Relaxing Assumptions, Improving Inference: Utilizing Machine Learning for Valid Causal Inference*

Marc Ratkovic†

May 31, 2020

Abstract

Despite its ubiquity, the linear regression suffers from several well-known flaws. First, specification choices can affect inference. Second, the regression is a correlative tool, generally not returning an average causal effect. Third, the method assumes no interference among observations. We introduce a method that addresses these shortcomings. First, the method combines a machine learning approach to control for background covariates and learn patterns of interference with a regression to estimate the coefficient on the treatment variable of theoretical interest. Second, the method models the treatment variable as well as the outcome, allowing recovery of a causal effect. The method applies whether the treatment variable is a binary, continuous, or a count variable. We give assumptions under which the method's estimate is consistent for the average causal effect, and we provide diagnostics to assess these conditions. We show that its standard errors are asymptotically valid and semiparametrically efficient. A simulation study shows that the proposed method performs favorably relative to several recent machine learning methods, and application to real-world datasets illustrates the method's utility.

Key Words: average treatment effect, causal inference, partially linear regression, machine learning

*I would like to thank Soichiro Yamauchi and Max Gopelrud for work on developing the software, John Londregan, Scott de Marchi, Brandon Stewart, Kevin Munger, Curtis Signorino, Christopher Lucas, Matt Blackwell, Dean Knox, Neal Beck, Cyrus Samii, Matias Cataneo, Rod Little, Walter Mebane, and Jonathan Katz, for helpful comments; Camille DeJarnett for excellent research assistance; and Stefan Wager for guidance in implementing his software. Presented at the Midwest Political Science Association Annual Meeting, April 7, 2018 and at the Duke University Methods Seminar. Not for citation or distribution without permission from the author.

†Assistant Professor, Department of Politics, Princeton University, Princeton NJ 08544. Phone: 608-658-9665, Email: ratkovic@princeton.edu, URL: http://scholar.princeton.edu/ratkovic

1 Introduction

The standard linear regression is the field's most commonly encountered quantitative tool, used to estimate effect sizes, adjust for background covariates, and conduct inference. At the same time, the method requires a set of assumptions that have been long-acknowledged as problematic (e.g. Lenz and Sahn, 2017; Samii, 2016; Achen, 2002; Leamer, 1983). The fear that our inference will reflect these assumptions, rather than the design of the study and the data, has led our field to explore alternatives including estimation via machine learning (e.g., Hill and Jones., 2014; Grimmer, Messing and Westwood, 2017; Beck, King and Zeng, 2000; Beck and Jackman, 1998) and identification using the analytic tools of causal inference (e.g. Acharya, Blackwell and Sen, 2016; Imai et al., 2011; Sekhon, 2009). We integrate these two literatures tightly, formally, and practically, thereby generating a method and set of diagnostics that can improve the reliability of our inference. In doing so, we contribute to a growing literature combining machine learning and causal inference (Athey, Tibshirani and Wager, 2019; Chernozhukov et al., 2018; Fong, Hazlett and Imai, 2018; Hainmueller and Hazlett, 2013), while improving on these methods in several regards.

Our contributions are twofold. First, we introduce a method that produces a valid causal estimate and confidence interval whether the treatment variable is binary, continuous, or count. We use a machine learning method to learn nonlinearities and interactions in the control variables, and also adjusts for interference that can be explained by the covariates. Producing a valid confidence interval in this setting requires a *semiparametrically e cient* estimate, a property we explain below. Semiparametric efficiency guarantees a consistent and asymptotically normal estimate even when the control variables and patterns of interference

are not specified in advance but instead learned from the data. Second, we implement and illustrate a set of diagnostics that can be used by the applied researcher to assess the method's underlying assumptions. We plan to release accompanying software upon acceptance.

Specifically, the method utilizes machine learning in order to address each of three critiques of the regression. The first is that regression-based inference is model-dependent, meaning our inference may be sensitive to the inclusion or exclusion of control variables (Lenz and Sahn, 2017; Samii, 2016; Achen, 2002; King, Keohane and Verba, 1994; Leamer, 1983). In response, scholars have advocated for machine learning and nonparametric methods as means to avoid linearity and additivity assumptions (Montgomery and Olivella, 2018; Grimmer, Messing and Westwood, 2017; Hill and Jones., 2014; Hainmueller and Hazlett, 2013; Beck, King and Zeng, 2000; Beck and Jackman, 1998). These methods, though, are tuned for prediction and their naive use will lead to invalid inference (Chernozhukov et al., 2018).

A recent body of work has combined machine learning with the familiar regression, using machine learning to control for background variables by estimating a regression coefficient on the key variable of interest (Chernozhukov et al., 2018; Belloni, Chernozhukov and Hansen, 2014*a*,*b*). These methods fall prey to the second, and deeper, critique: the regression is fundamentally a correlative measure and does not estimate an average causal effect (Aronow and Samii, 2017; Angrist and Pischke, 2009). A causal effect is identified by random fluctuations in the treatment variable, and therefore causal estimation requires modeling not just the mean of the treatment variable but also its variance. Failing to model the variance of the treatment (for example, Chernozhukov et al., 2018; Fong, Hazlett and Imai, 2018; Belloni, Chernozhukov and Hansen, 2014*b*) will produce estimates that are not consistent for the

average causal effect of the treatment on the outcome.

Finally, each of these methods assumes that there is no interference among observations (see Athey and Imbens, 2016, for an overview). To the extent that the impact of one observation on another remains unmodeled, our point estimates may be biased and our inference invalid. Taken together, these three critiques carry deep implications for how our field accumulates knowledge, suggesting that the linear regression is not a reliable guide to how we assess our theories and test our hypotheses (Samii, 2016).

The proposed method works in three distinct steps: first, it learns patterns of interference in the data as well as heterogeneity in treatment assignment; second, it generates a set of covariates that adjust for the different biases in the data; and third, it then uses these covariates as controls regressing the outcome on the treatment. A key element of our approach is to conduct each of these steps in different subsets of the data. Our *sample-splitting* strategy involves splitting the data into thirds and conducting each step on a separate split. We then rotate the step done in each step, known as *cross- tting*, repeat this process, and average the results. Our three-split approach returns unbiased estimates with valid confidence intervals, while the repeated cross-fitting helps to restore the efficiency lost through splitting the data.

A substantively important proportion of the field's theory-testing depends on getting *p*-values and confidence intervals correct. Yet, for a variety of reasons, the standard regression and many cutting-edge recent methods fail to do so. We review several of these reasons next, and move on to our proposed alternative. We then illustrate the method in simulated data and in two replication studies. Using data from Mattes and Weeks (2019), we show that the method is as efficient as least squares on experimental data. Then, in the observational setting of Enos (2015), we show how the method can help deal with a continuous treatment

variable.

2 Preliminaries

We consider the situation where the researcher is interested in conducting inference on the average effect of a treatment, t_i, on an outcome y_i for observations $i \in \{1; 2; : : : ; n\}$, with $\mathbf{y}$ and $\mathbf{t}$ the vector of outcome and of treatments. We assume the researcher has p covariates associated with observation i, denoted $\mathbf{x}_i$, with x_{ij} the j^{th} element of this vector. We denote the full $n \times p$ covariate matrix as $\mathbf{X}$, with the intercept in the first column, $\mathbf{X}_j$ the j^{th} column, $\mathbf{X}_{\ i}$ the matrix with the i^{th} row deleted, and $\mathbf{t}_{\ i}$ the treatment vector with the i^{th} observation deleted.

2.1 The Average Partial Effect

The impact of a one-unit movement of the treatment variable, t_i, on the predicted value of the outcome, y_i, for observations with covariate value $\mathbf{x}_i$ is referred to as the *partial e ect*. It is the slope from regressing the outcome on the treatment, but only for observations with covariate value $\mathbf{x}_i$ (e.g., Wooldridge, 2013, pg. 76):

$$(\mathbf{x}_i) = \frac{\mathrm{Cov}(y_i; t_i|\mathbf{x}_i)}{\mathrm{Var}(t_i|\mathbf{x}_i)}: \tag{1}$$

We may wish to allow the partial effect to vary not just with the value $\mathbf{x}_i$ but also by *other* observations' covariates. We then fix not only the observation's covariates $\mathbf{x}_i$ but also

all other observations' covariates, $\mathbf{X}_{\ i}$,

$$(\mathbf{x}_i; \mathbf{X}_{\ i}) = \frac{\mathrm{Cov}(y_i; t_i | \mathbf{x}_i; \mathbf{X}_{\ i})}{\mathrm{Var}(t_i | \mathbf{x}_i; \mathbf{X}_{\ i})} \tag{2}$$

which we denote as $(\mathbf{x}_i; \mathbf{X}_{\ i}) = {}_i$ for parsimony.

The *average partial e ect*, , represents the average impact of the treatment variable on the outcome, after controlling for covariates,

$$= \mathsf{E}\{\ _i\}: \tag{3}$$

This parameter, , is the impact of a one-unit movement of the treatment on the outcome, after adjusting for covariates. As it is the parameter most often desired in empirical studies, we focus on the estimation of and inference on this parameter.

2.2 A Causal View.

Interpreting the average partial effect as a causal parameter requires three assumptions. We characterize these within the potential outcomes framework (Imbens and Rubin, 2015; Holland, 1986), where each observation i is equipped with a potential outcome function $y_i(t_i; \mathbf{t}_{\ i})$ that maps a vector of treatments for all observations to an observed outcome, y_i.

First, we assume there is a single version of a treatment, so the all potential outcomes are well-defined.[1] Second, we require that the treatment have some random component, so other potential outcomes can be observed with positive probability. These two assumptions are standard. Third, we require that we have observed a sufficiently rich set of covariates

[1] See Imbens and Rubin (2015) Ch. 1 for a useful discussion.

to break two forms of confounding: the impact of one observation's treatment on another, and any confounding between a given observation's treatment and outcome. Importantly, we do not need to assume that observations act individualistically; instead, we require that the covariates can adjust for interference. Formally:

Assumption 1 *Causal Assumptions*

1. *Stable treatment value: There is a single version of each treatment value.*
2. *Positivity: The treatment is not deterministic,* $\text{Var}(t_i|\mathbf{x}_i; \mathbf{X}_{\ i}) > 0$ *for all observations* i.
3. *Ignorability:*[2]
   (a) $t_i \perp\!\!\!\perp \mathbf{t}_{\ i}|\mathbf{x}_i; \mathbf{X}_{\ i}$
   (b) $y_i(t_i; \mathbf{t}_{\ i}) \perp\!\!\!\perp t_i|\mathbf{x}_i; \mathbf{X}_{\ i}$

The assumptions clarify the nature of our estimand. We consider the average effect of manipulating each given observation's treatment on its outcome, and assume the covariates block any impact of other observations' treatments on the treatment or outcome. As we assume the covariates adjust for indirect effects, under these assumptions the average partial effect is an average direct effect of the treatment on the outcome.

Enumerating these assumptions also suggests avenues for diagnostics. We provide in Section 4.4 means to check for violations of the positivity and unconfoundedness assumptions.

First, though, we turn to estimating the average partial effect.

[2]Here, the notation $A \perp\!\!\!\perp B|C$ means that event A is conditionally independent of B given C.

2.3 Consistent, Asymptotic Normality

The central goal of this manuscript is to generate an estimate of the average partial effect, $\hat{\delta}$, that is consistent and asymptotically normal,

$$\sqrt{n}(\hat{\delta} - \delta) \xrightarrow{d} \mathcal{N}(0, \Sigma^2). \tag{4}$$

under the weakest available assumptions. Weakening assumptions helps to ensure that our inference reflects real patterns in the data rather than our modeling assumptions.

Consistent, asymptotic normality is desirable as it allows standard inference. Consistency guarantees that as we get more data, our estimate $\hat{\delta}$ will concentrate on the true value δ – which, as we show below, is by no means guaranteed by several existing methods in relatively simple settings. Second, the property allows us to use our familiar tools to conduct valid inference: we can take our point estimate and standard errors and use them to construct valid confidence intervals and p-values in the standard fashion. Our familiar critical values of 1.64 and 1.96 will generate confidence intervals and p-values with the expected properties.

When attempting to estimate and conduct inference on an average partial effect, we rarely observe multiple observations at each covariate profile. Since we cannot recover consistent estimates of the partial effect at each point (δ_i), we have no way of recovering an estimate of the average partial effect (δ) without some set of modeling assumptions. We will move through various sets of assumptions that can be made, from the standard regression to recent cutting-edge methods.

2.4 The Linear Regression

The standard method for estimating an average partial effect is the linear regression model,

$$y_i = \ ^{LS} t_i + \mathbf{x}_i^{>} \ + \ _i; \ \ \mathsf{E}(\ _i | t_i; \mathbf{X}_i) = 0: \tag{5}$$

where the parameter $\ ^{LS}$ is the slope parameter estimated by the regression. We will denote as the *nuisance parameters*, which are parameters not of direct inference but that must be estimated well in order to recover a consistent, asymptotically normal estimate of . In the regression setting, $= \{ \ \}$, the parameters on the control variables.

Under several assumptions (e.g. Assumptions MLR 1-MLR 5 Wooldridge, 2013), the least squares estimate b^{LS} will in fact be consistent, unbiased, and asymptotically normal for the average partial effect, , in that

$$\sqrt{n}(\mathsf{b}^{LS} - \ ) \quad \mathcal{N}(0; \ ^2):$$

Key to achieving consistent, asymptotic normality on b^{LS} is that our estimates of the nuisance parameters, b, themselves converge at the *parametric rate* of $n^{\ 1=2}$, as

$$\sqrt{n}(\mathsf{b}^{LS} - \ ) \quad \mathcal{N}(\mathbf{0}_k; \Omega \ ):$$

with Ω a finite variance matrix.

This parameter $\ ^{LS}$ can also be recovered through the *partialing out* approach to regression (e.g. Wooldridge, 2013, 3.22). Recovering $\ ^{LS}$ can be done through a two-step process:

in the first, the outcome and treatment are regressed on the covariates and, in the second, the residuals from the outcome regression are regressed on the residuals from the treatment regression. Formally, denote as

$$\tilde{y}_i^{LS} = y_i - \mathsf{E}^{LS}(y_i|\mathbf{x}_i); \quad \tilde{t}_i^{LS} = t_i - \mathsf{E}^{LS}(t_i|\mathbf{x}_i);$$

where $\mathsf{E}^{LS}()$ denotes using a linear regression to calculate the expectation. Taking $\sigma_T^2(\mathbf{x}_i) = \text{Var}(t_i|\mathbf{x}_i)$, the regression parameter can be expressed as

$$\begin{aligned} \tau^{LS} &= \frac{\mathsf{E}(\tilde{y}_i \tilde{t}_i)}{\mathsf{E}(\tilde{t}_i^2)} \\ &= \frac{\mathsf{E}\{\text{Cov}(y_i, t_i|\mathbf{x}_i)\}}{\mathsf{E}\{\sigma_T^2(\mathbf{x}_i)\}}. \end{aligned}$$

If we denote as $\hat{y}_i^{LS}$ and $\hat{t}_i^{LS}$ as the least squares estimates of y_i, t_i given $\mathbf{x}_i$, we can estimate τ^{LS} with

$$\hat{\tau}^{LS} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i^{LS})(t_i - \hat{t}_i^{LS})}{\sum_{i=1}^n (t_i - \hat{t}_i^{LS})^2}$$

This basic intuition of subtracting off the effect of the covariates and then regressing the resultant residuals will carry through to more complex settings, with important caveats. To set these up, we first discuss to several critiques of the regression.

3 Critiques of the Regression and Existing Work

The proposed method addresses three critiques of the regression: the *Misspeci cation Critique*, that our regression may not be in-truth linear in the covariates; the *Heterogeneity Critique*, that ignoring the random element of the treatment variable can introduce bias; and the *Interference Critique*, that unmodeled interference can bias the estimate.

3.1 The Misspecification Critique

The misspecification critique takes on the linear specification, suggesting that we instead estimate a more flexible model connecting the treatment and the outcome, as

$$y_i = \quad (t_i; \mathbf{x}_i) + \quad _i:$$

where is a flexible function allowing for interactions and nonlinearities in the data. Any number of machine learning methods have been suggested (e.g., Montgomery and Olivella, 2018; Hill and Jones., 2014; Beck, King and Zeng, 2000; Mohanty and Shaffer, 2018; Hainmueller and Hazlett, 2013).

Using this machine learning (ML) method, the average partial effect by estimating the average partial effect can be estimated as the average slope at each observed datum,

$$\mathsf{b}^{ML} = \frac{1}{n} \sum_{i=1}^{N} \frac{@}{@t} \mathsf{b}(t; \mathbf{x}_i) \Big|_{t=t_i}$$

generating an estimate that does not rely on linearity assumptions.

A machine learning method can reduce model dependence by learning, rather than as-

suming, how the control variables enter the model. Yet, using the same data to learn how the covariates impact the treatment and outcome *and* use this model to adjust for the covariates *and* to conduct inference induces a subtle form of over-fitting.

This over-fitting will generate incorrect inference, meaning misleading p-values and confidence intervals, a fact which has remained largely unacknowledged in our field.[3] In fact, an estimate $\hat{\theta}^{ML}$ will, in general, be biased such that

$$\lim_{n \to \infty} \sqrt{n}|\hat{\theta}^{ML} - \theta| = \infty; \tag{6}$$

which means that the estimate $\hat{\theta}^{ML}$ is not consistent and asymptotically normal. This implies that both analytic and bootstrap confidence intervals will contain the true value with probability zero in the limit, generating invalid inference.

3.2 Split-Sampling, Cross-Fitting, and Semiparametric Efficiency in the Partially Linear Model

We can restore consistent, asymptotical normality–and hence valid inference–by using a *split-sample* approach. Chernozhukov et al. (2018) term this "double machine learning" (DML) , using a machine learning method on half the data to model the confounders and the other half to estimate a treatment effect.[4] The authors work in the partially linear model (PLM, see, e.g. Belloni, Chernozhukov and Hansen, 2014*b*; Härdle et al., 2012; Donald and Newey,

[3] Notably, Mohanty and Shaffer (2018) acknowldege the miscalibration of the method's confidence intervals in Appendix C.2.

[4] See, e.g, Condition H' of Bickel (1982), Theorem 25.57 of van der Vaart (1998), as well as references in Chernozhukov et al. (2018) for earlier work advocating for a split-sample approach. Recent works in political science exploring this approach include Samii, Paler and Daly (2016); Egami et al. (2018); Fong (N.d.).

1994):[5]

$$y_i = {}^{PLM} t_i + f(\mathbf{x}_i) + {}_i; \; \mathsf{E}({}_i|\mathbf{x}_i; t_i) = 0; \tag{7}$$

$$t_i = g(\mathbf{x}_i) + v_i; \; \mathsf{E}(v_i|\mathbf{x}_i) = 0: \tag{8}$$

Here $f; g$ are nonparametric functions of the data, relaxing the linearity and additivity assumptions of the standard linear model. Because the partially linear model contains both nonparametric components ($\{f; g\}$) and a parametric component ( ), it is an example of a *semiparametric* model. The functions ($\{f; g\}$) are *nuisance functions* since they are not of direct interest but must be adjusted for in order to properly estimate .

A semiparametrically efficient estimate b^{PLM} allows for valid inference on ^{PLM}. A semiparametrically efficient estimate is one asymptotically indistinguishable from one where the true nuisance functions $\{f; g\}$ were known in advance. For example, if we knew these functions we could simply evaluate them at the observed data and enter them as controls in our model,

$$y_i = {}^{PLM} t_i + [f(\mathbf{x}_i); g(\mathbf{x}_i)]^{>} + e_i: \tag{9}$$

This model is known as a *parametric submodel* in that it is a parametric model that perfectly adjusts for the nuisance components. The proof strategy, then, involves showing that any gap between the proposed estimator and that from the parametric submodel is asymptotically negligible.

[5]Note that ^{PLM} is not, in general, the average partial effect , for reasons discussed below.

Two assumptions ensure a semiparametrically efficient estimate of θ^{PLM}.[6] First, the estimate must be constructed using a split-sample strategy. Half the data is used to model the effect of the covariates on the outcome and treatment that we generate models of the treatment and outcome given the covariates on half the data. Next, the nuisance functions are evaluted on the other half and included as controls when regressing the outcome on the treatment. Creating a wall between learning the model and conducting inference sidesteps the overfitting bias given above. The second assumption is that the nuisance estimates converge uniformly at a rate of $n^{1/4}$. Importantly, this rate is below the parametric rate of $n^{1/2}$, and can be achieved by a wide range of machine learning methods such as random forests or high-dimensional regressions (Chernozhukov et al., 2018).

Formally, denote as $\mathcal{S}_0, \mathcal{S}_1$ the two splits of the data of size $n_0 + n_1 = n$ and $\hat{E}^{\mathcal{S}_0}$ an estimated conditional mean using a machine learning method on the data in split $\mathcal{S}_0$. The double-machine learning estimate is then a partialed out, split-sample estimate,[7]

$$\hat{\theta}^{DML} = \frac{\sum_{i \in \mathcal{S}_1} \left(y_i - \hat{E}^{\mathcal{S}_0}(y_i|\mathbf{x}_i)\right)\left(t_i - \hat{E}^{\mathcal{S}_0}(t_i|\mathbf{x}_i)\right)}{\sum_{i \in \mathcal{S}_1} \left(t_i - \hat{E}^{\mathcal{S}_0}(t_i|\mathbf{x}_i)\right)^2}$$

If the nuisance estimate achieves the $n^{1/4}$ uniform rate, the DML estimate satisfies

$$\sqrt{n_1}(\hat{\theta}^{DML} - \theta^{PLM}) \quad \mathcal{N}(0, \sigma^2_{PLM})$$

[6] Formally we are describing conditions 25.55-25.56 in van der Vaart (1998) in the context of the PLM, using the approximately least favorable formulation of Sec 25.11, since we use moment conditions and not a full probability model. SPE follows under an assumption of equivariant errors; for SPE under heteroskedasticity in the outcome model, see Chernozhukov et al. (Section 2.2.4 2018). For replacing the split-sample approach with Donsker restrictions on the nuisance functions, see Theorem 25.54.

[7] This is the "DML-2" estimate; replacing the denominator with $\sum_{i \in \mathcal{S}_1} t_i \hat{e}_i$ gives the "DML-1" estimate. The two are first-order equivalent and, in the limiting case of the linear regression, numerically equivalent.

which allows for valid inference.

The most glaring shortcoming of the double machine learning approach is the e ciency loss from splitting the data. We would expect the standard errors to be about $\sqrt{2}$ 1:4 times larger than that of a full sample analysis. Chernozhukov et al. (2018) introduce repeated cross- tting as a means to restore e ciency. Cross- tting swaps the roles of the two splits, and averages the two subsequent estimates. This process is then averaged over repeated splits. Doing so restores e ciency of the estimator, as we illustrate in Section 6.1.

3.3 The Heterogeneity Critique

While promising, double machine learning and related approaches fail to recover consistent estimate of even if we properly specify $f; g$. Intuitively, the average partial e ect for a single observation is the expected impact of random uctuations in the treatment on the outcome, for that observation. The larger the variance in these uctuations, the more informative the observation is for the regression estimate. If heteroskedasticity in the treatment assignment covaries with the treatment e ect across observations, the regression e ect and average partial e ect will diverge. Even if we knew $f; g$ in advance, the estimated coe cient will not in general be unbiased for an average partial e ect (Aronow and Samii, 2016).

The gap between the regression estimate and the average partial e ect is driven by not modeling the randomness in the treatment variable (see, e.g. Angrist and Pischke, 2009; Hirano and Imbens, 2005). Formally, rearranging terms in ^{LS} gives[8]

$$^{LS} = \frac{E\{ \ ^2_T(x_i; X \ _i) \ _i g}{E\{ \ ^2_T(x_i; X \ _i)g} : \tag{10}$$

[8] Substitute $_i = Cov(y_i; t_i | x_i; X \ _i) = \ ^2_T(x_i; X \ _i)$ and the equality follows.

while the average partial e ect is $= E(\ _i)$. If we denote the uctuation of the partial e ect around its average as $(x_i; X\ _i) = E(\ _i \quad jx_i; X\ _i)$, then the gap between the two can be characterized as

$$^{LS} \quad = \frac{E f\ ^2_T(x_i; X\ _i)\ (x_i; X\ _i)g}{E f\ ^2_T(x_i; X\ _i)g}; \tag{11}$$

which we will refer to astreatment variance bias Inspection reveals that either one of two conditions are su cient to guarantee that the treatment variance bias is zero. The rst occurs when there is no treatment e ect heterogeneity $(x_i; X\ _i) = 0$ for all observations), the second when there is no treatment assignment heteroskedasticity $^2_T(x_i; X\ _i)$ is constant across observations). As observational studies rarely justify either assumption (see Samii, 2016, for a more complete discussion), we are left with a gap between the average partial e ect and the parameter estimated by the regression, ^{LS}.

Methods that work in the partially linear model are susceptible to this bias. We adjust for treatment variance bias by explicitly modeling the e ect heterogeneity term and treatment heteroskedasticity term (g_2).

3.4 The Interference Critique

The works we have discussed so far assume that observations are conditionally independent given each observation's covariates. While a useful assumption for both analytic and pedagogical reasons, there may be a reasonable concern that our observations may in fact be inter-related even after conditioning on the unit's observed covariates. For example, observations that are geographically proximal may behave similarly (Ward and O'Loughlin, 2002;

Ripley, 1988), or social actors may react to ideologues on the other end of the political divide (Hall and Thompson, 2018). In each setting, some part of an observation's outcome may be attributable to the behavior of other observations.

Modeling interference has required a careful specification of the nature and types of interference being considered and adjusted for. In seminal work, Manski (1993, 2013) characterizes different types of interference (see also, Bramoulle, Djebbari and Fortin, 2007; Young, 1998) based on how a given observation's treatment and outcome is impact by others. A second body of work comes from public health and statistics, working in a causal inference framework (Hudgens and Halloran, 2008; Sobel, 2006; Aronow and Samii, 2017). These works decompose the total effect of a binary treatment on an outcome into direct and indirect effects, but they apply to a binary or categorical treatment where the interference structure is known.

These approaches, though, require *a priori* knowledge over the type of interference—namely, what variables drive the interference, and how the interference affects both the treatment variable and the outcome. Our contribution here is using a machine learning method to learn the type of interference in the data: what variables are driving interference, and in what manner. In settings where we can determine how "close" observations are, such as by latitude or longitude (e.g. Braumoeller et al., 2018) or degrees apart in a network (e.g. Egami, 2018), we may assume *homophily*, such that nearby observations behave similarly. Unlike these methods, we also allow for *heterophily*, where observations "far apart" behave similarly.

The problem involves two components: a measure of proximity and an interferent. The

rst addresses which variables are driving how close two observations are.[9] In the spatial setting, for example, these may be latitude and longitude. More generally, observations closer in age may be behaving similarly (homophily) or observations with very di erent education levels may be be behaving similarly (heterophily). These measures require both a variable driving proximity as well as a bandwidth parameter, which, characterizes the radius of impact of observations on proximal observations. For example, with a larger bandwidth, interference may be measurable between people with a ten-year age gap, but with a smaller bandwidth, only a three year age gap. The second, the interferent, is the variable that impacts other observations. For example, the treatment level of a given observation may be driven in part by the income levels (the interferent) of other observations with a similar age (the proximity measure). The problem is that the researcher may not know which variables are determining proximity, which variables are moved by the interference, and the bandwidth parameters must be estimated. Our method addresses each of these problems within the estimation procedure.

We next move to the proposed method.

4 The Proposed Method

The model expands the partially linear model to include both interactions and heterogeneity in the treatment assignment mechanism. We refer to it as the partially linear causal e ect (PLCE) since, under an ignorability and positivity assumption, it returns a causal estimate of the treatment on the outcome. Like the partially linear model, it allows the covariates to enter the model parametrically. Unlike the partially linear model, the PLCE model uses the

[9] Manski (2013, 1993) refers to this as the reference group

covariates to adjust for heterogeneity bias and interference.

Our treatment and outcome models are

$$y_i = \tau t_i + f(x_i) + \gamma(x_i; X_i)t_i + \psi_y(x_i; X_i; t_i; \eta_y) + \varepsilon_i \tag{12}$$

$$t_i = g_1(x_i) + g_2(x_i; X_i)\varepsilon_i + \psi_t(x_i; X_i; \eta_t) + v_i \tag{13}$$

The slope in the first term of the outcome model τ, is the parameter of interest. The first nuisance functions in each model $(f(x_i); g_1(x_i))$ are inherited from the partially linear model. The next terms address treatment variance bias: $\gamma(x_i; X_i)t_i$ captures treatment effect heterogeneity while the term $g_2(x_i; X_i)$ captures any systematic variance in the treatment assignment. The last two terms, $\psi_t(x_i; X_i; \eta_t)$; $\psi_y(x_i; X_i; t_i; \eta_y)$, capture any biases attributable to interference. The parameters η_y and η_t are vectors of bandwidths, that governs how the impact of one observation on the other decays in terms of their proximity.

4.1 An Overview of the Theory and Estimation Strategy

Before presenting the formalities of our model, we outline our estimation strategy and the theoretical questions it raises. The details of the approach, both theoretically and in practice, are many; we defer what we can to the Appendices.

Heuristically, we note that the nuisance components fall into two sets. The first, $f; \varepsilon; g_1; \eta_y; \eta_t$, enter the model linearly. If these were the only nuisance functions, the model would resemble that considered by (Chernozhukov et al., 2018). The second set of parameters, $g_2; \eta_y; \eta_t$ do not enter the model linearly, forcing us to adjust the split-sample approach.

We present our strategy in Figure 1. We split the data into three splits, $S_0; S_1; S_2$. In

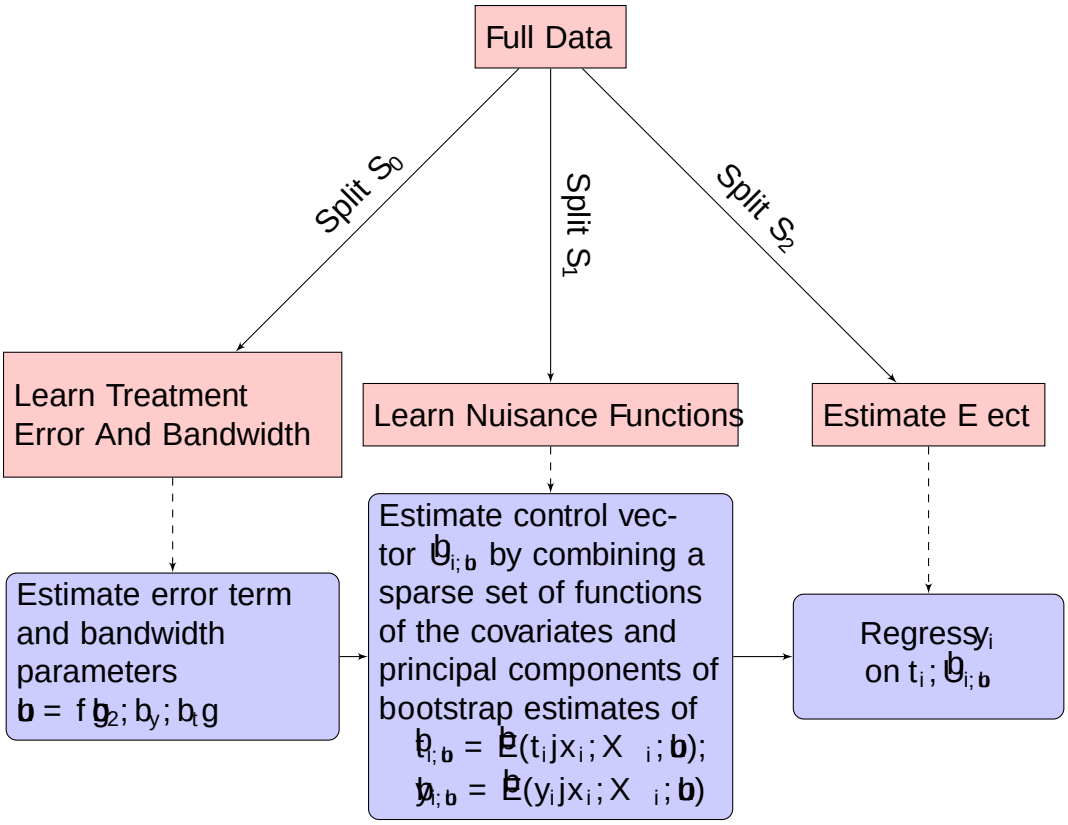


Figure 1: Our estimation strategy.

the rst, we learn the components in the model that enter nonlinearly,

$$u = \{g_2, \ _y, \ _t\}$$

These components are the heteroskedasticity term g_2 and the interference bandwidth parameters for the outcome and treatment, $_y, \ _t$. Then $\hat{b}$ will be our estimates of these components, using data from split S_0.

In the second split, we take these models of from S_0, and evaluate $\hat{b}$ in split S_1. In this split, we then model

$$\hat{y}_i = \hat{E}^{S_1}(y_i|x_i, X_{-i}; \hat{b}); \ \hat{t}_i = \hat{E}^{S_1}(t_i|x_i, X_{-i}; \hat{b})$$

Existing methods enter these point estimates $\hat{\gamma}(x_i, \hat{\theta})$ as controls (Chernozhukov et al., 2018), or use this model to select and include covariates (Belloni, Chernozhukov and Hansen, 2014a). We follow a different strategy by including two sets of covariates. First, as with Belloni, Chernozhukov and Hansen (2014a), we include a set of functions of the variables selected in set S_1 using a sparse regression model. Second, in addition we include principal components of the bootstrapped variance of the fitted values.[10] The estimated vector $\hat{U}_{i;\mathbf{b}}$, constructed on the first two splits, is then evaluated on split S_2. We then use this vector as controls when regressing y_i on t_i. From this regression, we recover an estimate $\hat{\alpha}$ and generate robust standard errors. We then cross-fit the procedure, swapping the roles of the three samples, and then average over repeated cross-fits.

If we only include the selected relevant variables, as in Belloni, Chernozhukov and Hansen (2014a), we must rely on a sparsity assumption (see Chernozhukov et al., 2018, Remark 4.3 for details). By including the principal components, we relax the sparsity assumption to assuming that the functions be well-approximated by a sparse weighted average of the bases. Under this assumption, every basis may in fact contribute to adjusting the nuisance function. We show in Appendix D that doing so does indeed produce performance gains.

The advantages of the proposed method are several. First, unlike several recent methods (Fong, Hazlett and Imai, 2018; Mohanty and Shaffer, 2018; Hainmueller and Hazlett, 2013), our split-sample approach allows for valid inference on the average partial effect. Second, to our knowledge, we are the first to estimate an average partial effect while both allowing for a continuous treatment variable and learning patterns of interference across observations.

[10] We select the number of principal componenents by cross-validating in split S_1, so the process does not touch splits S_0 or S_2

Figure 2: Nonlinear transformations of each variable used to construct basis functions.

This overview of our approach raises several important questions, theoretical and practical. We move onto these next.

4.2 Modeling the Conditional Mean and Interference

In order to model the nuisance functions, we generate a large set of nonlinear and interactive transformations of our covariates, and then use these in a high-dimensional regression. We give the full description of the algorithm in Appendix C, but we summarize our modeling strategy here.

First, the original data are standardized, and we append the principal components of the original data. If we start with p covariates, the principal components generate p additional

covariates that are linear combinations of the original covariates. All variables are then rank-transformed, to reduce the impact of outliers, and rescaled to run from 0 to 1.

We do not assume that any of the nuisance functions are linear in the covariates. Instead, we assume the nuisance functions are linear in a set of nonlinear, interactive transformations of the covariates. Each function of the covariate is referred to as a *basis function*. We will denote the k^{th} of these transformations applied to covariate X_j as

$$X_j \mapsto \phi_k(X_j)$$

The first two basis functions are the intercept and the linear term,

$$\phi_0(X_j) = 1_n; \quad \phi_1(X_j) = X_j$$

For the nonlinear functions, we use a class of functions known as "B-splines," and the exact transformations are given in Figure 4.2.[11] Counting the intercept and linear term with the five nonlinear transformation, we have seven terms generated from each covariate and principal component.

We then include with all two-way interactions between these terms,

$$\phi_{k;k'}(X_j; X_{j'}) = \phi_k(X_j)\ \phi_{k'}(X_{j'})$$

The set $\{\phi_{k;k'}(X_j; X_{j'})\}$ will constitute the basis functions we use in our models.

For p original covariates, this method generates $1 + 4p + (4p - 1)/2$ total basis functions.

[11] We use cubic smoothing splines with knots at each quartile.

For $p \in \{5, 10\}$, this gives 2,415 and 9,730 basis functions, respectively, for each of $\{g_1, g_2\}$. Importantly, these functions have the capacity to capture a wide variety of nonlinear and interactive trends in the data, without needing to be specified by the researcher. We give the details for how we handle all of these bases in our regression in Appendix C.

We also use these basis functions to model interference. The interference terms split into two components, proximity and an interferent. The proximity measure captures how the strength of one observation on another. Specifically, we consider how close observation i' is to observation i, as a function of how close $_k(x_{ij})$ is to $_k(x_{i'j})$, with bandwidth $_{jk}$ as

$$\text{Proximity: } p_{i;i'}(\ _{jk}) = \frac{e^{\frac{1}{jk}(\ _k(x_{ij}) \quad _k(x_{i'j}))^2}}{\sum_{i=1}^{n} e^{\frac{1}{jk}(\ _k(x_{ij}) \quad _k(x_{i'j}))^2}}$$

This measure accounts for homophily, in that two observations are close if they are similar on the observed covariate. Due to the nonlinearities in the basis functions, it also accounts for heterophily, where observations that are dissimilar or at the extremes may have some impact on each other.

We then include an interferent which may be driven by an entirely different function as the proximity measure $_{k'}$, as well as a different variable, $X_{j'}$,

$$\text{Interferent: } \quad _{k'}(x_{i'j'})$$

We combine these two into our interference function, the total effect on observation i

with proximity $p_{i;i'}(\ _{jk})$ and interferent $\ _{k'}(x_{i'j'})$

$$\ _{j;k;j'\ ;k'}(x_i; X\ _i) = \sum_{\substack{i'=1 \\ i' \neq i}}^{n} \underbrace{p_{i;i'}(\ _{jk})}_{\text{Proximity}} \underbrace{\ _{k'}(x_{i'j'})}_{\text{Interferent}}$$

Note that the summation is taken over all observations except i, so we are capturing the effect of all observations but i on observation i. These basis functions are used in a high-dimensional regression for estimation; details can be found in Appendix C.

4.3 Assumptions

We make three sets of assumptions on the PLCE. The first follows the standard least squares assumptions (see, e.g. Wooldridge, 2013, Assumptions MLR 1-5 in ch. 3):

Assumption 2 (PLCE Assumptions)

1. Representative Sample. We observe a representative sample of size n, $\{y_i; t_i; x_i\}_{i=1}^{n}$.

2. Linear in Functions of Covariates. The population model is given by Equations (12)-(13).

   (a) Structure of Functions. The functions $f; g_1; g_2;$ are linear in the basis functions $\{\ _{kk'}(X_j; X_{j'})\}$, the interference function $\ _t$ is linear in the interference terms $\ _{j;k;j'\ ;k'}(x_i; X\ _i)$ and $\ _y$ is linear in $\ _{j;k;j'\ ;k'}(x_i; X\ _i; t\ _i)$

   (b) Constraints on Functions. Each function $\{f; g_1; g_2; \ ; \ _y; \ _t\}$ has finite variance.

3. Positivity. The treatment variable has positive variance $Var(t_i | x_i; X\ _i) > 0$ for every observation.

4. Zero Conditional Mean. The errors are uncorrelated with all basis and interference functions.

   (a) No Omitted Confounders in Outcome Error. $E(\varepsilon_i | t_i, x_i, X_{\varepsilon i}, t_{\varepsilon i}) = 0$

   (b) No Omitted Confounders in Treatment Error. $E(v_i | x_i, X_{\varepsilon i}) = E(v_i \varepsilon_i | x_i, X_{\varepsilon i}) = E(v_i \varepsilon_i | x_i, X_{\varepsilon i}) = 0$

   (c) Ignorable Interference. $E(\varepsilon_i \varepsilon_{i'} | t, x_i, X_{\varepsilon i}, t) = E(v_i v_{i'} | x_i, X_{\varepsilon i}) = E(v_i \varepsilon_i v_{i'} \varepsilon_{i'} | x_i, X_{\varepsilon i}) = 0$ for observations i, i'.

5. Asymptotic Normality of Errors. All fourth moments and cross-moments of the error terms are finite.

4.3.1 Scope Conditions and Discussion of Assumptions

We hew closely to the simple regression for our first set of assumptions, as many of our intuitions carry through. Assumptions 1, 3 and 5 are either identical to or straightforward extensions of the standard least squares assumptions. Assumption 1 requires that our sample be representative of the population of interest. Assumption 5 allows a central limit theorem, and can be assessed using quantile plots of the errors relative to a normal, as with any regression. We provide a means for assessing Assumption 3 in Section 4.4.

The first part of Assumption 2 is an extension of the standard assumption, that the nuisance functions are linear in known functions of the covariates. While our method subsumes the linear model, we also allow for a large number of possible nonlinear and interactive terms. Requiring a finite variance for these terms is sufficient to ensure that the functions are not too erratic as to preclude inference on.[12]

[12] For a formal discussion, van der Vaart (1998, ch. 19.4).

Assumption 4 simply structures the error terms. The conditional independence assumption is standard in the semiparametric literature (see, e.g. Chernozhukov et al., 2018; Donald and Newey, 1994; Robinson, 1988). An advantage over the regression is that the researcher can include a large number of potential covariates and the proposed method will learn any nonlinearities or interactions that must be adjusted for. Crucially, these covariates must be pre-treatment: the inclusion of post-treatment variables will induce bias in the e ect estimate (see, e.g. Acharya, Blackwell and Sen, 2016).

Assumption 4 a and 4 b assume no omitted variables in the outcome or treatment model. Assumption 4 c assumes that there is no unmodeled interference, after conditioning on the observed covariates. The assumption of no omitted variables should be made only hesitantly outside of experimental settings, so we implement the sensitivity analysis of Cinelli and Hazlett (2020) as explained in order to characterize how sensitive any results are to an omitted variable.

4.3.2 Representation and Estimation Accuracy

Our second round of assumptions come in two parts. First, we require that the PLCE model admits a parametric representation, such that there are a set of control vectors that could adjust for all the biases in the model. Second, we characterize the accuracy with which we must estimate the nuisance functions in order to recover a semiparametrically e cient estimate of the average partial e ect.

Central to our approach is a function,U, which will adjust for all of nuisance functions that may bias our estimate of . This function takes as its arguments the data and the

nuisance components u and returns a vector of length k:

$$U_{i;u} = U(x_i; X_{\ i}; t_{\ i}; u): \tag{14}$$

We assume that we can represent these nuisance functions with this vector and adjust for all of the sources of confounding:

Assumption 3 Representation The PLCE model can be represented as the model

$$y_i = t_i + U_{i;u}^{>} + e_i \tag{15}$$

such that the least squares estimate from this regression is consistent and asymptotically normal for the average partial e ect, .

This model is infeasible, in that we do not observe $U_{i;u}$ but instead must estimate $\hat{U}_{i;b}$. Before describing how we construct this estimate, we rst characterize how accurate our estimate must be:

Assumption 4 Approximation All nuisance components can be estimated at an $n^{1=4}$ uniform rate.

Our theoretical work has gone into, e ectively, condensing all of the nuisance functions in the PLCE to a single control vector, $\hat{U}_{i;b}$. This vector will adjust for nonlinearities and interactions, as well as any uncovered interference. This then gives:

Proposition 1 The Feasible Estimatender Assumptions 2, 3 and 4, and using the split-

sample strategy described in Figure 1, the model

$$y_i = \tau_i + \mathbf{U}_{i;\mathbf{b}}^{>} \boldsymbol{\theta}_{\mathbf{b}} + e_i$$

will generate a semiparametrically e cient estimate of . Proof: See Appendix B.

Our primary theoretical contribution comes in developing a general method for generating semiparametrically e cient causal e ect methods that do not rely on a sparsity assumption, while adjusting for interference and heterogeneity bias.

4.3.3 Scope Conditions and Discussion of Assumptions

Our assumptions are general, and compare favorably to many in the literature, though they do contain caveats. One set of works require that the conditional mean be represented by a nite number of basis functions, such as interactions and square terms (see, e.g., Belloni, Chernozhukov and Hansen, 2014). [13] We do not rely on a sparsity assumption, as the principal components we recover may be an average of a large number of covariates. We require our constructed control vector $\mathbf{U}$ to be nite-dimensional but note that, with added assumptions, we could allow the dimensionality in $\mathbf{U}$ to grow, though we save this for further work (see, e.g. Cattaneo, Jansson and Newey, 2018).

The idea that we can express the nonparametric nuisance functions as a parametric control vector dates back to Stein (1956). He introduced the idea of a parametric approximation to the nonparametric function constructed from the variance of the estimates, characterizing the semiparametric e ciency bound as the minimal variance achieved by a parametric

[13] The authors rely on an \approximate sparsity" assumption where the model is sparse up to an error tending to zero in sample size.

submodel. Our approach is, e ectively, generating a feasible estimate of Stein's parametric submodel.

Our use of principal components is a form of \su cient dimension reduction" (Li, 2018; Hsing and Ren, 2009), where we assume that the covariates can be reduced to a set that fully captures any systematic variance in the outcome. Our primary theoretical contribution comes from sidestepping the analytic issues in characterizing the covariance function of the observations analytically (see, e.g. Wahba, 1990) by instead taking principal components of the bootstrap sample. Our sample-splitting strategy is also original. As well, we are the only method to learn and adjust for covariate-driven interference.

Our representation assumption, the $U_{i;u}$ is nite dimensional, is the major constraint on the complexity of the models we can estimate. This sort of assumption is standard, for example Belloni, Chernozhukov and Hansen (2014) make a sparsity assumption and Savje, Aronow and Hudgens (2019) assume that interference patterns do not grow too quickly in sample size. This assumption rules out several important classes of models. First, the method is designed for cross-sectional data rather than panel models. Second, we can only account for interference that can be adjusted for in $U_{i;u}$, so if the variance attributable to interference grows in sample size, our model would no longer prove semiparametrically e cient.

In not requiring distributional assumptions on the treatment variable, we can push past a causal inference literature that is most developed with a binary treatment. Many of the problems we address have been resolved in the binary treatment setting (Robins, Rotnitzky and Zhao, 1994; Glynn and Quinn, 2010; van der Laan and Rose, 2011) or where the treatment density is assumed (Fong, Hazlett and Imai, 2018), so estimating inverse density weights is easier. Nonparametric estimates of inverse density weights are inherently unstable, so we

do not pursue this approach but see Kennedy (in press). Rather, we mean-adjust for confounding by constructing a set of control variates. We show below, through simulation and empirical examples, that the method generates reliable estimates.

4.4 Diagnostics.

Assumptions suggest diagnostics. We implement a sensitivity analysis, used to assess how strong an unobserved confounder must be in order to overturn our results. Since our method is, in effect, a regression in split S_2, diagnostics for the regression are applicable. We utilize the recent method of Cinelli and Hazlett (2020), reporting the statistics suggested those authors. The statistics are calculated on split S_2, and averaged over cross fitting. Along with point estimates and standard errors, we follow the authors' suggestion and report three statistics. The first two, robustness values RV and $RV_{0.05}$, range from 0 to 1 and characterize how strong an unobserved confounder must be in order to reduce the observed effect to 0 (RV) or to make it no longer significant at the 95% level ($RV_{0.05}$). Larger numbers indicate a more robust result. The second, the extreme value statistic $R^2_{Y \sim D|X}$, assumes a "worst-case" confounder that perfectly explains the residuals in the regression, and characterizes how much of the variance in the treatment this confounder must explain in order to eliminate the estimated effect, again ranging from 0 to 1 with larger values preferred. We illustrate the use of these statistics in Section 6.

We also assess the positivity assumption. Positivity is violated when the treatment variable is a deterministic function of the covariates. We assess this by measuring how much variability in each residual using a measure known as kurtosis (see, e.g. Wooldridge, 2013,

Figure 3: Excess kurtosis plot for diagnosing positivity.

Appendix B, p. 737). Given a mean-zero random variable X, the kurtosis is defined as

$$_X = E(X^4)=E(X^2)^2$$

Since the kurtosis is constructed from a fourth moment, and can be written as $E(Z^2); Z = X^2$, the kurtosis captures the variance of the variance, and can guide in assessing the viability of the positivity assumption. A clustering of the treatment error near zero, above what we would expect from a normal density, suggests positivity concerns.

To assess this, we construct for each observation a statistic b_i. Denoting as $b_{i;s;t}$ the residual from estimating the treatment for observation i on split s, we estimate the kurtosis as

$$b_i = \frac{\frac{1}{S}\sum_{i_s=1}^{S} b_{i;t}^4}{(\frac{1}{S}\sum_{i_s=1}^{S} b_{i;s;t}^2)^2}: \qquad (16)$$

We then plot sort the values and plot the excess kurtosis, which is how much above or below they are above an in-truth normal variable.

The use of the diagnostic plot is illustrated in Figure 4.4. The lefthand side contains ve di erent error densities. The rst one is thin-tailed and raises the deepest concerns about violating positivity, since the residuals are tightly clustered near zero. The next is normal, then a fat-tailed and skewed density follow. The last density may also be of interest, as it is a combination of the thin-tailed density, where some observations may violate positivity, and a normal density, where some do not.

The righthand side presents the diagnostic plot. The normal density falls on the 0 line. The thin-tailed distribution falls everywhere above 0 and ares up to the right. The fat-tailed distribution falls below 0, aring down. The skewed distribution agrees with the normal close to zero, but then ares up above 0 as the thin-tailed distribution. The mixture of the normal and thin-tailed creates aU-shape, going down below 0 then up again.

A positive excess kurtosis statistic, everywhere above zero, suggests that the researcher should examine the data for violations of positivity. This method is diagnostic and, it must be emphasized, needs to be combined with substantive knowledge. If a violation is found, the researcher should discern for what observations the residuals are pooling near zero, and consider trimming them from the analysis. This will change the estimand from the average e ect to a local average e ect on the trimmed sample.

We next move onto a simulation study.

5 An Illustrative Simulation.

The simulations assess performance across three dimensions: treatment effect heterogeneity, treatment variance bias, and interference. Considering the impact of the presence or absence of each across these three dimensions results in eight different simulation settings.

In each setting, we draw a standard normal covariate x_{i1}, error terms u_i and ε_i, each standard normal, with the covariate standardized so that $\frac{1}{n}\sum_{i=1}^{n} x_i = 0$ and $\frac{1}{n}\sum_{i=1}^{n} x_i^2 = 1$. Four additional normal noise covariates are included, with pairwise correlations among all covariates 0.5, but only the first is used to generate the treatment and the outcome.

In assessing treatment effect heterogeneity, we vary whether the treatment interacts with the covariate x_i. The two models without interference are given below:

	The Additive Model	The Interactive Model
Outcome Model:	$y_i = y_i + x_{i1}^2 + \varepsilon_i$	$y_i = t_i \; x_{i1}^2 + \varepsilon_i$
Treatment Model:	$t_i = x_{i1} + v_i$	$t_i = x_{i1} + v_i$

In the first model, x_{i1}^2 enters additively, so the average partial effect is $= 1$ effect is homogeneous. In the second, it enters interactively, inducing a simple, but nonlinear, heterogeneity. In every setting, the sample average causal effect is 1, due to the homogenous effect in the additive model and the scaling of x_i in the interactive model.

We then allow for two models of the allow for two speci cations of treatment variance,

$$\text{No Treatment Heteroskedasticity: } \varepsilon_i \sim N(0,1) \quad v_i \sim N(0,1)$$

$$\text{Treatment Heteroskedasticity: } \varepsilon_i \sim N\left(0, \frac{x_{i1}^2+1}{2}\right) \quad u_i \sim N\left(0, \frac{x_{i1}^2+1}{2}\right)$$

Because we have scaled x_{i1}^2 to have mean one, the errors have the same unconditional variance, $E(v_i^2) = E(\varepsilon_i^2) = 1$ across both settings.

We then transform the covariates, as

$$x_i = x_{i1}, \frac{1}{2}x_{i2}, x_{i2}, \frac{1}{2}x_{i1}, x_{i3}, x_{i4}, x_{i5}$$

and give each method the data $\{y_i, t_i, x_i\}_{i=1}^n$. We vary the sample size $n \in \{250, 500, 750, 1000\}$, but report results for $n = 1000$ in the body and the remainder in Appendix D.[14]

Along with the proposed method (PLCE), we have discussed four different machine learning methods that purport to offer causal estimates in a general regression setting: Kernel Regularized Least Squares ((KRLS) Mohanty and Shaffer, 2018; Hainmueller and Hazlett, 2013), CBPS for continuous treatments ((CBPS) Fong, Hazlett and Imai, 2018),[15] Double Machine Learning (DML) of Chernozhukov et al. (2018), and the generalized random forest (GRF) of Athey, Tibshirani and Wager (2019).

We have implemented a rather simple set of simulations. The transformations are all

[14] At smaller sample sizes, we find the method performs similarly in terms of point estimation, and $n = 250$ the confidence intervals are valid but a bit conservative, while for $n = 500$ and above, the results looks similar to the results in the body.

[15] In this simulation, we use parametric CBPS, so that we can recover standard error estimates. So as not to handicap the method, we give it both the covariates and their square terms, so the true generative model is being balanced.

straightforward: an interaction, a square term, and a linear transformation of the covariates. As well, the true model is in-truth parametric, with only a single covariate driving the treatment and outcome. To the extent that we observe systematic errors in point estimation or coverage of con dence intervals, it is because a method is not looking for what is in fact a rather simple nonlinearity or interaction.

5.1 Results for the Setting Without Interference

Results for the simulations without interference are in Figure 5. The rst column shows the distribution of point estimates, where the true value is 1, in gray. We are equally concerned with whether the method allows for valid inference. The \coverage rate" is the proportion of simulations for which the constructed con dence interval contains the true value of 1 (see, e.g. Wooldridge, 2013, Sec. 4.3). The second column shows the coverage rates: expected coverage is on the x-axis and actual coverage is on the y-axis. For example, consider in the top right plot the point marked \ " at $(0:90; 0:83)$, which is on the CBPS curve. Here, we construct a 90% con dence interval of the form $[\hat{\beta} - 1:64\hat{\sigma}_{\hat{\beta}}; \hat{\beta} + 1:64\hat{\sigma}_{\hat{\beta}}]$ and measure the proportion of simulations where the con dence interval contains the true value. In this case, for CBPS, this value is 0.83, so the 90% con dence interval is invalid, albeit only slightly too narrow. More generally, if a curve falls below the 45 line, the con dence intervals are too narrow and hence invalid. If the curve falls above the 45 line, the con dence intervals are valid but wide.

Simulation settings increase in complexity going down the rows. The rst row contains contains the additive model with no heteroskedasticity; the second, the interactive model with no heteroskedasticity; the third, the additive model with heteroskedasticity; and the

Figure 4: Results for simulations without interference. The first column shows the distribution of point estimates, where the true value is 1, in gray. The second column shows the coverage rates: expected coverage is on the x-axis and actual coverage is on the y-axis. If a curve falls below the 45, the confidence intervals are too narrow and hence invalid. If the curve falls above the 45 line, the confidence intervals are valid but wide. The proposed method, PLCE, is the only one to perform well across all settings.

fourth, the interactive model with heteroskedasticity.

Starting in the simplest setting in the first row, if the data generating process is in-truth linear, additive, and homoskedastic with a homogeneous treatment effect, every method per-

forms well. Moving to the interactive model, in row 2, GRF, DML, and CBPS all show discernible bias. Only least squares and the proposed method have valid coverage. In the third row, with an additive effect but treatment assignment heteroskedasticity, again only the proposed method and least squares provide unbiased estimates with valid confidence intervals. In the final row, with both treatment effect and treatment assignment heteroskedasticity, only the proposed method and KRLS provide unbiased estimates and valid intervals.

Several machine learning methods fail to provide unbiased estimation and inference in the presence of a simple interaction or heteroskedasticity. Across all settings, the proposed method is the only one that that allows for valid inference.

5.2 Results for the Setting With Interference

We next consider the same setup as above, except we add a term for interference. In constructing the interference, we take our proximity measure as

$$\text{Proximity: } p_{i;i'0} = \frac{e^{\frac{1}{2}(x_{i1} - x_{i'0_1})^2}}{\sum_{i=1}^{n} e^{\frac{1}{2}(x_{i1} - x_{i'0_1})^2}}:$$

The bandwidth was selected so the kernel would follow a standard normal.

The interferent in the outcome model is the treatment, while in the propensity model it is the term x_{i1}^2. Using to denote the previous terms from the earlier simulation, we generate

Figure 5: Results for simulations without interference. The first column shows the distribution of point estimates, where the true value is 1, in gray. The second column shows the coverage rates: expected coverage is on the x-axis and actual coverage is on the y-axis. If a curve falls below the 45, the confidence intervals are too narrow and hence invalid. If the curve falls above the 45 line, the confidence intervals are valid but wide. The proposed method, PLCE, is the only one to perform well across all settings.

the outcome and treatment as

$$y_i = \sum_{i' \neq i}^{n} p_{i;i'} t_{i'}$$

$$t_i = \sum_{i' \neq i}^{n} p_{i;i'} x_{i'1}^2$$

The error structure stays the same as the previous setting.

The results from this setting are presented in Figure 5.1. Going down the rows, all methods save least squares return accurate point estimates in the simplest setting, but only the proposed method provides valid con dence intervals. We see a similar pattern in the next three rows: point estimates are reasonable, especially for the proposed method, KRLS, and to some extent GRF. KRLS returns valid intervals in the fourth setting, but not the rst three. CBPS, OLS, and GRF provide reasonable con dence intervals in the second setting, but not the others. As above, only the proposed method provides both reliable point estimates and valid con dence intervals across each of the settings.

We next move on to empirical applications.

6 Empirical Applications

The proposed method allows for several advances over both the standard regression and cutting-edge machine learning methods. We illustrate using two datasets. In the rst, we address concerns that the split-sample approach embedded in repeated cross- tting may in ate the variance of our estimated e ects. We consider a survey experiment, where we know least squares provides an unbiased estimate of the average treatment e ect, and show that the proposed method returns practically identical point estimates and standard errors. In the second application, we reanalyze data from a study where a continuous treatment variable was arti cially dichotomized in order to estimate a causal e ect. The proposed method handles continuous treatment variables, allowing us to maintain the treatment variable on its original scale.

6.1 Maintaining E ciency

Mattes and Weeks (2019) conduct a survey experiment in the United States, asking respondents about a hypothetical foreign a airs crisis involving China and military presence in the Arctic. Varied is whether the hypothetical President is a hawk or dove, whether the policy is conciliatory or maintains status quo military levels, the party of the President, and whether the policy is e ective in reducing Chinese military presence in the Arctic. The outcome is whether the respondent disapproves of the President's behavior; controls consist of measures of the respondent's hawkishness, views on internationalism, trust in other nations, previous vote, age, gender, education, party ID, ideology, interest in news, and importance of religion in their life.

We focus on how the estimated causal e ect of conciliation varies between hawks and doves, as reported in Table 2 of the original work. Results appear in Figure 6.2. We begin by diagnosing positivity. As expected, since the treatment is in-truth randomized and hence there are no actual positivity concerns, the excess kurtosis lines stays below zero and trend downwards.

For the two estimated e ects, we nd that PLCE returns results quite similar to least squares in an experimental setting. Importantly, the standard errors are comparable between the two methods, suggesting that the proposed method does not result in a loss of e ciency relative to standard methods. Also, for all three robustness measures ($R^2_Y(\ _{D\,jX}$; RV; RV $_{=0:05}$), we nd the proposed method to be comparable to least squares with covariates and, by these measures, slightly more robust.

	PLCE	Di -in-Mean	OLS w Covariates
Hawks			
Estimate	11.55	11.98	11.97
s.e.	3.51	3.80	3.80
$R^2_{Y\ DjX}$ (%)	2.21	1.65	1.68
RV (%)	12.68	12.13	12.27
$RV_{=0:05}$(%)	5.74	4.79	4.87
Doves			
Estimate	35.09	35.43	35.19
s.e.	2.77	3.12	2.85
$R^2_{Y\ DjX}$ (%)	19.88	16.14	18.69
RV (%)	38.65	35.30	37.81
$RV_{=0:05}$(%)	33.46	29.99	32.67

Figure 6: Comparing PLCE and Least Squares Regression in an Experimental Setting. The excess kurtosis curve (top) falls entirely below zero, which is to be expected when a treatment is known to be randomized. The table at the bottom presents estimated e ects and the sensitivity analysis results. Across all statistics, the PLCE performs comparably to least squares on this data. Point estimates, standard errors, and the sensitivity analysis statistics are all similar.

6.2 Estimating a Causal E ect in the Presence of a Continuous Treatment

We next reanalyze data from a recent study that estimated the causal e ect of racial threat on voter turnout (Enos, 2015). The author operationalizes racial threat by distance to a public

housing project, a continuous measure, and measures its impact on voting behavior. The demolishment of a subset of the projects in the early 2000s in Chicago provides a natural experiment used for identifying the causal e ect. The author implements a di erence-in-di erence analysis (e.g Angrist and Pischke, 2009) which, unfortunately, requires a binary treatment. To accommodate the method, the author arti cially dichotomizes the continuous treatment variable, considering all observations closer than some threshold distance to the projects as exposed to racial threat and observations further away as not.

The threshold is not actually known, or even estimable, given the data. There is no reason to suspect that racial threat only extends, say, 1.3 kilometers, and drops o precipitously after.[16] The proposed method allows estimation of the average causal e ect of distance on the outcome.

We conduct four separate analyses. For the rst, we estimate the causal e ect of distance on change in turnout for white residents within one kilometer of a demolished housing project. The treatment variable is Euclidean distance to the housing project, and the control variables consist of turnout in the previous two elections (1998, 1996), age, squared age, gender, median income for the Census block, value of dwelling place, and whether the deed for the residence is in the name of the voter.[17] We next generate three matched samples for further analysis.[18] The rst are black voters within one kilometer of a demolished housing project. As argued in the original piece (p. 11), this group will not face racial threat, and so it provides a

[16] We emphasize that this problem was not an issue with the original author's design or insights, but instead of the methodological standing of the day.

[17] See the supplemental materials of Enos (2015) for more details.

[18] We estimate distance as a function of all covariates for white residents within one kilometer of a demolished project using a random forest. We then use this model to predict the treatment level, using black residents within one kilometer and then white and black residents greater than one kilometer away. Nearest neighbor matching is implemented to construct the three additional datasets.

	Estimate	s.e.	$R^2_{Y\sim D\,|\,X}$ (%)	RV (%)	$RV_{=0:05}$ (%)
Whites < 1km from Demolished Project	0.10	0.040	0.16	3.37	1.32
Blacks < 1km from Demolished Project	-0.11	0.046	0.21	4.03	1.79

Figure 7: Positivity Diagnostic, Causal E ect Estimate, and Sensitivity Analysis.
In the lefthand plot, the excess kurtosis curve falls below zero, indicating little concerns over positivity. The righthand plot gives the estimated causal e ect of distance from a demolished public housing unit, in kilometers, on turnout for di erent subsets of the data. Whites closer to the projects are more likely to vote, while blacks closer to the projects are less likely to vote. For blacks and whites further than 1km from any projects, the e ects are substantively and statistically insigni cant.

measure of the secular trend in turnout absent racial threat. The next two samples consist of white and black voters, but both further than one kilometer from any housing project, either demolished or not. The latter two groups serve as placebo groups, since they are su ciently far from a demolished project that any threat should be muted.

Figure 7 presents the diagnostic results and e ect estimates. In the lefthand plot, the excess kurtosis curve falls below zero, indicating little concerns over positivity. The righthand plot gives the estimated causal e ect of distance from a demolished public housing unit, in kilometers, on turnout for di erent subsets of the data. In the top two rows, we consider

residents within 1 km of a demolished unit; in the bottom two rows, we consider residents greater than 1 km from any public housing unit, demolished or not.

We estimate that living adjacent to a public housing unit, rather than 1 km away, causes a decrease in turnout of about 9.7 percentage points, an effect in line with the results from the original analysis (see Figure 1 there). The remaining results suggest that race may be a factor. The estimated effect -0.11 (0.046) for blacks near the projects suggests a form of racial threat for blacks: as they move closer to the projects, they are less likely to vote. The bottom two lines consider distal blacks and whites, providing a placebo test. In both cases, we estimate no substantively or statistically significant effects of distance on turnout. The top two results, though, appears to be sensitive to omitted confounders. The sensitivity analysis statistics are an order of magnitude smaller than in the experimental setting above, suggesting that these results are sensitive to the no omitted confounders assumption. Any omitted confounder that explains more than 1.32% or 1.79% of the variance in both the treatment and the result can leave the result for whites and blacks not significant, respectively. If a confounder explains 3.37% or 4.03% of the variance, it can zero out the point estimate for whites and blacks, respectively. An omitted "worst-case" predictor of y that perfectly explains the regression's residuals need only explain 0.16% and 0.21% of the treatment to eliminate this effect.

7 Conclusion

The naive use of the regression, particularly on observational data, has been under attack from several different angles, and for quite a while (e.g. Leamer, 1983). We both suspect, and in some cases have proven, that our standard regression-based approach is not the exact right way to conduct inference, but the familiar tools of the regression coefficient, standard

error, and p-value return such sensible results that we move ahead regardless. Our goal has been to improve inference for political scientists.

The proposed method utilizes recent advances in political methodology, econometrics, and statistics to correct several of these issues. Importantly, we are not trying to replace the regression with a fancy machine-learning method, but to keep the parts we like (coe cients, standard errors, p-values) while using a machine learning method to improve the parts we do not like (control speci cations for the treatment and outcome). The method learns how the covariates impact the treatment variable and outcome, before attempting inference on the treatment variable. Necessary for the approach is a split-sample strategy, using di erent subsets of the data for modeling randomness in the treatment assignment, how the covariates impact the treatment and outcome, and conducting inference, respectively. We provide diagnostics for several assumptions, and plan to extend this into the panel, instrumental variables, and mediation settings next.

References

Acharya, Avidit, Matthew Blackwell and Maya Sen. 2016. "Explaining Causal Findings Without Bias: Detecting and Assessing Direct Effects." American Political Science Review 110(3).

Achen, Christopher. 2002. "Toward a new political methodology: Microfoundations and ART." Annual Review of Political Science 5:423–450.

Angrist, Joshua D. and Jörn-Steffen Pischke. 2009. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton: Princeton University Press.

Aronow, Peter and Cyrus Samii. 2016. "Does Regression Produce Representative Estimates of Causal Effects?" American Journal of Political Science 60(1):250–267.

Aronow, Peter and Cyrus Samii. 2017. "Estimating Average Causal Effects Under General Interference." Annals of Applied Statistics p. Forthcoming.

Athey, Susan and Guido W Imbens. 2016. "The Econometrics of Randomized Experiments.".

Athey, Susan, Julie Tibshirani and Stefan Wager. 2019. "Generalized Random Forests." Annals of Statistics. Forthcoming.

Beck, Nathaniel, Gary King and Langche Zeng. 2000. "Improving Quantitative Studies of International Conflict: A Conjecture." American Political Science Review 94(1):21–35.

Beck, Nathaniel and Simon Jackman. 1998. "Beyond linearity by default: Generalized additive models." American Journal of Political Science pp. 596–627.

Belloni, Alexandre, Victor Chernozhukov and Christian Hansen. 2014a. "High-Dimensional Methods and Inference on Structural and Treatment Effects." Journal of Economic Perspectives 28(2):29–50.

Belloni, Alexandre, Victor Chernozhukov and Christian Hansen. 2014b. "Inference on Treatment Effects after Selection among High-Dimensional Controls." Review of Economic Studies 81(2):608–650.

Bickel, P. J. 1982. "On Adaptive Estimation." Annals of Statisitcs 10(3):647–671.

Bramoulle, Yann, Habiba Djebbari and Bernard Fortin. 2007. "Identification of Peer Effects through Social Networks." Journal of Econometrics 150(1):41–555.

Braumoeller, Bear F, Giampiero Marra, Rosalba Radice, and Aisha E. Bradshaw. 2018. "Flexible Causal Inference for Political Science." Political Anaysis 26(1):54–71.

Cattaneo, Matias D., Michael Jansson and Whitney K. Newey. 2018. "Alternative Asymptotics and the Partially Linear Model with Many Regressors." Econometric Theory 34(2):277–301.

Chernozhukov, Victor, Denis Chetverikov, Esther Demirer, Mertand Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/Debiased Machine Learning for Treatment and Structural Parameters." The Econometrics Journal.

Cinelli, Carlos and Chad Hazlett. 2020. "Making sense of sensitivity: extending omitted variable bias." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 82(1):39–67.

Donald, S. G. and W. K. Newey. 1994. \"Series Estimation of Semilinear Models."Journal of Multivariate Analysis 50(1):30{40.

Egami, Naoki. 2018. \Unbiased Estimation and Sensitivity Analysis for Network-Speci c Spillover E ects: Application to An Online Network Experiment." Working Paper.

Egami, Naoki, Christian J. Fong, Justin Grimmer, Margaret E. Roberts and Brandon M. Stewart. 2018. \How to Make Causal Inferences Using Textsa"rXiv:1802.02163 [cs, stat] . arXiv: 1802.02163.

Enos, Ryan D. 2015. \What the Demolition of Public Housing Teaches Usabout the Impact of Racial Threaton Political Behavior." American Journal of Political Scince60(1):12.

Fan, Jianqing and Jinchi Lv. 2008. \Sure independence screening for ultrahigh dimensional feature space."Journal of the Royal Statistical Society: Series B70:849{911.

Fong, Christian. N.d. \Machine Learning Predictions as Regression Covariates." . Forthcoming.

Fong, Christian, Chad Hazlett and Kosuke Imai. 2018. \Covariate balancing propensity score for a continuous treatment: Application to the e cacy of political advertisements." The Annals of Applied Statistics12(1):156{177.

Glynn, Adam N. and Kevin M. Quinn. 2010. \An Introduction to the Augmented Inverse Propensity Weighted Estimator." Political Analysis 18(1):36{56.

Grimmer, Justin, Solomon Messing and Sean J Westwood. 2017. \Estimating heterogeneous

treatment e ects and the e ects of heterogeneous treatments with ensemble methods." Political Analysis 25(4):1{22.

Hainmueller, Jens and Chad Hazlett. 2013. \Kernel Regularized Least Squares: Reducing Misspeci cation Bias with a Flexible and Interpretable Machine Learning Approach." Political Analysis 22(2):143{168.

Hall, Andres B. and Daniel M. Thompson. 2018. \Who Punishes Extremist Nominees? Candidate Ideology and Turning Out the Base in US Elections." American Political Science Review 112(3):509{52.

Hardle, Wolfgang Karl, Marlene Muller, Stefan Sperlich and Axel Werwatz. 2012. Nonparametric and semiparametric models. Springer Science & Business Media.

Hill, Daniel and Zachary Jones. 2014. \An Empirical Evaluation of Explanations for State Repression." American Political Science Review 108(3):661{687.

Hirano, Keisuke and Guido Imbens. 2005. Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives. John Wiley and Sons, Ltd, Chichester, UK chapter The Propensity Score with Continuous Treatments.

Holland, Paul W. 1986. \Statistics and Causal Inference (with Discussion)." Journal of the American Statistical Association 81:945{960.

Hsing, Tailen and Haobo Ren. 2009. \An RKHS formulation of the inverse regression dimension-reduction problem." Annals of Statistics 37(2):726{755. Zbl: 1162.62053.

Hudgens, Michael G. and Elizabeth Halloran. 2008. \Toward Causal Inference with Interference." Journal of the American Statistical Association 103(482):832{842.

Imai, Kosuke, Luke Keele, Dustin Tingley and Teppei Yamamoto. 2011. \Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies." American Political Science Review 105(4):765{789.

Imbens, Guido W. and Donald B. Rubin. 2015. Causal inference for statistics, social, and biometical sciences. Cambridge University Press.

Kennedy, Edward, Ma Zongming McHugh Matthew Small Dylan. in press. \Nonparametric Methods for Doubly Robust Estimation of Continuous Treatment E ects." Journal of the Royal Statistical Society, Series B

King, Gary, Robert Keohane and Sidney Verba. 1994. Designing Social Inquiry. Princeton: Princeton University Press.

Leamer, Edward E. 1983. \Let's Take the Con Out of Econometrics." The American Economic Review 73(1):31{42.

Lenz, Gabriel and Alexander Sahn. 2017. \Achieving Statistical Signi cance with Covariates and without Transparency." Working Paper.

Li, Bing. 2018. Su cient Dimension Reduction: Methods and Applications with R. 1 edition ed. Chapman and Hall/CRC.

Manski, Charles F. 1993. \Identi cation of Endogenous Social E ects: The Re ection Problem." The Review of Economic Studies 60(3).

Manski, Charles F. 2013. \Identi cation of treatment response with social interactions." The Econometrics Journal 16(1):S1{S23.

Mattes, Michaela and Jessica L. P. Weeks. 2019. \Hawks, Doves, and Peace: An Experimental Approach." American Journal of Political Science 63(1):53{66.

Mohanty, Pete and Robert Sha er. 2018. \bigKRLS: Optimized Kernel Regularized Least Squares." Political Analysis Forthcoming. R package version 2.0.4.

Montgomery, Jacob M. and Santiago Olivella. 2018. \Tree-based models for political science data." American Journal of Political Science.

Murphy, Kevin P. 2012. Machine learning: a probabilistic perspective. MIT press.

Ratkovic, Marc and Dustin Tingley. 2017. \Sparse Estimation and Uncertainty with Application to Subgroup Analysis." Political Analysis 1(25):1{40.

Ripley, Brian D. 1988. Statistical Inference for Spatial Processes. Cambridge: Cambridge University Press.

Robins, James M., Andrea Rotnitzky and Lue Ping Zhao. 1994. \Estimation of Regression Coe cients When Some Regressors are Not Always Observed." Journal of the American Statistical Association 89(427):846{866.

Robinson, Peter. 1988. \Root-N Consistent Semiparametric Regression." Econometrica 56(4):931{954.

Samii, Cyrus. 2016. \Causal Empricism in Quantitative Research." Journal of Politics 78(3):941{955.

Samii, Cyrus, Laura Paler and Sarah Zukerman Daly. 2016. \Retrospective Causal Inference with Machine Learning Ensembles: An Application to Anti-recidivism Policies in Colombia." Political Analysis 24(4):434{456.

Savje, Fredrik, Peter M. Aronow and Michael G. Hudgens. 2019. \Average treatment effects in the presence of unknown interference." arXiv:1711.06399 [math, stat]. arXiv: 1711.06399.

Sekhon, Jasjeet S. 2009. \Opiates for the Matches: Matching Methods for Causal Inference." Annual Review of Political Science 12(1):487{508.

Sobel, Michael E. 2006. \What Do Randomized Studies of Housing Mobility Demonstrate?: Causal Inference in the Face of Interference." Journal of the American Statistical Association 101(476):1398{1407.

Stein, Charles. 1956. E cient Nonparametric Testing and Estimation. The Regents of the University of California.

van der Laan, Mark J. and Sherri Rose. 2011. Targeted Learning Causal Inference for Observational and Experimental Data. Springer.

van der Vaart, A. W. and J. H. van Zanten. 2008. Reproducing kernel Hilbert spaces of Gaussian priors. In Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh. Vol. 3 of IMS Collections. pp. 200{222.

van der Vaart, Aad. 1998. Asymptotic Statistics. Vol. 3 of Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press.

Wahba, Grace. 1990. Spline Models for Observational Data Society for Industrial and Applied Mathematics.

Ward, Michael and John O'Loughlin. 2002. \Special Issue on Spatial Methods in Political Science."Political Analysis 10(3):211{216.

Wooldridge, Je rey M. 2013. Introductory Econometrics: A Modern Approach. 6 ed. Cincinnati, OH: South-Western College Publishing.

Young, H. Peyton. 1998. Individual Strategy and Social Structure Princeton University Press, Princeton, NJ.

A Causal Identification

The partial effect of the treatment on the outcome at a given point x_i can be conceptualized as the limit of an estimated slope coefficient from regressing the y_i on t_i for all with the same covariate value x_i and also fixing X_{-i}. The causal interpretation of this parameter involves considering all possible combinations $y_i((t_i; t_{-i}); t_i; t_{-i})$ for all values of t and regressing $y_i(t_i; t_{-i})$ on t_i.

In order for these two parameters to be the same, we need three things. First, $y_i(t_i; t_{-i})$ must equal y_i when the treatment takes the value t. This is the first assumption. Second, the variance of the treatment variable must be positive, so that the denominator of the regression coefficient is nonzero. Third, restricting ourselves to observations with covariate values $x_i; X_{-i}$ should allow us to move t_i freely of the other treatment and of any unobserved confounders. This is our third assumption.

Formally, denote as Cov, Var the sample covariances, and Cov_T, Var_T these operators for a given observation taken with respect to the treatment. Then, under the causal identification assumptions, we can equate the partial effect and causal effect as

$$\beta_i = \underbrace{\frac{\text{Cov}(y_i; t_i | x_i; X_{-i})}{\text{Var}(t_i | x_i; X_{-i})}}_{\text{Partial Effect}} = \underbrace{\frac{\text{Cov}_T(y_i(t_i; t_{-i}); t_i | t_{-i})}{\text{Var}_T(t_i | t_{-i})}}_{\text{Causal Effect}}. \tag{17}$$

The estimand is well-defined by the stable treatment assumption; its denominator is nonzero by the postivity assumption; and the ignorability assumption allows us to equate the numerators and denominators.

Equating the partial effect and observation-level effect for each observation equates their

averages.

B Proof of SPE

The proof proceeds in two steps. First, we establish that a $n^{1/4}$ estimation rate on the nuisance functions gives an $n^{1/4}$ rate on the control vector $\theta_{i,b}$. Then, we show that this $n^{1/4}$ rate on $\theta_{i,b}$ combined with the split-sample approach gives us a semparametrically efficient estimate b.

B.1 Convergence Rates, Principal Components of the Variance, and a Feasible Estimator

We are now ready to construct the control set U_i and its feasible counterpart $\hat{U}_i$. In order to do so, we first describe the nonparametric function space within which we estimate.

B.1.1 Preliminaries: The Structure of the Nonparametric Function Space

In order to develop a feasible estimation strategy, we need to add some structure to the nonparametric nuisance functions. We assume the conditional means of the outcome and treatment live in a Reproducing Kernel Hilbert Space (RKHS, see Murphy, 2012; Wahba, 1990; van der Vaart and van Zanten, 2008, for a refresher on RKHS theory and Gaussian processes).

We use two key results from RKHS theory. Each involves the "reproducing kernel," (RK) a function that measures the similarity between two points. The key insight is that the basis representation may involve thousands or more possible bases, but the function can be equivalently expressed in terms of this $n \times n$ RK matrix. The RK is better thought of as a

generalized covariance matrix rather than a kernel, as in kernel density estimation. Modeling in terms of the RK instead of the basis representation is known as the \kernel trick," a well-established tool in machine learning (e.g. Murphy, 2012; Hainmueller and Hazlett, 2013). Our chief innovation here is using a basis representation to estimate the reproducing kernel. Our assumption that $U_{i;u_i}$ is nite dimensional is equivalent to assuming the RK is nite dimensional, or in our setting, the covariance of the basis-representation coe cient tted values is nite dimensional. This is a \su cient dimension reduction" assumption (Li, 2018; Hsing and Ren, 2009). Our results build on Hsing and Ren (2009) by providing a feasible means of estimating a RK, rather than assuming it follows some known functional form, such as being composed of Gaussian radial basis functions (as in Hainmueller and Hazlett, 2013).

By estimating and including principal components of the RK as covariates, we are able to adjust for a broad class of functions in our regression. The split sample, as described above, will guard against over tting, ensuring a valid con dence interval.

B.1.2 Formalities

An RKHS is a space of functions denoted H_m with generic element m and inner product between between elements $m; m^0$ as $\langle m; m^0 \rangle_{H_m}$. Functions in an RKHS are bounded and continuous and have nite norm, $0 < \langle m; m \rangle_{H_m} < C_m < 1$, for all $m \in H_m$, and $\langle m; m \rangle_{H_m} = 0$ for all m not in H_m.

Every RKHS has a function called the \reproducing kernel" (RK) a symmetric function that takes as its arguments two covariates from space X and returns a real number, K :

X X 7! < .[19].

We utilize two properties of the reproducing kernel. First, the Riesz Representation Theorem guarantees that any function $m \in H_m$ can be expressed in two equivalent ways. The first is linear in a set of (potentially infinite) bases, b_k, the second in terms of the reproducing kernel:

$$m(x) = b(x)^{>} c_k \tag{18}$$

$$= \sum_{i=1}^{n} K(x_i; x) \; _k = K_i(x)^{>} \tag{19}$$

By this means, we can see that adjusting for the RK will adjust for any function with basis representation in $b(x)$. We will denote the matrix with $b(x_i)$ in rows as B and the $n \quad n$ reproducing kernel matrix with entry in $i; i^0$ as $K(x_i; x_{i^0})$.

Second, the RK has the property that it \reproduces" the inner product for any function in the space,

$$\langle m; m^0 \rangle_{H_m} = \int_{x \in X} \int_{x^0 \in X} m(x) m^0(x^0) K(x; x^0) dF_x dF_{x^0} \tag{20}$$

which, combined with the law of large numbers, will give us the estimate

$$\widehat{\langle m; m^0 \rangle}_{H_m} = \frac{1}{n} \sum_{i=1}^{n} \sum_{i^0=1}^{n} m(x_i) m^0(x_{i^0}) K(x_i; x_{i^0}) \tag{21}$$

Third, we connect these two using the celebrated Representation Theorem of Craven and

[19] For a gentle introduction to the intuition between an RKHS, see Hainmueller and Hazlett (2013)

Wahba (see Wahba, 1990). We assume, in this case, that the function m lives in the space

$$m \in H_0 \oplus H_m$$

where H_0 is a parametric space spanned by basis $S(x_i)$, known functions of the data known in advance and fixed in sample size, and H_m is an infinite dimensional RKHS. Then, considering the regression problem where $m(x) = E(z_i | x = x_i)$, the best estimate of $m(x)$ (in a mean square sense) can be found through solving

$$\{\hat{\alpha}_m, \hat{b}_m\} = \underset{\alpha, e_m}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} (z_i - S(x_i)^\top \alpha_m + b_{i,m}^\top e_m)^2 + \frac{1}{n} \lambda_m e_m^\top B_m^\top K_m B_m e_m \qquad (22)$$

for some $\lambda_m > 0$, where the first term is the residual sum of squares and the second is the estimate of the norm of the functional. Through this representation, we are able to derive an estimate of the principal components of the $n \times n$ RK evaluated on the data. We then use these principal components to generate a set of controls that will adjust for a broad range of functions.

B.1.3 Formal Statement of Results

First, we derive closed form representation of the span of the RK for each nuisance space, which will give us the true control set U_f and U_g, where these are the control set for the outcome and treatment, respectively. We use $\text{span}(A)$ to denote the span of matrix A, and we say $A \overset{sp}{=} A^0$ if matrices A and A^0 have the same span. All proofs are in the following section.

Lemma 1 (Connecting the Kernel, Error Variance, and Data) For $j \in \{f, g\}$

and $_j$ 2 f ; v g, de ne

$$M_{\theta;\mathbf{b};j} = \text{sp}\ \frac{1}{n}(\text{Var}(\mathbf{b}_j))\ \ B^{>}_{j;\mathbf{b}}\text{Var}(\mathbf{b}_j)B_{j;\mathbf{b}} \tag{23}$$

$$M_{U;u;j} = \lim_{n!1}\ \text{sp}\ \left\{M_{\theta;\mathbf{b};j}\right\} \tag{24}$$

Under Assumptions 2, 3, 4 we can expres$S_{j;u}$ $^{\text{sp}}$ $B_{j;u}M_{U;u;j}$.

These estimates are infeasible, since we observe neither variance M$_{\theta;\mathbf{b};j}$. We next derive a feasible estimate. Doing so requires sample-splitting subset S_1 into S_{1a} and S_{1b}, creating a crucial conditional independence between the coe cient estimates and estimated residuals. We use cross- tting to recover an M matrix. We denote as $\text{Var}^{;C}(A)$ the bootstrap estimated variance of vector A on split C. Note that $\mathbf{b}$ is estimated on split S_0 but evaluated on S_1.

Lemma 2 (Feasible Estimation of the Kernel) For j 2 f f; g g and $\mathbf{b}_j$ 2 f $\mathbf{b}$; $\mathbf{b}$g,

$$M_{\theta;\mathbf{b};j} = \text{sp}\ \frac{1}{n_1}\ \text{Var}^{;S_{1a}}(\mathbf{b}_j)\ \ B^{>}_{j;\mathbf{b}}\text{Var}^{;S_{1b}}(\mathbf{b}_j)B_{j;\mathbf{b}} \tag{25}$$

Denote a matrix $U_{j;\mathbf{b}}$ such that $U_{j;\mathbf{b}}$ $^{\text{sp}}$ $B_{j;\mathbf{b}}M_{\theta_{j;\mathbf{b}}}$, and $_{\theta_{j;\mathbf{b}}}$ and $_{U_{j;u}}$ the least-squares projection matrices of an arbitrary unit vector of length n_1 onto the matrices in the subscript.

Under Assumptions 3 and 4, and using the split-sample strategy described in Figure 1,

$$\lim_{n!1}\ n^{1=4}\ \left(_{\theta_{j;\mathbf{b}}} - _{U_{j;u}}\right) !^{u}\ 0 \tag{26}$$

in operator norm.

We utilize this result below, as it establishes our $n^{1/4}$ rate on $\hat{\theta}_{j,b}$.

B.1.4 Proof of Lemma 1.

To reduce notation, we suppress notation indicating that all calculations are done on split sample S_1, and, for the treatment model, the basis representation are evaluated using estimated errors from sample S_0. We denote as $b_j \in \{b_j, b_{-j}\}$ with i^{th} element $b_{j,i}$. Also, to reduce notation, we assume that the linear terms in $S(x_i)$ have been partialed out of the outcome.

Since the conditional means are in a reproducing kernel Hilbert space, the estimates $\hat{b}_j$ are sample minimizers of

$$L_{K_j,\lambda_j}(e_j) = \frac{1}{n}\sum_{i=1}^{n}(y_i - B_{j,i}^{\top} e_j)^2 + \frac{1}{n}\lambda_j e_j^{\top} B_j^{\top} K_j B_j e_j \tag{27}$$

$$\hat{b}_j = \underset{e_j}{\operatorname{argmin}}\, L_{K_j,\lambda_j}(e_j) \tag{28}$$

for some K_j, λ_j. The estimating equations give

$$\frac{\partial}{\partial e_j} L_{K_j,\lambda_j}(e_j)\Big|_{e_j = \hat{b}_j} = 0 \tag{29}$$

$$\frac{1}{n}\sum_{i=1}^{n} B_{j,i}\hat{b}_{j,i} = \frac{1}{n}\lambda_j B_j^{\top} K_j B_j \hat{b}_j \tag{30}$$

and taking the variance of both sides

$$\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} B_{j,i}\hat{b}_{j,i}\right) = \frac{\lambda_j^2}{n^2} B_j^{\top} K_j B_j \mathrm{Var}(\hat{b}_j) B_j^{\top} K_j B_j \tag{31}$$

Noting that $\frac{1}{n}B_j^{\top}B_j$ converges to a Gram matrix, since the bases are bounded, a nondegenerate solution requires $\lambda_j = O_p(1/\sqrt{n})$

Let $U_{j;k}$ be an arbitrary unit vector in the span of K_j. Right-multiplying by $B_j^{>} U_{j;k}$ and rearranging, with the inverse with respect to the range of K_j, gives

$$\frac{1}{n^2} B_j^{>} K_j B_j \, \mathrm{Var}(b_j) B_j^{>} K_j B_j \quad \mathrm{Var}\left[p\frac{1}{n} \sum_{i=1}^{n} B_{j;i} b_{j;i} \right] B_j^{>} U_{j;k} = N \; {}^2 B_j^{>} U_{j;k}; \qquad (32)$$

showing that $U_{j;k}$ is a singular vector of the matrix on the lefthand side.

We now show that $\mathrm{span}(U_j) = \mathrm{span}\{ \mathrm{Var}(b_j) \;\; B_j^{>} \mathrm{Var}(b) \}$. First, denote V_1 as an arbitrary vector in $\mathrm{span}\{ \mathrm{Var}(b_j) \;\; B_j^{>} \mathrm{Var}(b_j) \}$. V_1 is, therefore, in the span of the lefthandside of Equation 32, and hence in the span of U_j. Next, denote V_2 an arbitrary element of $\mathrm{span}(U_j)$. V_2 is therefore additive in the singular vectors $U_{j;k}$, and therefore satisfies equation 32. Therefore, the two spaces are equal.

B.2 Proof of Lemma 2.

First, the bootstrap variance estimate is valid by the uniform convergence of b_j and van der Vaart (1998, ch. 23.2). The second claim follows through three steps. First, the error attributable to b_i in $B_{i;j;\, b_i}$ is $o_p(n^{\;1=4})$ uniformly. Second, the projection matrix is a continuous function of b_j, and b_i, both of which converges at $n^{\;1=4}$. Convergence is then controlled by the slower of these, and both are $n^{\;1=4}$. Using a Bonferroni bound over the finite dimensions of $U_{i;u}$, the result gives uniform convergence at the $n^{\;1=4}$ rate in the projection matrices.

B.3 Semiparametric Efficiency under an $n^{\;1=4}$ rate on $b_{i;\, b}$

We start with the consequence of Lemma 2, clarifying what rates we require and on which nuisance components.

Assumption 5 The treatment and outcome are well-approximated such that the estimates converge as

$$\lim_{n\to\infty} n^{1/4} \left| \hat{\gamma}_{\theta,u} - \hat{\gamma}_{U,u} \right| \xrightarrow{u} 0 \tag{33}$$

$$\lim_{n\to\infty} n^{1/4} \left| \hat{\gamma}_{\theta,\hat{\mathbf{b}}} - \hat{\gamma}_{\theta,u} \right| \xrightarrow{u} 0 \tag{34}$$

$$\lim_{n\to\infty} n^{1/4} \left| \hat{\tau}_{\theta,u} - \hat{\tau}_{U,u} \right| \xrightarrow{u} 0 \tag{35}$$

$$\lim_{n\to\infty} n^{1/4} \left| \hat{\tau}_{\theta,\hat{\mathbf{b}}} - \hat{\tau}_{\theta,u} \right| \xrightarrow{u} 0. \tag{36}$$

We are interested in analyzing the empirical process,

$$\sqrt{n_2}\left(\hat{\tau}_{\theta,\hat{\mathbf{b}}} - \quad\right) \tag{37}$$

taken on split S_2. We proceed by decomposing

$$\sqrt{n_2}\left(\hat{\tau}_{\theta,\hat{\mathbf{b}}} - \quad\right) = \sqrt{n_2}\left(\hat{\tau}_{U,u} - \quad\right) + \sqrt{n_2}\left(\hat{\tau}_{\theta,\hat{\mathbf{b}}} - \hat{\tau}_{U,u}\right). \tag{38}$$

The first term on the righthand side in the previous display is the infeasible estimator, were U and u known. The strategy involves showing that the sample splitting and assumption on the error rates sends the second term to zero uniformly. Establishing the second term goes to zero uniformly requires expanding the difference between the feasible and infeasible estimator, to which we now turn. We start with $\sqrt{n_2}\hat{\tau}_{\theta,\hat{\mathbf{b}}}$

$$\sqrt{n_2}\hat{\tau}_{\theta,\hat{\mathbf{b}}} = \sqrt{n_2}\frac{\frac{1}{n_2}\sum_{i\in S_2}(y_i - \hat{\gamma}_{i;\theta,\hat{\mathbf{b}}_i})(t_i - \hat{\tau}_{i;\theta,\hat{\mathbf{b}}_i})}{\frac{1}{n_2}\sum_{i\in S_2}(t_i - \hat{\tau}_{i;\theta,\hat{\mathbf{b}}_i})^2} \tag{39}$$

Starting with the denominator,

$$\frac{1}{n_2}\sum_{i2S_2}(t_i - \hat{t}_{i;\hat{\theta}_i;\hat{\boldsymbol{\beta}}_i})^2 \xrightarrow{u} E(u_i^2) \tag{40}$$

by convergence in $\hat{\theta}$ and u, linearity in u, and the continuous mapping theorem, which is the same limit as the denominator of $\hat{\theta}_{U;u}$. We will combine this component with the numerator via the uniform Slutsky's theorem later.

Moving on to the numerator, we expand it as

$$\sqrt{n_2}\frac{1}{n_2}\sum_{i2S_2}(y_i - \hat{y}_{i;\hat{\theta}_i;\hat{\boldsymbol{\beta}}_i})(t_i - \hat{t}_{i;\hat{\theta}_i;\hat{\boldsymbol{\beta}}_i}) \tag{41}$$

$$\begin{aligned}=\sqrt{n_2}\frac{1}{n_2}\sum_{i2S_2} & (y_i - \hat{y}_{i;U_i;u_i} + \hat{y}_{i;U_i;u_i} - \hat{y}_{i;\hat{\theta}_i;u_i} + \hat{y}_{i;\hat{\theta}_i;u_i} - \hat{y}_{i;\hat{\theta}_i;\hat{\boldsymbol{\beta}}_i}) \\ & (t_i - \hat{t}_{i;U_i;u_i} + \hat{t}_{i;U_i;u_i} - \hat{t}_{i;\hat{\theta}_i;u_i} + \hat{t}_{i;\hat{\theta}_i;u_i} - \hat{t}_{i;\hat{\theta}_i;\hat{\boldsymbol{\beta}}_i})\end{aligned} \tag{42}$$

which splits into nine terms

$$= \sqrt{n_2}\,\frac{1}{n_2}\sum_{i\in S_2}(y_i - \hat{y}_{I;U_i;u_i})(t_i - \hat{t}_{I;U_i;u_i}) + \frac{1}{n_2}\sum_{i\in S_2}(y_i - \hat{y}_{I;U_i;u_i})(\hat{t}_{I;U_i;u_i} - \hat{t}_{I;\hat{\theta}_i;u_i}) + \tag{43}$$
$$\underbrace{\qquad\qquad}_{=(a)} \qquad\qquad \underbrace{\qquad\qquad\qquad\qquad}_{=(b)}$$
$$\frac{1}{n_2}\sum_{i\in S_2}(y_i - \hat{y}_{I;U_i;u_i})(\hat{t}_{I;\hat{\theta}_i;u_i} - \hat{t}_{I;\hat{\theta}_i;\hat{\mathbf{b}}_i})$$
$$\underbrace{\qquad\qquad\qquad\qquad}_{=(c)}$$

$$+ \frac{1}{n_2}\sum_{i\in S_2}(\hat{y}_{I;U_i;u_i} - \hat{y}_{I;\hat{\theta}_i;u_i})(t_i - \hat{t}_{I;U_i;u_i}) + \frac{1}{n_2}\sum_{i\in S_2}(\hat{y}_{I;U_i;u_i} - \hat{y}_{I;\hat{\theta}_i;u_i})(\hat{t}_{I;U_i;u_i} - \hat{t}_{I;\hat{\theta}_i;u_i}) + \tag{44}$$
$$\underbrace{\qquad\qquad}_{=(d)} \qquad\qquad \underbrace{\qquad\qquad\qquad\qquad}_{=(e)}$$
$$\frac{1}{n_2}\sum_{i\in S_2}(\hat{y}_{I;U_i;u_i} - \hat{y}_{I;\hat{\theta}_i;u_i})(\hat{t}_{I;\hat{\theta}_i;u_i} - \hat{t}_{I;\hat{\theta}_i;\hat{\mathbf{b}}_i})$$
$$\underbrace{\qquad\qquad\qquad\qquad}_{=(f)}$$

$$+ \frac{1}{n_2}\sum_{i\in S_2}(\hat{y}_{I;\hat{\theta}_i;u_i} - \hat{y}_{I;\hat{\theta}_i;\hat{\mathbf{b}}_i})(t_i - \hat{t}_{I;U_i;u_i}) + \frac{1}{n_2}\sum_{i\in S_2}(\hat{y}_{I;\hat{\theta}_i;u_i} - \hat{y}_{I;\hat{\theta}_i;\hat{\mathbf{b}}_i})(\hat{t}_{I;U_i;u_i} - \hat{t}_{I;\hat{\theta}_i;u_i}) + \tag{45}$$
$$\underbrace{\qquad\qquad}_{=(g)} \qquad\qquad \underbrace{\qquad\qquad\qquad\qquad}_{=(h)}$$
$$\frac{1}{n_2}\sum_{i\in S_2}(\hat{y}_{I;\hat{\theta}_i;u_i} - \hat{y}_{I;\hat{\theta}_i;\hat{\mathbf{b}}_i})(\hat{t}_{I;\hat{\theta}_i;u_i} - \hat{t}_{I;\hat{\theta}_i;\hat{\mathbf{b}}_i})$$
$$\underbrace{\qquad\qquad\qquad\qquad}_{=(i)}$$

Term (a) is exactly the numerator in $\hat{b}_{U;u}$, so we next show that the remaining terms go to zero uniformly. These will do so for two reasons: either by the split-sample approach or the assumption on the $n^{1=4}$ rate on the convergence.

To illustrate the logic of sample splitting, we give the sources of variance for each of the six terms and the split from which its variance depends below:

Term	Source fo Variance
$y_i - \hat{\mu}_{1;U_i;u_i}, t_i - \hat{\tau}_{1;U_i;u_i}$	S_2
$\hat{\mu}_{1;U_i;u_i} - \hat{\mu}_{1;\hat{\theta}_i;u_i}, \hat{\tau}_{1;U_i;u_i} - \hat{\tau}_{1;\hat{\theta}_i;u_i}$	S_1
$\hat{\mu}_{1;\hat{\theta}_i;u_i} - \hat{\mu}_{1;\hat{\theta}_i;\hat{\alpha}_i}, \hat{\tau}_{1;\hat{\theta}_i;u_i} - \hat{\tau}_{1;\hat{\theta}_i;\hat{\alpha}_i}$	S_0

The first cross term is attributable to variance in S_2, which is what we want: this is the term that will drive our estimation and inference. For the remaining terms, the cross-product terms containing variance generated by di erent splits are zero, since the two forms of variance come from di erent data and are conditionally independent. This includes terms (b), (c), (d), (f), (g), (h) above. We illustrate with term (b):

$$\lim_{n\to\infty} p\sqrt{n_2}\left(\frac{1}{n_2}\sum_{i\in S_2}(y_i - \hat{\mu}_{1;U_i;u_i})(\hat{\tau}_{1;U_i;u_i} - \hat{\tau}_{1;\hat{\theta}_i;u_i})\right) \tag{46}$$

$$= \lim_{n\to\infty} p\sqrt{n_2}E\left\{(y_i - \hat{\mu}_{1;U_i;u_i})(\hat{\tau}_{1;U_i;u_i} - \hat{\tau}_{1;\hat{\theta}_i;u_i})\right\} \tag{47}$$

$$= \lim_{n\to\infty} p\sqrt{n_2}E\left\{E\left[(y_i - \hat{\mu}_{1;U_i;u_i})(\hat{\tau}_{1;U_i;u_i} - \hat{\tau}_{1;\hat{\theta}_i;u_i})\middle| i \in S_1\right]\right\} \tag{48}$$

$$= \lim_{n\to\infty} p\sqrt{n_2}E\left\{(y_i - \hat{\mu}_{1;U_i;u_i})E\left[(\hat{\tau}_{1;U_i;u_i} - \hat{\tau}_{1;\hat{\theta}_i;u_i})\middle| i \in S_1\right]\right\} \tag{49}$$

$$= \lim_{n\to\infty} p\sqrt{n_2}E\left\{(y_i - \hat{\mu}_{1;U_i;u_i})E(\hat{\tau}_{1;U_i;u_i} - \hat{\tau}_{1;\hat{\theta}_i;u_i})\right\} \tag{50}$$

$$= \lim_{n\to\infty} p\sqrt{n_2}E(y_i - \hat{\mu}_{1;U_i;u_i})E(\hat{\tau}_{1;U_i;u_i} - \hat{\tau}_{1;\hat{\theta}_i;u_i}) \tag{51}$$

$$= \lim_{n\to\infty} p\sqrt{n_2} \quad 0 \quad 0 \tag{52}$$

$$!^{u} \ 0 \tag{53}$$

where the lines follow from the law of large numbers, law of iterated expectations, the fact that any variance in $y_i - \hat{\mu}_{1;U_i;u_i}$ is due to S_2, the random splitting of the sample into the three

splits, taking the constant expectation outside of the outer expectation, uniform convergence, and then multiplying by zero.

The remaining cross-terms where both terms contain variance from the same sample are terms (e) and (i), and these go away by assumption. We illustrate with (e):

$$\lim_{n\to\infty} \sqrt{n_2}\frac{1}{n_2}\sum_{i\in S_2}(\hat{y}_{l;U_i;u_i} - \hat{y}_{l;\theta_i;u_i})(\hat{t}_{l;U_i;u_i} - \hat{t}_{l;\theta_i;u_i}) \tag{54}$$

$$\lim_{n\to\infty} \sqrt{n}\,(\hat{y}_{U;u} - \hat{y}_{\theta;u})(\hat{t}_{U;u} - \hat{t}_{\theta;u}) \tag{55}$$

$$\lim_{n\to\infty} \sqrt{n}\,\hat{y}_{U;u} - \hat{y}_{\theta;u}\quad \hat{t}_{U;u} - \hat{t}_{\theta;u} \tag{56}$$

$$= \lim_{n\to\infty} n^{1/4}\,\hat{y}_{U;u} - \hat{y}_{\theta;u}\quad n^{1/4}\,\hat{t}_{U;u} - \hat{t}_{\theta;u} \tag{57}$$

$$\to^{u} 0 \tag{58}$$

where the lines follow because $n_2 = n/3$ and the definition of uniform convergence, the Cauchy-Schwarz inequality, distributing $\sqrt{n}$, and then Assumption 5.

With this argument, we have completed our argument that terms (b) - (i) are uniformly $\sqrt{n}$-negligible. Since (a) is the only term that does not disappear, it follows that

$$\lim_{n\to\infty} \sqrt{n_2}\,\frac{1}{n_2}\sum_{i\in S_2}(y_i - \hat{y}_{l;\theta_i;\mathbf{b}_i})(t_i - \hat{t}_{l;\theta_i;\mathbf{b}_i}) \quad \frac{1}{n_2}\sum_{i\in S_2}(y_i - \hat{y}_{l;U_i;u_i})(t_i - \hat{t}_{l;U_i;u_i}) \to^{u} 0. \tag{59}$$

Combined with Slutsky's theorem for the denominator, it follows that

$$\lim_{n\to 0} \sqrt{n}\,|\hat{b}_{\theta;\mathbf{b}} - b_{U;u}| \to^{u} 0 \tag{60}$$

meaning any difference between our feasible estimator and the infeasible SPE estimator is uniformly $\sqrt{n}$-negligible. This implies our estimator is SPE.

C Estimation Details

The algorithm splits into two components. If the first, a full sample analysis, we select a viable set of basis functions. In the second, we implement the split-sample approach given in the text. Regression in this section will refer to the sparse regression of Ratkovic and Tingley (2017).

C.1 Full Sample Analysis

In the full sample, we construct a set of viable basis functions for each nuisance element. It is important that we err in favor of over-selecting bases at this stage, since we are using the full data.

Our screening comes in two steps: we construct all possible bases, as described in Section 4.2, and save the $\min(\max(25 \cdot (1 + n^{-.2}), 400), 50)$ with the highest correlation with the outcome (Fan and Lv, 2008). We then use a sparse regression to select a subset of these, and then repeat this process on thirty bootstrapped samples, saving all bases that are selected in any bootstrapped samples. We then take the maintained bases, regress the outcome on this maintained sample, and repeat the selection process on the residuals. These bases are saved.

We apply this screening process to the outcome and the treatment, to model f_0, g_1. To model g_2, the conditional heteroskedasticity term, we apply the basis selection process to the absolute value of the treatment residuals, after regressing the treatment on the selected set.

In selecting the interference bases, we begin with the full-sample residuals for the outcome and treatment, after regressing each on the selected bases. We construct all sets of interference bases, as given in Section 4.2, and keep the 30 with the highest correlation with the outcome. In this step, we set the bandwidth equal to the "rule-of-thumb" bandwidth for computational reasons, though we will optimize this parameter in the split-sample component below. We iterate the process, giving us 60 bases for the outcome to model, ℓ_y, and 60 for the treatment to model ℓ_t. We maintain all 120 interference bases throughout.

C.2 Split Sample Analysis

We next outline our split sample algorithm.

C.2.1 Split S_0

In split S_0, we regress the treatment on the selected bases for g_1 from above to generate a model for the treatment error. We regress the outcome on the maintained bases for ℓ_t. We take the residuals from each of these, and select an optimal bandwidth to maximize the correlation between the residuals and the interference basis for each maintained interference basis, again only using the data in S_0.

We now have an estimated treatment residual and estimated bandwidth parameters, which we carry to S_1.

C.2.2 Split S_1

We now generate our covariate control vector. To get the selected bases, we use only data on S_1, and regress the outcome on the bases and interference bases in ℓ_y, maintaining ℓ_t selected bases. We then regress the treatment on the bases in g_1, the estimated treatment

error times the bases in g_2, and the bases in $_t$. The selected bases are pooled.

Next, we take the full set of bases for the outcome and generate the estimated reproducing kernel following the procedure in Lemma 2, using 30 bootstraps. We take the principal components of the estimated reproducing kernel, and select the number of principal components using ordinary cross-validation. We repeat this process for the treatment using the full set of treatment covariates.

C.2.3 Split S_2

The sparse bases selected in S_1 and the maintained principal components are included as controls for the outcome on the treatment. The full set in this $\theta_{i;b}$ has some linear dependence, so we standardize each variable and take principal components, in order to eliminate redundant covariates, and include all principal components with a non-zero weight (the diagonal d element in the SVD decomposition).

These are entered as controls from regressing the outcome on the treatment. Robust standard errors are calculated.[20] Standard errors estimated o S_2 are divided by $\sqrt{3}$, since we will be averaging over splits of size $\frac{1}{3}$ (see, e.g. Theorem 3.2 of Chernozhukov et al., 2018).

(?) We found that averaging over cross- ts was far more important than cross- tting within a given split. In our implementation, we cross- t once, xing S_0, then swapping the roles of $S_1; S_2$, getting two point estimates and standard error estimates for each split. We recommend 25 splits for 50 total estimates, unless the estimate still appears unstable, at which point this number should be increased.

[20] In the Mattes and Weeks (2019), the original authors used the \HC0" standard errors so we use those there but the standard errors we introduce below everywhere else.

We next average over cross- ts. Chernozhukov et al. (2018) recommend taking the median over repeated cross- ts, as it is robust to outliers. We were concerned about the e ciency loss{the variance of the median is about 57% larger than the mean if the data is normally distributed. Instead, we implemented the Hodges-Lehmann estimator, which is the median of all pairwise averages. This estimate has variance only 5% larger than the mean but is still robust to outliers (the breakdown point for the Hodges-Lehmann estimator is 27% versus 50% for the median). We found the estimator to be more robust than the mean, but also more e cient than the median.

We confront two di erent consistent variance estimates that seem natural and are easily recovered from the method. The rst is found by averaging the variance from each split and cross- t, using the law of total variance. We call this $\hat{b}_1^2$. The second comes from noting that we can construct a jackknife variance estimate constructed from the point estimates. This is the sample variance of the point estimates scaled by $n=(3)=(n \quad n=3) = 1=2$ since we are constructing the variance estimate from $n=3$ observations, which we denote as $\hat{s}_{HC}^2$. With no clear reason to favor either, since both are consistent under heteroskedasticity, we take their average.

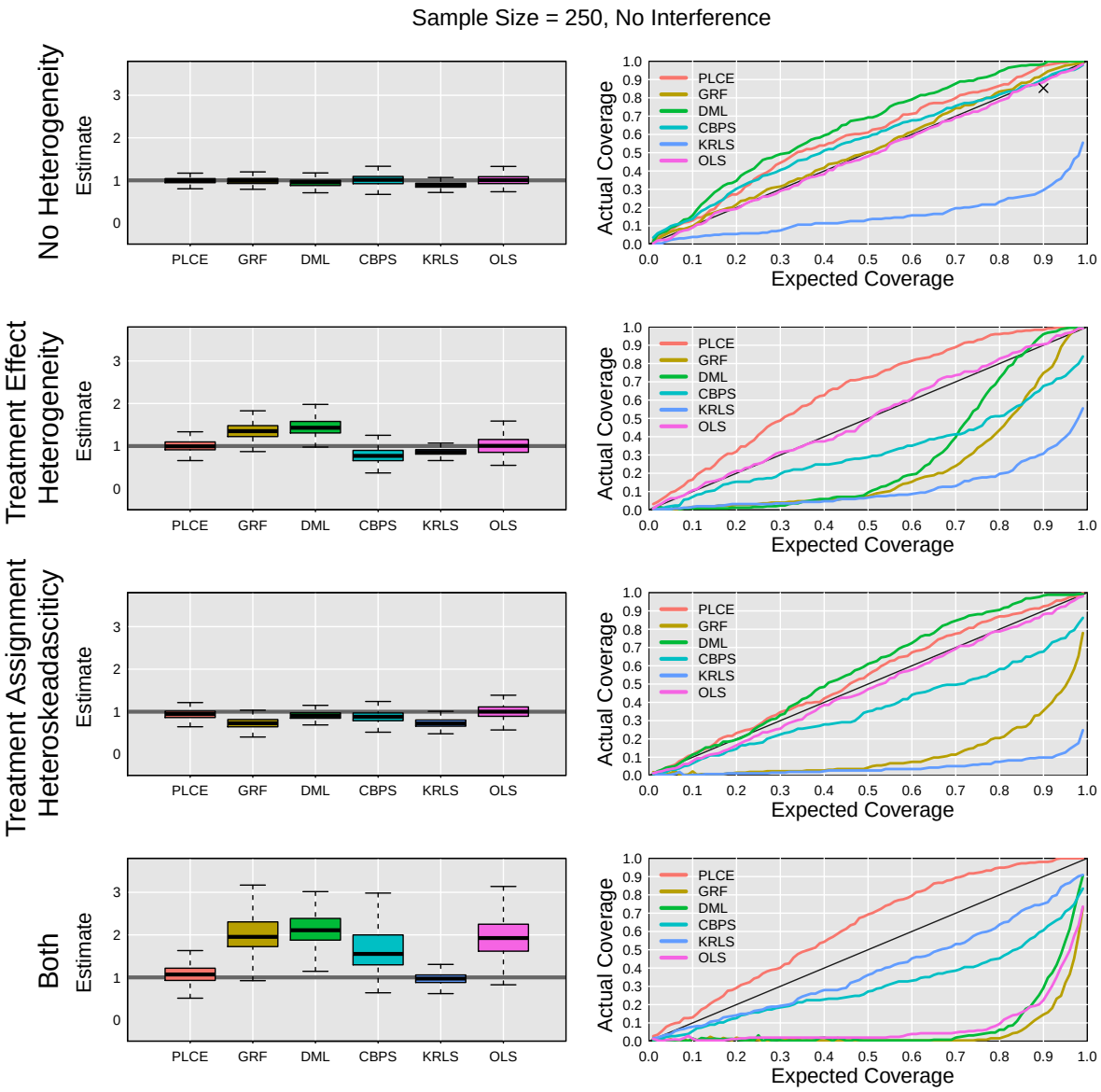
We denote as $HL()$ the Hodges-Lehman mean of a vector. Given vectors of estimates

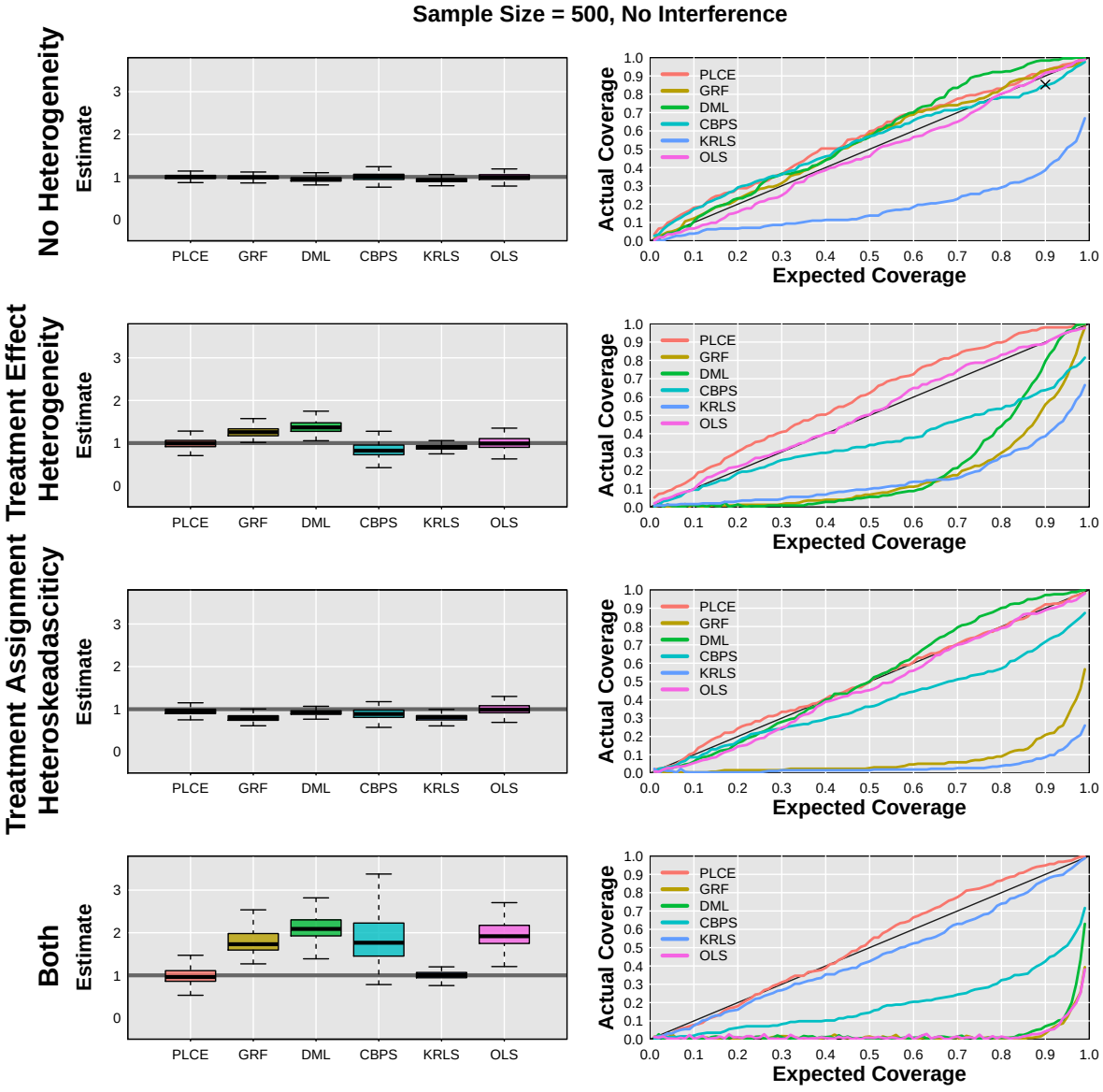
$\tilde{\theta}$ and S variance estimates $\hat{\sigma}_1^2$ over splits and cross- ts, we then return an estimate,

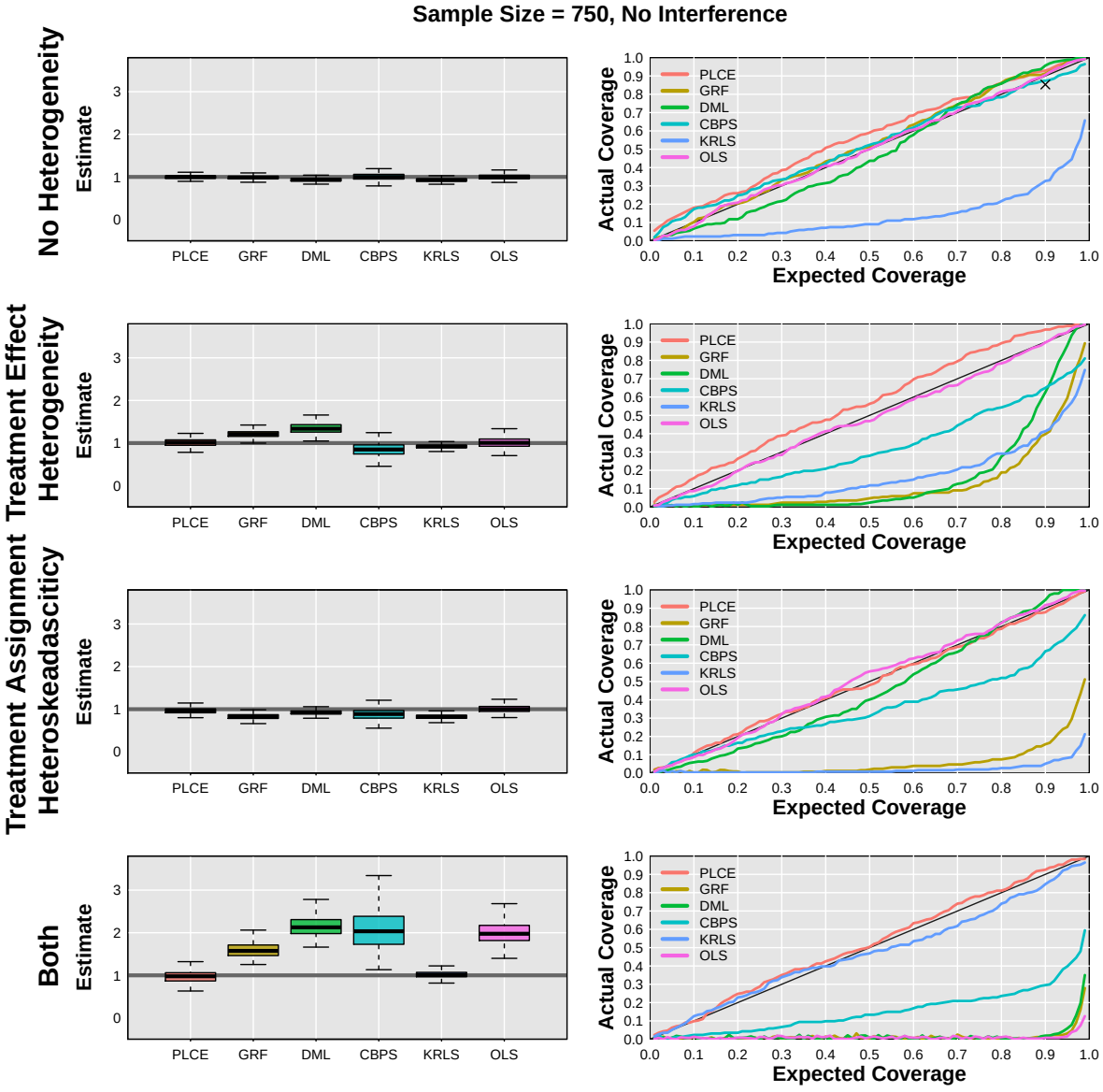
$$\begin{aligned}
\hat{\theta}^{PLCE} &= HL \quad \tilde{\theta} \\
\hat{\sigma}_1^2 &= HL \quad \hat{\sigma}^2 \quad + \frac{1}{S} HL \quad \tilde{\theta} \quad \hat{\theta}^{PLCE} \quad {}_2 \Big)^2 \\
\hat{\sigma}_2^2 &= \frac{1}{2} HL \quad \tilde{\theta} \quad \hat{\theta}^{PLCE} \quad {}_2 \Big)^2 \\
{}^{2;PLCE} &= \frac{1}{2}\hat{\sigma}_1^2 + \frac{1}{2}\hat{\sigma}_2^2
\end{aligned}$$

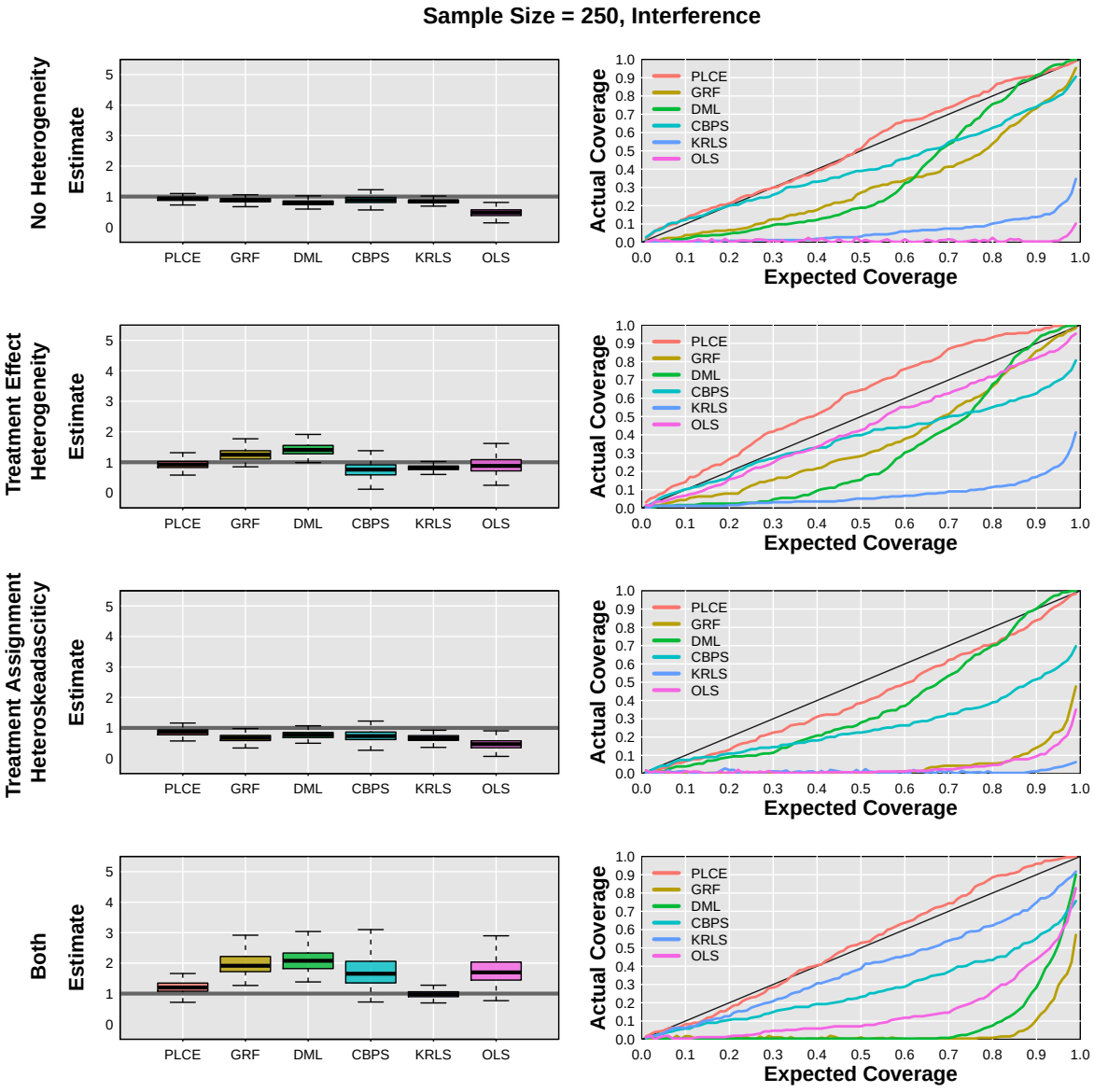
D Additional Simulations

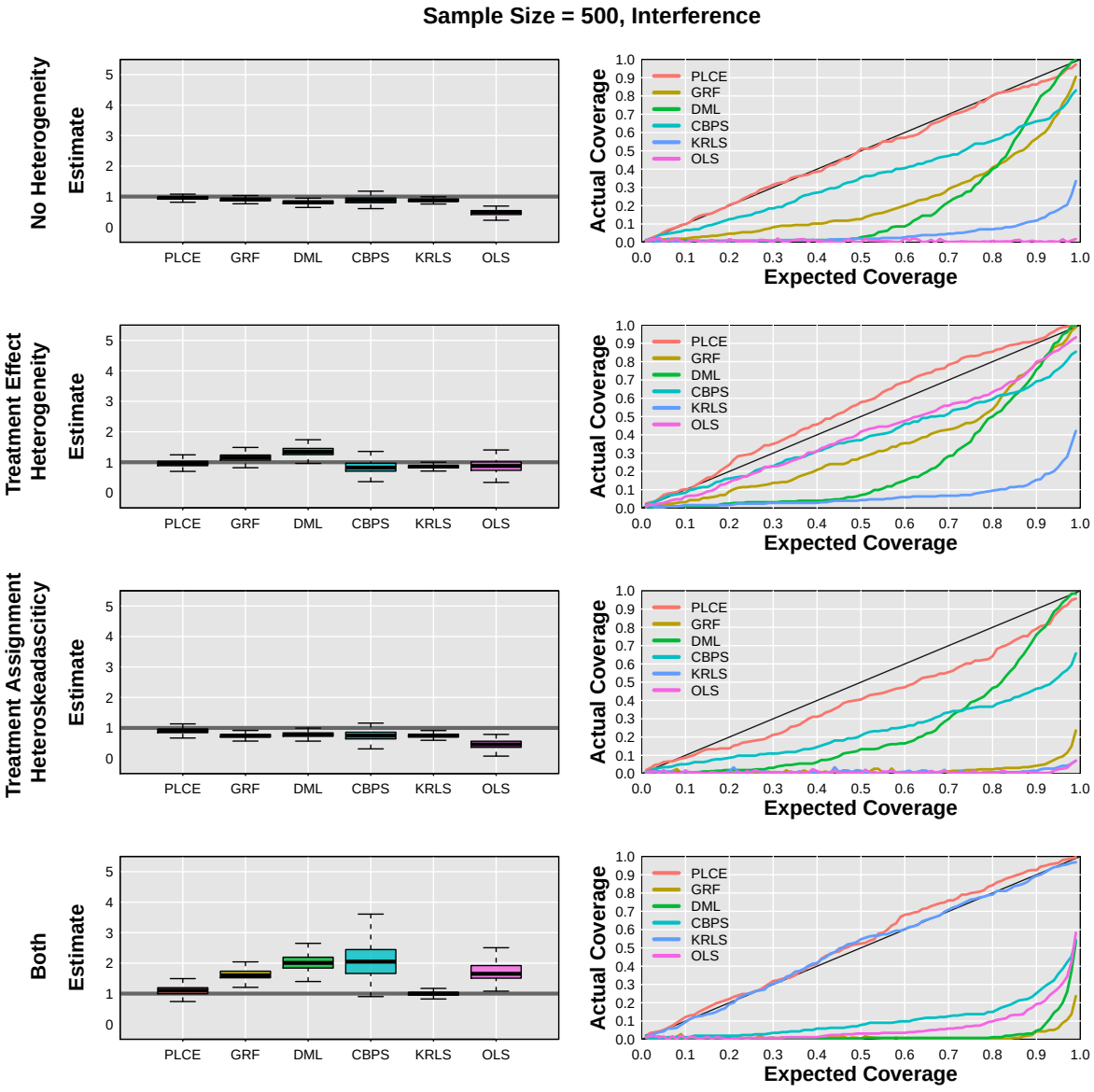
This appendix presents simulations for sample sizes $n \in \{100, 250, 500\}$ to supplement those in the text at $n = 1000$. It also presents results comparing the first and second order efficient estimates, where the latter is the PLCE estimator and the former does not include the bootstrapped principal components.

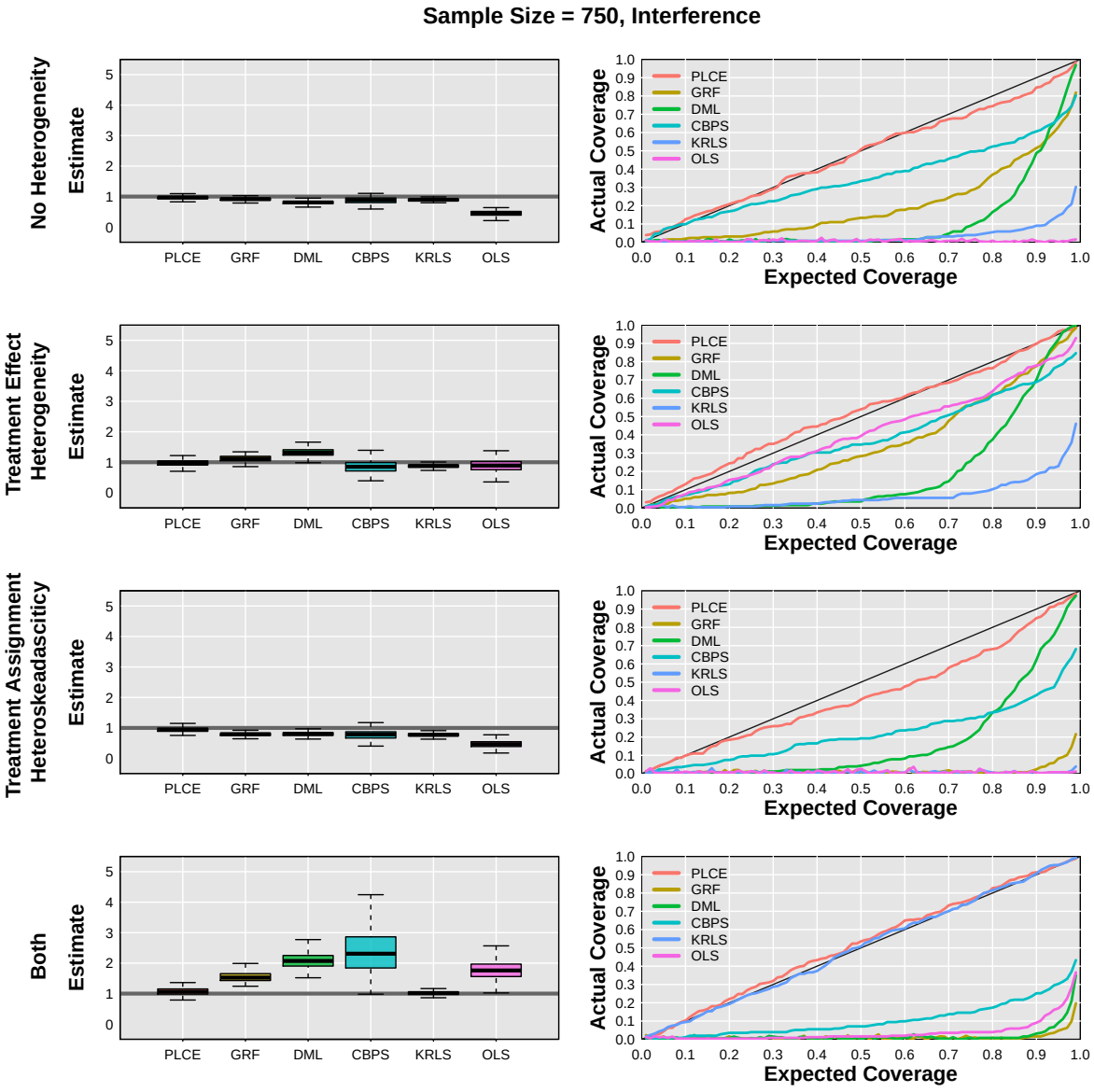

Sample Size = 250, No Interference
No Heterogeneity
Treatment Effect Heterogeneity
Treatment Assignment Heteroskeadasciticy
Both
Estimate
PLCE
GRF
DML
CBPS
KRLS
OLS
Actual Coverage
Expected Coverage


Sample Size = 500, No Interference
No Heterogeneity
Treatment Effect Heterogeneity
Treatment Assignment Heteroskeadasciticy
Both
Estimate
PLCE
GRF
DML
CBPS
KRLS
OLS
Actual Coverage
Expected Coverage


Sample Size = 750, No Interference
Treatment Assignment Heteroskeadasciticy
Treatment Effect Heterogeneity
No Heterogeneity
Both
Estimate
PLCE
GRF
DML
CBPS
KRLS
OLS
Actual Coverage
Expected Coverage


Sample Size = 250, Interference
No Heterogeneity
Treatment Effect Heterogeneity
Treatment Assignment Heteroskeadasciticy
Both
Estimate
PLCE
GRF
DML
CBPS
KRLS
OLS
Actual Coverage
Expected Coverage


Sample Size = 500, Interference
No Heterogeneity
Treatment Effect Heterogeneity
Treatment Assignment Heteroskeadasciticy
Both
Estimate
PLCE
GRF
DML
CBPS
KRLS
OLS
Actual Coverage
Expected Coverage


Sample Size = 750, Interference
No Heterogeneity
Treatment Effect Heterogeneity
Treatment Assignment Heteroskeadasciticy
Both
Estimate
PLCE
GRF
DML
CBPS
KRLS
OLS
Actual Coverage
Expected Coverage

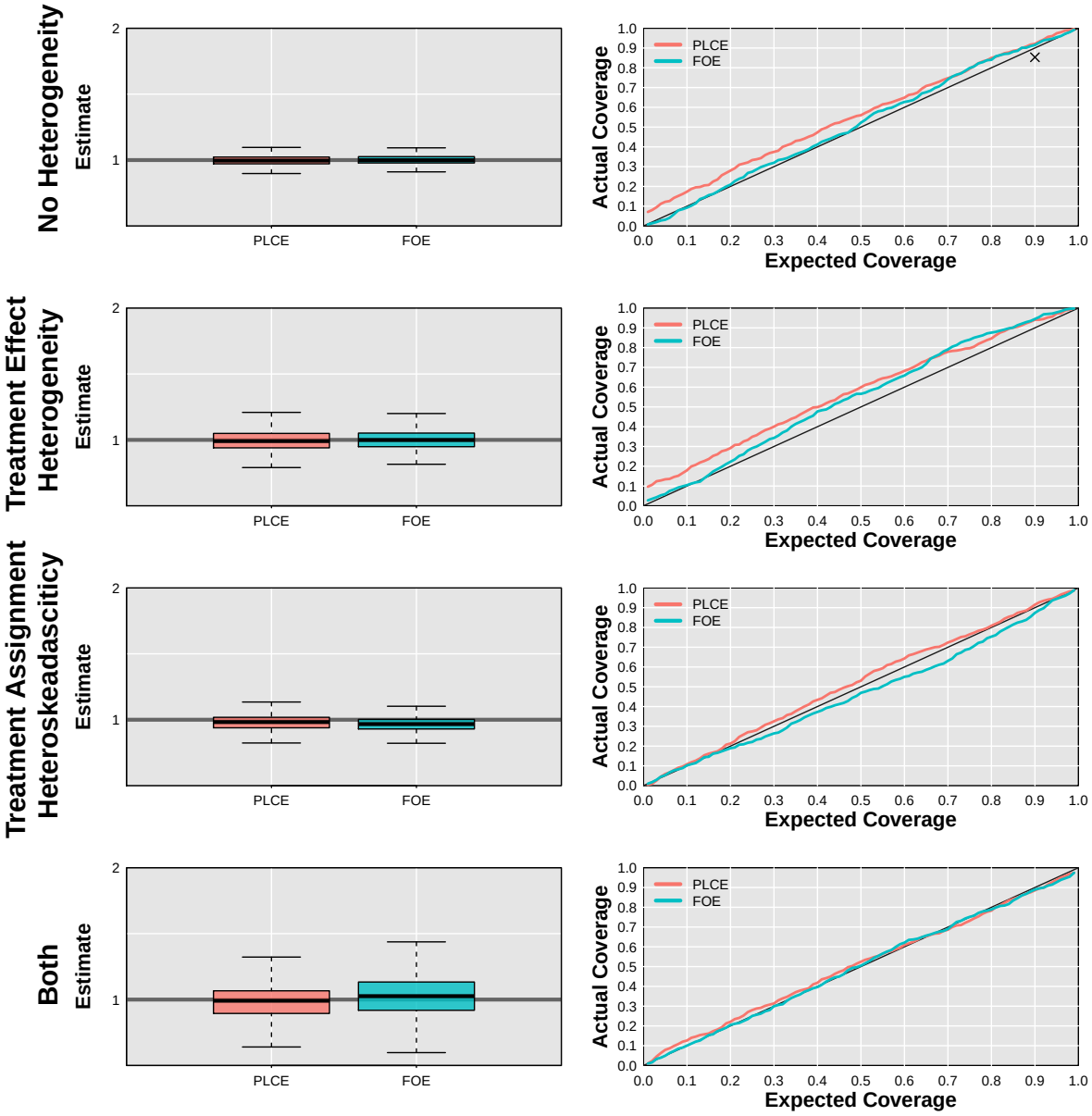


Figure 8: **Comparing estimates with (PLCE) and without (FOE) bootstrapped principal components.** The setup follows that in the body. The two perform comparably in this setting without interference.

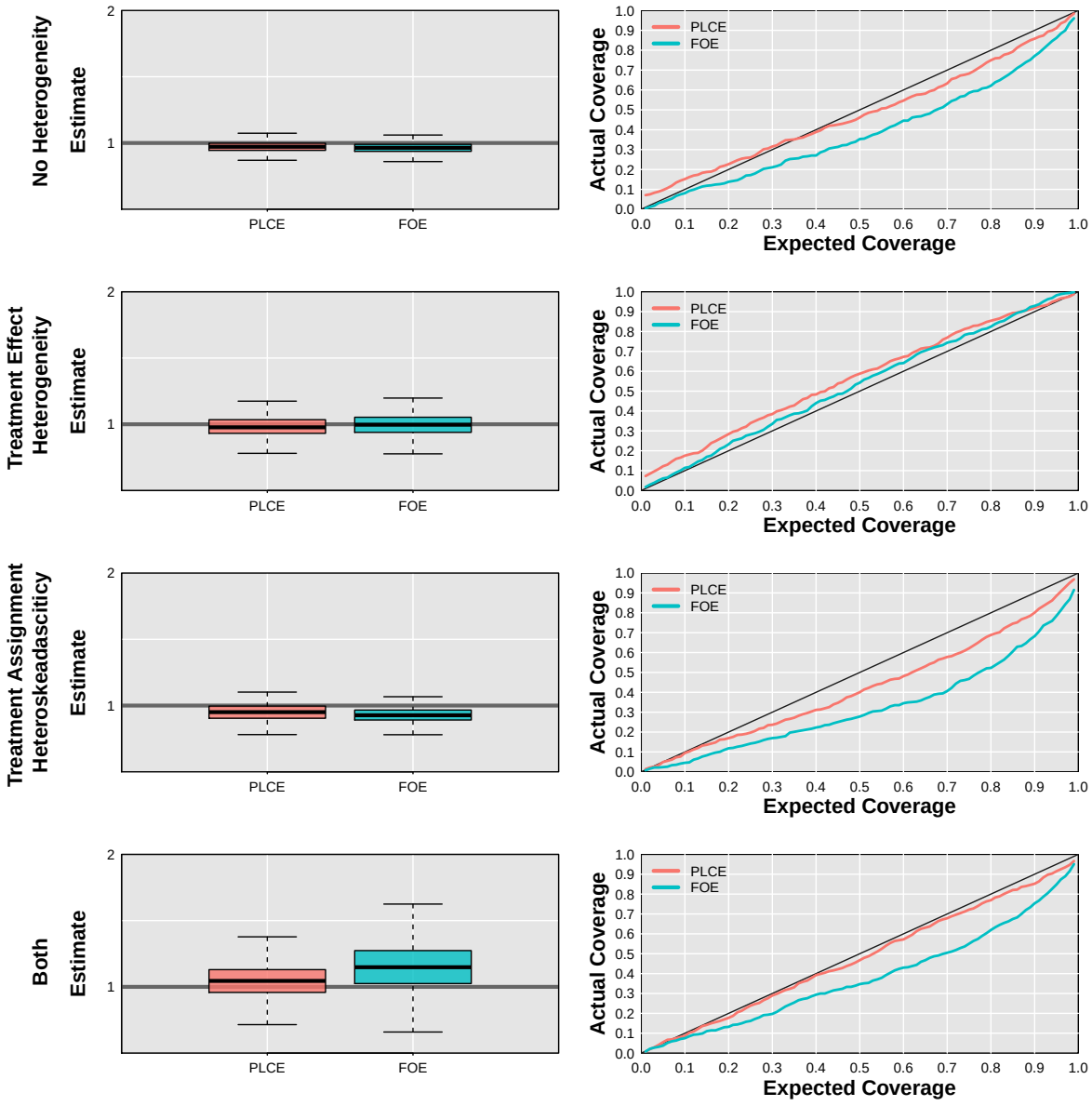


Figure 9: **Comparing estimates with (PLCE) and without (FOE) bootstrapped principal components.** The setup follows that in the body. PLCE outperforms the FOE efficient estimate in point estimation in the final row, and in all save the second row outperforms in terms of coverage.