further details on the interpretation of restrictions on covering numbers.

Under our assumptions, we obtain the following result about the large sample properties of the KPS estimator and the corresponding variance estimate.

Theorem 1. *Suppose that Assumptions 1–3 hold. Then*

$$\sqrt{n}(\hat{\tau}_{\text{kps}} - \tau) \xrightarrow{d} N(0, \mathbb{V}(\psi_i(\pi, \bar{\mu}))) \quad \text{and} \quad \hat{V}_{\text{kps}} = \mathbb{V}(\psi_i(\pi, \bar{\mu})) + o_P(1).$$

If in addition $\bar{\mu} = \mu$, then $\mathbb{V}(\psi_i(\pi, \bar{\mu})) = V_{\text{eff}}$ and thus $\hat{\tau}_{\text{kps}}$ reaches the semiparametric efficiency bound.

The theorem shows that $\hat{\tau}_{\text{kps}}$ is $\sqrt{n}$-consistent for τ, asymptotically normal, asymptotically unbiased, and that it reaches the semiparametric efficiency bound if $\hat{\mu}$ consistently estimates μ. It also shows that the simple variance estimator $\hat{V}_{\text{kps}}$ is consistent, which implies the validity of standard large sample methods for inference. What is remarkable relative to analogous results for other efficient estimators of τ, however, is that Theorem 1 holds without any restrictions on the rate b_n at which the bias of $\hat{\mu}$ tends to zero. We can therefore satisfy Assumptions 2 and 3 by choosing an estimator $\hat{\mu}$ that sufficiently "over-smooths"

[5] The assumption also allows $\mathcal{M}_n$ to consist of the sum of one potentially non-smooth function and other functions from a smoothness class. This extension allows us to deal with settings where the bias of $\hat{\mu}$ is not a smooth function itself.

the data, since for all commonly used nonparametric regression estimators increasing the amount of smoothing increases the rate of convergence of the stochastic part and decreases the "complexity" of the estimated function.[6] The fact that more smoothing also slows down the rate at which the bias tends to zero, which would be a problem for other estimators, is of no concern in our setting. The following example further illustrates this point.

Example 1. *Suppose that $\mathbb{X} \subset \mathbb{R}^d$ is compact, and let $\mathcal{M}_n$ be the collection of all functions m defined over $\{0,1\} \times \mathbb{X}$ such that the partial derivatives of $m(t,\cdot)$ up to order $l > d/2$ are uniformly bounded by c_n for $t = 0,1$. Then $N_2(\epsilon, \mathcal{M}_n^*) \lesssim \exp(\epsilon^{-d/l} c_n)$; see van der Vaart (1998, Example 19.9). Now suppose that there is a single continuously distributed covariate, and that $\hat{\mu}$ is the local linear regression estimator with bandwidth h. Then by arguing as in Ojeda (2008) and Portier and Segers (2015) it follows that under standard regularity conditions one can choose a function μ_n such that*

$$\|\partial_x^s \hat{\mu} - \partial_x^s \mu_n\|_\infty = O_P\left(\left(\frac{\log(n)}{nh^{1+2s}}\right)^{1/2}\right) \quad \text{for } s = 0,1,2, \quad \text{and} \quad \|\mu_n - \mu\|_\infty = O(h^2)$$

if $\mu(t,\cdot)$ has bounded second-order partial derivatives. Assumptions 2 and 3 then hold for example with $l = 2$, $\alpha = 1/2$, $a_n = (nh^5/\log(n))^{-1/2}$, $b_n = h^2$ and $c_n = (nh/\log(n))^{-1/2}$ if $h \propto n^{-\theta}$ with $0 < \theta < 5/13$. That is, we only need that h does not tend to zero too quickly, but it is allowed to vanish arbitrarily slowly.[7]

Theorem 1 thus implies that a estimator of the nuisance parameter μ with very slowly vanishing bias suffices for constructing an efficient estimator of τ when the propensity score is known, whereas efficiency of the alternative estimators described in Section 2.2 generally

[6] With local polynomial regression, for example, increasing the bandwidth decreases the variance and leads to a more regular estimate. Similarly, the variance of a series estimator decreases if a smaller number of series terms is chosen, and the complexity of the estimated function is being reduced.

[7] A similar reasoning applies to higher-dimensional settings and local polynomial estimators of different order. It also implies to other estimation procedures like series estimators, for which Assumptions 2 and 3 can be shown to be satisfied whenever the number of series terms increases sufficiently slowly with the sample size.

requires the stochastic part *and* the bias of the respective nonparametric components to converge with a rate that is $o(n^{-1/4})$. As pointed out for example by Linton (1995) or Robins and Ritov (1997), asymptotic approximations to the distribution of some estimator that rely on the assumption of a very accurate nonparametric estimate of a nuisance function can be fragile in practice. For this reason, we expect inference based on Theorem 1 to more accurate in finite samples than inference based on alternative efficient estimators described in Section 2.2. This means for example that a standard confidence interval of the form

$$\left[\hat{\tau}_{\text{kps}} \pm 1.96 \times (\hat{V}_{\text{kps}}/n)^{1/2}\right] \tag{3.1}$$

should not only cover τ with 95% probability asymptotically, but also that in finite samples the coverage level should be approximately correct and rather robust to variations in how the estimation of μ is implemented. This is the type of improvement one can achieve from exploiting knowledge of the propensity score.

Theorem 1 also shows that the confidence interval (3.1) should remain approximately valid in finite samples if $\hat{\mu}$ is such that its finite sample bias is very large. This is not *a priori* obvious, and it might at first seem like an abuse of asymptotics to leverage an approximation based on the fact that a bias term tends to zero asymptotically when in finite samples its magnitude is substantial. However, Theorem 1 covers the case that $\bar{\mu} \neq \mu$, which means that $\hat{\mu}$ is inconsistent for μ. While $\hat{\tau}_{\text{kps}}$ is no longer efficient in such a setting, it remains $\sqrt{n}$-consistent for τ, asymptotically normal, and asymptotically unbiased in this case; and the simple variance estimator $\hat{V}_{\text{kps}}$ remains consistent for its asymptotic variance. We interpret this finding as a further robustness result regarding inference based on $\hat{\tau}_{\text{kps}}$. Again, this is a rather unique feature for inference based on a treatment effect estimator, and it is obtained by exploiting knowledge of the propensity score.

3.2. **Local Polynomial Regression.** The result in Theorem 1 is based on deriving a bound on the difference between $\sqrt{n}(\hat{\tau}_{\text{kps}} - \tau)$ and its asymptotically linear representation using techniques from empirical process theory. For the case that $\hat{\mu}$ consistently estimates μ, that is $\bar{\mu} = \mu$, this bound is given by

$$\sqrt{n}(\hat{\tau}_{\text{kps}} - \tau) - \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi_i(\pi, \mu) - \tau) = O_P(c_n^{1/2} a_n^{1-\alpha/2}) + O_P(b_n).$$

This is a "worst case" bound that is valid for all nonparametric estimators $\hat{\mu}$, including hypothetical or infeasible ones, as long as they satisfy the high level conditions laid out in Assumptions 2 and 3. For a specific nonparametric estimation procedure, this bound can generally be improved. This is illustrated in this subsection for the case that $\hat{\mu}$ is obtained by local polynomial regression, carried out separately in the subpopulations of treated and untreated units.

Local polynomial regression is a class of kernel-based smoothers that has been studied extensively by Fan (1993), Ruppert and Wand (1994), Fan and Gijbels (1996) and others. It is well-known to have attractive bias properties relative to other kernel-based methods, such as the Nadaraya-Watson estimator. We use the following notation. For generic vectors $b = (b_1, \ldots, b_d)$ and $a = (a_{(0,\ldots,0)}, a_{(1,0,\ldots,0)}, \ldots, a_{(0,\ldots,0,l)})$, let $\mathcal{P}_{l,a}(b) = \sum_{0 \le |s| \le l} a_s b^s$ be a polynomial of order l, where $\sum_{0 \le |s| \le l}$ denotes the summation over all d-vectors s of positive integers with $0 \le |s| \le l$. Also let $\mathcal{K}$ be a univariate probability density function, put $K_h(b) = \prod_{j=1}^{d} \mathcal{K}(b_j/h)/h$ for any bandwidth $h \in \mathbb{R}_+$, and define

$$\hat{a}(t, x) = \operatorname*{argmin}_{\alpha} \sum_{i=1}^{n} \left(Y_i - \mathcal{P}_{l,\alpha}(X_i - x)\right)^2 K_h(X_i - x) \mathbb{I}\{T_i = t\}.$$

Then the lth order local polynomial estimator of $\mu(t, x)$ is given by

$$\hat{\mu}(t, x) = \hat{a}_{(0,\ldots,0)}(t, x).$$

Note that we are using the same bandwidth for each component of the covariate vector X_i for notational convenience only, and that more general bandwidth choices are possible. The following assumption collects some regularity conditions that are standard in the literature on local polynomial regression.

Assumption 4. *(i) X_i is continuously distributed given $T_i = t$ with compact and convex support for $t = 0, 1$; (ii) the corresponding conditional density functions are bounded, have bounded first order derivatives, and are bounded away from zero, uniformly over the respective support; (iii) $\mu(t, \cdot)$ is $(l+1)$-times continuously differentiable for $t = 0, 1$; (iv) $\sup_x \mathbb{E}(|Y_i|^{2+\delta}|T_i = t, X = x) < \infty$ for some constant $\delta > 0$ and $t = 0, 1$; (v) the kernel $\mathcal{K}$ is twice continuously differentiable, and such that $\int \mathcal{K}(u)du = 1$, $\int u\mathcal{K}(u)du = 0$, $\int |u^2\mathcal{K}(u)|du < \infty$, and $\mathcal{K}(u) = 0$ for u not contained in some compact set.*

Under this assumption, we obtain the following characterization of the second order terms of the estimator $\hat{\tau}_{\text{kps}}$ if μ is estimated by local polynomial regression.

Theorem 2. *Suppose that Assumptions 1 and 4 hold, and that the bandwidth h is such that $h \to 0$ and $n^2h^{3d}/\log(n)^3 \to \infty$ as $n \to \infty$. Then*

$$\sqrt{n}(\hat{\tau}_{\text{kps}} - \tau) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\psi_i(\pi, \mu) - \tau) = O_P(h^{l+1}) + O_P(n^{-1/2}h^{-d/2}) = o_P(1).$$

Theorem 2 substantially improves upon the result in Example 1. In particular, it implies that for $\hat{\tau}_{\text{kps}}$ to reach the semiparametric efficiency bound it is not necessary to require a degree of smoothness of $\mu(t, \cdot)$ that depends on the dimensionality of the covariates. If $\hat{\mu}$ is estimated by local linear regression, for example, the theorem shows that $\sqrt{n}(\hat{\tau}_{\text{kps}} - \tau) \xrightarrow{d} N(0, V_{\text{eff}})$ irrespective of the value of d when $\mu(t, \cdot)$ is only twice continuously differentiable and the bandwidth satisfies $h \propto n^{-\theta}$ with $0 < \theta < 2/(3d)$. In this regard, the KPS estimator differs markedly from the other procedures discussed in Section 2.2, which all require higher-order differentiability conditions on the respective nuisance function in settings with many covariates in order to control the magnitude of the asymptotic bias of the respective nonparametric estimate. Knowledge of the propensity score therefore acts like a "dimension reduction" device.

An inspection of the proof of Theorem 2, which follows directly from a result in Rothe and Firpo (2016), shows that both terms on the right-hand side of the previous equation have mean zero, which means that $\hat{\tau}_{\text{kps}}$ is second-order unbiased. The orders of magnitude of the two terms are the same as those of the asymptotic bias and the pointwise asymptotic standard deviation of $\hat{\mu}$, respectively. The sum of the second order terms is minimized if the bandwidth is chosen such that $h \propto n^{-1/(2(l+1)+d)}$. As long as $d < 4l + 4$, such a choice is compatible with the restrictions on the bandwidth required by the theorem. Such a choice would also minimize the order of the integrated mean squared error of $\hat{\mu}$, and hence a bandwidth satisfying this property could be estimated via cross-validation. When such a bandwidth choice is feasible, we get that

$$\sqrt{n}(\hat{\tau}_{\text{kps}} - \tau) - \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\psi_i(\pi,\mu) - \tau) = O_P(n^{-(l+1)/(2(l+1)+d)}).$$

For example, in the case of local linear regression with a single covariate, where $l = d = 1$, the second order terms achieve their smallest possible magnitude $O_P(n^{-2/5})$ if we choose $h \propto n^{-1/5}$.

It is instructive to compare the rate of $O_P(n^{-2/5})$ to analogous ones for other efficient estimators of τ. Rothe and Firpo (2016) show that for $l = d = 1$ the difference between local polynomial versions of the estimators reviewed in Section 2.2 and their asymptotic linear representation is at best of the order $O_P(n^{-1/6})$ for inverse probability weighting, $O_P(n^{-3/10})$ for regression adjustment, and $O_P(n^{-7/18})$ for the doubly robust estimator. These rates are all slower than the one that can be achieved with knowledge of the propensity score; and the difference would become even more pronounced in higher-dimensional settings. This underscores how knowledge of the propensity score allows the construction of an estimator with superior second order properties even in settings with a single covariate.

3.3. **Parametric First-Stage Estimators.** An additional nice feature of the estimator $\hat{\tau}_{\text{kps}}$ is that it remains efficient if the estimator $\hat{\mu}$ is not obtained by nonparametric regression, but based on a correctly specified parametric model. This is a consequence of general results for double robust estimators derived for example in Robins, Rotnitzky, and Zhao (1994) or Scharfstein, Rotnitzky, and Robins (1999). The KPS estimator differs in this regard from regression adjustment and inverse probability weighting, which lose efficiency if correct parametric restrictions are imposed on the respective nuisance function. As in the nonparametric case, the estimator $\hat{\tau}_{\text{kps}}$ remains $\sqrt{n}$-consistent for τ, asymptotically normal, and asymptotically unbiased; and the simple variance estimator $\hat{V}_{\text{kps}}$ remains consistent for its asymptotic variance. To formally show this, we let $\mathcal{M} = \{m_\theta(t,x) : \theta \in \Theta\}$ be a class of candidates for μ that is indexed by $\Theta \subset \mathbb{R}^{2d}$, let $\hat{\mu} = m_{\hat{\theta}}$, where $\hat{\theta} \in \Theta$ is a function of the data, and impose the following standard regularity condition.

Assumption 5. *(i) $\mathcal{M}$ is such that $|m_{\theta_1}(t,x) - m_{\theta_2}(t,x)| \leq h(t,x)\|\theta_1 - \theta_2\|$ for all θ_1, θ_2 and some function h with $\mathbb{E}(|h(T_i, X_i)|^2) < \infty$; (ii) $\hat{\theta} = \theta^* + o_P(1)$ for some $\theta^* \in \Theta$.*

This structure covers most parametric procedures that are typically used for estimating conditional expectation functions in practice. Examples include Ordinary Least Squares (OLS) estimates of linear regression models, and Maximum Likelihood (ML) estimates of Probit/Logit specifications in the case of a binary outcome variable. Note that the assumption does not require that $\hat{\mu} = m_{\hat{\theta}}$ is a consistent estimator of μ, but only that it converges to a fixed probability limit m_{θ^*}. Moreover, the condition allows for parameter estimators with "irregular" rates of convergence, as it only requires that $\hat{\theta}$ is consistent for θ^*. The assumption yields the following result.

Theorem 3. *Suppose that Assumption 1 and 5 hold. Then*

$$\sqrt{n}(\hat{\tau}_{\text{kps}} - \tau) \xrightarrow{d} N(0, \mathbb{V}(\psi_i(\pi, m_{\theta^*}))) \quad \text{and} \quad \hat{V}_{\text{kps}} = \mathbb{V}(\psi_i(\pi, m_{\theta^*})) + o_P(1).$$

If in addition $\mu = m_{\theta^*}$, then $\mathbb{V}(\psi_i(\pi, m_{\theta^*})) = V_{\text{eff}}$ and thus $\hat{\tau}_{\text{kps}}$ reaches the semiparametric efficiency bound.

This result has an interesting implication for the analysis of settings where the treatment probability does not depend on the covariates, and thus $\pi(x)$ is a constant function. In this case, which occurs for example in the context of simple randomized experiments, empirical studies often estimate τ by a linear regression of the outcome variable on the treatment indicator and the covariates. That is, an estimate of τ is given by $\hat{\tau}_{\text{ols}}$, where

$$(\hat{\alpha}_{\text{ols}}, \hat{\tau}_{\text{ols}}, \hat{\beta}_{\text{ols}}) = \operatorname*{argmin}_{\alpha,\tau,\beta} \sum_{i=1}^{n} (Y_i - \alpha - T_i \cdot \tau - X_i'\beta)^2.$$

Simple algebra shows that $\hat{\tau}_{\text{ols}}$ in fact coincides with the estimator $\hat{\tau}_{\text{kps}}$ if a linear specification for μ is assumed and estimated by OLS. That is, if we define $\hat{\mu}(t,x) = \hat{\alpha}_{\text{ols}} + t\hat{\tau}_{\text{ols}} + x'\hat{\beta}_{\text{ols}}$, and have that $\pi(X_i) \equiv \pi^*$, then $\hat{\tau}_{\text{kps}} = \hat{\tau}_{\text{ols}}$. By Theorem 3, $\hat{\tau}_{\text{ols}}$ is thus fully efficient if the true conditional expectation of the outcome variable is linear in the treatment indicator and the covariates. This equivalence result does not carry over to experiments with more complicated randomization schemes where the propensity score varies across units. In such settings $\hat{\tau}_{\text{ols}}$ is generally inconsistent for τ, whereas the KPS estimator retains the properties described in Theorem 3.

4. MONTE CARLO EVIDENCE

In this section, we report the results of a Monte Carlo experiment that despite its simplicity nevertheless nicely illustrates the relevance of the theoretical results obtained above for finite sample settings. We consider a data generating process (DPG) under which the potential outcomes are generated as $Y_i(1) = \mu(1, X_i) + \varepsilon_i$ and $Y_i(0) = 0$, where $\mu(1,x) = (3x-1)^2$, $X_i \sim U(0,1)$, $\varepsilon_i \sim N(0, 1/2)$, and X_i and ε_i are stochastically independent; and propensity score is $\pi(x) = 1 - \mu(1,x)^2/5$. The setup is therefore such that $\tau = .5$ and that $V_{\text{eff}} \approx 1.917$. We then estimate the ATE using simulated data sets of size $n = 500$ from this DGP for the KPS estimator, the inverse probability weighting estimator (IPW), the regression adjustment estimator (REG), and the doubly robust estimator (DR). For each estimator, the

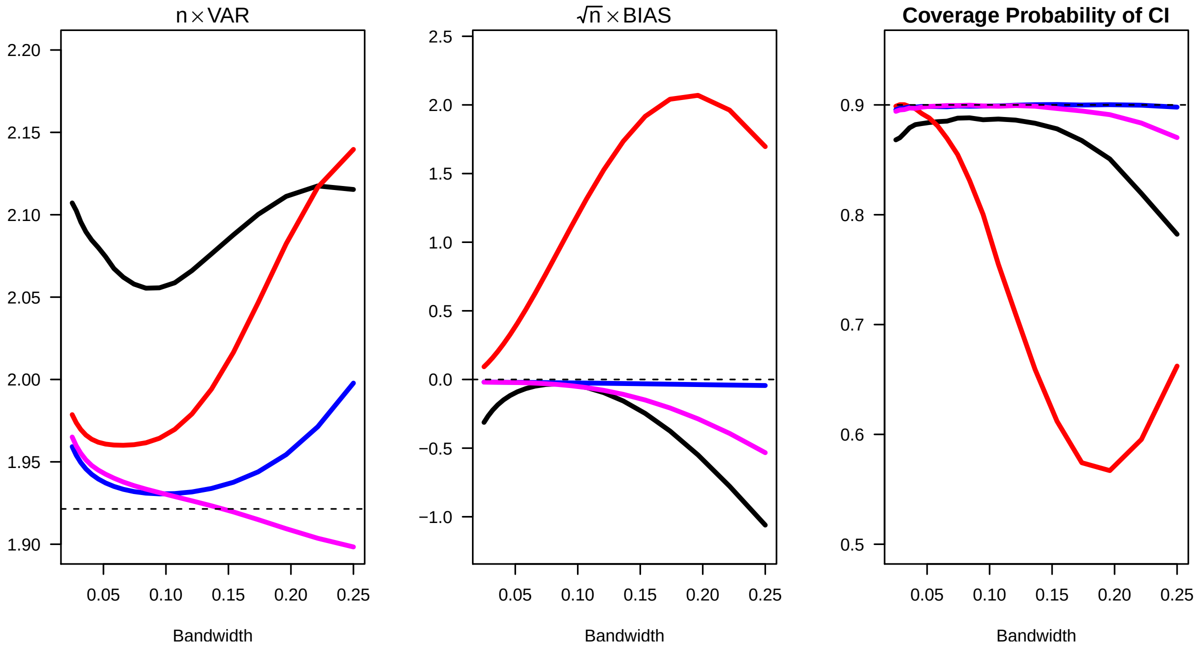

Figure 4.1: Simulation Results: Variance, bias and coverage probability of associated confidence interval with 90% nominal coverage rate for the KPS (blue line), IPW (black line), REG (red line) and DR (purple line) estimators as a function of the bandwidth used to estimate the respective nonparametric component(s).

respective nonparametric component is estimated via local linear regression with bandwidth $h \in \{.025, .05, \ldots, .25\}$. In each simulation run and each of the aforementioned procedures, we compute both the point estimator and the corresponding standard confidence interval for τ of the form (3.1) with nominal level $\alpha = .9$.

Figure 4.1 shows the results from 10,000 replications of our Monte Carlo experiment. The three panels display the estimators' variance, the estimators' bias, and the coverage probability of the corresponding confidence intervals, as a function of the bandwidth used to estimate the respective nonparametric component(s).[8] We can see that the KPS estimator is virtually unbiased for all values of h, that its variance is very close to the semiparametric efficiency bound for all but the largest values of h under consideration, and that the confi-

[8] The IPW estimator occasionally produces some extreme outliers for smaller values of h in our simulation setup. We exclude these outliers when calculating the summary statistics presented in Figure 4.1.

dence interval has approximately correct coverage probability uniformly across all bandwidth values we consider. These observations are exactly in line with our theoretical results.

The three remaining estimators all show deviations from the predictions of classical first order asymptotic theory, albeit to a different extent. REG and IPW are both substantially biased outside of a very narrow range of bandwidth values, and their finite sample variance is also well above V_{eff}. In consequence, the corresponding confidence intervals tend to under-cover the true parameter, with coverage probabilities as low as 55% for the REG estimator and 77% for IPW. The DR estimator does better in terms of bias and variance than REG and IPW, but nevertheless exhibits a noticeable bias for larger bandwidth values, which results in moderate under-coverage of the corresponding confidence interval.

5. CONCLUSIONS

This paper shows that knowledge of the propensity score can be exploited for the construction of an estimator of the ATE that achieves the semiparametric efficiency bound and has attractive theoretical and practical properties relative to other first order efficient procedures that have been proposed in the literature. It thereby clarifies that while imposing knowledge of the propensity score decreases the efficiency of many treatment effect estimators, this is not the case universally. Indeed, the paper shows that knowledge of the propensity score is immensely useful if it is exploited appropriately. Our results suggest that the KPS estimator should be used in empirical practice whenever the propensity score can reasonably be modeled as known, and they contribute more generally to the understanding of the role of the propensity score for ATE estimation.

A. MATHEMATICAL APPENDIX

A.1. **Proof of Theorem 1.** Let $\lambda_n(m) = n^{-1/2}\sum_{i=1}^n(\psi_i(\pi, m) - \mathbb{E}(\psi_i(\pi, m)))$ for any generic function $m(t, x)$ defined over $\{0, 1\} \times \mathcal{X}$ such that $\mathbb{E}(\psi_i(\pi, m))$ exists and is finite. Simple algebra shows that $\mathbb{E}(\psi_i(\pi, m)) = \tau$ for *any* such generic function $m(t, x)$, and thus $\lambda_n(m) =$

$n^{-1/2}\sum_{i=1}^{n}(\psi_i(\pi,m)-\tau)$. The first statement of the theorem follows from an application of the Central Limit Theorem to $\lambda_n(\bar{\mu})$ if $\lambda_n(\hat{\mu})-\lambda_n(\bar{\mu})=o_P(1)$. By linearity, we have that $\lambda_n(\hat{\mu})-\lambda_n(\bar{\mu})=\lambda_n(\hat{\mu}-\mu_n)+\lambda_n(\mu_n-\bar{\mu})$; and Assumption 2 implies that $\lambda_n(\mu_n-\bar{\mu})=O_P(b_n)=o_P(1)$. Next, for any fixed $m^*\in\mathcal{M}_n^*$ and any $\epsilon>0$ it holds that

$$\begin{aligned}
P(|\lambda_n(m^*)|>\epsilon) &\leq \frac{1}{\epsilon}\sup_{m\in\mathcal{M}_n^*}\mathbb{E}(|\lambda_n(m)|)\\
&\lesssim \frac{1}{\epsilon}\int_0^{a_n}\sqrt{\log(N_2(s,\mathcal{M}_n^*))}ds\\
&= \frac{a_n^{1-\alpha/2}c_n^{1/2}}{\epsilon},
\end{aligned}$$

using Markov's inequality, the maximal inequality in Corollary 19.35 in van der Vaart (1998), and our Assumption 3. Assumption 2 and 3 together also imply that $P(\hat{\mu}-\mu_n\in\mathcal{M}_n^*)=1+o(1)$, and thus we find that $\lambda_n(\hat{\mu}-\mu_n)=O_P(a_n^{1-\alpha/2}c_n^{1/2})=o_P(1)$, since $c_n=o(a_n^{\alpha-2})$ by Assumption 3. Taken together, we thus have that

$$\lambda_n(\hat{\mu})-\lambda_n(\bar{\mu})=O_P(a_n^{1-\alpha/2}c_n^{1/2})+O_p(b_n)=o_P(1),$$

as claimed. The remaining statements of the theorem follow from standard arguments. □

A.2. **Proof of Theorem 2.** The proof of this result follows directly from Lemma 3 in Rothe and Firpo (2016), who to study the properties of double robust estimators when both the propensity score and the conditional expectation function are estimated by local polynomial regression. □

A.3. **Proof of Theorem 3.** By Assumption 5, there exists a sequence $d_n=o(1)$ such that $\|\hat{\theta}-\theta^*\|=O_P(d_n)$. Now let $\bar{\mathcal{M}}_n=\{m_\theta\in\mathcal{M}:|\theta-\theta^*|\leq d_n\}$, and note that it follows from Example 19.7 in van der Vaart (1998) that $\log(N_2(\epsilon,\bar{\mathcal{M}}_n))\lesssim\log(1/\epsilon)$. By arguing as in the proof of Theorem 1, we also find that for any fixed $m^*\in\bar{\mathcal{M}}_n$ and any $\epsilon>0$ it holds that $P(|\lambda_n(m^*)|>\epsilon)\lesssim d_n/\epsilon$, and thus $\lambda_n(m_{\hat{\theta}})-\lambda_n(m_{\theta^*})=O_P(d_n)=o_P(1)$, which implies the statement of the theorem. □

REFERENCES

Abadie, A., and G. W. Imbens (2016): “Matching on the Estimated Propensity Score,” *Econometrica*, 84(2), 781–807.

Angrist, J., and J. Hahn (2004): “When to control for covariates? Panel asymptotics for estimates of treatment effects,” *Review of Economics and statistics*, 86(1), 58–72.

Angrist, J. D., and J.-S. Pischke (2008): *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

Cattaneo, M. (2010): “Efficient semiparametric estimation of multi-valued treatment effects under ignorability,” *Journal of Econometrics*, 155(2), 138–154.

Chen, X., H. Hong, and A. Tarozzi (2008): “Semiparametric Efficiency in GMM Models with Auxiliary Data,” *Annals of Statistics*, 36(2), 808–843.

Chen, X., O. Linton, and I. Van Keilegom (2003): “Estimation of semiparametric models when the criterion function is not smooth,” *Econometrica*, 71(5), 1591–1608.

Fan, J. (1993): “Local linear regression smoothers and their minimax efficiencies,” *Annals of Statistics*, 21(1), 196–216.

Fan, J., and I. Gijbels (1996): *Local polynomial modelling and its applications*. Chapman & Hall/CRC.

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

Hahn, J., and G. Ridder (2013): “Asymptotic variance of semiparametric estimators with generated regressors,” *Econometrica*, 81(1), 315–340.

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

Ichimura, H., and S. Lee (2010): “Characterization of the asymptotic distribution of semiparametric M-estimators,” *Journal of Econometrics*, 159(2), 252–266.

Imai, K., and D. A. van Dyk (2004): “Causal inference with general treatment regimes: generalizing the propensity score,” *Journal of the American Statistical Association*, 99(467), 854–867.

Imbens, G. (2004): “Nonparametric estimation of average treatment effects under exogeneity: A review,” *Review of Economics and Statistics*, 86(1), 4–29.

Imbens, G., W. Newey, and G. Ridder (2007): “Mean-square-error calculations for average treatment effects,” *Working Paper.*

Imbens, G., and J. Wooldridge (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47, 5–86.

Khan, S., and E. Tamer (2010): "Irregular identification, support conditions, and inverse weight estimation," *Econometrica*, 78(6), 2021–2042.

Linton, O. (1995): "Second order approximation in the partially linear regression model," *Econometrica*, 63(5), 1079–1112.

Mammen, E., C. Rothe, and M. Schienle (2016): "Semiparametric estimation with generated covariates," *Econometric Theory*, p. to appear.

Masry, E. (1996): "Multivariate local polynomial regression for time series: uniform strong consistency and rates," *Journal of Time Series Analysis*, 17(6), 571–599.

Newey, W. (1994): "The Asymptotic Variance of Semiparametric Estimators," *Econometrica*, 62, 1349–1382.

Newey, W., and D. McFadden (1994): "Large sample estimation and hypothesis testing," *Handbook of Econometrics*, 4, 2111–2245.

Newey, W. K. (1997): "Convergence rates and asymptotic normality for series estimators," *Journal of Econometrics*, 79(1), 147–168.

Ojeda, J. (2008): "Hölder continuity properties of the local polynomial estimator," *Working Paper*.

Portier, F., and J. Segers (2015): "On the weak convergence of the empirical conditional copula under a simplifying assumption," *Working Paper*.

Robins, J., and Y. Ritov (1997): "Toward a Curse of Dimensionality Appropriate (CODA) Asymptotic Theroy for Semi-Parametric Models," *Statistics in Medicine*, 16(3), 285–319.

Robins, J., and A. Rotnitzky (1995): "Semiparametric efficiency in multivariate regression models with missing data," *Journal of the American Statistical Association*, 90(429), 122–129.

Robins, J., A. Rotnitzky, and L. Zhao (1994): "Estimation of regression coefficients when some regressors are not always observed," *Journal of the American Statistical Association*, 89(427), 846–866.

Rosenbaum, P., and D. Rubin (1983): "The central role of the propensity score in observational studies for causal effects," *Biometrika*, 70(1), 41–55.

Rothe, C., and S. Firpo (2016): "Semiparametric estimation and inference using doubly robust moment conditions," *Working Paper*.

Ruppert, D., and M. Wand (1994): "Multivariate locally weighted least squares regression," *Annals of Atatistics*, 22(3), 1346–1370.

Scharfstein, D., A. Rotnitzky, and J. Robins (1999): "Adjusting for nonignorable drop-out using semiparametric nonresponse models," *Journal of the American Statistical Association*, 94(448), 1096–1120.

van der Vaart, A. (1998): *Asymptotic Statistics*. Cambridge University Press.