Relaxing Assumptions, Improving Inference: Integrating Machine Learning and the Linear Regression*

Marc Ratkovic†

November 8, 2021

Abstract

Valid inference in an observational study requires a correct control specification, but a correct specification is never known. We introduce a method that constructs a control vector from the observed data that, when included in a linear regression, adjusts for several forms of bias. These include: nonlinearities and interactions in the background covariates, biases induced by heterogeneous treatment effects, and specific forms of interference. The first is new to our field, that latter two are original contributions. We incorporate random effects, a set of diagnostics, and robust standard errors. With additional assumptions, the estimates allow for causal inference on both binary and continuous treatment variables. In total, the model provides a flexible means to adjust for biases commonly encountered in our data, makes minimal assumptions, returns efficient estimates, and can be implemented using publicly available software.

Key Words: semiparametric inference, partially linear model, causal inference, machine learning

*I would like to thank Soichiro Yamauchi and Max Goplerud for work on developing the software, Dustin Tingley, Max Goplerud, John Londregan, Scott de Marchi, Brandon Stewart, Kevin Munger, Curtis Signorino, Christopher Lucas, Matt Blackwell, Dean Knox, Neal Beck, Cyrus Samii, Matias Cattaneo, 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 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 Not for citation or distribution without permission from the author.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, 2021; 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).

We integrate these two literatures tightly, formally, and practically, with a method and associated software that can improve the reliability of quantitative inference in our field. In doing so, we make two contributions. First, we introduce to our field the concepts and strategies necessary to integrate machine learning with the standard linear regression model (Athey, Tibshirani and Wager, 2019; Chernozhukov et al., 2018). Second, we extend this class of models to address two forms of bias of concern to our field. Specifically, we adjust for a bias induced by unmodeled treatment effect heterogeneity highlighted by Aronow and Samii (2016). In correcting this bias, and under additional assumptions on the data, the proposed method allows for causal effect estimation whether the treatment variable is continuous or binary. We also adjust for biases induced by exogenous interference, which occurs when an observation's outcome or treatment is impacted by the characteristics of other observations (Manski, 1993).

The goal is to allow for valid inference that does not rely on a researcher-selected control specification. Our method, as with several in this literature, uses a machine learning method to adjust for background variables while returning a linear regression coefficient and standard error for the treatment variable of theoretical interest. Following Chernozhukov et al. (2018), we implement a split-sample strategy. We use a machine learning method on one part of the data to construct a control vector that can adjust for nonlinearities and heterogeneities in the background covariates as well as the two biases described above. Then, on the remainder of the data, we include this control vector in a linear regression of the outcome on the treatment. The split-sample strategy serves as a crucial guard against over-fitting. By alternating which subsample is used for constructing control variables from the background covariates and which subsample is used for inference, and then aggregating the separate estimates, the efficiency lost by splitting the sample can be regained. We illustrate this on experimental data, showing that the proposed method generates point estimates and standard errors no different than a full-sample linear regression model.

Our primary audience is the applied researcher currently using a linear regression for inference, but who may be unsettled by the underlying assumptions. We develop the method first as a tool for descriptive inference, generating a slope coefficient and a standard error on a variable of theoretical interest but relying on machine learning to adjust for background covariates. We then discuss the assumptions necessary to interpret the coefficient as a causal estimate. In order to encourage adoption of the proposed method, we will publish software that implements the proposed method and diagnostics described in this manuscript.[1]

[1]The software is publicly available on GitHub but will be published to the Comprehensive R Archive Network upon acceptance.

2 Implications and Applications of the Proposed Method

Quantitative inference in an observational setting requires a properly specified model, meaning the control variables must be observed and entered correctly by the researcher, in order to recover an unbiased estimate of the effect of interest. A correct specification, of course, is never known, raising concerns over "model-dependent" inference (King and Zeng, 2006).

Contrary advice on how to specify controls in a linear regression remains unresolved. This advice ranges from including all the relevant covariates but none of the irrelevant ones (King, Keohane and Verba, 1994, Sec. 5.2–5.3), which states rather than resolves the issue; including at most three variables (Achen, 2002); or at least not all of them (Achen, 2005); or maybe none of them (Lenz and Sahn, 2021). Others have advocated for adopting machine learning methods including neural nets (Beck, King and Zeng, 2000), smoothing splines (Beck and Jackman, 1998), nonparametric regression (Hainmueller and Hazlett, 2013), tree-based methods (Hill and Jones, 2014; Montgomery and Olivella, 2018), or an average of methods (Grimmer, Messing and Westwood, 2017).

None of this advice has found wide use. The advice on the linear regression is largely untenable, given that we normally have a reasonable idea of which background covariates to include, but cannot guarantee that an additive, linear control specification is correct. Machine learning methods offer several important uses, including prediction (Hill and Jones, 2014) and uncovering nonlinearities and heterogeneities (Beck, King and Zeng, 2000; Imai and Ratkovic, 2013). Estimating these sorts of conditional effects and complex models are useful in problems that involve prediction or discovery. For problems of inference, where the researcher desires a confidence interval or p-value on a regression coefficient, these methods

will generally lead to invalid inference (a point we return to below in Section 4 and illustrate in Section 8).

Providing a reliable and flexible means of controlling for background covariates, and clarifying when and whether the estimated effect admits a causal interpretation, is essential to the accumulation of knowledge in our field. We provide such a strategy here.

2.1 Turning Towards Machine Learning

Conducting valid inference with a linear regression coefficient without specifying how the control variables enter the model has been long-studied in the fields of econometrics and statistics (See, e.g., Robinson, 1988; Newey, 1994; Bickel et al., 1998; van der Vaart, 1998, esp. ch. 25.). Recent methods have brought these theoretical results to widespread attention by combining machine learning methods to adjust for background covariates with a linear regression for the variable of interest (Chernozhukov et al., 2018; Athey, Tibshirani and Wager, 2019). We work in this same area, introducing key concepts to our field.

While well-developed in cognate fields, political methodologists have put forth several additional critiques of the linear regression left unaddressed by these aforementioned works. The first critique comes from King (1990) in the then-nascent subfield of political methodology. In a piece both historical and forward-looking, he argued that unmodeled geographic interference was a first-order concern of the field. More generally, political interactions are often such that interference and interaction across observations is the norm. Most quantitative analyses simply ignore interference. Existing methods that do address it rely on strong modeling assumptions requiring, for example, that interference is being driven by known covariates, or that observations only impact similar or nearby observations, or that

the interference is homogenous over the sample. We offer the first method that learns and adjusts for general patterns of exogenous interference (See Section 5.2 for a full discussion).

The second critique arises from a causal critique of observational studies. From this approach, we adopt two concerns. The first is a careful attention to modeling the treatment variable. The second is precision in defining our parameter of interest as an aggregate of observation-level effects. Aronow and Samii (2016) show that a correlation between treatment effect heterogeneity and variance in the treatment assignment will cause the linear regression coefficient to be biased for the causal effect—even if the background covariates are included properly. We offer the first method that explicitly adjusts for this bias. In doing so, we allow for causal effect estimation regardless of whether the treatment variable is continuous or binary (under Assumptions we give in Section 6).

2.2 Practical Considerations of the Proposed Method

The major critique of the linear regression, that its assumptions are untenable, is hardly new. Despite this critique, the linear regression has several positive attributes worthy of preservation. First is its transparency and ease of use. The method, its diagnostics, assumptions, and theoretical properties are well-understood, implemented in commonly available software, and allow for easy inference. Coefficients and standard errors can be used to generate confidence intervals and p-values, and a statistically significant result provides a necessary piece of evidence that a hypothesized effect is present in the data. Importantly, the proposed method maintains these advantages.

We illustrate these points in a simulation study designed to highlight blind spots of existing methods. We then reanalyze experimental data from Mattes and Weeks (2019),

showing that if the linear regression model is in fact correct, our method neither uncovers spurious relationships in the data nor comes at the cost of inflated standard errors. In the second application, we illustrate how to use the method with a continuous treatment. Enos (2016) was forced to dichotomize a continuous treatment, distance from public housing projects, in order to estimate the causal effect of racial threat. To show his results were not model-dependent, he presented results from dichotomizing the variable at ten different distances. We handle this situation more naturally, allowing a single estimate of the effect of distance from housing projects on voter turnout.

3 Anatomy of a Linear Regression

Our central focus is in improving estimation and inference on the *marginal effect*, which is the average effect of a one unit move in a variable of theoretical interest t_i on the predicted value of an outcome y_i, after adjusting for background covariates $\mathbf{x}_i$.[2] We will denote the marginal effect as θ.

Estimation of the marginal effect is generally done with a linear regression,

$$y_i = \theta t_i + \mathbf{x}_i^{>}\gamma + e_i; \quad \mathsf{E}(e_i|\mathbf{x}_i, t_i) = 0. \tag{1}$$

where the marginal effect θ is the *target parameter*, meaning the parameter on which we wish to conduct inference.

We will refer to terms constructed from the background covariates $\mathbf{x}_i$ and entered into the linear regression as *control variables*. For example, if we include a square term of the third

[2]The marginal effect is sometimes referred to as the *average partial effect*.

variable x_{3i}, the background covariate vector is $\mathbf{x}_i$ but the control vector is now $[\mathbf{x}_i^{\top}, x_{3i}^2]^{\top}$. We will reserve γ for slope parameters on control variables.

Inference on a parameter is *valid* if we can use a point estimate and standard error to construct confidence intervals and p-values with the expected theoretical properties. Formally, we say that $\widehat{\theta}$ and its estimated standard deviation $\widehat{\sigma}_{\widehat{\theta}}$ allow for valid inference if, for any θ,

$$\sqrt{n}\left(\frac{\widehat{\theta}-\theta}{\widehat{\sigma}_{\widehat{\theta}}}\right) \quad \mathcal{N}(0,1) \tag{2}$$

The *limiting distribution* of a statistic is the distribution to which its sampling distribution converges (see Wooldridge, 2013, App. C12), so in the previous display the limiting distribution of the z-statistic on the left is a standard normal distribution.

The remaining elements of the model, the control specification ($\mathbf{x}_i^{\top}\gamma$) and the distribution of the error term, are the *nuisance components*, meaning they are not of direct interest but need to be properly adjusted for in order to allow valid inference on θ. The component with the control variables is *specified*, in that its precise functional form is assumed by the researcher.

Heteroskedasticity-consistent (colloquially, "robust") standard errors allow valid inference on θ without requiring the error distribution to be normal, or even equivariant.[3] In this sense, we can leave the error distribution *unspecified*. This insight proves critical to the proposed

[3]For more on the use and misuse of heteroskedasticity-consistent standard errors, Freedman (2006) notes that they are not useful if the model is misspecified, King and Roberts (2015) propose using disagreement between analytic and heteroskedasticity-consistent as a model diagnostic, but note Aronow (2016)'s critique of this approach as over-reliant on modeling assumptions. Our view aligns most closely with Aronow (2016) and derives from the general approach in van der Vaart (1998).

method: valid inference in a statistical model is possible even when components of the model are unspecified.[4]

In statistical parlance, the linear regression model fit with heteroskedasticity-consistent standard errors is an example of a *semiparametric model*, as it combines both a specified component ($\theta t_i + \mathbf{x}_i^{\top}\gamma$) and an unspecified component, the error distribution.

This chain of reasoning then begs the question, can we specify even less? And, given recent advances in our field, can we interpret the estimated effect in a causal fashion? We turn to the first question and then address the second in Section 6.

4 Moving Beyond the Linear Regression

In moving beyond the standard linear regression, we utilize a machine learning method to construct a control vector that can be included in a linear regression of the outcome on the treatment. This vector will allow for valid inference on θ even in the presence of unspecified nonlinearities and interactions in the background covariates. This section relies on the discussion in Chernozhukov et al. (2018) and the textbook treatment of van der Vaart (1998). The presentation remains largely informal, with technical details deferred to Appendix A. We then extend this approach in Section 5.

[4]Unspecified does not mean arbitrary. Heteroskedasticity-consistent standard errors require that the residuals be mean zero given the covariates and treatment and that the estimated residuals follow a central limit theorem; see White (1980, Assumptions 2 and 3). This includes distributions commonly encountered in applied data, while excluding fat-tailed distributions like the Cauchy.

4.1 The Partially Linear Model

Rather than entering the background covariates in an additive, linear fashion, we could have them enter through unspecified functions, f, g:[5]

$$y_i = \theta t_i + f(\mathbf{x}_i) + e_i;\ \mathsf{E}(e_i|t_i, \mathbf{x}_i) = 0 \tag{3}$$

$$t_i = g(\mathbf{x}_i) + v_i;\ \ \mathsf{E}(v_i|t_i, \mathbf{x}_i) = \mathsf{E}(e_i v_i|\mathbf{x}_i) = 0. \tag{4}$$

The resulting model is termed the *partially linear model*, as it is still linear in the treatment variable but the researcher need not assume a particular control specification. This model subsumes the additive, linear specification, but the functions f, g also allow for non-linearities and interactions in the covariates.

4.2 Semiparametric Efficiency

With the linear regression, where the researcher assumes a control specification, the least squares estimates are the uniformly minimum variance unbiased estimate (e.g. Wooldridge, 2013, Sec. 2.3). This efficiency result does not immediately apply to the partially linear model, as we do not assume a particular form for f, g in advance but instead learn it from the data. We must instead rely on a different conceptualization of efficiency: semiparametric efficiency.

An estimate of θ in the partially linear model is semiparametrically efficient if, first, it allows for valid inference on θ and, second, its variance is asymptotically indistinguishable

[5]While our covariates can enter the model nonlinearly, our estimate will still be linear in the sense of being additive in the outcome variable (Wooldridge, 2013, Sec. 2.4.).

from an estimator constructed from the true, but unknown, nuisance functions f, g. Establishing this property proceeds in two broad steps.[6] The first step involves constructing an estimate of θ assuming the true functions f, g were known. This estimate is *infeasible*, since it is constructed from unknown functions. The second step then involves providing assumptions and an estimation strategy such that the feasible estimate constructed from the estimated functions $\widehat{f}, \widehat{g}$ share the same limiting distribution as the infeasible estimate constructed from f, g.

For the first step, consider the *reduced form* model that combines the two models in Equations 3-4,

$$y_i = \theta t_i + [f(\mathbf{x}_i), g(\mathbf{x}_i)]\gamma + e_i. \tag{5}$$

If f, g were known, θ could be estimated efficiently using least squares.[7] The estimate, $\widehat{\theta}$ will be efficient and allow for valid inference on θ.

Following Stein (1956), we would not expect any feasible estimator to outperform this infeasible estimator, so its limiting distribution is termed the *semiparametric efficiency bound*. With this bound established, we turn to generating a feasible estimate that shares a limiting distribution with this infeasible estimate.

[6] We include a complete, self-contained technical discussion in Appendix A.

[7] There are often multiple, and asymptotically equivalent, ways to estimate θ (see, e.g. Robins et al., 2007). Chernozhukov et al. (2018) propose regressing $y_i - f(\mathbf{x}_i)$ on $t_i - g(\mathbf{x}_i)$, whereas we instead estimate θ from the reduced form model. The two are asymptotically equivalent, but we favor the reduced form approach as it allows us to incorporate intuitions and diagnostics from the linear regression.

4.3 Double Machine Learning for Semiparametrically Efficient Estimation

Estimated functions $\widehat{f}, \widehat{g}$, presumably estimated using a machine learning method, can be used to construct control variables and entered into a linear regression as

$$y_i = \theta t_i + [\widehat{f}(\mathbf{x}_i), \widehat{g}(\mathbf{x}_i)]\gamma + e_i. \tag{6}$$

Semiparametric efficiency can be established by characterizing and eliminating the gap between the infeasible model in Equation 5 and the feasible model in Equation 6. We do so by introducing *approximation error* terms

$$\Delta_{\widehat{f},i} = \widehat{f}(\mathbf{x}_i) - f(\mathbf{x}_i);\ \Delta_{\widehat{g},i} = \widehat{g}(\mathbf{x}_i) - g(\mathbf{x}_i), \tag{7}$$

that capture the distance between the true functions f, g and their estimates $\widehat{f}, \widehat{g}$ at each $\mathbf{x}_i$.

Given these approximation errors, we can rewrite Equation 6 in the familiar form of a *measurement error* problem (Wooldridge, 2013, Ch. 9.4), where we consider the estimated functions $\widehat{f}, \widehat{g}$ as "mismeasuring" the true functions f, g:

$$y_i = \theta t_i + [f(\mathbf{x}_i), g(\mathbf{x}_i)]\gamma_1 + [\widehat{f}(\mathbf{x}_i) - f(\mathbf{x}_i), \widehat{g}(\mathbf{x}_i) - g(\mathbf{x}_i)]\gamma_2 + e_i \tag{8}$$

$$y_i = \theta t_i + [f(\mathbf{x}_i), g(\mathbf{x}_i)]\gamma_1 + [\Delta_{\widehat{f},i}, \Delta_{\widehat{g},i}]\gamma_2 + e_i \tag{9}$$

Establishing semiparametric efficiency of our feasible estimator, then, consists of establishing a set of assumptions and an estimation strategy that leaves the approximation error terms

asymptotically negligible.

There are two pathways by which the approximation error terms may bias our estimate. The first is if the approximation errors do not tend to zero. Eliminating this bias requires that the approximation errors vanish asymptotically, specifically at an $n^{1/4}$ rate.[8] Though seemingly technical, this assumption is actually liberating. Many modern machine learning methods utilized by our field provably achieve this rate (Chernozhukov et al., 2018), including random forests (Hill and Jones, 2014; Montgomery and Olivella, 2018), neural networks (Beck, King and Zeng, 2000), and sparse regression models (Ratkovic and Tingley, 2017). This assumption allows the researcher to condense all the background covariates into a control vector constructed from $\widehat{f}, \widehat{g}$, where these functions are estimated via a flexible machine learning method. Any nonlinearities and interactions in the background covariates are then learned from the data rather than specified by the researcher.

Eliminating the second pathway requires that any correlation between the approximation errors $\Delta_{\widehat{f},i}, \Delta_{\widehat{g},i}$ and the error terms e_i, v_i tend to zero.[9] Doing so requires addressing a subtle aspect of the approximation error: the estimates $\widehat{f}, \widehat{g}$ are themselves functions of e_i, v_i, as they are estimates constructed from a single observed sample. Even under the previous

[8] Formally, this requires

$$n^{1/4}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\Delta_{\widehat{f},i}^2} \xrightarrow{\text{u}} 0; \quad n^{1/4}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\Delta_{\widehat{g},i}^2} \xrightarrow{\text{u}} 0, \tag{10}$$

where $\xrightarrow{\text{u}}$ denotes converges uniformly, which is the notion of convergence needed for complex, nonparametric functions. We provide full details in Appendix A.

[9] Again, details appear in Appendix A, but valid inference will require that the terms

$$\sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\Delta_{\widehat{f},i}e_i\right\} \xrightarrow{\text{u}} 0; \quad \sqrt{n}\left\{\frac{1}{n}\sum_{i=1}^{n}\Delta_{\widehat{f},i}v_i\right\} \xrightarrow{\text{u}} 0 \tag{11}$$

as well as corresponding terms with $\Delta_{\widehat{g},i}$.

Algorithm 1: The Double Machine Learning Algorithm

Data: Outcome, treatment, and covariates $\{y_i, t_i, \mathbf{x}_i\}_{i=1}^{n}$
Result:
for r *in 1 to* R **do**
  Split the sample in half, generating $\mathcal{S}_1, \mathcal{S}_2$.
  for j *in 1 to 2* **do**
    Estimate f, g in subsample $\mathcal{S}_j$ using a machine learning method.
    Regress $y_i - \widehat{f}(\mathbf{x}_i)$ on $t_i - \widehat{g}(\mathbf{x}_i)$ using data from the other subsample.
  Aggregate the point estimate and standard error over splits.
Return the median point estimate and standard error over repeated cross-fits.

assumption on the convergence rate of $\widehat{f}, \widehat{g}$, this bias term may persist.

The most elegant, and direct, way to eliminate this bias is to employ a *split-sample* strategy.[10] First, the data is split in half into subsamples denoted $\mathcal{S}_1$ and $\mathcal{S}_2$ of size n_1 and n_2 such that $n_1 + n_2 = n$. Data from $\mathcal{S}_1$ is used to learn $\widehat{f}, \widehat{g}$ and data from $\mathcal{S}_2$ to conduct inference on θ. Since the nuisance functions are learned on data wholly separate from that on which inference is conducted, this bias term tends to zero.

Sample-splitting raises real efficiency concerns, as it only uses half the data for inference and thereby inflates standard errors by $\sqrt{2} \approx 1.4$. In order to restore efficiency, Chernozhukov et al. (2018) propose a *cross- tting* strategy, whereby the roles of the subsamples $\mathcal{S}_1, \mathcal{S}_2$ are swapped and the estimates combined. *Repeated cross- tting* consists of aggregating estimates over multiple cross-fits, allowing all the data to be used in estimation and returning results that are not sensitive to how the data is split. The resultant estimate is itself semiparametrically efficient. A description of the algorithm appears in Table 1.

[10] See Fong and Tyler (2021); Ratkovic (2021) for contemporary works in political science exploring a split-sample strategy.

4.4 Constructing Covariates and Second-Order Semiparametric Efficiency

Our first advance over the Double Machine Learning strategy is constructing a set of covariates that will further refine our estimates of the nuisance functions. Doing so gives more assurance that we will, in fact, adjust for the true nuisance functions f, g.

In order to do so, we make the approximation

$$\widehat{f}(\mathbf{x}_i) \approx f(\mathbf{x}_i) + U_{f,i}^{>}\gamma_f; \;\; \widehat{g}(\mathbf{x}_i) \approx g(\mathbf{x}_i) + U_{g,i}^{>}\gamma_g \tag{12}$$

or, equivalently,

$$\Delta_{\widehat{f},i} \approx U_{f,i}^{>}\gamma_f; \;\; \Delta_{\widehat{g},i} \approx U_{g,i}^{>}\gamma_g \tag{13}$$

for some vector of parameters γ_f, γ_g.

These new vectors of control variables $U_{f,i}, U_{g,i}$ capture the fluctuations of the estimated functions $\widehat{f}, \widehat{g}$ around the true values, f, g. The expected fluctuation of an estimate around its true value is measured by its standard error (Wooldridge, 2013, Sec. 2.5), so we construct these control variables from the variance matrix of the estimates themselves. Denoting as $\widehat{f}(\mathbf{X}), \widehat{g}(\mathbf{X})$ the vectors of estimated nuisance component $\widehat{f}, \widehat{g}$, we first construct the variance matrices $\widehat{\text{Var}}(\widehat{f}(\mathbf{X}))$ and $\widehat{\text{Var}}(\widehat{g}(\mathbf{X}))$. In order to summarize these matrices, we construct the control vectors $\widehat{U}_{\widehat{f},i}, \widehat{U}_{\widehat{g},i}$ from principal components of the square root of the variance matrices.[11]

[11] For technical and implementation details, including how this approach integrates with the split-sample

Including these constructed covariates as control variables offers advantages both practical and theoretical. As a practical matter, augmenting the control set $\widehat{f}, \widehat{g}$ with the constructed control vectors $\widehat{U}_{\widehat{f},i}, \widehat{U}_{\widehat{f},i}$ helps guard against misspecification or chance error in the estimates $\widehat{f}, \widehat{g}$, adding an extra layer of accuracy to our estimate and making it more likely that we can properly adjust for f, g.

As a theoretical matter, the method is an example of a *second-order semiparametrically e cient* estimator. Double machine learning is *rst-order semiparametrically e cient* as it only adjusts for the conditional means $\widehat{f}, \widehat{g}$. By including the estimated controls $\widehat{f}, \widehat{g}$ but also principal components $U_{\widehat{f},i}, U_{\widehat{g},i}$ constructed from the variance (the second moment, see Wooldridge (2013) App. D.7), we gain an extra order of efficiency and return a *second-order semiparametrically e cient* estimate.[12] The theoretical gain is that second-order efficiency requires only an $n^{1/8}$ order of convergence on the nuisance terms rather than $n^{1/4}$. While seemingly technical, this simply means that we can conduct valid inference on θ while demanding less accuracy from the machine learning method estimating the nuisance terms. At the most intuitive level, including these additional control vectors will make it more likely that we capture the nuisance terms.

5 Improving on the Partially Linear Model

Double Machine Learning addresses a particular issue, namely learning how the background covariates enter the model. Several issues of interest to our field remain unaddressed. We turn to these next, which comprise our central contributions.

strategy, see Appendices A and D-E.

[12] For more on second- and higher-order efficiency, see (van der Vaart, 2014; Li et al., 2011; Robins et al., 2008; Dalalyan, Golubev and Tsybakov, 2006, esp. Eq. 4.). While this theoretical literature is developed, we are the first to incorporate these ideas into software and allow their use in an applied setting.

5.1 Adjusting for Treatment Effect Heterogeneity Bias

Aronow and Samii (2016) show that the linear regression estimate of a coefficient on the treatment variable is biased for the marginal effect. The bias emerges through insufficient care in modeling the treatment variable and heterogeneity in the treatment effect, and the authors highlight this bias as a key difference between a linear regression estimate and a causal estimate.

To see this bias, denote as θ_i the effect of the treatment on the outcome for observation i. We can then define the marginal effect as $\theta = \mathsf{E}(\theta_i)$. To simplify matters, presume the true functions f, g are known, allowing us to isolate as-if random fluctuations of e_i, v_i. Incorporating the effect heterogeneity into the partially linear model gives

$$y_i = t_i\theta_i + [f(\mathbf{x}_i), g(\mathbf{x}_i)]\gamma + e_i \tag{14}$$

$$= t_i\theta + t_i(\theta_i - \theta) + [f(\mathbf{x}_i), g(\mathbf{x}_i)]\gamma + e_i. \tag{15}$$

The unmodeled effect heterogeneity introduces an omitted variable, $t_i(\theta_i - \theta)$, which gives a bias of[13]

$$\mathsf{E}(\widehat{\theta} - \theta) = \frac{\mathsf{E}\left\{\mathrm{Cov}(t_i, t_i(\theta_i - \theta)|\mathbf{x}_i)\right\}}{\mathsf{E}\left\{\mathrm{Var}(t_i|\mathbf{x}_i)\right\}} \tag{16}$$

$$= \frac{\mathsf{E}(v_i^2(\theta_i - \theta))}{\mathsf{E}(v_i^2)} \tag{17}$$

which we will refer to as *treatment effect heterogeneity bias*.

Inspection reveals that either one of two conditions are sufficient to guarantee that the

[13] See Wooldridge (2013) Eq. 5.4.

treatment heterogeneity variance bias is zero. The first occurs when there is no treatment effect heterogeneity ($\theta_i = \theta$ for all observations), the second when there is no treatment assignment heteroskedasticity ($\mathsf{E}(v_i^2)$ 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 marginal effect θ and the parameter estimated by the partially linear model. By doing so, any fluctuation of the partial effect around the marginal effect $(\theta_i - \theta)$ is ignorable and will not bias inference on θ.

We will address this form of bias through modeling the random component in the treatment assignment. Just as we model the conditional means through unspecified nuisance functions, we will introduce an additional function that will capture heteroskedasticity in the treatment variable.

5.2 Interference and Group-Level Effects

The proposed method also adjusts for group-level effects and interference. For the first, researchers commonly encounter data with some known grouping, say at the state, province, or country-level. To accommodate these studies, we incorporate random effect estimation in our model. The proposed method also adjusts for interference, where observations may be impacted by observations that are similar in some regards ("homophily") or different in some regards ("heterophily"). For example, observations that are geographically proximal may behave similarly (Ward and O'Loughlin, 2002; Ripley, 1988), actors may be connected via a social network (Sobel, 2006; Aronow and Samii, 2017), 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 characteristics of

other observations.

Existing approaches require *a priori* knowledge over what variables drive the interference as well as how the interference affects both the treatment variable and the outcome. Instead, we use a machine learning method to learn the type of interference in the data: what variables are driving interference, and in what manner.

The problem involves two components: a measure of proximity and an interferent. The first addresses which variables are driving how close two observations are.[14] In the spatial setting, for example, these may be latitude and longitude. More generally, observations closer in age may behave similarly (homophily) or observations with different education levels may behave similarly (heterophily). The strength of the interference is governed by a bandwidth parameter, which characterizes the radius of impact of proximal observations on a given observation. For example, with a larger bandwidth, interference may be measurable between people with a ten-year age range, but for a narrower bandwidth, it may only be discernible within a three-year range.

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 level (the interferent) of other observations with a similar age (the proximity measure). Our method implements a machine learning method that uses the background covariates to learn proximity variables, interferents, and estimate bandwidth parameters.

The proposed method learns and adjusts for two types of interference: that driven entirely by covariates, or the effect of one observations' treatment on other observations' outcomes. For example, if the interference among observations is driven entirely by exogenous covari-

[14] Manski (2013, 1993) refers to this as the *reference group*.

ates, such as age or geography, the method allows for valid inference on θ. Similarly, if there are spillovers such that one observation's treatment affects another's outcome, as with, say, vaccination, we can adjust for this form of spillover (e.g. Hudgens and Halloran, 2008).[15]

The proposed method does not adjust for what Manski (1993) terms *endogenous interference*, which occurs when an observation's outcome is driven by the behavior of some group that includes itself. This form of interference places the outcome variable on both the lefthand- and righthand-side of our model, inducing a simultaneity bias (see Wooldridge, 2013, ch.16). Similarly, we cannot adjust for the simultaneity bias in the treatment variable. The third form of interference we cannot account for is when an observation's treatment is affected by its own or others' outcomes, a form of post-treatment bias (Acharya, Blackwell and Sen, 2016).

6 Relation to Causal Effect Estimation

We have developed the method so far as a tool for descriptive inference, estimating a slope term on a treatment variable of interest. If the data and design allow, the researcher may be interested in a causal interpretation of her estimate.

Generating a valid causal effect estimate of the marginal effect requires two steps beyond the descriptive analysis. First, the estimate must be consistent for a parameter constructed from an average of observation-level causal effects. By correcting for the treatment effect heterogeneity bias described in Section 5.1, we accomplish this. In doing so, we are able to estimate causal effects regardless of whether the treatment variable is binary or continuous. Second, we must specify the conditions on the data that will allow us to recover this estimate.

[15] In this situation, we are estimating what is termed the "direct effect" of the treatment, since we are adjusting for indirect effects that come from other observations.

We discuss the assumptions here, with a formal presentation in Appendix C.

First, a *stable value assumption* requires a single version of each level of the treatment so that all potential outcomes are well-defined. Most existing studies include a non-interference assumption in this assumption, which we are able to avoid through modeling the interference in the data. Second, a *positivity assumption* requires that the treatment assignment be non-deterministic for every observation. These first two assumptions are standard. The first is a matter of design and conceptual clarity,[16] while our software implements a diagnostic for the second; see Appendix B.

The third assumption, the *ignorability assumption*, is where we diverge from most existing studies. As with existing studies, the assumption requires sufficient observation-level covariates to break confounding between treatment and outcome. An omitted confounder is always a threat to causal inference in an observational study.

Unlike most existing studies, we assume neither an absence of interference or that the structure of the interference is known. Causal estimation with the proposed method still requires that the observed covariates are sufficient to break confounding that may emerge through interference, in both the treatment and the outcome. When modeling the outcome, we can recover a causal estimate when other observations' treatments affect an observation's outcome.

As discussed in Section 5.2, violations may occur when after conditioning, an observation's treatment is impacted by other observations' treatment levels or an observation's outcome by other observations' outcome level. This simultaneity bias will invalidate causal inference. Another violation may occur if there is a direct effect of the outcome on the

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

treatment. While our software implements a diagnostic to assess the sensitivity of our results to these assumptions (see Appendix B), their plausibility must be established through substantive knowledge by the researcher.

The assumptions clarify the nature of our estimand. By assuming the covariates adjust for indirect effects that may be coming from other observations, our estimate is an average direct effect of the treatment on the outcome. Secondly, since we are adjusting for other observations' treatments at their observed level, we are estimating an average controlled direct effect. The causal effect we estimate is then the average effect of a one-unit move of a treatment on the outcome, given all observations' covariates and fixing their treatments at the realized value.

7 The Proposed Model

The proposed model expands the partially linear model to include exogenous interference, heteroskedasticity in the treatment assignment mechanism, and random effects. We refer to it as the partially linear causal effect (*PLCE*) model since, under the assumptions in Section 6, it returns a causal estimate of the treatment on the outcome.

The treatment and outcome models for the proposed method are

$$y_i = \theta t_i + f(\mathbf{x}_i) + \phi_y(\mathbf{x}_i, \mathbf{X}_{\ i}, \mathbf{t}_{\ i}, \mathbf{h}_y) + a_{j[i]} + e_i \tag{18}$$

$$t_i = g_1(\mathbf{x}_i) + g_2(\mathbf{x}_i, \mathbf{X}_{\ i})\widetilde{v}_i + \phi_t(\mathbf{x}_i, \mathbf{X}_{\ i}, \mathbf{t}_{\ i}, \mathbf{h}_t) + b_{j[i]} + v_i \tag{19}$$

where we require the following conditions on the error terms

$$a_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_a^2); \quad b_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_b^2) \tag{20}$$

$$\mathsf{E}(e_i|\mathbf{x}_i, \mathbf{X}_{\ i}, t_i, \mathbf{t}_{\ i}) = \mathsf{E}(v_i|\mathbf{x}_i, \mathbf{X}_{\ i}) = \mathsf{E}(\widetilde{v}_i|\mathbf{x}_i, \mathbf{X}_{\ i}) = 0 \tag{21}$$

$$\mathsf{E}(e_i v_i|\mathbf{x}_i, \mathbf{X}_{\ i}) = \mathsf{E}(e_i \widetilde{v}_i|\mathbf{x}_i, \mathbf{X}_{\ i}) = \mathsf{E}(v_i \widetilde{v}_i|\mathbf{x}_i, \mathbf{X}_{\ i}) = 0 \tag{22}$$

$$\mathsf{E}(e_i^2|\mathbf{x}_i, \mathbf{X}_{\ i}, t_i, \mathbf{t}_{\ i}) > 0; \quad \mathsf{E}(v_i^2|\mathbf{x}_i, \mathbf{X}_{\ i}) > 0 \tag{23}$$

Moving through the components of our model, θ, is the parameter of interest. The first set of nuisance functions $(f(\mathbf{x}_i), g_1(\mathbf{x}_i))$ are inherited from the partially linear model. The pure error terms e_i, v_i also follow directly from the partially linear model.

The interference components are denoted $\phi_y(\mathbf{x}_i, \mathbf{X}_{\ i}, \mathbf{t}_{\ i}, \mathbf{h}_y), \phi_t(\mathbf{x}_i, \mathbf{X}_{\ i}, \mathbf{h}_t)$. The vector of bandwidth parameters are denoted $\mathbf{h}_t, \mathbf{h}_y$, which will govern the radius for which one observation impacts others. Note that either the treatment or the covariates from one observation can affect another's outcome, but the only interference allowed in the treatment model comes from the background covariates (see Section 6).

The treatment variable has two error components. The term v_i is "pure noise," in that its variance is not a function of covariates. The term $\widetilde{v}_i$ is noise associated with heteroskedasticity in the treatment variable. The component $g_2(\mathbf{x}_i, \mathbf{X}_{\ i})\widetilde{v}_i$ will adjust for treatment effect heterogeneity bias. The term $\widetilde{v}_i$ is the error component in the treatment associated with the function $g_2(\mathbf{x}_i, \mathbf{X}_{\ i})$ which drives any systematic heteroskedasticity in the treatment variable.

The conditions on the error terms are also standard. The terms $a_{j[i]}, b_{j[i]}$ are random effects with observation i in group $j[i]$ (Gelman and Hill, 2007), and Condition 20 assumes the random effects are realizations from a common normal distribution. Equations 21 as-

sume no omitted variables that may bias our inference on θ. This conditional independence assumption is standard in the semiparametric literature (see, e.g. Chernozhukov et al., 2018; Donald and Newey, 1994; Robinson, 1988). Equations 22 ensures the error terms are all uncorrelated. Any correlation between e_i and either v_i or $\widetilde{v}_i$ would induce simultaneity bias. The absence of correlation between v_i and $\widetilde{v}_i$ allows us to fully isolate the heteroskedasticity in the treatment variable in order to eliminate treatment effect heterogeneity bias. The final assumptions in Expression 22 require that there be a random component in the outcome variable and the treatment for each observation. The latter is the positivity assumption in Section 6.

From the models in Equations 18-19, we can then construct the infeasible reduced form equation

$$\begin{aligned} y_i =& \theta t_i + \\ & [f(\mathbf{x}_i), \phi_y(\mathbf{x}_i, \mathbf{X}_{\ i}, \mathbf{t}_{\ i}, \mathbf{h}_y), g_1(\mathbf{x}_i), g_2(\mathbf{x}_i, \mathbf{X}_{\ i})\widetilde{v}_i, \phi_t(\mathbf{x}_i, \mathbf{X}_{\ i}, \mathbf{h}_t)]^{\intercal}\gamma_{PLCE} \\ & + c_{j[i]} + e_i \end{aligned} \tag{24}$$

where the random effect combines those from the treatment and outcome model, $c_{j[i]} = a_{j[i]} + b_{j[i]}$. We next extend the logic from Section 4.2, in order to construct a semiparametrically efficient estimate of θ.

7.1 Formal Assumptions

The following assumption will allow a semiparametrically efficient estimate of the marginal effect in the PLCE model.

Assumption 1 (PLCE Assumptions)

1. *Population Model. The population model is given in Equations 18 - 19, and all random components satisfy the conditions in Equation 20-23.*

2. *E cient Infeasible Estimate. Were all nuisance functions known, the least squares estimate from the reduced form model in Equation 24 would be e cient and allow for valid inference on* θ.

3. *Representation. There exists a nite dimensional control vector* $U_{u,i}$ *that allows for valid and e cient inference on* θ.

4. *Approximation Error. All nuisance components are estimated such that the approximation errors converge uniformly at the rate* $n^{1/8}$.

5. *Estimation Strategy. The split-sample strategy of Figure 1 is employed.*

The first assumption requires that the structure of the model and conditions on the error terms are correct. The second assumption serves two purposes. First, requires that the standard least squares assumptions (see, e.g., Wooldridge, 2013, Assumptions MLR 1-5 in ch. 3.) hold for the infeasible, reduced form model. This requires no unobserved confounders or unmodeled interference.[17] Second, it establishes the semiparametric efficiency bound, which is the limiting distribution of the infeasible estimate $\widehat{\theta}$ from this model.

The third assumption structures the control vector, $U_{u,i}$. This vector contains all estimates of each the nuisance functions in Equation 24, producing a first-order semipara-

[17] Crucial to our split-sample strategy is that the observations are conditionally independent, meaning we can recover a valid marginal effect estimate on any randomly-generated split. We assume that this aspect is not broken by unmodeled interference. Intuitively, we will condense all of the interference to the functions ϕ_y, ϕ_t such that, after conditioning on these, observations are independent.

metrically efficient estimate. This assumption guarantees that by including the second-order covariates as described in Section 4.4, we will still be able to estimate θ using least squares.[18]

The fourth and fifth assumptions are analogous to those implemented in the Double Machine Learning strategy. Including the constructed covariates allows us to relax the accuracy of the approximation error from $n^{1/4}$ to $n^{1/8}$, while the importance of the repeated cross-fitting strategy in eliminating biases between approximation errors and the error terms e_i, v_i is discussed in Section 4.2.

7.1.1 Scope Conditions and Discussion of Assumptions

The assumption that $U_{i,u}$ is finite dimensional is the primary constraint on our model. Effectively, this assumption allows us to condense all the nuisance functions into a single control vector, allowing us to run a linear regression in subsample $\mathcal{S}_2$. Our assumption compares favorably to many in the literature. Belloni, Chernozhukov and Hansen (2014) make a "sparsity assumption," that the conditional mean can be well-approximated by a subset of functions of the covariates.[19] We relax this assumption, since our estimated principal components may be an average of a large number of covariates and functions of covariates.

Our use of principal components is a form of "sufficient dimension reduction" (Li, 2018; Hsing and Ren, 2009), where we assume that the covariates and nonlinear functions of the covariates can be reduced to a set that fully captures any systematic variance in the outcome.[20] We are able to sidestep the analytic issues in characterizing the covariance function of the observations analytically (see, e.g., Wahba, 1990) by instead taking principal

[18] With added assumptions, the dimensionality of these covariates could grow on the order of $\sqrt{n}$, though we save this for further work (see, e.g. Chernozhukov et al., 2018; Cattaneo, Jansson and Newey, 2018).

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

[20] See Appendix D.1 for a discussion of how we model nonlinearities and interactions in our model.

components of the variance matrix. Our sample-splitting strategy is also original.

Restricting $U_{i,u}$ to be finite dimensional means that we cannot accommodate models where the dimensionality of the variance grows in sample size. To give two examples, in the panel setting, we can handle random effects for each unit but not arbitrary nonparametric functions per unit. Second, we can account for interference, but only if we do not allow the dimensionality of the interference to grow in the sample size. This assumption is in line with those made by other works on interference (Savje, Aronow and Hudgens, 2021).

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; van der Laan and Rose, 2011) or where the treatment density is assumed (Fong, Hazlett and Imai, 2018). Nonparametric estimates of inverse density weights are inherently unstable, so we do not pursue this approach but see Kennedy et al. (2017). Rather, we mean-adjust for confounding by constructing a set of control variables. We show below, through simulation and empirical examples, that the method generates reliable estimates.

7.2 Estimation Strategy: Three-Fold Split Sample

Heuristically, two sets of nuisance components enter our model. The first are used to construct nuisance functions: the treatment error $\widetilde{v}_i$ that interacts with g_2 and the bandwidth parameters $\mathbf{h}_y, \mathbf{h}_t$ that parameterize the interference terms. We denote these parameters as the set $u = \{\{\widetilde{v}_i\}_{i=1}^n, \mathbf{h}_y, \mathbf{h}_t\}$. The second set are those that, given the first set, enter additively into the model. These consist of the functions f, g_1, but also the functions g_2, ϕ_y, ϕ_t. If u were known, estimating these terms would collapse into the Double Machine Learning

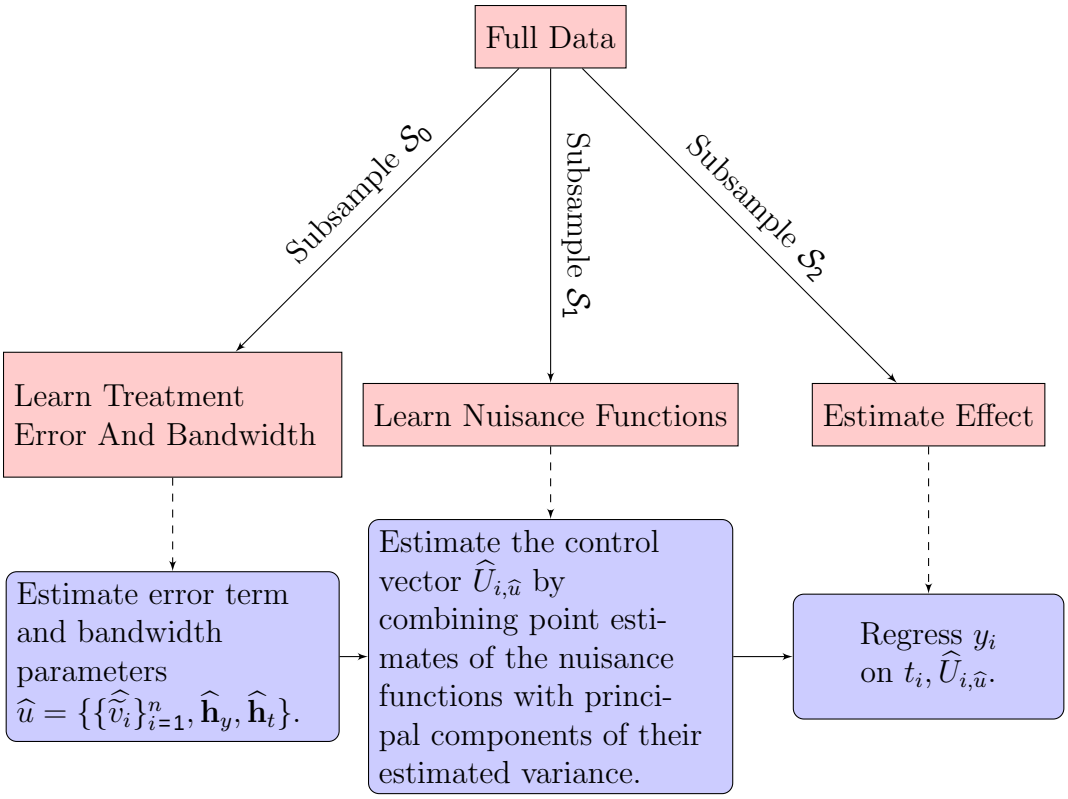


Figure 1: **Our estimation strategy.** To estimate the components in our model, we split our data into thirds. In the first subsample, $\mathcal{S}_0$, we estimate the interference bandwidths and treatment residuals. In the second, $\mathcal{S}_1$, given the estimates from the first, we construct the covariates that will adjust for all the biases in the model. In the third, $\mathcal{S}_2$, we run a linear regression using our constructed covariates.

strategy outlined in Section 4.3. Since u is not known, it must also be estimated in a separate step, necessitating a third split of our data.

We defer precise implementation details to Appendices D and E, but more important than our particular implementation choices is the general strategy for estimating the nuisance components such that the approximation errors do not bias our inference on θ. We outline this strategy here.

We begin by splitting the data into three subsamples, $\mathcal{S}_0$, $\mathcal{S}_1$, and $\mathcal{S}_2$, each containing a third of the data. Then, in subsample $\mathcal{S}_0$, we estimate all of the components in the models

in Equations 18 and 19, but only retain those marked below,

$$y_i = \theta t_i + f(\mathbf{x}_i) + \phi_y(\mathbf{x}_i, \mathbf{X}_{\ i}, \mathbf{t}_{\ i}, \underbrace{\mathbf{h}_y}_{S_0}) + a_{j[i]} + e_i \tag{25}$$

$$t_i = g_1(\mathbf{x}_i) + g_2(\mathbf{x}_i, \mathbf{X}_{\ i}) \underbrace{\widetilde{v}_i}_{S_0} + \phi_t(\mathbf{x}_i, \mathbf{X}_{\ i}, \mathbf{t}_{\ i}, \underbrace{\mathbf{h}_t}_{S_0}) + b_{j[i]} + v_i \tag{26}$$

We then take these retained components, $\widehat{\mathbf{h}}_y, \widehat{\mathbf{h}}_t$ and a model for estimating the error terms $\left\{\widehat{\widetilde{v}}_i\right\}_{i=1}^n$, and carry them to subsample $\mathcal{S}_1$.

Using the data in subsample $\mathcal{S}_1$, we can evaluate the bandwidth parameters and error term using the values from the previous subsample, and given these, estimate the terms marked below,

$$y_i = \theta t_i + \underbrace{f}_{S_1}(\mathbf{x}_i) + \underbrace{\phi_y}_{S_1}(\mathbf{x}_i, \mathbf{X}_{\ i}, \mathbf{t}_{\ i}, \widehat{\mathbf{h}}_y) + \underbrace{a_{j[i]}}_{S_1} + e_i \tag{27}$$

$$t_i = \underbrace{g_1}_{S_1}(\mathbf{x}_i) + \underbrace{g_2}_{S_1}(\mathbf{x}_i, \mathbf{X}_{\ i})\widehat{\widetilde{v}}_i + \phi_t(\mathbf{x}_i, \mathbf{X}_{\ i}, \mathbf{t}_{\ i}, \widehat{h}) + \underbrace{b_{j[i]}}_{S_1} + v_i. \tag{28}$$

Having now estimated all nuisance terms, including the random effects. we are ready to conduct valid inference. We construct our feasible control variable $\widehat{U}_{\widehat{u},i}$. This variable consists of two sets of covariates. The first is the point estimates of all of the nuisance components estimated from $\mathcal{S}_0$ and $\mathcal{S}_1$ but evaluated on $\mathcal{S}_2$. It also includes the second-order terms, also estimated on subsample $\mathcal{S}_1$ but evaluated on subsample $\mathcal{S}_2$. This control vector is then entered into the reduced form model

$$y_i = \theta t_i + \widehat{U}_{\widehat{u},i}^{>}\gamma + e_i. \tag{29}$$

which generates an estimate $\widehat{\theta}$ and its standard error.

We then implement a cross-fitting strategy, where we swap the roles of each subsample in our estimate, repeat this cross-fitting, and aggregate the results. Complete details appear in Appendix E.

We now turn to illustrate the performance of the proposed method in a simulation study.

8 Illustrative Simulations

The simulations assess performance across three dimensions: treatment effect heterogeneity bias, random effects, and interference, generating eight different simulation settings. In each, we draw a standard normal covariate x_{i1}, error terms v_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 each setting, the marginal effect is in-truth 1, the systematic component is driven entirely by the first covariate, and all covariates, random effects, and the error terms are normally distributed. Table 1 provides details. The first model is additive, non-interactive, and equivariant in all errors, serving as a baseline. The second model induces treatment effect heterogeneity bias by including an interaction between the treatment and squared covariate along with heteroskedasticity in the treatment residual. The third adds a fifty-leveled, standard normally distributed random effect as a confounder, and the fourth adds an interaction term. Note that summations are over all other observations, such that the outcome is a function of other observations' treatment level while the treatment is a function of other observations' squared covariate.

Model Specifications

$$\begin{aligned}
\text{Baseline: } & y_i = t_i + x_{i1}^2 + \epsilon_i & t_i = x_{i1} + v_i;\ v_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \\
\text{Treatment Effect Heterogeneity: } & y_i = t_i \times x_{i1}^2 + \epsilon_i & t_i = x_{i1} + v_i;\ v_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}\left(0, \frac{x_{i1}^2 + 1}{2}\right) \\
\text{Random Effects: } & y_i = \cdot + a_{j[i]} & t_i = \cdot + a_{j[i]};\ a_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1);\ \#j = 50 \\
\text{Interference: } & y_i = \cdot + \psi_{t,i} & t_i = \cdot + \psi_{x,i}
\end{aligned}$$

Constructing Interference Terms

$$\text{Interference Covariates: } \psi_{t,i} = \sum_{i^\vartheta \neq i} p_{i,i^\vartheta} \times t_{i^\vartheta};\ \psi_{x,i} = \sum_{i^\vartheta \neq i} p_{i,i^\vartheta} \times x_{i^\vartheta 1}^2$$

$$\text{where } p_{i,i^\vartheta} = \frac{e^{\ (x_{i1}\ \ x_{i^\vartheta 1})^2}}{\sum_{i^\vartheta \neq i} e^{\ (x_{i1}\ \ x_{i^\vartheta 1})^2}}$$

Table 1: **Simulation Specifications.** We begin with a baseline additive model, and the second adds treatment effect heterogeneity bias by introducing a correlation between effect heterogeneity and treatment assignment heteroskedasticity. We also include a fifty-leveled random effect as confounder, with other methods being given the indicator variables in their control set. The final specification adds an interference term, with the precise construction of the term at the bottom. The residual term ϵ_i follows a standard normal.

The covariates are then transformed as

$$\mathbf{x}_i = \left[x_{i1} - \frac{1}{2} x_{i2},\ x_{i2} - \frac{1}{2} x_{i1},\ x_{i3},\ x_{i4},\ x_{i5} \right].$$

and each method is given the outcome, treatment, transformed covariates and indicator variables for the random effects, regardless of whether the random effects are in the true data generating process. We report results for $n = 1000$ with additional sample sizes in Appendix F.[21]

Along with the proposed method (PLCE), we implement four different machine learning

[21] 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 appear similar to the results in the body.

methods. Kernel Regularized Least Squares ((KRLS) Hainmueller and Hazlett, 2013) fits a single, nonparametric regression, takes the partial derivative of the fitted model with respect to the treatment variable, and returns the average of these values as the marginal effect. The Covariate Balancing Propensity Score for continuous treatments ((CBPS) Fong, Hazlett and Imai, 2018)[22] generates a set of weights that eliminate the effect of confounders under the assumption that the treatment distribution is normal and equivariant. We also implement the Double Machine Learning (DML) of Chernozhukov et al. (2018) with use random forests used to learn $\widehat{f}, \widehat{g}$ and the generalized random forest (GRF) of Athey, Tibshirani and Wager (2019), which is similar to DML but uses a particular random forest algorithm tuned for efficient inference on a marginal effect. Lastly, we also include least squares (OLS) for comparison.

The simulations were designed to highlight our theoretical expectations in the simplest possible setting. KRLS is closest to our method, in that we also implement a nonparametric regression model. Like us, KRLS should handle nonlinearities well, but since it does not engage in a split-sample strategy, we expect undercoverage with its confidence intervals. DML and GRF do engage in a split-sample strategy, but, like KRLS, were not designed to handle random effects. We expect all three to perform poorly. OLS should handle the random effects well, since we simply enter them as covariates in the model, but should be particularly susceptible to treatment effect heterogeneity bias. None of the methods were constructed to adjust for interference. PLCE should do a reasonable job across all settings, as it was designed to handle random effects and adjust for both interference and treatment

[22] 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.

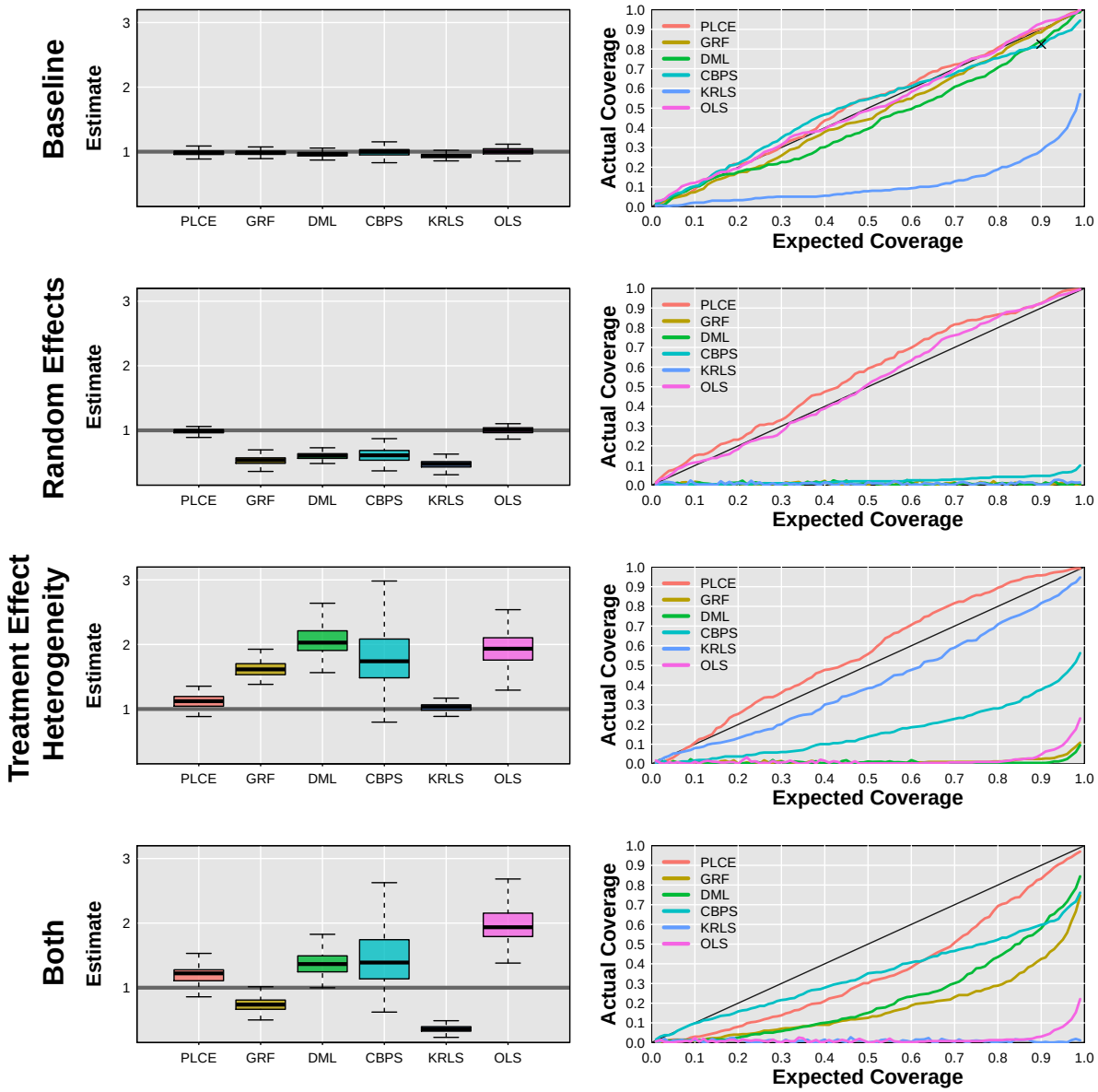


Figure 2: **Results for simulations without interference.** The first column shows the distribution of point estimates, where the true value is 1, in gray. The second column shows the coverage rates: expected coverage is on the x-axis and actual coverage is on the y-axis. If a curve falls below the 45 line, 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 compared against Generalized Random Forests (GRF), Double Machine Learning (DML), the Covariate Balancing Propensity Score for continuous treatments (CBPS), Kernel Regularized Least Squares (KRLS), and least squares (OLS). The proposed method, PLCE, is the only one to perform well across all settings.

effect heterogeneity bias.

8.1 Results for the Setting Without Interference

Results for the simulations without interference are in Figure 8. The first column shows the distribution of point estimates, with the true value of 1 in gray. We are equally concerned with whether the method allows for valid inference. The second column shows the coverage rates: expected coverage is on the x-axis and actual coverage is on the y-axis.[23] For example, consider in the top right plot the point marked "$\times$" at $(0.90, 0.82)$, which is on the CBPS curve. Here, we construct a 90% confidence interval of the form $[\widehat{\theta} - 1.64\widehat{\sigma}_{\widehat{\theta}}, \widehat{\theta} + 1.64\widehat{\sigma}_{\widehat{\theta}}]$ and measure the proportion of simulations where the confidence interval contains the true value of 1. In this case, for CBPS, this value is 0.82, so the 90% confidence interval is invalid, albeit only slightly too narrow. More generally, if a curve falls below the 45 line, the confidence intervals are too narrow and hence invalid. If the curve falls above the 45 line, the confidence intervals are valid but wide.

Simulation settings increase in complexity going down the rows. The first row contains contains the baseline model; the second, the model with group indicators added; the third, the baseline model with treatment effect heterogeneity; and the fourth, both treatment effect heterogeneity model and random effects.

Starting in the first row every method performs well in the baseline model, though KRLS exhibits under-coverage. In the second row, we add random effects, and only the proposed method and OLS provide unbiased estimates and valid intervals. In the third row, the proposed method and KRLS are unbiased with valid intervals. In the final row, with both random effects and treatment effect heterogeneity, every method shows discernible bias, though

[23] The "coverage rate" is the proportion of samples for which the constructed confidence interval contains the true value of 1 (see, e.g. Wooldridge, 2013, Sec. 4.3.).

the proposed method and KRLS have the lowest bias and the least misleading confidence errors.

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

8.2 Results for the Setting With Interference

Figure 8 presents results from the simulations in the presence of interference. All methods save least squares return accurate point estimates in the simplest setting, with the proposed method, DML, and CBPS providing narrow but reasonable confidence intervals. Coverage from GRF, while valid in the setting without interference, is now near-zero. In the remaining rows, point estimates are reasonable, particularly for the proposed method but also KRLS. The impact of interference shows up in the coverage rates. In the bottom three settings, coverage rates are near-zero for all methods. Only the proposed method provides both reliable point estimates and confidence intervals across each of the settings.

9 Empirical Applications

We illustrate the proposed method using data from two recent studies. First, we reanalyze experimental data to illustrate that the proposed method returns estimates and standard errors similar to a linear regression when the linear regression is the correct thing to do. Second, we show how the method can estimate a treatment effect with a continuous treatment variable. We use data from a study where the researcher was forced to dichotomize a continuous treatment in order to estimate a causal effect.

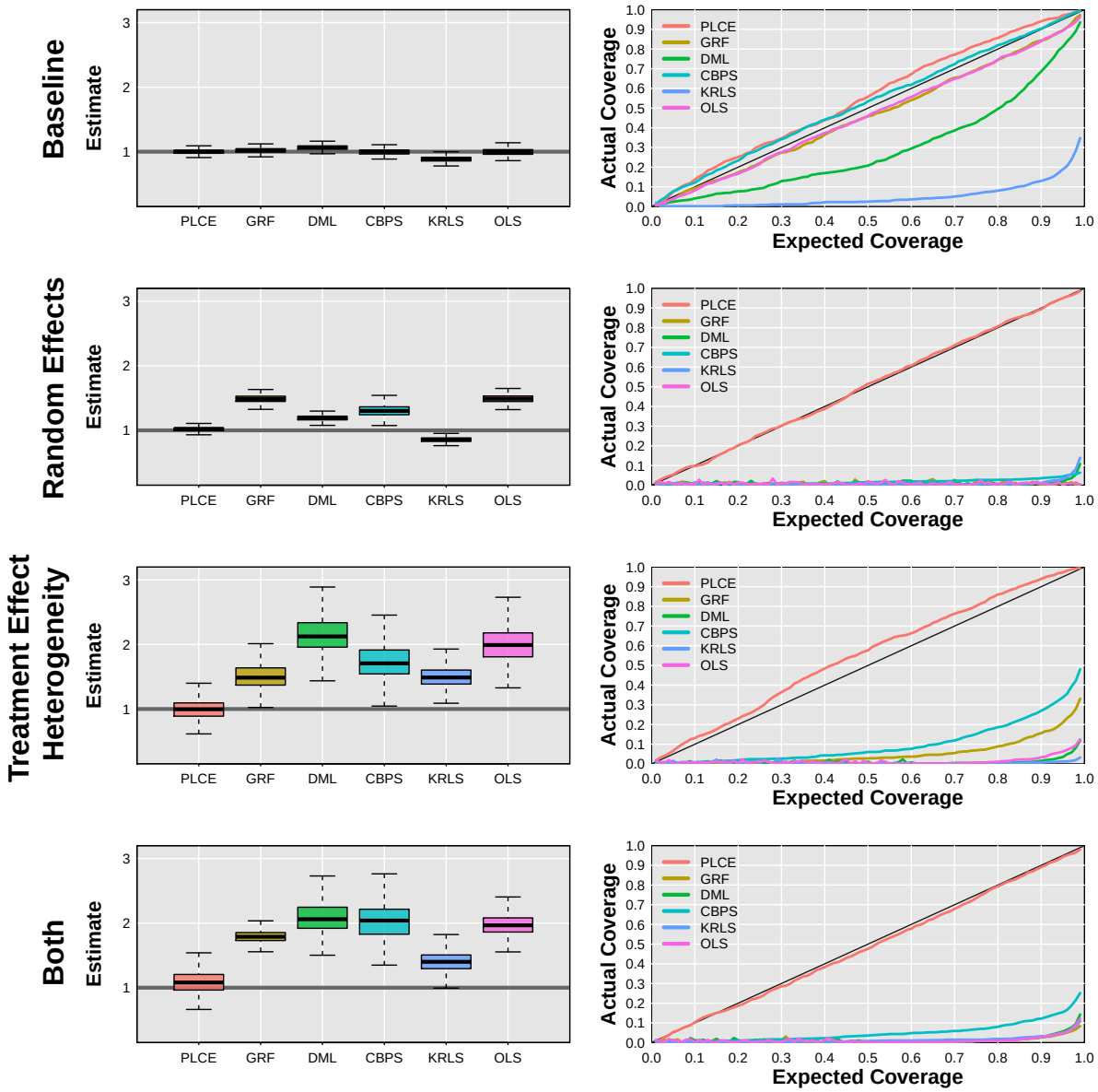


Figure 3: **Results for simulations with interference.** The first column shows the distribution of point estimates, where the true value is 1, in gray. The second column shows the coverage rates: expected coverage is on the x-axis and actual coverage is on the y-axis. If a curve falls below the 45 , 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 compared against Generalized Random Forests (GRF), Double Machine Learning (DML), the Covariate Balancing Propensity Score for continuous treatments (CBPS), Kernel Regularized Least Squares (KRLS), and least squares (OLS). The proposed method, PLCE, is the only one to perform well across all settings.

	Hawks			Doves		
	PLCE	Diff-in-Mean	OLS	PLCE	Diff-in-Mean	OLS
Estimate	11.72	11.98	11.97	36.24	35.43	35.19
s.e.	3.63	3.80	3.80	2.78	3.12	2.85

Figure 4: **Comparing PLCE and a Linear Regression in an Experimental Setting.** Across all statistics, the PLCE estimates perform comparably to least squares on this experimental data. The repeated cross-fitting strategy does not inflate standard errors in this setting.

9.1 Maintaining Efficiency

Mattes and Weeks (2019) conduct a survey experiment in the United States, asking respondents about a hypothetical foreign affairs 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 effective 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 effect of conciliation varies between hawks and doves, as reported in Table 2 of the original work. Results appear in Table 9.1. For the two estimated effects, we find that PLCE returns results quite similar to least squares. Importantly, the standard errors are comparable across the methods. This suggests no efficiency loss when employing our method in a situation where we know least squares is unbiased and efficient.

9.2 Estimating a Causal Effect in the Presence of a Continuous Treatment

We next reanalyze data from a recent study that estimated the causal effect of racial threat on voter turnout (Enos, 2016). 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 effect. The author implements a difference-in-difference analysis which, unfortunately, requires a binary treatment. To accommodate the method, the author artificially 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, 0.3 kilometers, and drops off precipitously after. The proposed method allows estimation of the average causal effect of distance on the outcome.

We conduct four separate analyses. For the first, we estimate the causal effect of distance on change in turnout for white residents within one kilometer of a demolished housing project. The treatment variable is 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. We also include a random effect for identifying the nearest housing

Figure 5: **Causal Effect Estimate of Racial Threat.** Revisiting the study by Enos (2016), we find a statistically and substantively relevant impact of racial threat on white voters (top row) and, as predicted by theory, not for black voters near the housing projects (second row) or for black and white voters further than 1 *km* from any projects.

project to each individual.[24] We next generate three matched samples for further analysis.[25] The first 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 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 sufficiently far from a demolished project that any threat should be muted.

Figure 5 presents the effect estimates. We estimate that living adjacent to a public housing unit, rather than 1 km away, causes a decrease in turnout of about 7.93 percentage

[24] See the supplemental materials of Enos (2016) for more details.

[25] 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.

points for white residents ($s.e. = 0.0390,\ z = 2.03$), an effect in line with the results from the original analysis (see Figure 1 there). The estimated effect for black voters near housing projects of 0.0054 ($s.e. = 0.035,\ z = 0.15$) is not significant. The bottom two lines consider distal blacks and whites, providing a placebo test. We find no effect of distance on turnout.

Along with not relying on a user-specified control set, the proposed method allows for causal effect estimation with a continuous treatment variable. We find results of a similar magnitude to the original study, but without needing to transform the data so that it is amenable to a framework that generally relies on a binary treatment.

10 Conclusion

Testing our intuitions and hypotheses against the data in a way that does not rely on strong assumptions is essential to a reliable accumulation of knowledge. Doing so builds faith that our results and theory are driven by actual trends in the data and not a particular set of choices made by the researcher.

To this end, we have introduced to the field a framework, taken from the field of semi-parametric inference, for conducting valid inference while allowing machine learning methods to construct a control vector that can account for a wide range of commonly encountered biases. Essential to this approach is a sample splitting strategy, where we never use the same data to both construct the control vector and conduct inference. We have extended this literature, allowing for inference robust to both heterogeneities in the treatment effect and particular patterns of interference amongst observations. The method extends causal inference, as well, accommodating continuous treatment variables. The accompanying software allows these analyses to be done in a line or two of code and allow for several diagnostics.

Ultimately, our aim is to allow for more believable, less assumption-driven inference. We move the field in this direction, where machine learning can be incorporated into our workaday research as a means of controlling for background covariates, freeing the researcher to develop and test theories with some confidence that the results are not driven by her ability to specify every element of a statistical model.

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):512–529.

Achen, Christopher. 2002. "Toward a New Political Methodology: Microfoundations and ART." *Annual Review of Political Science* 5:423–450.

Achen, Christopher H. 2005. "Let's Put Garbage-Can Regressions and Garbage-Can Probits Where They Belong." *Con ict Management and Peace Science* 22(4):327–339.

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* 11(4):1912–1947.

Aronow, Peter M. 2016. "A Note on "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It"." *arXiv:1609.01774 [stat]* .

Athey, Susan, Julie Tibshirani and Stefan Wager. 2019. "Generalized Random Forests." *Annals of Statistics* 47(2):1148–1178.

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* 42(2):596–627.

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

Bickel, Peter J., Chris A. J. Klaassen, Ya'acov Ritov and Jon A. Wellner. 1998. *E cient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag.

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, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey and James Robins. 2018. "Double/debiased machine learning for treatment and structural parameters." *The Econometrics Journal* 21(1):C1–C68.

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.

Dalalyan, A. S., G. K. Golubev and A. B. Tsybakov. 2006. "Penalized Maximum Likelihood and Semiparametric Second-Order Efficiency." *The Annals of Statistics* 34(1):169–201.

de Boor, C. 1978. *A Practical Guide to Splines*. New York: Springer.

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

Enos, Ryan D. 2016. "What the Demolition of Public Housing Teaches Us about the Impact of Racial Threat on Political Behavior." *American Journal of Political Science* 60(1):123–142.

Fong, Christian, Chad Hazlett and Kosuke Imai. 2018. "Covariate Balancing Propensity Score for a Continuous Treatment: Application to the Efficacy of Political Advertisements." *The Annals of Applied Statistics* 12(1):156–177.

Fong, Christian and Matthew Tyler. 2021. "Machine Learning Predictions as Regression Covariates." *Political Analysis* 29(4):467–484.

Freedman, David A. 2006. "On The So-Called 'Huber Sandwich Estimator' and 'Robust Standard Errors'." *The American Statistician* 60(4):299–302.

Gelman, Andrew and Jennifer Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambrdige University Press.

Grimmer, Justin, Solomon Messing and Sean J Westwood. 2017. "Estimating Heterogeneous Treatment Effects and the Effects of Heterogeneous Treatments with Ensemble Methods." *Political Analysis* 25(4):1–22.

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

Hall, Andrew 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):509–524.

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

Hsing, Tailen and Haobo Ren. 2009. "An RKHS Formulation of the Inverse Regression Dimension-Reduction Problem." *Annals of Statistics* 37(2):726–755.

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.

Imai, Kosuke and Marc Ratkovic. 2013. "Estimating treatment effect heterogeneity in randomized program evaluation." *The Annals of Applied Statistics* 7(1):443–470.

Imbens, Guido W. and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Kennedy, Edward, Zongming Ma, Matthew McHugh and Dylan Small. 2017. "Nonparametric Methods for Doubly Robust Estimation of Continuous Treatment Effects." *Journal of the Royal Statistical Society, Series B* 79(4):1229–1245.

King, Gary. 1990. "On Political Methodology." *Political Analysis* 2:1–29.

King, Gary and Langche Zeng. 2006. "The Dangers of Extreme Counterfactuals." *Political Analysis* 14(2):131–159.

King, Gary and Margaret E. Roberts. 2015. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It." *Political Analysis* 23(2):159–179.

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 S. and Alexander Sahn. 2021. "Achieving Statistical Significance with Control Variables and Without Transparency." *Political Analysis* 29(3):356–369.

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

Li, Lingling, Eric Tchetgen Tchetgen, Aad van der Vaart and James M. Robins. 2011. "Higher Order Inference on a Treatment Effect Under Low Regularity Conditions." *Statistics & probability letters* 81(7):821–828.

Long, L.S. and L.H. Ervin. 2000. "Using Heteroscedasticity Consistent Standard Errors in the Linear Regression Model." *The American Statistician* 54:217–224.

Manski, Charles F. 1993. "Identification of Endogenous Social Effects: The Reflection Problem." *The Review of Economic Studies* 60(3):531–542.

Manski, Charles F. 2013. "Identification 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.

Montgomery, Jacob M. and Santiago Olivella. 2018. "Tree-based Models for Political Science Data." *American Journal of Political Science* 62(3):729–744.

Newey, Whitney. 1994. "Kernel Estimation of Partial Means and a General Variance Estimator." *Econometric Theory* 10(2):233–253.

Ratkovic, Marc. 2021. Subgroup Analysis: Pitfalls, Promise, and Honesty. In *Advances in Experimental Political Science*, ed. James N. Druckman and Donald P. Green. pp. 271–288.

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, Lingling Li, Eric Tchetgen and Aad van der Vaart. 2008. *Higher order in uence functions and minimax estimation of nonlinear functionals*. Institute of Mathematical Statistics.

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

Robins, James, Mariela Sued, Quanhong Lei-Gomez and Andrea Rotnitzky. 2007. "Comment: Performance of Double-Robust Estimators When 'Inverse Probability' Weights Are Highly Variable." *Statistical Science* 22(4):544–559.

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.

Savje, Fredrik, Peter M. Aronow and Michael G. Hudgens. 2021. "Average Treatment Effects in the Presence of Unknown Interference." *Annals of Statistics* 49(2):673–701.

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. Efficient Nonparametric Testing and Estimation. In *Berkeley Symposium on Mathematical Statistics and Probability*, ed. Jerzy Neyman. pp. 187–195.

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

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

van der Vaart, Aad. 2014. "Higher Order Tangent Spaces and Influence Functions." *Statistical Science* 29(4):679–686.

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.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica* 48(4):817–838.

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

A Formal Derivation of a Semiparametrically Efficient Estimator in the Partially Linear Model

A.1 Notation

In order to integrate the technical discussion with the broader statistical literature, we adopt the standard empirical process notation, where P_n denotes the sample mean, $P_n x_i = \frac{1}{n}\sum_{i=1}^{n} x_i$, P the population mean $Px_i = \mathsf{E}(x_i)$ and G_n the empirical process $G_n\widehat{\beta} = \sqrt{n}(P_n\widehat{\beta} - \beta)$ where $\beta = P\widehat{\beta}$. The remaining notation is as in the body.

A.2 Characterizing the Semiparametric Efficiency Bound

The semiparametrically efficient estimate can be calculated treating the model where we assume the true nuisance functions were known, called the parametric submodel, as a linear regression,

$$\mathbf{y} = \theta\mathbf{t} + \mathbf{X}_\eta\gamma_y + \mathbf{e}$$

$$\mathbf{t} = \mathbf{X}_\eta\gamma_t + \mathbf{u}$$

with the i^{th} row of $\mathbf{X}_\eta$ is $\mathbf{x}_{\eta,i} = [f(\mathbf{x}_i), g(\mathbf{x}_i)]$. $\mathbf{H}_\eta$ is the projection matrix of $\mathbf{X}_\eta(\mathbf{X}_\eta^{>}\mathbf{X}_\eta)^{\ 1}\mathbf{X}_\eta^{>}$, where the matrix is assumed full rank, and $\mathbf{A}_\eta = \mathbf{I}_n - \mathbf{H}_\eta$, the annihilator matrix.

Denote as $\widetilde{\mathbf{y}}, \widetilde{\mathbf{t}}$ the residuals after regressing $\mathbf{y}, \mathbf{t}$ on the matrix $\mathbf{X}_\eta$, i.e.

$$\begin{aligned}
\widetilde{\mathbf{t}} &= \mathbf{A}_\eta \mathbf{t} = \underbrace{\mathbf{A}_\eta \mathbf{u}}_{:=\widetilde{\mathbf{u}}} \\
&= \widetilde{\mathbf{u}} \\
\widetilde{\mathbf{y}} &= \mathbf{A}_\eta y = \theta \mathbf{A}_\eta \mathbf{t} + \underbrace{\mathbf{A}_\eta \mathbf{e}}_{:=\widetilde{\mathbf{e}}} \\
&= \theta \widetilde{\mathbf{t}} + \widetilde{\mathbf{e}} \\
&= \theta \widetilde{\mathbf{u}} + \widetilde{\mathbf{e}}
\end{aligned}$$

We can then recover the semiparametrically efficient estimate

$$\widehat{\theta} = \frac{\widetilde{\mathbf{y}}^{\top} \widetilde{\mathbf{t}}}{\widetilde{\mathbf{t}}^{\top} \widetilde{\mathbf{t}}} \tag{30}$$

$$= \frac{P_n \widetilde{y}_i \widetilde{t}_i}{P_n \widetilde{t}_i^2} \tag{31}$$

$$= \frac{P_n (\theta \widetilde{u}_i^2 + \widetilde{u}_i \widetilde{e}_i)}{P_n \widetilde{u}_i^2} \tag{32}$$

$$= \theta + \frac{P_n \widetilde{u}_i \widetilde{e}_i}{P_n \widetilde{u}_i^2} \tag{33}$$

By the law of large numbers, the estimator is clearly consistent for θ. Then, to calculate its limiting distribution,

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) = G_n \widehat{\theta} = \sqrt{n}\left\{\frac{P_n \widetilde{u}_i \widetilde{e}_i}{P_n \widetilde{u}_i^2}\right\} \tag{34}$$

$$\mathcal{N}\left(0, \frac{\mathsf{E}\left(\widetilde{u}_i^2 \widetilde{e}_i^2\right)}{\mathsf{E}\left(\widetilde{u}_i^2\right)^2}\right). \tag{35}$$

This gives the semiparametric efficiency bound for our model.

Now, were f, g known, we could simply recover a point and variance estimate through the method of least squares, and we know this estimate is efficient. We do not know f, g, so we next construct an estimate that is asymptotically indistinguishable from the estimate calculated from the parametric submodel. This estimate will be semiparametrically efficient.

A.3 Deviations Between the Feasible Estimate and the Semiparametrically Efficient Estimator

Now we do not know f, g, but instead estimate $\widehat{f}, \widehat{g}$, introducing $\Delta_{\widehat{f}}, \Delta_{\widehat{g}}$ into the linear regression. The argument follows exactly as above, except we now must account for these approximation error terms.

We write the models in terms of $\widehat{f}, \widehat{g}$, and hence in terms of f, g and $\Delta_{\widehat{f}}, \Delta_{\widehat{g}}$, giving the models

$$\mathbf{y} = \theta\mathbf{t} + \mathbf{X}_\eta\gamma_y + \mathbf{X} \ \ \beta_y + \mathbf{e}$$

$$\mathbf{t} = \mathbf{X}_\eta\gamma_t + \mathbf{X} \ \ \beta_t + \mathbf{u}$$

with the i^{th} row of $\mathbf{X}$ \ \ is $\mathbf{x}_{\ ,i} = [\Delta_{\widehat{f},i}, \Delta_{\widehat{g},i}]$.

We can then construct

$$\widetilde{\mathbf{t}} = \mathbf{A}_\eta \mathbf{t} = \underbrace{\mathbf{A}_\eta \mathbf{X} \quad \beta_y}_{:= \widetilde{\ }_t} + \underbrace{\mathbf{A}_\eta \mathbf{u}}_{:= \widetilde{\mathbf{u}}} \tag{36}$$

$$= \widetilde{\mathbf{u}} + \widetilde{\Delta}_t \tag{37}$$

$$\widetilde{\mathbf{y}} = \mathbf{A}_\eta y = \theta \mathbf{A}_\eta \mathbf{t} + \underbrace{\mathbf{A}_\eta \mathbf{X} \quad \beta_y}_{:= \widetilde{\ }_y} + \underbrace{\mathbf{A}_\eta \mathbf{e}}_{:= \widetilde{\mathbf{e}}} \tag{38}$$

$$= \theta \widetilde{\mathbf{t}} + \widetilde{\mathbf{e}} \tag{39}$$

$$= \theta \widetilde{\mathbf{u}} + \theta \widetilde{\Delta}_t + \widetilde{\Delta}_y + \widetilde{\mathbf{e}} \tag{40}$$

Given these partialed-out values, we can construct

$$\sqrt{n}(\widehat{\theta} - \theta) = G_n \widehat{\theta} = \sqrt{n} \left(\frac{P_n \widetilde{t}_i \widetilde{y}_i}{P_n \widetilde{t}_i^2} - \theta \right) \tag{41}$$

Beginning with the denominator, we get

$$P_n \widetilde{t}_i^2 = P_n \left\{ \widetilde{u}_i^2 + 2 \widetilde{u}_i \widetilde{\Delta}_{\widehat{g},i} + \widetilde{\Delta}_{\widehat{g},i}^2 \right\} \xrightarrow{\mathsf{u}} \mathsf{E}(u_i^2) \tag{42}$$

by the uniform consistency of $\widehat{f}, \widehat{g}$. We can then use a uniform Slutsky's theorem and characterize the limiting behavior as

$$\sqrt{n}(\widehat{\theta} - \theta) = G_n \widehat{\theta} = \sqrt{n} \left(\frac{P_n \widetilde{t}_i \widetilde{y}_i}{\mathsf{E}(\widetilde{u}_i^2)} - \theta \right) \tag{43}$$

and expanding the numerator gives us,

$$\sqrt{n}(\widehat{\theta}-\theta) = G_n\widehat{\theta} = \sqrt{n}\left(\underbrace{\frac{\theta P_n\widetilde{u}_i^2 + P_n\widetilde{u}_i\widetilde{e}_i}{P\widetilde{u}_i^2}}_{\text{e cient estimate}} - \theta\right) + \underbrace{\sqrt{n}\frac{B}{P\widetilde{u}_i^2}}_{\text{bias terms}} \tag{44}$$

$$\text{where } B = P_n\left\{2\theta\widetilde{u}_i\widetilde{\Delta}_{\widehat{g},i} + \widetilde{u}_i\widetilde{\Delta}_{\widehat{f},i} + \widetilde{\Delta}_{\widehat{g},i}\widetilde{e}_i + \theta\widetilde{\Delta}_{\widehat{g},i}^2 + \widetilde{\Delta}_{\widehat{g},i}\widetilde{\Delta}_{\widehat{f},i}\right\} \tag{45}$$

The first element of the sum shares a limiting distribution with the estimate given above, and hence achieves the semiparametric efficiency bound.

Establishing semiparametric efficiency of an estimate is, at its simplest, deriving a set of assumptions under which $\sqrt{n}B \xrightarrow{u} 0$. Recall that $\widetilde{\Delta}_{\widehat{f},i}, \widetilde{\Delta}_{\widehat{g},i}$ are each an arbitrary linear combination of the approximation error terms and $\widetilde{u}_i$ and $\widetilde{e}_i$ are a linear combination of the errors. Zeroing out the first three terms can be guaranteed when

$$\sqrt{n}P_n u_i\Delta_{\widehat{f},i} \xrightarrow{u} 0, \quad \sqrt{n}P_n u_i\Delta_{\widehat{g},i} \xrightarrow{u} 0 \tag{46}$$

and the third when

$$\sqrt{n}P_n e_i\Delta_{\widehat{f},i} \xrightarrow{u} 0, \quad \sqrt{n}P_n e_i\Delta_{\widehat{g},i} \xrightarrow{u} 0. \tag{47}$$

This is accomplished through a split-sample strategy, as the split sample approach guarantees that the random element in the approximation error is conditionally independent of that in the inference sample.[26] See van der Vaart (1998, ch. 25) for more.

[26] An alternative approach is to assume that the functions f, g are sufficiently simple that this bias term is negligible. This is referred to as a *Donsker-class* assumption; see (van der Vaart, 1998, esp. Ch. 19) for details.

The last two bias terms involve square and cross-products of the approximation error terms,

$$\sqrt{n}P_n\Delta^2_{\widehat{f},i}, \quad \sqrt{n}P_n\Delta^2_{\widehat{g},i}; \quad \sqrt{n}P_n\Delta_{\widehat{f},i}\Delta_{\widehat{g},i}. \tag{48}$$

By taking the square roots of the square terms and apply Cauchy-Schwarz to the cross-product, these terms go to zero when

$$n^{1/4}\sqrt{P_n\Delta^2_{\widehat{g},i}} \xrightarrow{\mathsf{u}} 0; \quad n^{1/4}\sqrt{P_n\Delta^2_{\widehat{f},i}} \xrightarrow{\mathsf{u}} 0, \tag{49}$$

which gives the $n^{1/4}$ rate described in the text. Under these conditions, B tends to zero uniformly and the estimate is semiparamterically efficient.

A.4 Second-Order Semiparametric Efficiency

So long as the covariate vectors $\widehat{U}_{\widehat{f},i}, \widehat{U}_{\widehat{g},i}$ are finite dimensional, which they are by assumption, then the argument above establishes their first-order semiparametric efficiency.

As to why the required convergence rate drops from $n^{1/4}$ to $n^{1/8}$, we have to examine the convergence of these two covariates. Consider the convergence of $\widehat{U}_{\widehat{f},i}$, with an analogous argument for $\widehat{U}_{\widehat{g},i}$. In this case, the principal components are constructed from cross-observation covariances,

$$f(\mathbf{x}_i) \approx \widehat{f}(\mathbf{x}_i) + \widehat{U}^{\succ}_{\widehat{f},i}\gamma_f \tag{50}$$

$$f(\mathbf{x}_i) = \widehat{f}(\mathbf{x}_i) + \sum_{i^\theta=1}^{n} \widehat{\mathrm{Cov}}(\widehat{f}(\mathbf{x}_i), \widehat{f}(\mathbf{x}_{i^\theta}))w_{i^\theta} \tag{51}$$

for scalars $\widehat{w}_{i^\emptyset}$.[27] The finite-dimensional assumption of the $U_{f,i}$ constrains $w_{i^\emptyset}$, since these are linear combinations of the finite terms in γ_f, and the split-sample approach will ensure any approximation errors in $\widehat{\gamma}_f$ and $\widehat{f}$ be uncorrelated.

As to the gain in efficiency, note that we need to consider convergence of terms like

$$\sqrt{n}P_n \left\{ \widehat{\text{Cov}}(\widehat{f}(\mathbf{x}_i), \widehat{f}(\mathbf{x}_{i^\emptyset})) - \text{Cov}(\widehat{f}(\mathbf{x}_i), \widehat{f}(\mathbf{x}_{i^\emptyset})) \right\}^2 \tag{52}$$

in our regression. For first-order semiparametric efficiency, the $n^{1/4}$ rate is recovered from taking the square root of this term. For the second order calculation, though, note that since $\widehat{f}(\mathbf{x}_i), \widehat{f}(\mathbf{x}_{i^\emptyset})$ if both are converging at $n^{1/8}$, their product in the covariance is converging at $n^{1/4}$.

So what is making this rate gain happen? The finite-dimensional assumption of the covariance matrix is doing a ***lot*** of theoretical work. It makes convergence of the sums described above tractable, allowing us to argue dimension-by-dimension by a Cramer-Wold device van der Vaart (1998). For a more general theoretical discussion, see Li et al. (2011) and Robins et al. (2008).

As a practical matter, our stance on second-order efficiency is that if our finite dimensional assumption is correct then we achieve second-order efficiency by fully capturing the variance in the approximation errors. If this assumption is not correct, though, we still recover a semiparametrically efficient estimate and–due to our split sample strategy–the principal components should help or, at worst, increase the variance of our estimates. Our simulations and applied examples provide compelling evidence that our approach is reasonable.

[27] This is an example of a *second-order U-statistic*, see van der Vaart (1998) Ch. 12 for more.

B Diagnostics

We implement a sensitivity analysis in order to assess how strong an unobserved confounder must be in order to overturn our results. Since our method is, in effect, a linear regression in subsample $\mathcal{S}_2$, diagnostics for the linear regression are applicable.

We implement the recent method of Cinelli and Hazlett (2020) in our software. Following the authors' suggestion, we report three statistics. The statistics are calculated on subsample $\mathcal{S}_2$, and averaged over cross-fits. 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 $R^2_{Y\ DjX}$, assumes a "worst-case" confounder that perfectly explains the residuals, 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 also assess the positivity assumption. Positivity is violated when the treatment variable is a deterministic function of the covariates. We do so by graphically assessing the kurtosis of the residuals (Wooldridge, 2013, Appendix B, p. 737.).[28]

Denoting as $\widehat{\epsilon}_{i,s}$ the residual from estimating the treatment for observation i on repeated cross-fit iteration s, we estimate the kurtosis $\widehat{\kappa}_i$ as

$$\widehat{\kappa}_i = \frac{\frac{1}{S}\sum_{s=1}^{S} \widehat{\epsilon}_{i,s}^4}{(\frac{1}{S}\sum_{s=1}^{S} \widehat{\epsilon}_{i,s}^2)^2}. \tag{53}$$

[28] For a random variable X, its kurtosis is $\mathsf{E}(X^4)/\mathsf{E}(X^2)^2$ Since the kurtosis is constructed from a fourth moment, and can be written as $\mathsf{E}(Z^2); Z = X^2$, the kurtosis captures the variance of the variance.

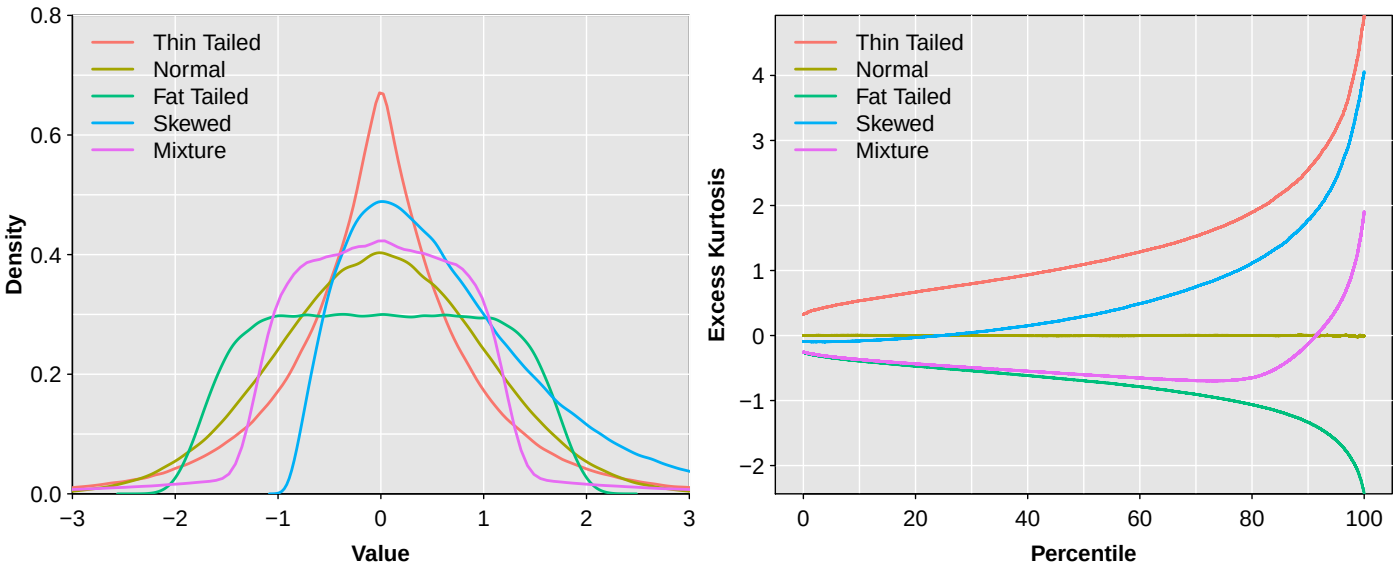


Figure 6: **Excess kurtosis plot for diagnosing positivity.** The lefthand side presents the five different error densities assessed on the right. The first one is thin-tailed and raises the deepest concerns about violating positivity. The next is normal, then a fat-tailed and skewed density. The last density combines a the thin-tailed density, where some observations may violate positivity, and a normal density. Consulting the righthand figure, a positive excess kurtosis statistic everywhere above zero suggests that the researcher should examine the data for violations of positivity.

The excess kurtosis is the extent that this statistic falls above that expected from a normal distribution, and we plot them from high to low. The lefthand side of Figure B contains five possible error densities. The first 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 combines a thin-tailed density, where some observations may violate positivity, and a normal density.

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

and thin-tailed creates a U-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 identify observations for which the residuals are pooling near zero and consider trimming them from the analysis. This will change the estimand from the average effect to a local average effect on the trimmed sample.

These statistical diagnostics and the excess kurtosis plot are all generated by our software.

C Causal Assumptions

In giving these assumptions, we utilize the *potential outcomes notation* (Imbens and Rubin, 2015), where each observation is equipped with a potential outcome function $y_i(t)$ which deterministically maps an arbitrary treatment level t to the outcome for observation i under that treatment, $y_i(t)$. We will denote as $\mathbf{X}_{\ i}, \mathbf{t}_{\ i}$ the background covariates and treatments for all observations except observation i.

ASSUMPTION 2 *Causal Assumptions*

1. *Stable treatment value: There is a single version of each treatment value.*
2. *Positivity: The treatment is not deterministic,* $\mathrm{Var}(t_i|\mathbf{x}_i, \mathbf{X}_{\ i}) > 0$ *for all observations* i*.*
3. *Ignorability:*[29]
   (a) $t_i \perp\!\!\!\perp \mathbf{t}_{\ i} | \mathbf{x}_i, \mathbf{X}_{\ i}$

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

(b) $y_i(t_i, \mathbf{t}_{\ i}) \perp\!\!\!\perp t_i | \mathbf{t}_{\ i}, \mathbf{x}_i, \mathbf{X}_{\ i}$

The discussion of these assumptions appears in the text. We next show that the these assumptions identify the marginal effect.

The partial effect of the treatment on the outcome at a given point $\mathbf{x}_i$ 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 $\mathbf{x}_i$ and also fixing $\mathbf{X}_{\ i}$. The causal interpretation of this parameter involves considering all possible combinations $(y_i(t_i, \mathbf{t}_{\ i}), t_i, \mathbf{t}_{\ i})$ for all values of $\mathbf{t}$ and regressing $y_i(t_i, \mathbf{t}_{\ i})$ on t_i.

In order for these two parameters to be the same, we need three things. First, $y_i(t_i, \mathbf{t}_{\ i})$ must equal y_i when the treatment takes the value $\mathbf{t}$. This is the first assumption. Second, the variance of the treatment variable must be positive, so that the denominator of the coefficient is nonzero. Third, restricting ourselves to observation i with covariate values $\mathbf{x}_i, \mathbf{X}_{\ i}, \mathbf{t}_{\ i}$ 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 $\text{Cov}_{\mathcal{T}}, \text{Var}_{\mathcal{T}}$ these operators for a given observation taken with respect to the treatment. Then, under the causal identificaton assumptions, we can equate the marginal effect and causal effect as

$$\theta_i = \underbrace{\frac{\text{Cov}(y_i, t_i | \mathbf{x}_i, \mathbf{X}_{\ i}, \mathbf{t}_{\ i})}{\text{Var}(t_i | \mathbf{x}_i, \mathbf{X}_{\ i}, \mathbf{t}_{\ i})}}_{\text{Partial E ect}} = \underbrace{\frac{\text{Cov}_{\mathcal{T}}\left(y_i(t_i, \mathbf{t}_{\ i}), t_i | \mathbf{t}_{\ i}\right)}{\text{Var}_{\mathcal{T}}\left(t_i | \mathbf{t}_{\ i}\right)}}_{\text{Causal E ect}}. \tag{54}$$

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 numer-

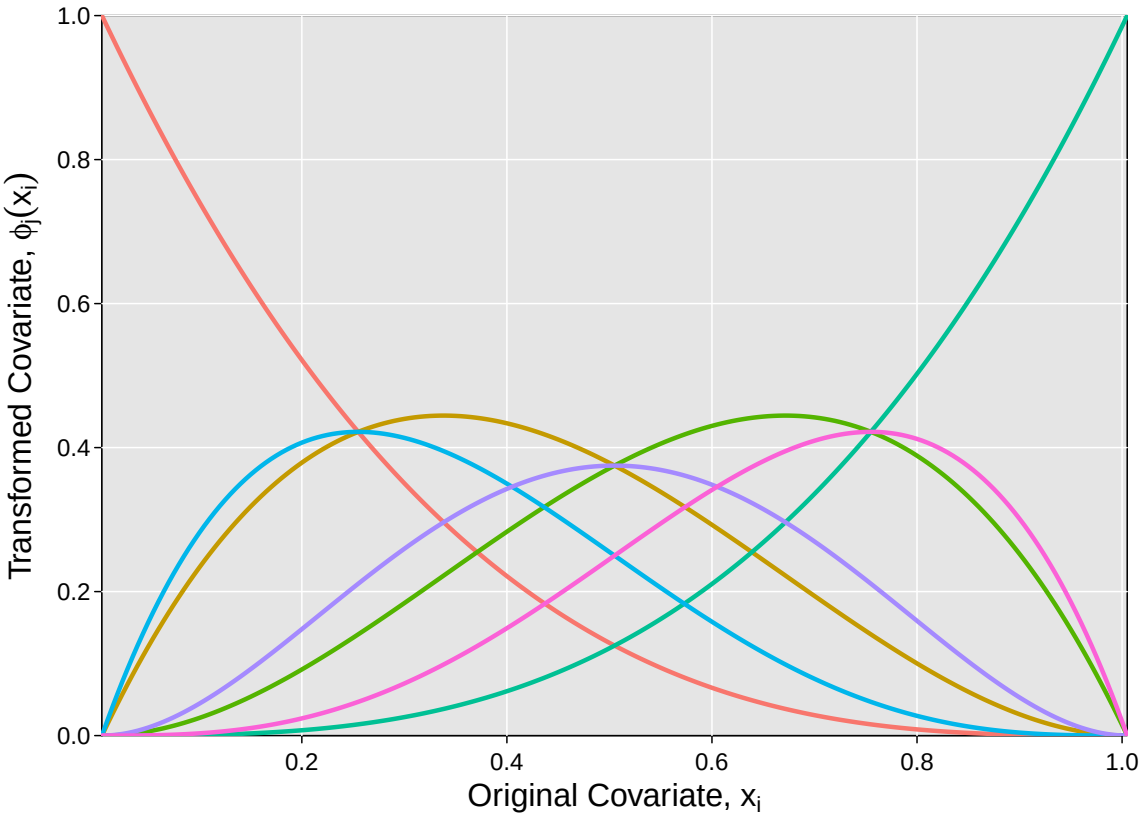


Figure 7: **Nonlinear transformations of each variable used to construct basis functions.**

ators and denominators.

Equating the marginal effect and observation-level effect for each observation equates their averages.

D Implementation Details: Preliminaries

D.1 Preliminaries: Basis Functions

D.1.1 B-Spline Basis Functions

We adjust for nonlinearities in the control variables by transforming them into a set of basis functions knowns as "B-splines" (de Boor, 1978). The original control variables are rank-

transformed and rescaled to run from 0 to 1. Then for each covariate, we generate B-splines of degrees 3 and 4 for each variable, see Figure D.1.1. In this way, we augment the control variables with nonlinear transformations of each covariate.

We will denote the k^{th} of these transformations applied to covariate $\mathbf{X}_j$ as

$$\mathbf{X}_j \mapsto \phi_k(\mathbf{X}_j)$$

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

$$\phi_0(\mathbf{X}_j) = \mathbf{1}_n; \quad \phi_1(\mathbf{X}_j) = \mathbf{X}_j$$

Counting the intercept and linear term with the eight nonlinear transformation, we have nine terms generated from each covariate.

D.1.2 Constructing Basis Functions for our Nuisance Functions

Modeling Conditional Mean Components We model the nuisance functions f, g_1 in terms of all two-way interactions between the basis functions,

$$\phi_{k,k^\theta}(\mathbf{X}_j, \mathbf{X}_{j^\theta}) = \phi_k(\mathbf{X}_j)\phi_{k^\theta}(\mathbf{X}_{j^\theta})$$

Modeling Treatment Heteroskedasticity For the treatment heteroskedsticity term, we use the three-way interaction

$$\phi_{k,k^\theta}(\mathbf{X}_j, \mathbf{X}_{j^\theta}, \widehat{v}) = \widehat{v}\phi_k(\mathbf{X}_j)\phi_{k^\theta}(\mathbf{X}_{j^\theta})$$

where $\widehat{v} = \mathbf{t} - \widehat{E}(\mathbf{t}|\mathbf{X})$, an estimated residual. We return to how we estimate the residuals below, for now we note that the basis functions for g_2 are interactions between an estimated treatment residual and the two-way basis interactions.

Modeling Interference Components Each interference basis function is a function of two basis functions and a bandwidth parameter. We consider how close observation $i^{\emptyset}$ is to observation i, as a function of how close $\psi_k(x_{ij})$ is to $\psi_k(x_{i^{\emptyset}j})$, with bandwidth ν_{jk} as

$$\text{Proximity: } p_{i,i^{\emptyset}}(\nu_{jk}) = \frac{e^{\ \frac{1}{\nu_{jk}}(\phi_k(x_{ij}) \ \ \phi_k(x_{i^{\emptyset}j}))^2}}{\sum_{i^{\emptyset} \neq i} e^{\ \frac{1}{\nu_{jk}}(\phi_k(x_{ij}) \ \ \phi_k(x_{i^{\emptyset}j}))^2}}$$

This measure accounts for homophily and heterophily due to nonlinearity in the bases.

The interferent may be driven by an entirely different basis funciton and variable,$\phi_{k^{\emptyset}}$ and $\mathbf{X}_{j^{\emptyset}}$,

$$\text{Interferent: } \psi_{k^{\emptyset}}(x_{i^{\emptyset}j^{\emptyset}})$$

We combine these two into our interference function, the total effect on observation i with proximity $p_{i,i^{\emptyset}}(\nu_{jk})$ and interferent $\phi_{k^{\emptyset}}(x_{i^{\emptyset}j^{\emptyset}})$

$$\psi_{j,k,j^{\emptyset},k^{\emptyset}}(\mathbf{x}_i, \mathbf{X}_{\ i}) = \sum_{i^{\emptyset} \neq i} \underbrace{p_{i,i^{\emptyset}}(\nu_{jk})}_{\text{Proximity}} \times \underbrace{\phi_{k^{\emptyset}}(x_{i^{\emptyset}j^{\emptyset}})}_{\text{Interferent}}$$

The summation is taken over all observations except i, so we are capturing the effect of all observations but i on observation i, creating the interference bases used in our model.

We need to reduce the bases above down to a reasonable number for a linear regression.

We do so in two ways: through a correlation screen and then we fit a high-dimensional regression to these selected bases. We give specifics below, but provide an overview of the strategies here.

D.2 Screening Mean Basis Functions

We denote our first screening function as

$$screenmean(y, basis\ vectors, split) \tag{55}$$

which takes as its argument an outcome, and the basis vectors. The screening process constructs all interactions, finding the 200 bases with the largest correlation, then bootstraps on $\mathcal{S}_0$ thirty times, and maintaining all bases selected by our high-dimensional regression in the original sample or any bootstrap.

For example, when constructing bases for f, we would use the bases

$$bases_f = screenmean(\mathbf{y}, \{\phi_{k,k'}(\mathbf{X}_j, \mathbf{X}_{j'})\}, \mathcal{S}_0) \tag{56}$$

D.3 Screening Interference Basis Functions

We then implement a screen for the interference basis functions.

$$constructinterference(y, basis\ vectors, split) \tag{57}$$

Here, it constructs all possible interferent-proximity bases using data in the split. At a first pass, it uses a rule-of-thumb bandwidth to reduce down the total number of combinations

down to 200. After this, it optimizes the bandwidth for every remaining pair (as this is computationally costly), and then follows the bootstrap trimming provided above.

For example, in the outcome model, we would generate these terms using

$$bases_{\phi_y} = constructinterference(y - \widehat{\mathsf{E}}(y|bases_f), \{\phi_{k,k^\vartheta}(\mathbf{X}_j, \mathbf{X}_{j^\vartheta})\}, \mathcal{S}_0) \tag{58}$$

where the conditional expectation is evaluated using only data in $\mathcal{S}_0$.

D.4 The High-Dimensional Regression

High-dimensional regression in this section will refer to the sparse regression of Ratkovic and Tingley (2017). We use the hierarchy

$$y_i|\mathbf{x}_i, \beta, \mathbf{z}_i, b\sigma^2 \sim \mathcal{N}(\mathbf{x}_i^> \beta + \mathbf{z}_i^> b, \sigma^2) \tag{59}$$

$$\beta_k|\lambda, w_k, \sigma \sim DE\left(\lambda w_k/\sigma\right) \tag{60}$$

$$\lambda^2|N, K \sim \Gamma\left(\alpha, 1\right) \tag{61}$$

$$w_k|\gamma \sim \text{generalizedGamma}\left(1, 1, \gamma\right) \tag{62}$$

$$\gamma \sim \exp(1) \tag{63}$$

$$b|\sigma_g^2, \sim \mathcal{N}(0_{jbj}, \sigma_g^2 I_{jbj}) \tag{64}$$

$$\sigma_g^2 \sim InverseGamma(0, 1) \tag{65}$$

where in this case $\mathbf{x}_i$ includes the covariates augmented by the basis functions while $\mathbf{z}_i$ is a vector for the random effect, σ_g^2 is its variance, and $|b|$ is the number of random effects.

The model is fit via EM, with the tuning parameter α picked to maximize a BIC statistic.

Importantly, this gives us an estimate of $\widehat{\text{Var}}(\widehat{\beta}|\cdot)$, which we can use to calculate $\widehat{\text{Var}}(\widehat{y})$, and it is these principal components we enter as controls.

D.5 The Hodges-Lehmann Estimator

We will combine estimates over repeated cross-fits using the Hodges-Lehmann estimator. Chernozhukov et al. (2018) suggest the median, which is not efficient, while the mean is susceptible to outliers. The Hodges-Lehmann estimator, which we denote $HL()$, is the median of pairwise averages. It has nice robustness, with a breakdown point of 0.27 (with 0 for the mean and .5 for the median), at little loss of efficiency (5% less efficient than the mean if the data are i.i.d. gaussian, as opposed to 57% for the median). We will denote as $HL()$ the Hodges-Lehmann estimate of a vector.

E Implementation Details: Split Sample

E.1 Split $\mathcal{S}_0$

In this split, we generate a set of candidate bases for each nuisance component, estimates $\widehat{v}$, and estimate bandwidth parameters for the interference components.

Specifically, we generate the following sets of nuisance function bases:

$$bases_f = screenmean(\mathbf{y}, \{\phi_{k,k^\vartheta}(\mathbf{X}_j, \mathbf{X}_{j^\vartheta})\}, \mathcal{S}_0) \tag{66}$$

$$bases_{\phi_y} = constructinterference(y - \widehat{\mathsf{E}}(y|bases_f), \{\phi_{k,k^\vartheta}(\mathbf{X}_j, \mathbf{X}_{j^\vartheta})\}, \mathcal{S}_0) \tag{67}$$

$$bases_{g_1} = screenmean(\mathbf{t}, \{\phi_{k,k^\vartheta}(\mathbf{X}_j, \mathbf{X}_{j^\vartheta})\}, \mathcal{S}_0) \tag{68}$$

$$\widehat{\mathbf{v}} = \mathbf{t} - \widehat{E}(\mathbf{t}|bases_{g_1}) \tag{69}$$

$$bases_{g_2} = screenmean(|\widehat{\mathbf{v}}|, \{\phi_{k,k^\vartheta}(\mathbf{X}_j, \mathbf{X}_{j^\vartheta})\}, \mathcal{S}_0) \tag{70}$$

$$bases_{\phi_t} = constructinterference(\mathbf{t} - \widehat{\mathsf{E}}(\mathbf{t}|bases_{g_1}, \widehat{\mathbf{v}} \odot bases_{g_2}), \{\phi_{k,k^\vartheta}(\mathbf{X}_j, \mathbf{X}_{j^\vartheta})\}, \mathcal{S}_0) \tag{71}$$

$$\widehat{\mathbf{v}}_2 = \widehat{\mathbf{v}} - \widehat{\mathsf{E}}(\widehat{\mathbf{v}}|bases_{\phi_t}) \tag{72}$$

Going through these, the first two $bases_f$ and $bases_{\phi_y}$ follow come from above, and $bases_{g_1}$ is similar to $bases_f$. Next, we need model the treatment heteroskedasticity and interference in the treatment. To do so, we want to look for any systematic trends in $|\mathbf{v}|$, the absolute value of the error residual, which gives us the bases in g_2. Here, $\widehat{\mathsf{E}}$ denotes the high-dimensional regression given above. Then we construct the interference bases using the residuals to regress the treatment variable on mean bases $bases_{g_1}$ and interactions between the treatment residuals $\widehat{\mathbf{v}}$ and the bases $bases_{g_2}$ (where $\odot$ denotes the elementwise interaction), using data in $\mathcal{S}_0$. We then update $\widehat{\mathbf{v}}$ by using instead the residuals after regressing using our high-dimensional regression on $bases_{\phi_t}$, giving us $\mathbf{v}_2$.

At this point, we have what we need to move to the next split: estimated treatment residuals $\widehat{\mathbf{v}}_2$ and interference bases $bases_{\phi_y}, bases_{\phi_t}$ where the bandwidth parameters have been estimated on subsample $\mathcal{S}_0$.

E.2 Split $\mathcal{S}_1$

The algorithm now effectively condenses into that of Chernozhukov et al. (2018). Here, all estimation is done only using data in $\mathcal{S}_1$.

We regress $\mathbf{y}$ on $\{bases_f, bases_{\phi_y}\}$, retaining the point estimate, selected bases, and principal components of $\widehat{\text{Var}}(\widehat{\mathbf{y}}|\cdot)$. We select the number of principal comoponents so as 90% of the variance in $\widehat{\text{Var}}(\widehat{\mathbf{y}}|\cdot)$. Specifically, if we denote as $\widehat{\beta}$ the estimated coefficients from this model and B the full bases set, we take the matrix

$$\widehat{\text{Var}}(\widehat{\mathbf{y}}) = B\widehat{\text{Var}}(\widehat{\beta}|B)B^{\top} \tag{73}$$

and a sufficient number of principal components to explain 90% of the variance (i.e. 90% of the explained variance, as you would find in a scree plot).

We then follow the same strategy for regressing $\widehat{\mathbf{t}}$ on $\{bases_{g_1}, \widehat{\mathbf{v}}_2 \odot bases_{g_2}, bases_{\phi_2}\}$.

We combine the point estimates, selected bases, and principal components into our matrix $\widehat{U}_{\widehat{u}}$. This matrix tends to be over-packed, so we use this subsample to regress $\mathbf{y}$ and $\mathbf{t}$ on $\widehat{U}_{\widehat{u}}$ and remove any unidentified columns due to collinearity. This matrix has been constructed in its entirety without touching any observations in $\mathcal{S}_2$, meaning we can take it to that sample for inference.

E.3 Split $\mathcal{S}_2$

We now regress $\mathbf{y}$ on $\mathbf{t}$ and $\widehat{U}_{\widehat{u}}$ using data on $\mathcal{S}_2$. The point estimate and standard error are saved. At default, the standard errors are $HC3$ standard errors, where the residuals in the standard error calculations are replaced by their leave-one-out estimates (Long and Ervin,

2000).[30]

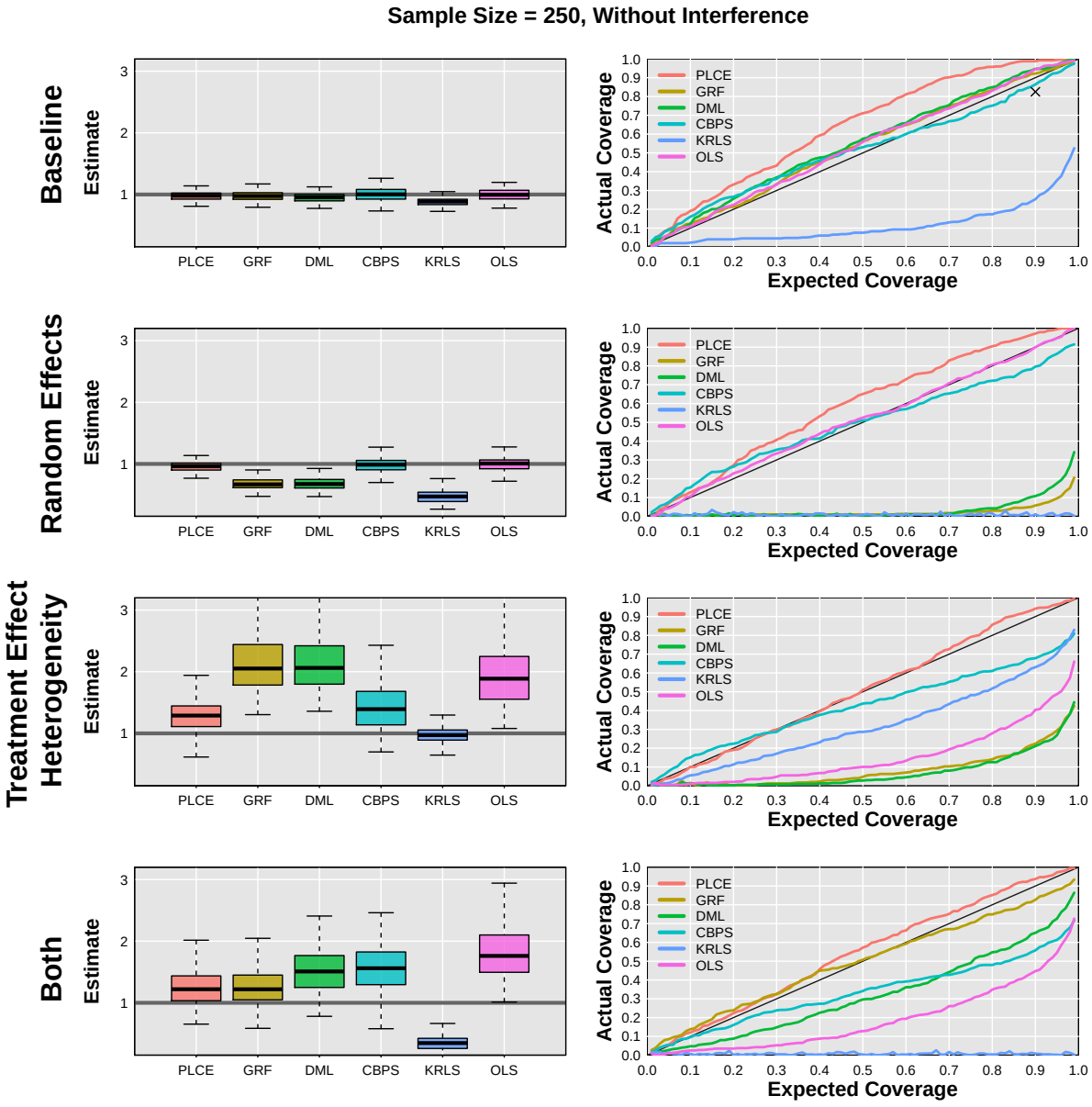
E.4 Cross-fitting and Repeated Cross-fitting

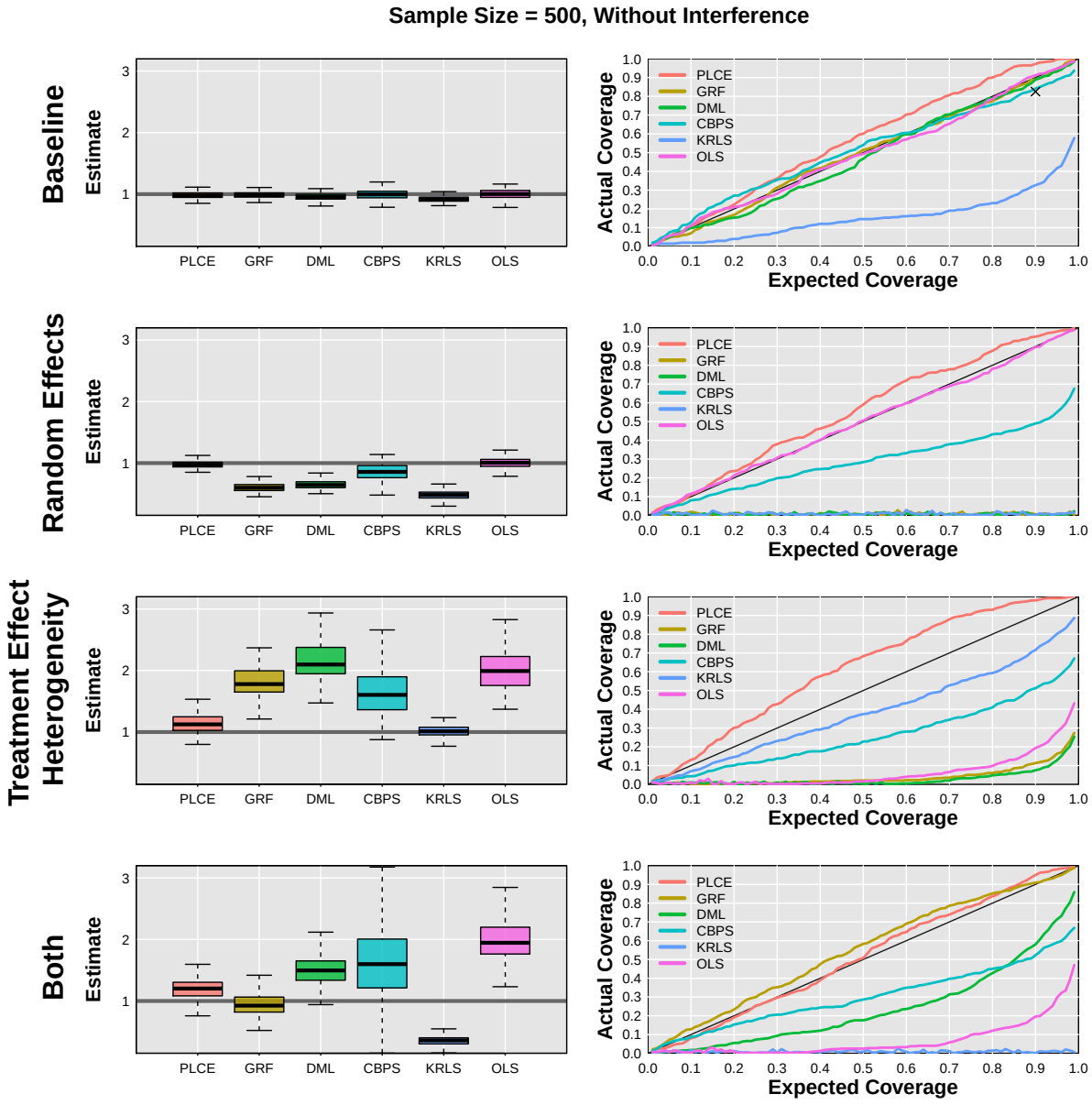
We then cross-fit once, swapping the roles of $\mathcal{S}_1$ and $\mathcal{S}_2$, as most of the computational time occurs in subsample $\mathcal{S}_0$. We average the point estimates and their variances for a given cross-fit, and then take the Hodges-Lehmann mean of each over the repeated cross-fits.

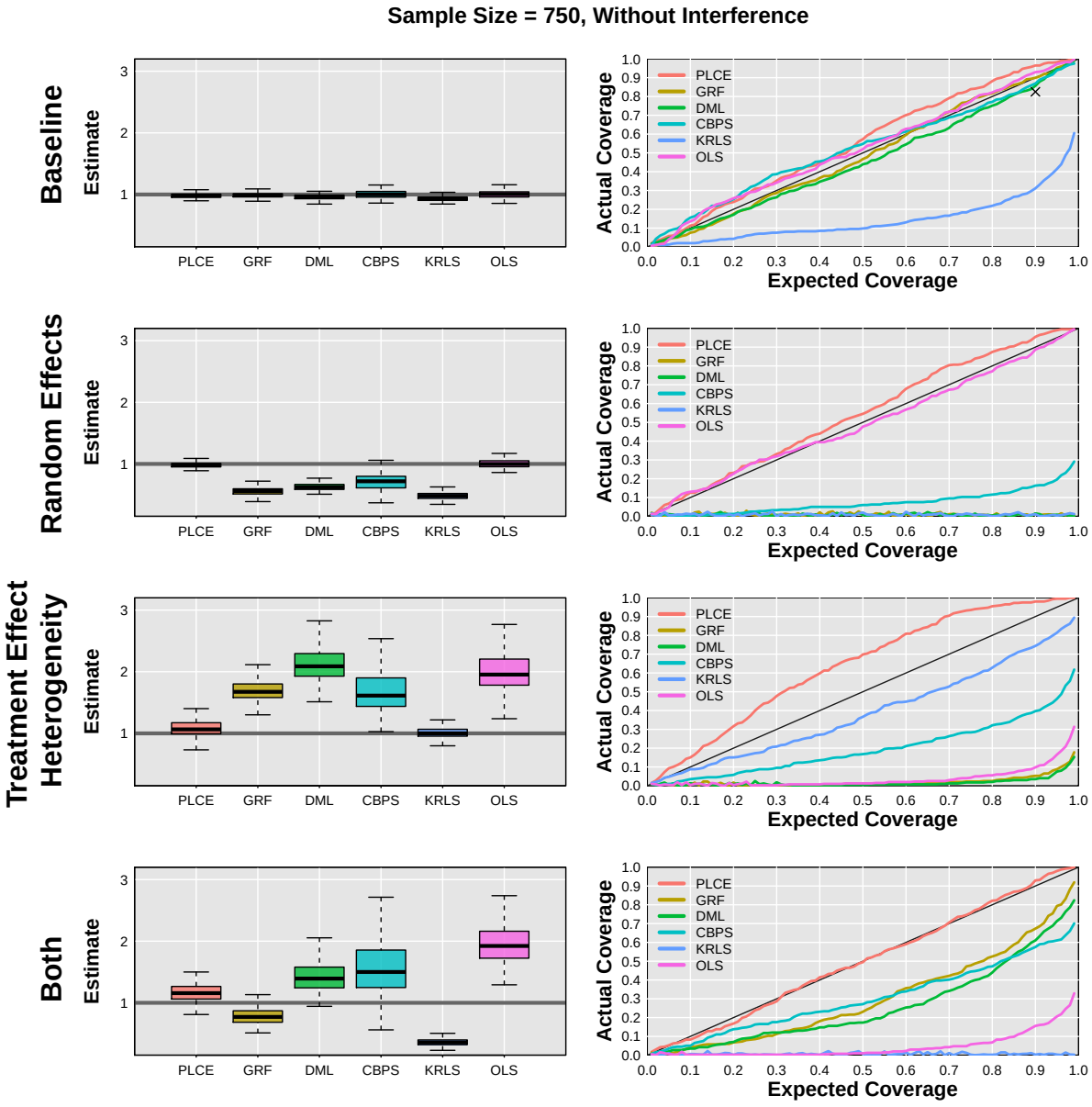
F Additional Simulations

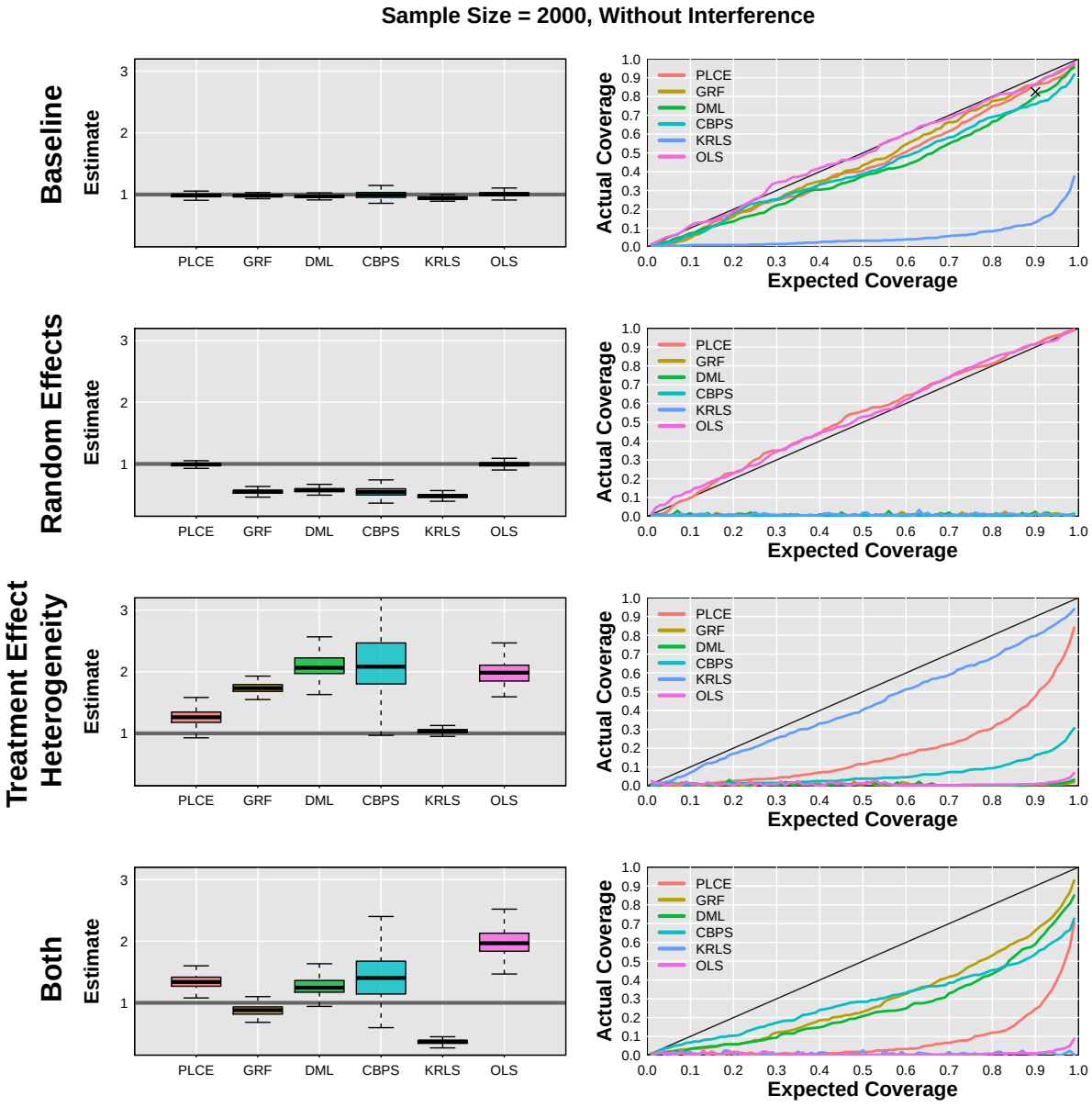
This appendix presents simulations for sample sizes $n \in \{250, 500, 750, 2000\}$ to supplement those in the text at $n = 1000$.

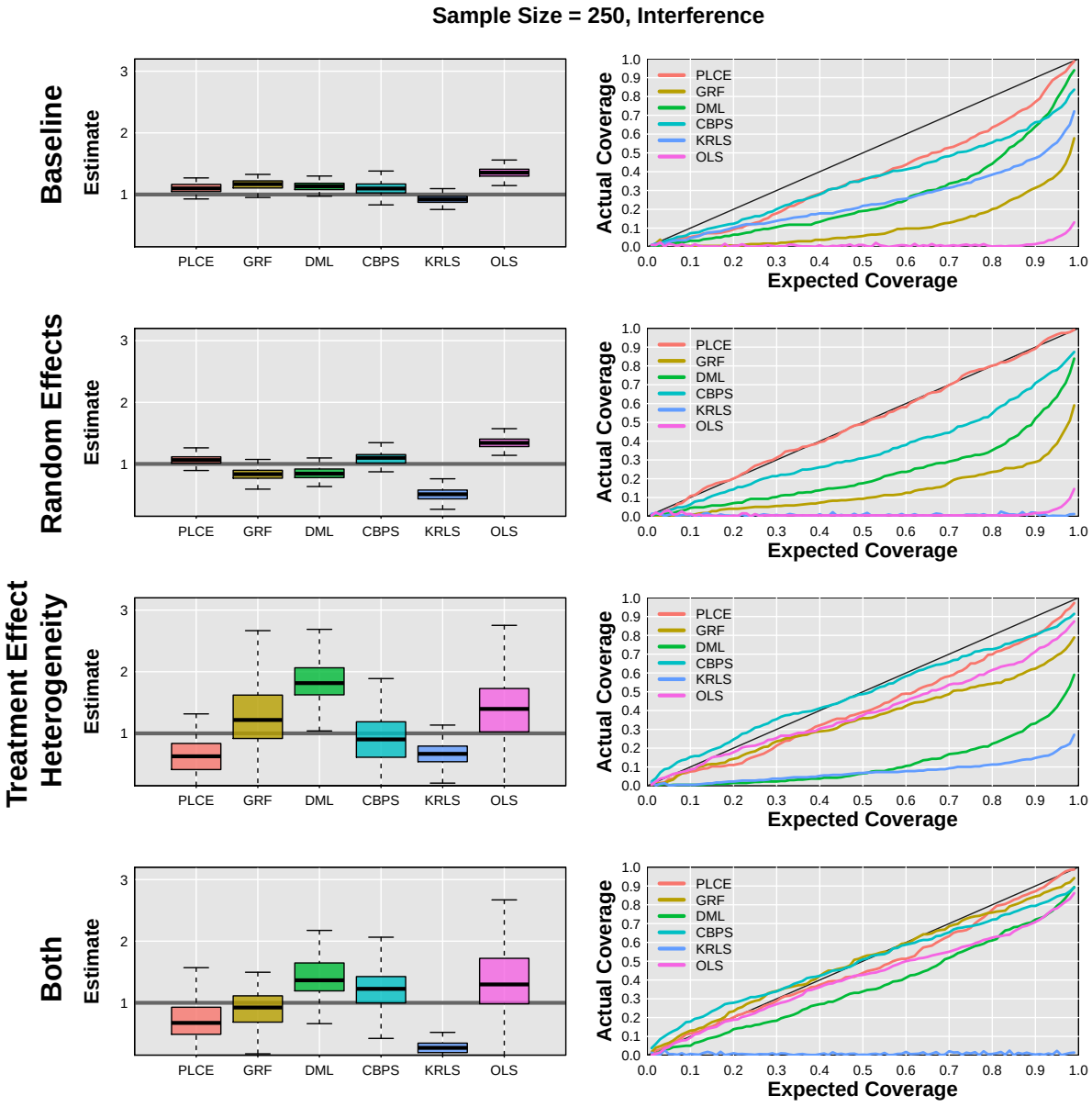
[30] We do so due to the uncertainty over the covariate set. We note that we implemented $HC0$, or the standard robust standard errors, in the Mattes and Weeks (2019) replication, to match their specification.

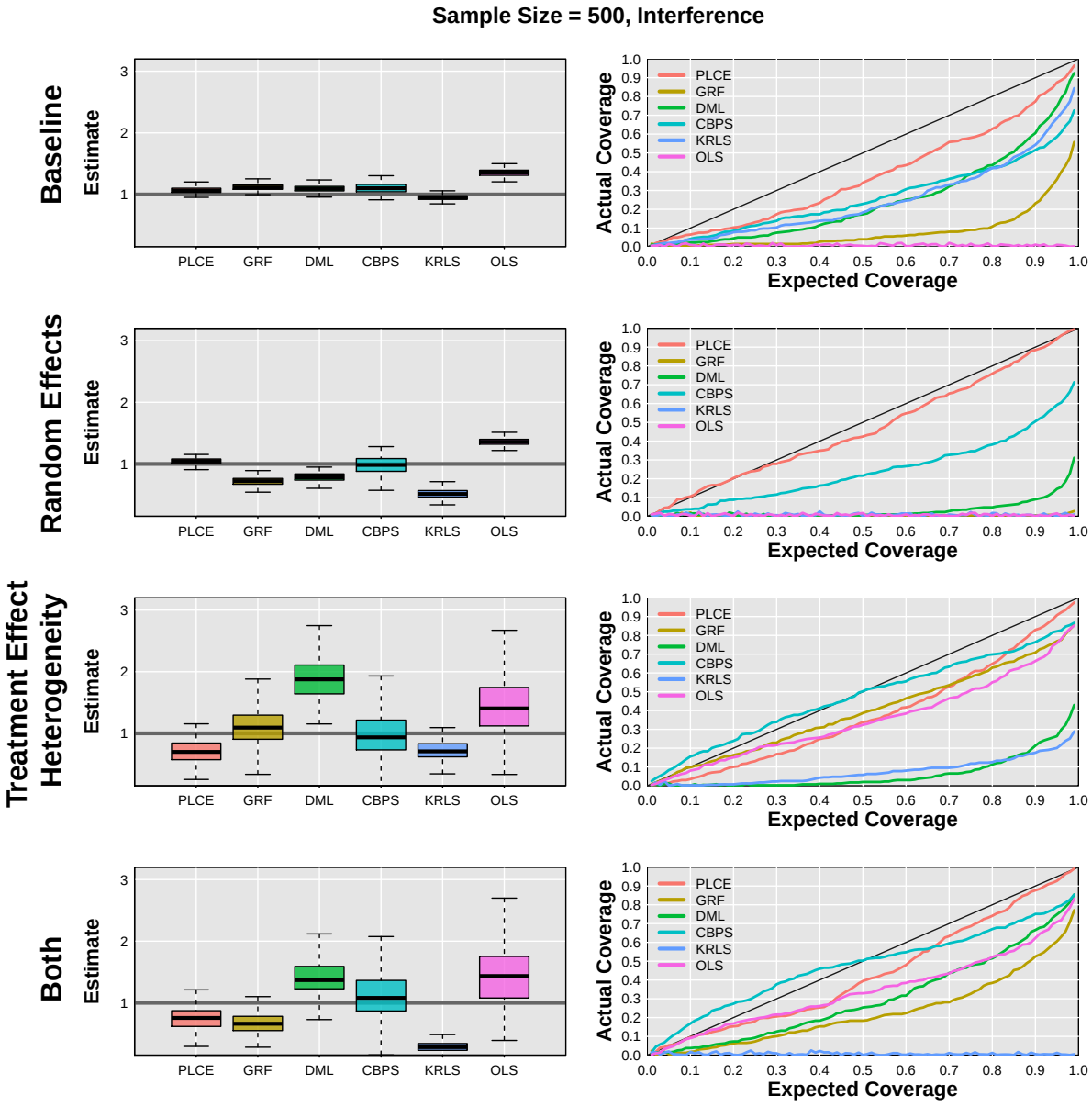

Sample Size = 250, Without Interference
Baseline
Random Effects
Treatment Effect Heterogeneity
Both
Estimate
PLCE
GRF
DML
CBPS
KRLS
OLS
Actual Coverage
Expected Coverage

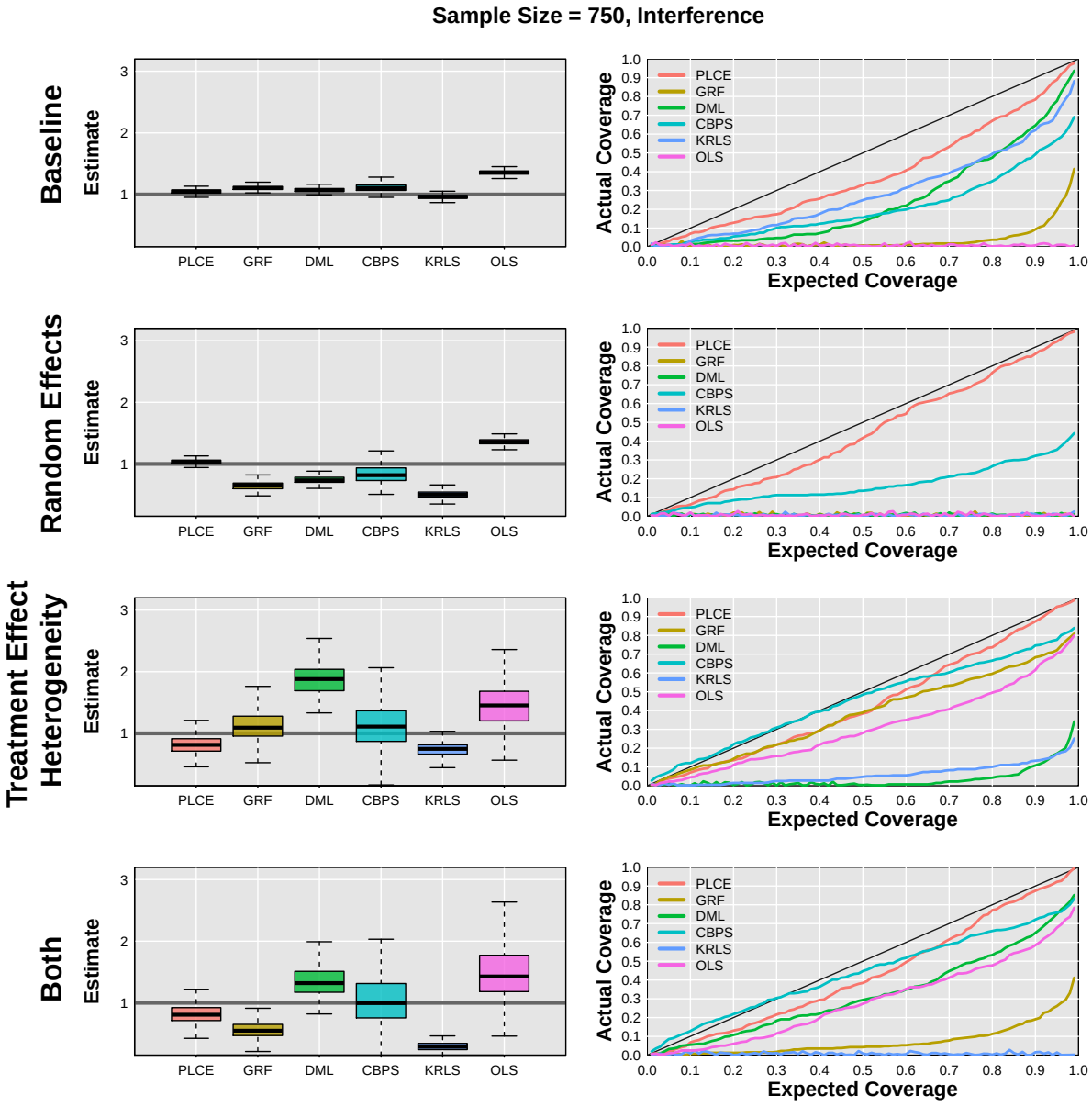

Sample Size = 500, Without Interference
Baseline
Random Effects
Treatment Effect Heterogeneity
Both
Estimate
PLCE
GRF
DML
CBPS
KRLS
OLS
Actual Coverage
Expected Coverage

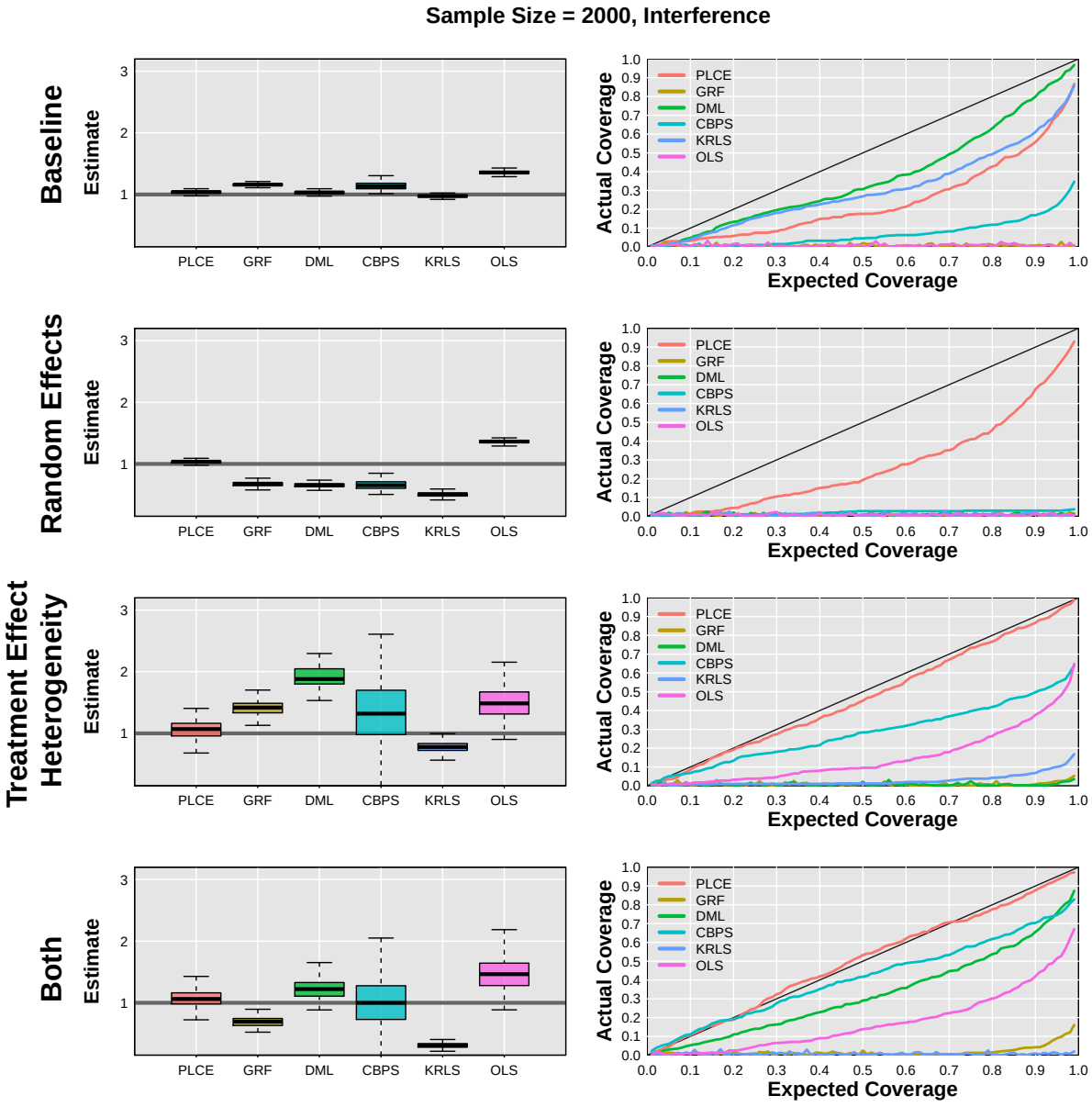

Sample Size = 750, Without Interference
Baseline
Random Effects
Treatment Effect Heterogeneity
Both
Estimate
PLCE
GRF
DML
CBPS
KRLS
OLS
Actual Coverage
Expected Coverage


Sample Size = 2000, Without Interference
Baseline
Random Effects
Treatment Effect Heterogeneity
Both
Estimate
PLCE
GRF
DML
CBPS
KRLS
OLS
Actual Coverage
Expected Coverage


Sample Size = 250, Interference
Baseline
Random Effects
Treatment Effect Heterogeneity
Both
Estimate
PLCE
GRF
DML
CBPS
KRLS
OLS
Actual Coverage
Expected Coverage


Sample Size = 500, Interference
Baseline
Random Effects
Treatment Effect Heterogeneity
Both
Estimate
PLCE
GRF
DML
CBPS
KRLS
OLS
Actual Coverage
Expected Coverage


Sample Size = 750, Interference
Baseline
Random Effects
Treatment Effect Heterogeneity
Both
Estimate
PLCE
GRF
DML
CBPS
KRLS
OLS
Actual Coverage
Expected Coverage


Sample Size = 2000, Interference
Baseline
Random Effects
Treatment Effect Heterogeneity
Both
Estimate
PLCE
GRF
DML
CBPS
KRLS
OLS
Actual Coverage
Expected Coverage