Flexible Covariate Adjustments in Randomized Experiments

Christoph Rothe

Abstract: Linear regression adjustments for pre-treatment covariates are widely used in economics to lower the variance of treatment effect estimates when analyzing data from randomized experiments. This method is robust to misspecification, and delivers reliable confidence intervals even in relatively small samples. More flexible covariate adjustments, using nonlinear parametric or fully nonparametric methods, have the potential to improve efficiency. They are rather uncommon in practice, however, because they can introduce bias or require very large samples in order for asymptotic inference to be reliable. This paper shows that with a simple modification of the treatment effect estimator, it is possible to alleviate these issues substantially. For a large class of covariate adjustments, estimation and inference in randomized experiments is possible without sacrificing the robustness properties of linear regressions. Full efficiency can be achieved through nonparametric adjustments under minimal conditions, in particular without imposing high-order smoothness restrictions in settings with many covariates.

JEL Classification: C13, C14, C21

Keywords: *Treatment effects, semiparametric efficiency, randomized experiment, nonparametric estimation.*

This Version: May 18, 2018. This paper replaces an earlier working paper titled "The Value of Knowing the Propensity Score for Estimating Average Treatment Effects", which was first circulated in April, 2015. I thank Joshua Angrist, Miikka Rokkanen, Stefan Wager and attendants of the 2016 Econometric Society Winter Meeting in San Francisco for helpful comments and discussions. Author's contact information: Department of Economics, University of Mannheim, Email: rothe@vwl.uni-mannheim.de, Web: http://www.christophrothe.net.

1. Introduction

Randomized experiments have become increasingly popular in many areas of applied economics, as they allow for straightforward inference on causal effects. With a binary treatment, for instance, the difference in average outcomes among treated and untreated units constitutes an unbiased estimator of the average treatment effect. It is also easy to form a confidence interval based on this estimator that, while formally justified by asymptotic theory, is known to work well even with rather moderate sample sizes.

The "difference-in-means" estimator is not statistically efficient, however, if one observes a vector of pre-treatment covariates in addition to units' outcomes and treatment assignments. Such data are frequently available in economic applications, and often incorporated into the analysis as additional control variables in a linear regression of outcomes on a treatment dummy. Standard regression theory implies that this (weakly) improves the efficiency of the treatment effect estimator without introducing bias, irrespective of whether the linear model is correctly specified or not. Moreover, it is straightforward to form asymptotically valid confidence intervals by using heteroscedasticity-robust standard errors.[1]

While simple linear regression adjustments lead to improvements over a "difference-in-means" analysis, they are generally not optimal. That is, unless the linear regression model is correctly specified, the resulting estimator of the average treatment effect does not have the smallest possible asymptotic variance that a regular estimator could have in such a setup.

More precise estimates can in principle be obtained by postulating more flexible parametric specifications of the relationship between covariates and outcomes. Moreover, full statistical efficiency can also be achieved by adjusting for covariates nonparametrically, using methods

[1] Throughout the paper, the parameter of interest is the population-level average treatment effect, and sampling from the population is the source of uncertainty about its value. This framework is commonly used in economics. An alternative framework, used frequently in the statistics literature, considers the sample average treatment effect of the observed units as the parameter of interest, and random assignment of the treatment as the only source of uncertainty. Linear regression adjustments can be biased in the latter framework, and potentially increase the variance of the treatment effect estimator. See Freedman (2008) and Lin (2013) for further discussion.

like series regression or local linear smoothing. Both kinds of adjustments are not widely used in practice, however, since they lack the robustness properties of the linear regression approach. Since there is no simple analogue of the "omitted variable bias formula" in nonlinear models, nonlinear parametric adjustments can lead to biased treatment effect estimates under misspecification (e.g. Gail, Wieand, and Piantadosi, 1984). Nonparametric adjustments on the other hand can be sensitive to the choice of tuning parameters in finite samples, can also introduce bias, and are only guaranteed to lead to full efficiency under delicate regularity conditions that are generally considered unrealistic in settings with several covariates (e.g. Robins and Ritov, 1997).

In this paper, we show that such concerns can be substantially alleviated by choosing an appropriately modified treatment effect estimator. The modification makes it is possible to conduct simple and robust inference in randomized experiments with nonlinear parametric or nonparametric covariate adjustments under very general conditions. Building on ideas from the double robustness literature (e.g. Robins, Rotnitzky, and Zhao, 1995; Robins and Rotnitzky, 2001; Van der Laan and Robins, 2003; Bang and Robins, 2005), we estimate the average treatment effect by the sample average of an empirical analogue of the efficient influence function. The latter is a simple transformation of the data, the treatment assignment probabilities, and the covariate-adjusted outcomes.

Using techniques from empirical process theory, we show that the resulting treatment effect estimator is $\sqrt{n}$-consistent (where n denotes the sample size), asymptotically unbiased, and asymptotically normal for a generic class of covariate adjustments that includes misspecified parametric and slowly converging nonparametric ones.[2] The estimator achieves full statistical efficiency if the covariate-adjustment procedure chosen by the researcher consistently estimates

[2]We focus on a traditional setup in which the number of covariates is fixed as the sample size increases. See Wager, Du, Taylor, and Tibshirani (2016), and Wu and Gagnon-Bartsch (2017) for results on estimators similar to ours in high-dimensional settings when variable selection methods or machine learning algorithms are used to adjust for covariates.

the conditional expectation of outcomes given treatment status and covariates. Valid inference is not contingent on achieving full efficiency: using a simple estimator of the treatment effect estimator's asymptotic variance, based on the average "squared" efficient influence function, one can construct confidence intervals that are asymptotically valid for any kind of covariate adjustment contained within our generic class. These results show that flexible covariate adjustments can be used in practice for inference on average treatment effects without sacrificing any of the robustness properties of linear regression adjustments.

With nonparametric covariate adjustments, for example via local linear or series regression, our treatment effect estimator falls into the class of semiparametric two-step (STS) estimators.[3] Conditions for STS estimators to be $\sqrt{n}$-consistent and asymptotically normal typically include that both the bias and the stochastic part of the nonparametric component are of smaller order than $n^{-1/4}$. In settings with several covariates, these two objectives can generally only be achieved under strong smoothness conditions on the function that is being estimated nonparametrically. Such conditions formally justify the use of bias-reducing nonparametric estimators, such as higher-order local polynomial regression. However, in practice the properties of the resulting STS estimators are often poorly approximated by the corresponding asymptotic theory (e.g. Robins and Ritov, 1997). Our main result differs substantially from those typically found in this literature in that it does not require any assumptions about the rate at which the bias of the nonparametric component tends to zero. This is because the structure of the efficient influence function, paired with randomized treatment assignment, removes the influence of first stage bias on the final estimator. Efficient estimation can be carried out without any higher-order differentiability conditions: randomized assignment effectively removes the "curse of dimensionality".

This paper is also related to literature on treatment effect estimation from observational

[3]STS estimators are estimators of a finite-dimensional parameter that depend on a nonparametrically estimated nuisance function. See Newey (1994), Andrews (1995), or Chen, Linton, and Van Keilegom (2003), among many others, for general results regarding their theoretical properties.

data. This is because the framework is formally equivalent to a setup with unconfounded treatment assignment and a known propensity score (cf. Imbens, 2004). Knowledge of the propensity score does not affect the efficiency bound for estimating the average treatment effect under unconfoundedness, and efficient estimation is possible if such knowledge is simply ignored (Hahn, 1998). Any treatment effect estimator that is efficient under unconfounded assignment is therefore also efficient when the propensity score is known. This raises the question to what extent knowledge of the propensity score is of practical or theoretical value in setups with unconfounded treatment assignment. The results of this paper can be seen as an answer to this question: knowledge of the propensity score makes it possible to construct estimators that achieve the efficiency bound under substantially weaker regularity conditions, and that have superior finite-sample properties.[4]

The remainder of the paper is structured as follows. In the following section, we introduce the model and our proposed estimation procedure. In Section 3, we derive its theoretical properties under general conditions, and propose a simple method for inference. Sections 4 and 5 specifically consider parametric and nonparametric covariate adjustments, respectively. Section 6 presents some simulation results, and Section 7 concludes. All proofs are collected in the Appendix.

2. Setup

This section describes the model, reviews some results on efficiency bounds for treatment effect estimation, and introduces the proposed estimation procedure.

2.1. **Model.** In our model, the researcher observes an i.i.d. random sample $\{(Y_i, T_i, X_i)\}_{i=1}^n$ of n units from a large population. Here $Y_i \in \mathcal{Y} \subset \mathbb{R}$ is the outcome variable, $T_i \in \{0, 1\}$ is a a treatment indicator, with $T_i = 1$ if unit i is treated and $T_i = 0$ otherwise, and $X_i \in \mathcal{X} \subset \mathbb{R}^d$ is

[4]This point of view had been emphasized in Rothe (2016), a working paper that has been superseded by the present paper.

a vector of pre-treatment covariates. Following the usual Neyman-Rubin framework, each unit has potential outcomes $Y_i(1)$ and $Y_i(0)$ with and without receiving the treatment, respectively, so that $Y_i = Y_i(T_i)$. The parameter of interest is the population average treatment effect

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

Treatment assignment is independent of the potential outcomes, but the probability of receiving the treatment is allowed to depend on the value of the covariates. Formally, we assume that

$$Y_i(1), Y_i(0) \perp T_i | X_i \quad \text{and} \quad \epsilon \leq \pi(X_i) \leq 1 - \epsilon \tag{2.1}$$

almost surely, for some constant $\epsilon > 0$ and $\pi(x) = P(T_i = 1 | X_i = x)$, the propensity score function, chosen by the analyst. In many randomized experiments the propensity score function is constant, and also equal to 1/2. This corresponds to a setup in which units are assigned to the treatment and control group with equal probability, irrespective of the value of their covariates. Non-constant propensity score functions might be chosen, for example, to improve statistical efficiency (if outcomes are known to be less variable among units with certain covariate values), or to make the data collection process less costly (if the cost of sampling units is related to their covariate values). Assuming that the propensity score is bounded away from zero and one ensures the existence of a regular estimator of τ (Khan and Tamer, 2010). Since the propensity score is chosen by the analyst, this condition is easy to fulfill in practice.

2.2. **Efficiency Bounds.** 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 the conditional variance function, respectively, of Y_i given $T_i = t$ and $X_i = x$. Hahn (1998) shows that under assumption (2.1) the asymptotic

variance of any regular estimator of τ is bounded from below by

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

and that any regular estimator $\widetilde{\tau}$ of τ whose asymptotic variance achieves the bound V_{eff} can be written as $\widetilde{\tau} = n^{-1} \sum_{i=1}^{n} \psi_i(\mu) + o_P(n^{-1/2})$, where

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

is the efficient influence function for estimating τ. It thus holds that

$$V_{\text{eff}} = \mathbb{V}(\psi_i(\mu)).$$

Our notation $\psi_i(\mu)$ for the efficient influence function emphasizes the fact that this quantity is known up to the conditional expectation function μ.

2.3. **Estimators.** In our model, the researcher postulates some specification of the conditional expectation function μ, and then generates an appropriate estimate $\hat{\mu}$. This could be a parametric specification of μ that is estimated by methods like least squares or maximum likelihood, or a nonparametric specification estimated by methods like local linear or series regression. Irrespective of which procedure is used by the researcher, the usual "covariate-adjusted" estimator of τ is given by

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

Important special cases of this estimator include the simple difference-in-means estimator that ignores the covariates, which can be obtained by fitting the specification $\mu(t, x) = \beta_0 + \beta_1 t$ by ordinary least squares, and the linear-regression-adjustment estimator, which is based

on the ordinary least squares estimate of the specification $\mu(t,x) = \beta_0 + \beta_1 t + \beta_2' x$. Both of the aforementioned versions of $\hat{\tau}_{adj}$ are consistent and unbiased irrespective of whether the chosen specification is correct or not if the propensity score function is constant (with a non-constant propensity score, the respective specification of μ has to be estimated by propensity-score-weighted least squares to retain this feature).

In general, consistency of $\hat{\tau}_{adj}$ requires correct specification, and consistent estimation, of the conditional expectation function μ. Suppose for example that the outcome is binary, and that the propensity score function is constant. Then it is natural to consider a Logit specification of the conditional expectation function, like $\mu(t,x) = \Lambda(\beta_0 + \beta_1 t + \beta_2' x)$, with Λ the cumulative distribution function of the Logistic distribution. If this model is misspecified, the resulting treatment effect estimator $\hat{\tau}_{adj}$ is generally biased and inconsistent. Versions of $\hat{\tau}_{adj}$ based on nonparametric estimates of μ are consistent under general conditions, but can be biased in finite samples and exhibit stochastic behavior that is not well approximated by conventional first-order asymptotic theory (Imbens, Newey, and Ridder, 2007).

These issues make it unattractive to use $\hat{\tau}_{adj}$ in practice with general covariate adjustments. We therefore consider a sample average of the natural estimate of the efficient influence function, namely

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

as an alternative estimator of τ.[5] We refer to $\hat{\tau}$ as the EIF estimator in the following. It can

[5]Versions of this estimator are considered by Wager et al. (2016) and Wu and Gagnon-Bartsch (2017) in different frameworks in the context of high-dimensional covariate adjustments. If the true value of the propensity score is replaced with an estimate, $\hat{\tau}$ becomes a special case of a doubly robust estimator. Such estimators are studied in classical setups by Robins et al. (1995), Robins and Rotnitzky (2001), Van der Laan and Robins (2003), or Bang and Robins (2005), among many others; and in high-dimensional or nonparametric setups by Belloni, Chernozhukov, and Hansen (2014), Farrell (2015), or Rothe and Firpo (2018), again among many others.

be understood as an "appropriately corrected" version of $\hat{\tau}_{adj}$, in the sense that

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

The rationale for considering $\hat{\tau}$ is the observation that $\mathbb{E}(\psi_i(m)) = \tau$ for any non-random function m for which the last expectation exists and is finite. Since $\mathbb{E}(\psi_i(m)) \approx (1/n)\sum_{i=1}^{n}\psi_i(m)$ in large samples, we expect $\hat{\tau}$ to be consistent for τ irrespective of whether $\hat{\mu}$ is consistent for μ; and that sampling variation in $\hat{\mu}$ has a relatively minor impact on the sampling variation of $\hat{\tau}$. Specifically, with $\bar{\mu}$ denoting the probability limit of $\hat{\mu}$, we show below that in large samples

$$\sqrt{n}(\hat{\tau} - \tau) \approx \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\psi_i(\bar{\mu}) - \tau)$$

with high accuracy under remarkably weak restrictions on the estimator $\hat{\mu}$. Here the term on the right-hand side of the last equation is commonly referred to as the asymptotically linear representation of $\hat{\tau}$. This relationship implies that $\hat{\tau}$ is approximately unbiased for τ, $\sqrt{n}$-consistent and asymptotically normal with limiting variance $\mathbb{V}(\psi_i(\bar{\mu}))$, and fully efficient if $\hat{\mu}$ consistently estimates μ; i.e. $\bar{\mu} = \mu$. Moreover, the sample variance of the $\psi_i(\hat{\mu})$, $i = 1, \ldots, n$, is a simple and natural estimate of $\mathbb{V}(\psi_i(\bar{\mu}))$; and this estimate can be used to construct confidence intervals for τ that are valid irrespective of whether $\bar{\mu} = \mu$.

3. Theoretical Results

In this section, we formally study the properties of the estimator $\hat{\tau}$ under general conditions on the properties of the estimator $\hat{\mu}$. We also propose methods to estimate the asymptotic variance, and to conduct inference.

3.1. **Regularity Conditions.**

The following notation is helpful for presenting our regularity conditions. 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 and b if $a(\eta) \leq Cb(\eta)$ for some constant C not depending on η. All limits are taken as $n \to \infty$. We impose two "high level" assumptions about the estimator $\hat{\mu}$.

Assumption 1. *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 $\bar{\mu}$ equal to the probability limit of $\hat{\mu}$, and to define the function μ_n as the sum of $\bar{\mu}$ and the asymptotic bias of the respective estimator $\hat{\mu}$ with respect to $\bar{\mu}$. Put differently, μ_n is chosen such that $\hat{\mu} - \mu_n$ is approximately mean zero. This allows, but does not require, $\hat{\mu}$ to be a consistent estimator of μ since is possible, but not necessary, that $\bar{\mu} = \mu$. With such choices of $\bar{\mu}$ and μ_n, Assumption 1 simply requires that the stochastic part and the bias of $\hat{\mu}$ converge to zero uniformly. It also 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 parametric models, such results generally follow if $\hat{\mu}$ is not "too volatile" in the estimated parameter (Andrews, 1992)

Assumption 2. *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. They ensure that the estimator

$\hat{\mu}$ does not overfit 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). The presence of the sequence c_n allows for function classes whose complexity increases with the sample size. 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.[6] For parametric models, the assumption again requires that $\hat{\mu}$ is not "too volatile" in the estimated parameter, in some appropriate sense.

We discuss explicit "low level" conditions under which Assumption 1–2 are satisfied for conventional parametric and nonparametric methods in Sections 4 and 5 below, respectively. For the moment, we would like to stress two important points. First, our two assumptions only restrict the rate a_n at which the stochastic component of $\hat{\mu}$ tends to zero, but not the rate b_n of the bias component. In our setup the estimator $\hat{\mu}$ can therefore not only be inconsistent for μ; it is also allowed to have an arbitrarily slowly vanishing asymptotic bias. Second, while our analysis is geared towards classical parametric and nonparametric estimation procedures for μ, our setup allows for all kinds of estimators $\hat{\mu}$. In particular, our assumptions in principle allow for estimators of μ based data dependent model specifications, or data dependent choices of tuning parameters, including those of the type used in modern machine learning methods. The only restriction is that in order to satisfy our Assumptions 1–2, such data dependencies cannot be arbitrary, but have to be subject to sufficient "discipline".[7]

3.2. **Treatment Effect Estimation.** Our main result regarding the properties of the treatment effect estimator we consider in this paper is the following bound on the difference between $\sqrt{n}(\hat{\tau} - \tau)$ and its asymptotically linear representation, which is obtained using techniques from empirical process theory.

[6] Assumption 2 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.

[7] For example, see Benkeser and Van Der Laan (2016) and Van Der Laan and Bibau (2017) for conditions under which entropy and uniform convergence results hold for LASSO-type estimators.

Theorem 1. *Suppose that Assumptions 1–2 hold. Then*

$$\sqrt{n}(\hat{\tau} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi_i(\bar{\mu}) - \tau) + O_P(c_n^{1/2} a_n^{1-\alpha/2}) + O_P(b_n). \tag{3.1}$$

Moreover, the term of order $O_P(b_n)$ in the previous equation has mean zero.

Theorem 1 gives a "worst case" bound that is valid for *any* estimator $\hat{\mu}$, including hypothetical or infeasible ones, as long as they satisfy the assumed high-level regularity conditions. We remark that when considering a specific estimator of μ, the result of this theorem can potentially be improved by exploiting special features of the respective procedure. We illustrate this point in the context of local linear regression in Section 5 below.

A direct implication of Theorem 1 is the following result.

Corollary 1. *Suppose that Assumptions 1–2 hold. Then (i) $\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, \mathbb{V}(\psi_i(\bar{\mu})))$; and (ii) $\mathbb{V}(\psi_i(\bar{\mu})) = V_{\text{eff}}$ if $\bar{\mu} = \mu$.*

The corollary shows that under our regularity conditions $\hat{\tau}$ is $\sqrt{n}$-consistent for τ, asymptotically normal, and asymptotically unbiased, irrespective of whether $\hat{\mu}$ consistently estimates μ. Moreover, it shows that the asymptotic variance of $\hat{\tau}$ reaches the efficiency bound V_{eff} if $\hat{\mu}$ consistently estimates μ, irrespective of the exact estimation procedure that is used to construct $\hat{\mu}$. This result is remarkable relative other asymptotic normality results for estimators of a finite-dimensional parameter that depend on flexible, possibly nonparametric estimates of a nuisance function, as it does not require any restrictions on the rate b_n at which the bias of $\hat{\mu}$ tends to zero. While the details depend on the exact specification, generally speaking this means that we can often satisfy Assumptions 1–2 by choosing an estimator $\hat{\mu}$ that sufficiently "over-smooths", or "over-regularizes", the data. This is because for all commonly used estimation procedures increased regularization speeds up the rate of convergence of the stochastic part and decreases the "complexity" of the estimated function.[8]

[8]To see this, consider the case of classical nonparametric regression models. With local polynomial

On the other hand, $\sqrt{n}$-consistency and asymptotic normality of a generic estimator that depends on a flexibly estimated nuisance function often requires that the stochastic part *and* the bias of the estimated nuisance function are of smaller order than $n^{-1/4}$. 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 accurately estimated nuisance function can be fragile in practice. The fact that $\hat{\tau}$ does not require such conditions make it attractive for empirical work.

3.3. **Variance Estimation and Confidence Intervals.** For empirical practice, it is important to obtain a consistent estimate of the asymptotic variance of $\hat{\tau}$, since this can be transformed into valid standard errors and confidence intervals for the parameter of interest. Since $\mathbb{V}(\psi_i(\bar{\mu})) = \mathbb{E}((\psi_i(\bar{\mu}) - \tau)^2)$, the natural estimate of the asymptotic variance of $\hat{\tau}$ is

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

irrespective of how $\hat{\mu}$ is obtained. This variance estimate is straightforward to compute, as it does not require estimating any nuisance parameters in addition to $\hat{\mu}$. It can be shown to be consistent under the conditions of Theorem 1. Together with the asymptotic normality result in Corollary 1, this then implies the validity of standard large sample methods for inference. In particular, it means that a confidence interval for τ with asymptotic coverage $1-\alpha$ is given by

$$C_{1-\alpha} = \left(\hat{\tau} \pm c_{1-\alpha} \times \sqrt{\hat{V}/n}\right),$$

where $c_{1-\alpha} = \Phi^{-1}(1-\alpha/2)$ with $\Phi^{-1}(\cdot)$ the standard normal quantile function is the critical value. The following corollary formally states these results.

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.

Corollary 2. *Suppose that Assumptions 1–2 hold. Then (i)* $\hat{V} = \mathbb{V}(\psi_i(\bar{\mu})) + o_P(1)$*; (ii)* $P(\tau \in C_{1-\alpha}) = 1 - \alpha + o(1)$*; and (iii)* $P(\bar{\tau} \in C_{1-\alpha}) = o(1)$ *for all* $\bar{\tau} \neq \tau$*.*

3.4. **Leave-One-Out Estimation.** An interesting variant of the estimator $\hat{\tau}$ is one where for $t = 0, 1$ and $i = 1, \ldots, n$ the estimate of $\mu(t, X_i)$ is computed without using the ith data point. Denoting the corresponding estimator by $\hat{\mu}_{(-i)}(t, X_i)$, the resulting "leave-one-out" treatment effect estimator is

$$\hat{\tau}_{LOO} = \frac{1}{n} \sum_{i=1}^{n} \psi_i(\hat{\mu}_{(-i)}).$$

"Leave-one-out" estimation is well-known to reduce the risk of over-fitting the data, and to reduce bias (e.g. Powell, Stock, and Stoker, 1989). In the present context, assume that for $t = 0, 1$ and $i = 1, \ldots, n$

$$\mathbb{E}(\hat{\mu}_{(-i)}(t, x) | Y_i, T_i, X_i) = \mu_n(t, x) \tag{3.2}$$

exists, is finite, and non-random. Then, by the law of iterated expectations and the linearity of the efficient influence function $\psi_i(\mu)$ in μ, we have that

$$\mathbb{E}(\hat{\tau}_{LOO}) = \mathbb{E}(\psi_i(\mu_n)) = \tau,$$

which means that $\hat{\tau}_{LOO}$ is exactly unbiased.[9] The following corollary shows that under a weak regularity condition, which basically states that one must be able to interpolate the estimates $\hat{\mu}_{(-i)}(t, X_i)$ with a sufficiently regular function, the conclusions of Theorem 1 hold for "leave-one-out" treatment effect estimation as well.

Corollary 3. *Suppose there exists a function* $\hat{\mu}$ *that satisfies Assumption 1–2, and is such that* $\hat{\mu}(t, X_i) = \hat{\mu}_{(-i)}(t, X_i)$ *for* $t = 0, 1$ *and* $i = 1, \ldots, n$*. Then the conclusion of Theorem 1*

[9]For many estimators the expectation in 3.2 does formally not exist. In this case we can obtain an analogous "conditional unbiasedness" result by assuming that the conditional expectation of $\hat{\mu}_{(-i)}(t, x)$ given all treatment assignments and covariates values exists.

holds with $\hat{\tau}_{LOO}$ replacing $\hat{\tau}$.

4. Parametric Covariate Adjustments

In this section, we give primitive conditions under which our Assumptions 1–2 are satisfied when $\hat{\mu}$ is a parametric estimation procedure. Suppose that the researcher uses the specification $\mu(t,x) = m_\theta(t,x)$, where m_θ is a function that is known up $\theta \in \Theta \subset \mathbb{R}^s$ for some $s \in \mathbb{N}$. This specification may be correct or incorrect. That is, there may or may not be some $\theta \in \Theta$ such that $\mu = m_\theta$. Also put $\hat{\mu} = m_{\hat{\theta}}$, where $\hat{\theta} \in \Theta$ is some estimator. We then impose the following regularity condition.

Assumption 3. *(i) The function $m_\theta(t,x)$ is such that $|m_{\theta_1}(t,x) - m_{\theta_2}(t,x)| \leq h(t,x)\|\theta_1 - \theta_2\|$ for all $\theta_1, \theta_2 \in \Theta$ and some function h with $\mathbb{E}(|h(T_i,X_i)|^2) < \infty$; (ii) there exists $\theta^* \in \Theta$ and a deterministic sequence θ_n^* taking values in Θ such that $\|\hat{\theta} - \theta_n^*\| = O_P(a_n)$ and $\|\theta_n^* - \theta^*\| = O(b_n)$ for $a_n = o(1)$ and $b_n = o(1)$.*

The first part of this assumption is a continuity condition on the candidate functions for μ with respect to the unknown parameter; and the second part prescribes that $\hat{\theta}$ has a non-random probability limit, and that its stochastic and bias component vanish with rate a_n and b_n, respectively. 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^*}. The assumption also allows for parameter estimators with "irregular" rates of convergence, i.e. ones that differ from the usual rate of $n^{-1/2}$, as it only requires that $\hat{\theta}$ is consistent for some θ^*.

Corollary 4. *Suppose that Assumption 3 holds. Then Assumptions 1–2 are satisfied with*

$\mathcal{M}_n = \{m_\theta(t,x) : \theta \in \Theta\}$, $\bar{\mu} = m_{\theta^*}$, $\mu_n = m_{\theta_n^*}$, $c_n = 1$, *and* $\alpha > 0$ *arbitrarily small.*

Through a minor variation of the proof of Theorem 1, one can show that its conclusion also holds under the conditions of Corollary 4 with $\alpha = 0$; see the appendix for details. This implies that the difference between $\sqrt{n}(\hat{\tau} - \tau)$ and its asymptotically linear representation is of the same order as the rate at which $\hat{\theta}$ approaches θ^*. That is, since $\|\hat{\theta} - \theta^*\| = O_P(a_n) + O(b_n)$ under our assumptions, it holds that

$$\sqrt{n}(\hat{\tau} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi_i(\bar{\mu}) - \tau) + O_P(\|\hat{\theta} - \theta^*\|).$$

In the most widely used parametric models one can typically choose θ_n^* such that $\|\hat{\theta} - \theta_n^*\| = O_P(n^{-1/2})$ and $\|\theta_n^* - \theta^*\| = O(n^{-1})$ under standard conditions, but, as mentioned above, setups with slower converging parametric estimators exist.

5. Nonparametric Covariate Adjustments

In this section, we give primitive conditions under which our Assumptions 1–2 are satisfied when $\hat{\mu}$ is obtained by classical nonparametric regression techniques. For concreteness, we focus on the case of local linear regression; but similar results can be obtained for other methods, such as series regression. We also illustrate how direct arguments can be used to derive results that improve upon the general finding of Theorem 1.

5.1. **Notation and Assumptions.** Local linear 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. Let $\mathcal{K}$ be a univariate probability density function, and put $K_h(b) = \prod_{j=1}^{d} \mathcal{K}(b_j/h)/h$ for any bandwidth $h \in \mathbb{R}_+$. Then the local linear estimator $\hat{\mu}(t,x)$ of $\mu(t,x)$ is given by the first

component of

$$(\hat{\mu}(t,x), \hat{\beta}(t,x)) = \underset{(m,b)}{\operatorname{argmin}} \sum_{j=1}^{n} \left(Y_j - m - b'(X_j - x)\right)^2 K_h(X_j - x)\mathbb{I}\{T_j = t\}.$$

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 linear regression.

Assumption 4. *(i) X_i is continuously distributed given $T_i = t$ with compact and convex support $\mathbb{X} \subset \mathbb{R}^d$ 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 twice 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 l times 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.*

5.2. **Application of General Results.** Let $\mathcal{M}(l, c_n)$ 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 are uniformly bounded by c_n for $t = 0, 1$. We then have the following result.

Corollary 5. *Suppose that Assumption 4 holds with $l > \max\{1, d/2\}$, and that $h \propto n^{-\theta}$ with*

$$0 < \theta < \frac{3 - d/l}{(3 + 2l - d/l)d}.$$

Then Assumptions 1–2 are satisfied with $\mathcal{M}_n = \mathcal{M}(l, c_n)$, $\alpha = d/l$, $a_n = (nh^d/\log(n))^{-1/2}$, $b_n = h^2$, and $c_n = (nh^{d(1+2l)}/\log(n))^{-1/2}$.

Together with our main results, the corollary implies that $\hat{\tau}$ reaches the semiparametric

efficiency bound with local linear covariate adjustments if the bandwidth h does not tend to zero too quickly. However, the bandwidth is allowed to vanish arbitrarily slowly, and thus asymptotic efficiency can alway be achieved by choosing a sufficiently slowly vanishing value of h.[10]

Note that Corollary 5 only requires that the kernel function $\mathcal{K}$ has derivatives up to some order that increases with the number of covariates. This condition ensures that $\hat{\mu}$ is sufficiently often differentiable for taking values in $\mathcal{M}(l, c_n)$ with high probability. No higher-order differentiability conditions on μ are needed, as it is not necessary for our results that the bias of $\hat{\mu}$ vanishes quickly. This differs markedly from generic STS estimators, which typically require higher-order differentiability conditions on the respective nuisance function in settings with many covariates in order to justify the use of methods that control the magnitude of the asymptotic bias of the respective nonparametric estimate.

5.3. **Improved Results Using Direct Arguments.** It is possible to improve upon the result of Corollary 5 and Theorem 1 in the present context by using direct arguments that exploit the specific structure of the local linear estimator. Consider the leave-one-out (LOO) version of the treatment effect estimator, which uses an estimate of μ that uses every observation but the ith in order to estimate $\mu(t, X_i)$, for $t = 0, 1$. Specifically, let

$$\hat{\tau}_{LOO} = \frac{1}{n} \sum_{i=1}^{n} \psi_i(\hat{\mu}_{(-i)})$$

where, for $i = 1, \ldots, n$ and $t = 0, 1$, the estimator $\hat{\mu}_{(-i)}(t, X_i)$ is the first component of

$$(\hat{\mu}_{(-i)}(t, X_i), \hat{\beta}_{(-i)}(t, X_i)) = \underset{(m,b)}{\operatorname{argmin}} \sum_{j=1, j\neq i}^{n} \left(Y_j - m - b'(X_j - x)\right)^2 K_h(X_j - x) \mathbb{I}\{T_j = t\}.$$

[10] In the next subsection, we show that one can actually allow for a wider range of bandwidth values than suggested by the corollary by exploiting the structure of the local linear regression estimator.

Using results in Rothe and Firpo (2018), we then obtain the following result regarding the properties of the corresponding treatment effect estimator.

Corollary 6. *Suppose that Assumption 4 holds with $l = 2$. Then*

$$\begin{aligned}\sqrt{n}(\hat{\tau}_{LOO} - \tau) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\psi_i(\mu) - \tau) + O_P(h^2) + O_P(n^{-1/2} h^{-d/2}) \\ + O_P(\log(n)^{3/2} n^{-1} h^{-3d/2}).\end{aligned} \tag{5.1}$$

Moreover, if $h \propto n^{-\theta}$ with $0 < \theta < 2/(3d)$, then $\sqrt{n}(\hat{\tau} - \tau) \xrightarrow{d} N(0, V_{\text{eff}})$.

This shows that the treatment effect estimator reaches the semiparametric efficiency bound for a much wider range of bandwidths than those found using our general result in the previous subsection. This is because the second and third remainder term in (5.1) are substantially smaller than their counterpart under Theorem 1.[11] The corollary also shows that it is not necessary to impose any smoothness condition that depends on the dimensionality of the covariates in order to construct an estimator of the average treatment effect that reaches the semiparametric efficiency bound.

An inspection of the proof of Corollary 6 shows that the first two second-order terms on right-hand side of (5.1) have mean zero, which is in line with the argument in Section 3.4. The orders of magnitude of these two terms are the same as those of the asymptotic bias and the pointwise asymptotic standard deviation of $\hat{\mu}_{(-i)}$, respectively. The final term on the right-hand side of (5.1) is due to presence of the inverse of the estimated (local) second moments of the covariates in the explicit expression of the local linear regression estimator.[12]

As long as $d < 8$, the treatment effect estimator $\hat{\tau}_{LOO}$ is $\sqrt{n}$-consistent if we choose

[11] If we did not use the "leave-one-out" version of the local linear regression estimator, there would be an additional term of order $O_P(n^{-1/2}h^{-d})$ on the right-hand side of equation (5.1). The use of a "leave-one-out" first-step estimator thus leads to a sizable improvement of the accuracy of the asymptotically linear approximation relative to the estimator studied in the previous subsection.

[12] We conjecture that this rate is actually not sharp. However, in view of Rothe and Firpo (2018), improving the rate would require lots of tedious calculations, and hence we do not investigate this issue any further here.

$h \propto n^{1/(4+d)}$. Such a choice minimizes the order of the integrated mean squared error of $\hat{\mu}_{(-i)}$, and hence a bandwidth satisfying this property can be estimated via cross-validation.

6. Simulations

In this section, we study the finite sample properties of the EIF estimator $\hat{\tau}$ through a Monte Carlo experiment, and compare them to those of other treatment effect estimators. Our aim is to illustrate that the theoretical results obtained above provide a realistic picture of the behavior of the EIF estimator in practice. For simplicity, we focus on the case of nonparametric covariate adjustments via local linear regression. We simulate potential outcomes as $Y(1) = \lambda(X_1, \ldots, X_5) + 2\cos(\pi X_1 X_2) + \varepsilon_1$ and $Y(0) = \lambda(X_5, \ldots, X_1) + \varepsilon_0$. Here $\lambda(a) = \sin(\pi a_1 a_2) + (a_3 + a_4 - 1)^2 + a_5/2$, the covariates $(X_1, \ldots X_5)$ are independent random variables following uniform distributions on $[0, 1]$, and the error terms $(\varepsilon_1, \varepsilon_0)$ are independent of each other and the covariates, normally distributed with mean 0 and standard deviation $1/5$. We also simulate the treatment dummy T as equal to 1 with probability $1/2$ independent of the covariates and the error terms. We thus have that $\tau \approx .589$ and $V_{\text{eff}} \approx 1.205$.

We then compare the performance of the following procedures: (i) a simple difference-in-means estimator, (ii) conventional linear regression adjustments, (iii) "direct" nonparametric covariate adjustments (i.e., the estimator $\hat{\tau}_{adj}$), and (iv) nonparametric covariate adjustments through the efficient influence function (i.e., the EIF estimator $\hat{\tau}$). For the last two estimators, nonparametric covariate adjustments are carried out via leave-one-out local linear regression,[13] and the bandwidth is chosen via cross-validation.

Table 1 presents the results of our simulation study based on 10,000 replications for each of the sample sizes $n = 100, 200, 400, 800$. For each estimator, we report its bias scaled by $\sqrt{n}$, its variance scaled by n (so that it can easily be compared to the efficiency bound

[13] We also considered the properties of estimators based on conventional "leave-in" local linear regression. This did not affect the empirical bias and variance properties of $\hat{\tau}_{adj}$ and $\hat{\tau}$ in a meaningful way, but lead to downward-biased estimates of the corresponding variance in smaller samples, and thus to under-coverage of the corresponding confidence intervals.

Table 1: Simulation results

	$\sqrt{n}\times$Bias				$n\times$Variance			
n	DIM	LR	DNP	EIF	DIM	LR	DNP	EIF
100	.014	.018	.459	.036	2.082	1.769	1.828	1.511
200	.024	.006	.508	.039	2.096	1.715	1.568	1.345
400	.004	.012	.463	.003	2.095	1.664	1.422	1.283
800	.009	.018	.458	.003	2.088	1.582	1.315	1.208
	Avg Variance Estimate				CI Coverage Probability			
n	DIM	LR	DNP	EIF	DIM	LR	DNP	EIF
100	2.066	1.747	1.319	1.319	.934	.936	.890	.928
200	2.061	1.680	1.276	1.276	.943	.943	.901	.938
400	2.055	1.643	1.247	1.247	.945	.948	.922	.943
800	2.049	1.624	1.231	1.231	.954	.951	.932	.953

Results for difference-in-means estimator (DIM), linear regression adjustments (LR), "direct" nonparametric covariate adjustments (DNP), and nonparametric covariate adjustments through the efficient influence function (EIF); based on 10,000 replications.

$V_{\text{eff}} \approx 1.205$), the average value of the corresponding variance estimator,[14] and the coverage probability of the corresponding confidence interval with nominal level 95%.

Both the difference-in-means estimator and linear regression adjustments perform as expected. We know that both estimators are exactly unbiased in our setup, so the non-zero bias figures reported in Table 1 are due to simulation noise. Linear regression leads to a roughly 19% reduction in asymptotic variance relative difference-in-means, but the resulting scaled variance still far exceeds the efficiency bound. The corresponding confidence intervals show minor deviations from nominal coverage for $n = 100$, but are correct up to simulation noise for larger sample sizes.

Direct nonparametric covariate adjustments turn out to be substantially biased, with the scaled bias being roughly constant over the various sample sizes under consideration. In terms of sampling variability, for $n = 100$ the variance actually exceeds that of linear

[14] We use the HC1 heteroscedasticity-robust variance estimator for linear regression adjustments, and the estimator $\hat{V}$ described above for both the "direct" nonparametric covariate adjustments and the EIF estimator. The latter choice is appropriate since both estimator have the same asymptotically linear representation.

regression adjustments. For larger samples we see improvements over linear regression, but even for $n = 800$ the variance still exceeds the efficiency bound by about 17%. Moreover, the corresponding variance estimator tends to be downward biased, which together with the bias issue leads to confidence intervals that undercover the parameter of interest. These problems illustrate why direct nonparametric covariate adjustments are rarely used in the context of randomized experiments.

In contrast, the modified treatment effect estimator based on the efficient influence function performs very well. The reported bias figures are very close to those of the difference-in-means estimator and linear regression adjustments, which we know are exactly unbiased. The variance is well-below that of the other estimators we consider here for all sample sizes, and reaches the efficiency bound for $n = 800$. Moreover, with the exception of a small deviation for $= 100$, the corresponding variance estimator accurately captures the finite-sample variance, and the resulting confidence intervals have excellent coverage. Overall, this confirms that our theoretical results in this paper provide a realistic approximation to the EIF estimators finite-sample properties even in settings with multiple continuously distributed covariates and relatively small sample sizes.

7. Conclusions

This paper shows that the scope for flexible covariate adjustments in randomized experiments is much bigger than generally considered in the empirical literature. By estimating average treatment effects through a sample analogue of the efficient influence function, one can improve upon linear regression adjustments in terms of efficiency without having to sacrifice any of their robustness properties. Moreover, fully efficient estimation is possible by using nonparametric covariate adjustments under conditions that are substantially weaker than those generally required in the literature on semiparametric two-step estimation.

A. Proofs

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 2. Assumption 1 and 2 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

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

as claimed. □

A.2. **Proof of Corollary 1.** The first statement follows from the Central Limit Theorem, since $O_P(a_n^{1-\alpha/2}c_n^{1/2}) + O_p(b_n) = o_P(1)$ under our assumptions. The second statement is obvious. □

A.3. **Proof of Corollary 2.** This result follows from standard arguments. □

A.4. **Proof of Corollary 3.** This result follows since $\hat{\mu}$ as defined in the Corollary satisfies the assumptions made for Theorem 1. □

A.5. **Proof of Corollary 4.** It follows from van der Vaart (1998, Example 19.7) that $N_2(\epsilon, \mathcal{M}_n) \lesssim \epsilon^{-\dim(\Theta)} \lesssim \exp(\epsilon^{-a})$ for all $a > 0$, which implies that Assumption 2 is satisfied under the conditions of the Corollary. From the Glivenko-Cantelli Theorem, it also follows that Assumption 1 is satisfied. However, it also follows from a slight modification of the proof of Theorem 1 that $\lambda_n(m_{\hat{\theta}}) - \lambda_n(m_{\theta^*}) = O_P(\|\hat{\theta} - \theta^*\|)$, as claimed in the comment after the Corollary. Specifically, since $N_2(\epsilon, \mathcal{M}_n)$ is actually of "smaller-than-exponential" order here, it follows that

$$P(|\lambda_n(m^*)| > \epsilon) \lesssim \frac{1}{\epsilon} \int_0^{a_n} \sqrt{\log(N_2(s, \mathcal{M}_n^*))} ds = \frac{a_n - a_n \log(a_n)}{\epsilon},$$

for any fixed $m^* \in \mathcal{M}_n^*$ and any $\epsilon > 0$. □

A.6. **Proof of Corollary 5.** For this proof, we use the notation that for $s = (s_1, \ldots, s_d)$ a vector of non-negative integers let $\partial_x^s m(t,x) = \partial_{x_1}^{s_1} \ldots \partial_{x_d}^{s_d} m(t,x)$ denotes the partial derivatives with respect to x of a generic function m. It then follows from Masry (1996) that we can choose a sequence of functions μ_n, equal to the sum of μ and an asymptotic bias term, such that

$$\|\hat{\mu} - \mu_n\|_\infty = O_P\left(\left(\frac{\log(n)}{nh^d}\right)^{1/2}\right) \quad \text{and} \quad \|\mu_n - \mu\|_\infty = O(h^2).$$

Hence Assumption 1 is satisfied. Moreover, van der Vaart (1998, Example 19.9) shows that $N_2(\epsilon, \mathcal{M}_n) \lesssim \exp(\epsilon^{-d/l} c_n)$. Differentiability of the kernel function then implies that $\hat{\mu}$ is continuously differentiable up to order l; and by arguing as in Masry (1996) and Portier and

Segers (2017) one can also show that for all s with $|s| \equiv \sum_{j=1}^{d} s_d \leq l$ we have

$$\|\partial_x^s \hat{\mu} - \partial_x^s \mu_n\|_\infty = O_P\left(\left(\frac{\log(n)}{nh^{d+2|s|}}\right)^{1/2}\right).$$

It then follows from simple algebra that Assumption 2 is satisfied given the restrictions on the bandwidth.

A.7. **Proof of Corollary 6.** This result follows from the proof of Lemma 3 in Rothe and Firpo (2018). □

References

Andrews, D. (1995): "Nonparametric kernel estimation for semiparametric models," *Econometric Theory*, 11, 560–586.

Bang, H. and J. Robins (2005): "Doubly robust estimation in missing data and causal inference models," *Biometrics*, 61, 962–973.

Belloni, A., V. Chernozhukov, and C. Hansen (2014): "Inference on Treatment Effects after Selection among High-Dimensional Controls," *Review of Economic Studies*, 81, 608–650.

Benkeser, D. and M. Van Der Laan (2016): "The highly adaptive lasso estimator," in *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, IEEE, 689–696.

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

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

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

Farrell, M. H. (2015): "Robust inference on average treatment effects with possibly more covariates than observations," *Journal of Econometrics*, 189, 1–23.

Freedman, D. A. (2008): "On regression adjustments to experimental data," *Advances in Applied Mathematics*, 40, 180–193.

Gail, M. H., S. Wieand, and S. Piantadosi (1984): "Biased estimates of treatment effect in randomized experiments with nonlinear regressions and omitted covariates," *Biometrika*, 71, 431–444.

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

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

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

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

Lin, W. (2013): "Agnostic notes on regression adjustments to experimental data: Reexamining Freedman's critique," *Annals of Applied Statistics*, 7, 295–318.

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

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

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

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

Portier, F. and J. Segers (2017): "On the weak convergence of the empirical conditional copula under a simplifying assumption," *Working Paper*.

Powell, J. L., J. H. Stock, and T. M. Stoker (1989): "Semiparametric estimation of index coefficients," *Econometrica*, 57, 1403–1430.

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

Robins, J. M. and A. Rotnitzky (2001): "Comment on "Inference for semiparametric models: some questions and an answer" by P. Bickel and J. Kwon," *Statistica Sinica*, 11, 920–936.

Robins, J. M., A. Rotnitzky, and L. P. Zhao (1995): "Analysis of semiparametric regression models for repeated outcomes in the presence of missing data," *Journal of the american statistical association*, 90, 106–121.

Rothe, C. (2016): "The Value of Knowing the Propensity Score for Estimating Average Treatment Effects," *IZA Working Paper*.

Rothe, C. and S. Firpo (2018): "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, 1346–1370.

Van Der Laan, M. and A. Bibau (2017): "Uniform Consistency of the Highly Adaptive Lasso Estimator of Infinite Dimensional Parameters," *Working Paper*.

Van der Laan, M. and J. Robins (2003): *Unified methods for censored longitudinal data and causality*, Springer.

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

Wager, S., W. Du, J. Taylor, and R. J. Tibshirani (2016): "High-dimensional regression adjustments in randomized experiments," *Proceedings of the National Academy of Sciences*, 113, 12673–12678.

Wu, E. and J. Gagnon-Bartsch (2017): "The LOOP Estimator: Adjusting for Covariates in Randomized Experiments," *Working Paper*.