Supplemental Materials for "Relaxing Assumptions, Improving Inference: Integrating Machine Learning and the Linear Regression"

August 30, 2022

The appendices in this document provide a self-contained introduction to semiparametric efficiency and the proposed method (Appendix A); a discussion of the necessary assumptions required to give the estimate a causal interpretation (Appendix B); a set of diagnostic tests implemented with the **PLCE** software (Appendix C); implementation details, both preliminary and detailed (Appendices D, E); and a set of simulation results extending those in the manuscript F.

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

In order to integrate the technical discussion with the broader statistical literature, I 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 $Px_i = \mathbb{E}(x_i)$ and G_n the empirical process $G_n x_i = \sqrt{n}(P_n x_i - Px_i)$. The remaining notation is as in the body.

A.1 Characterizing the Semiparametric Efficiency Bound

The semiparametrically efficient estimate can be calculated treating the model where I 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 $\mathbf{X}_\eta(\mathbf{X}_\eta^\top\mathbf{X}_\eta)^{-1}\mathbf{X}_\eta^\top$, 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.

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

The semiparametrically efficient estimate is then

$$\widehat{\theta} = \frac{\widetilde{\mathbf{y}}^{\top}\widetilde{\mathbf{t}}}{\widetilde{\mathbf{t}}^{\top}\widetilde{\mathbf{t}}} \tag{1}$$

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

$$= \frac{P_n(\theta \widetilde{u}_i^2 + \widetilde{u}_i \widetilde{e}_i)}{P_n \widetilde{u}_i^2} \tag{3}$$

$$= \theta + \frac{P_n \widetilde{u}_i \widetilde{e}_i}{P_n \widetilde{u}_i^2} \tag{4}$$

The estimator is clearly consistent for θ by the law of large numbers with limiting distribution

$$\sqrt{n}\left(\widehat{\theta} - \theta\right) = G_n \widehat{\theta} = \sqrt{n}\left\{\frac{P_n \widetilde{u}_i \widetilde{e}_i}{P_n \widetilde{u}_i^2}\right\} \tag{5}$$

$$\rightsquigarrow \mathcal{N}\left(0, \frac{\mathbb{E}\left(\widetilde{u}_i^2 \widetilde{e}_i^2\right)}{\mathbb{E}\left(\widetilde{u}_i^2\right)^2}\right). \tag{6}$$

This gives the semiparametric efficiency bound for the model. Now, were f, g known, least squares could recover a point and variance estimate through the method of least squares, and it would be efficient. Since f, g are not known, I next construct an estimate that is asymptotically indistinguishable from the estimate calculated from the parametric submodel. This estimate will be semiparametrically efficient.

A.2 Deviations Between the Feasible Estimate and the Semiparametrically Efficient Estimator

The functions f, g are not known but instead estimated as $\widehat{f}, \widehat{g}$, introducing $\Delta_{\widehat{f}}, \Delta_{\widehat{g}}$ into the linear regression. The argument follows exactly as above, except these approximation error

terms must be accounted for. Writing the models in terms of $\widehat{f}, \widehat{g}$, and hence in terms of f, g and $\Delta_{\widehat{f}}, \Delta_{\widehat{g}}$, gives

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

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

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

$$\widetilde{\mathbf{t}} = \mathbf{A}_\eta\mathbf{t} = \underbrace{\mathbf{A}_\eta\mathbf{X}_\Delta\beta_y}_{:=\widetilde{\Delta}_t} + \underbrace{\mathbf{A}_\eta\mathbf{u}}_{:=\widetilde{\mathbf{u}}} \tag{7}$$

$$= \widetilde{\mathbf{u}} + \widetilde{\Delta}_t \tag{8}$$

$$\widetilde{\mathbf{y}} = \mathbf{A}_\eta y = \theta\mathbf{A}_\eta\mathbf{t} + \underbrace{\mathbf{A}_\eta\mathbf{X}_\Delta\beta_y}_{:=\widetilde{\Delta}_y} + \underbrace{\mathbf{A}_\eta\mathbf{e}}_{:=\widetilde{\mathbf{e}}} \tag{9}$$

$$= \theta\widetilde{\mathbf{t}} + \widetilde{\mathbf{e}} \tag{10}$$

$$= \theta\widetilde{\mathbf{u}} + \theta\widetilde{\Delta}_t + \widetilde{\Delta}_y + \widetilde{\mathbf{e}} \tag{11}$$

These partialed-out values can be used to construct

$$\sqrt{n}(\widehat{\theta} - \theta) = G_n\widehat{\theta} = \sqrt{n}\left(\frac{P_n\widetilde{t}_i\widetilde{y}_i}{P_n\widetilde{t}_i^2} - \theta\right) \tag{12}$$

Beginning with the denominator,

$$P_n\widetilde{t}_i^2 = P_n\left\{\widetilde{u}_i^2 + 2\widetilde{u}_i\widetilde{\Delta}_{\widehat{g},i} + \widetilde{\Delta}_{\widehat{g},i}^2\right\} \xrightarrow{\text{u}} \mathbb{E}(u_i^2) \tag{13}$$

by the uniform consistency of $\widehat{f}, \widehat{g}$. A uniform Slutsky's theorem can then be used to characterize the limiting behavior as

$$\sqrt{n}(\widehat{\theta} - \theta) = G_n\widehat{\theta} = \sqrt{n}\left(\frac{P_n\widetilde{t}_i\widetilde{y}_i}{\mathbb{E}(\widetilde{u}_i^2)} - \theta\right) \tag{14}$$

and expanding the numerator gives,

$$\sqrt{n}(\widehat{\theta} - \theta) = G_n\widehat{\theta} = \sqrt{n}\left(\underbrace{\frac{\theta P_n\widetilde{u}_i^2 + P_n\widetilde{u}_i\widetilde{e}_i}{P\widetilde{u}_i^2}}_{\text{efficient estimate}} - \theta\right) + \underbrace{\sqrt{n}\frac{B}{P\widetilde{u}_i^2}}_{\text{bias terms}} \tag{15}$$

$$\text{where } B = P_n\left\{2\theta\widetilde{u}_i\widetilde{\Delta}_{\widehat{g},i} + \widetilde{u}_i\widetilde{\Delta}_{\widehat{f},i} + \widetilde{\Delta}_{\widehat{g},i}\widetilde{e}_i + \theta\widetilde{\Delta}_{\widehat{g},i}^2 + \widetilde{\Delta}_{\widehat{g},i}\widetilde{\Delta}_{\widehat{f},i}\right\} \tag{16}$$

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 $\sqrt{n}B \xrightarrow{\text{u}} 0$. 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

$$\sqrt{n}P_n u_i\Delta_{\widehat{f},i} \xrightarrow{\text{u}} 0, \quad \sqrt{n}P_n u_i\Delta_{\widehat{g},i} \xrightarrow{\text{u}} 0 \tag{17}$$

and the third when

$$\sqrt{n}P_n e_i\Delta_{\widehat{f},i} \xrightarrow{\text{u}} 0, \quad \sqrt{n}P_n e_i\Delta_{\widehat{g},i} \xrightarrow{\text{u}} 0. \tag{18}$$

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.[1] See van der Vaart (1998, ch. 25) for more. The last two bias terms involve square and cross-products of the approximation error terms,

$$\sqrt{n}P_n\Delta^2_{\widehat{f},i}, \quad \sqrt{n}P_n\Delta^2_{\widehat{g},i}; \quad \sqrt{n}P_n\Delta_{\widehat{f},i}\Delta_{\widehat{g},i}. \tag{19}$$

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

$$n^{1/4}\sqrt{P_n\Delta^2_{\widehat{g},i}} \xrightarrow{\text{u}} 0; \quad n^{1/4}\sqrt{P_n\Delta^2_{\widehat{f},i}} \xrightarrow{\text{u}} 0, \tag{20}$$

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.3 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.

Reducing the rate from $n^{1/4}$ to $n^{1/8}$ requires 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,

[1] 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 principal components are constructed from cross-observation covariances,

$$f(\mathbf{x}_i) \approx \widehat{f}(\mathbf{x}_i) + \widehat{U}_{\widehat{f},i}^{\top}\gamma_f \tag{21}$$

$$f(\mathbf{x}_i) = \widehat{f}(\mathbf{x}_i) + \sum_{i'=1}^{n} \widehat{\mathrm{Cov}}(\widehat{f}(\mathbf{x}_i), \widehat{f}(\mathbf{x}_{i'}))w_{i'} \tag{22}$$

for scalars $\widehat{w}_{i'}$.[2] The finite-dimensional assumption of the $U_{f,i}$ constrains $w_{i'}$, 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 estimating the covariates from principal components will introduce terms like

$$\sqrt{n}P_n \left\{ \widehat{\mathrm{Cov}}(\widehat{f}(\mathbf{x}_i), \widehat{f}(\mathbf{x}_{i'})) - \mathrm{Cov}(\widehat{f}(\mathbf{x}_i), \widehat{f}(\mathbf{x}_{i'})) \right\}^2 \tag{23}$$

into the 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 $\widehat{f}(\mathbf{x}_i), \widehat{f}(\mathbf{x}_{i'})$ are both converging at $n^{1/8}$, their product in the covariance is converging at $n^{1/4}$.

Essential to this argument is the finite-dimensional assumption on the covariance matrix. This particular assumption, sometimes termed "sufficient dimension reduction," (see Section 7.1.1 of the main body), makes convergence of the sums described above tractable. Under this assumption, argument can follow dimension-by-dimension by a Cramer-Wold device van der Vaart (1998). This finite-dimensional assumption, as a theoretical matter, is crude but as a practical matter aligns with the method's approach, where a control vector is simply

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

entered into a reduced form regression. For a more general theoretical discussion, see the citations in the main manuscript.

If the finite dimensional assumption is correct then the method achieves second-order efficiency by fully capturing the variance in the approximation errors. If this assumption is not correct, though, the method still recovers a semiparametrically efficient estimate and–due to the split sample strategy–the principal components should help or, at worst, increase the variance of the estimates. The simulations and applied examples provide compelling evidence that the approach is reasonable.

B Causal Assumptions

In giving these assumptions, I utilize the *potential outcomes notation*, 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, $y_i(t)$. I will denote as $\mathbf{X}_{-i}, \mathbf{t}_{-i}$ the background covariates and treatments for all observations except observation i.

ASSUMPTION 1 *Causal Assumptions*

1. *Stable treatment value: There is a single version of each treatment value.*

2. *Positivity: The treatment is not deterministic,* $\text{Var}(t_i|\mathbf{x}_i, \mathbf{X}_{-i}) > 0$ *for all observations.*

3. *Ignorability:*[3] $t_i \perp\!\!\!\perp \mathbf{t}_{-i}|\mathbf{x}_i, \mathbf{X}_{-i}$ *and* $y_i(t_i, \mathbf{t}_{-i}) \perp\!\!\!\perp t_i|\mathbf{t}_{-i}, \mathbf{x}_i, \mathbf{X}_{-i}$

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

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

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.

Equating the causal and descriptive parameters requires 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, only considering observation i with covariate values $\mathbf{x}_i, \mathbf{X}_{-i}, \mathbf{t}_{-i}$ allows t_i to move freely of the other treatment and of any unobserved confounders. This is the 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, the marginal effect and causal effect can be equated as

$$\theta_i = \underbrace{\frac{\text{Cov}(y_i, t_i | \mathbf{x}_i, \mathbf{X}_{-i}, \mathbf{t}_{-i})}{\text{Var}(t_i | \mathbf{x}_i, \mathbf{X}_{-i}, \mathbf{t}_{-i})}}_{\text{Partial Effect}} = \underbrace{\frac{\text{Cov}_{\mathcal{T}}(y_i(t_i, \mathbf{t}_{-i}), t_i | \mathbf{t}_{-i})}{\text{Var}_{\mathcal{T}}(t_i | \mathbf{t}_{-i})}}_{\text{Causal Effect}}. \tag{24}$$

The estimand is well-defined by the stable treatment assumption; its denominator is nonzero by the postivity assumption; and the ignorability assumption equates the numerators and denominators. Equating the marginal effect and observation-level effect for each observation equates their averages.

C Diagnostics

I implement a sensitivity analysis in order to assess how strong an unobserved confounder must be in order to overturn any results. Since the method is, in effect, a linear regression

in subsample $\mathcal{S}_2$, diagnostics for the linear regression are applicable.

I implement the recent method of Cinelli and Hazlett (2020) in the software. Following the authors' suggestion, the software 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 $R^2_{Y\sim D|X}$, 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.

I also assess the positivity assumption. Positivity is violated when the treatment variable is a deterministic function of the covariates. I do so by graphically assessing the kurtosis of the residuals (Wooldridge, 2013, Appendix B, p. 737.).[4] Denoting as $\widehat{\epsilon}_{i,s}$ the residual from estimating the treatment for observation i on repeated cross-fit iteration s, the method estimates the kurtosis $\widehat{\kappa}_i$ as

$$\widehat{\kappa}_i = \frac{\frac{1}{S}\sum_{s=1}^{S}\widehat{\epsilon}_{i,s}^4}{(\frac{1}{S}\sum_{s=1}^{S}\widehat{\epsilon}_{i,s}^2)^2}. \tag{25}$$

The excess kurtosis is the extent that this statistic falls above that expected from a normal distribution, and the software plots them from high to low. The lefthand side of Figure 1 contains five possible error densities. The first one is thin-tailed and raises the deepest

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

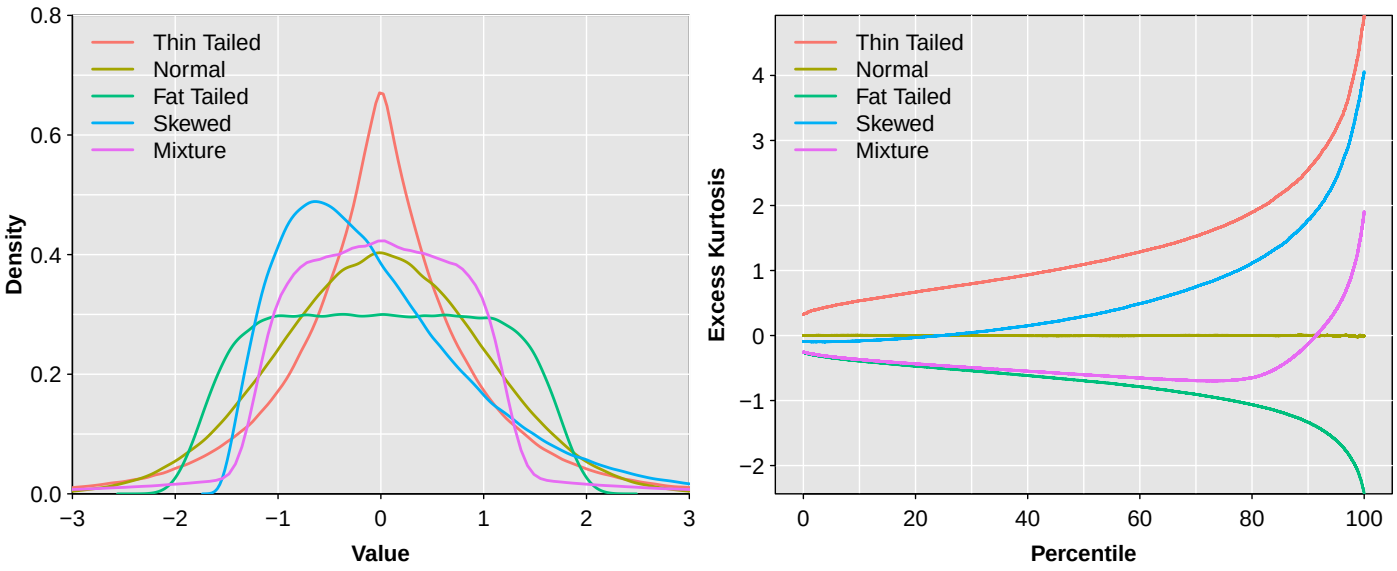

Figure 1: **Excess kurtosis plot for diagnosing positivity.** The lefthand side presents the five different error densities assessed on the right. The first one is thin-tailed and raises the deepest concerns about violating positivity. The next is normal, then a fat-tailed and skewed density. The last 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.

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 returned by my software.

D Implementation Details: Preliminaries

In this section, I give an overview of the estimation strategy. The software itself is public and can be expected. The results presented here were generated using the software submitted in the replication file and will be available via the `APSR` branch of the software's github at `https://github.com/ratkovic/PLCE/tree/APSR`. The publicly available version via `CRAN` may be updated over time and results may not match exactly those reported in this work.

D.1 Preliminaries: Basis Functions

D.1.1 B-Spline Basis Functions

The software adjusts 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, include along with the original variable is a set of degree 3 B-splines with different knots along its range, see Figure 2.

I denote the k^{th} of these transformations applied to covariate $\mathbf{X}_{\cdot j}$ as

$$\mathbf{X}_{\cdot j} \mapsto \phi_k(\mathbf{X}_{\cdot j})$$

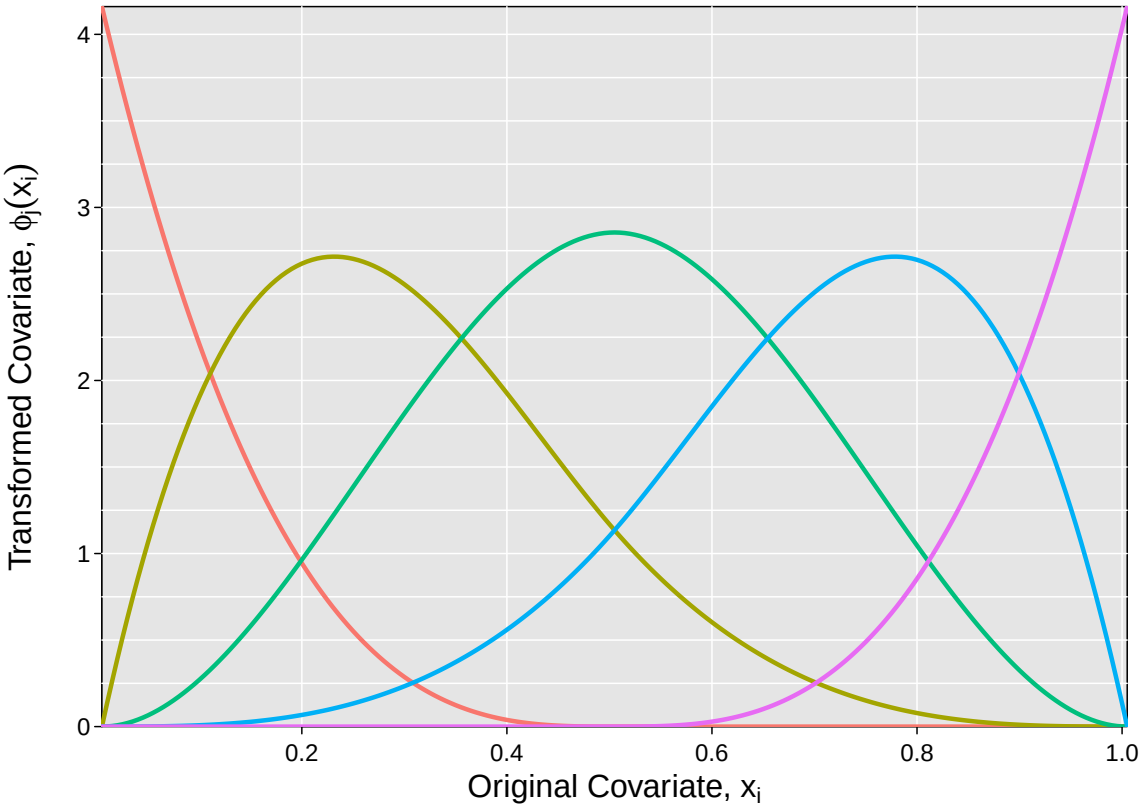

Figure 2: **Nonlinear transformations of each variable used to construct basis functions.**

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

$$\phi_0(\mathbf{X}_{\cdot j}) = \mathbf{1}_n; \quad \phi_1(\mathbf{X}_{\cdot j}) = \mathbf{X}_{\cdot j}$$

Including the intercept and linear term with the five nonlinear transformation, seven terms are generated from each covariate.

D.1.2 Constructing Basis Functions for the Nuisance Functions

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

$$\phi_{k,k'}(\mathbf{X}_{\cdot j}, \mathbf{X}_{\cdot j'}) = \phi_k(\mathbf{X}_{\cdot j})\phi_{k'}(\mathbf{X}_{\cdot j'})$$

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

$$\phi_{k,k'}(\mathbf{X}_{\cdot j}, \mathbf{X}_{\cdot j'}, \widehat{v}) = \widehat{v}\phi_k(\mathbf{X}_{\cdot j})\phi_{k'}(\mathbf{X}_{\cdot j'})$$

where $\widehat{v} = \mathbf{t} - \widehat{E}(\mathbf{t}|\mathbf{X})$, an estimated residual. I return to how I estimate the residuals below, but for now 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. I consider how close observation i' is to observation i, as a function of how close $\psi_k(x_{ij})$ is to $\psi_k(x_{i'j})$, with bandwidth ν_{jk} as

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

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

The interferent may be driven by an entirely different basis function and variable,$\phi_{k'}$ and

$\mathbf{X}_{\cdot j'}$,

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

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

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

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

I reduce the bases above down to a reasonable number for a linear regression in two ways: through a correlation screen and then fitting a high-dimensional regression to these selected bases. I give specifics below, but provide an overview of the strategies here.

D.2 Screening Mean Basis Functions

I denote the first screening function as

$$screenmean(y, basis\ vectors, split) \tag{26}$$

which takes as its argument an outcome, and the basis vectors.

For example constructing bases for f uses the bases

$$bases_f = screenmean(\mathbf{y}, \{\phi_{k,k'}(\mathbf{X}_{\cdot j}, \mathbf{X}_{\cdot j'})\}, \mathcal{S}_0) \tag{27}$$

The screening process constructs all interactions, finding the between 50 and 400 bases (growing in sample size) with the largest correlation, then uses a call to glmnet (the LASSO) to maintain a subset of these.

D.3 Screening Interference Basis Functions

I then implement a screen for the interference basis functions.

$$constructinterference(y, basis\ vectors, split) \tag{28}$$

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 glmnet/LASSO trimming provided above.

For example, in the outcome model, the software generates these terms using

$$bases_{\phi_y} = constructinterference(y - \widehat{\mathbb{E}}(y|bases_f), \{\phi_{k,k'}(\mathbf{X}_{\cdot j}, \mathbf{X}_{\cdot j'})\}, \mathcal{S}_0) \tag{29}$$

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). I use the hierarchy

$$y_i|\mathbf{x}_i, \beta, \mathbf{z}_i, b\sigma^2 \sim \mathcal{N}(\mathbf{x}_i^\top \beta + \mathbf{z}_i^\top b, \sigma^2) \tag{30}$$

$$\beta_k|\lambda, w_k, \sigma \sim DE\left(\lambda w_k/\sigma\right) \tag{31}$$

$$\lambda^2|N, K \sim \Gamma\left(\alpha, 1\right) \tag{32}$$

$$w_k|\gamma \sim \text{generalizedGamma}\left(1, 1, \gamma\right) \tag{33}$$

$$\gamma \sim \exp(1) \tag{34}$$

$$b|\sigma_g^2, \sim \mathcal{N}(0_{|b|}, \sigma_g^2 I_{|b|}) \tag{35}$$

$$\sigma_g^2 \sim InverseGamma(0, 1) \tag{36}$$

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 an estimate of $\widehat{\text{Var}}(\widehat{\beta}|\cdot)$, which is then used to calculate $\widehat{\text{Var}}(\widehat{y})$, and it is these principal components that are entered as controls.

D.5 The Hodges-Lehmann Estimator

I combine estimates over repeated cross-fits using the Hodges-Lehmann estimator. Double Machine Learning implements the median, which is not efficient, while the mean is not robust to outliers. The Hodges-Lehmann estimator, which I 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). I 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, I generate a set of candidate bases for each nuisance component, estimates $\widehat{v}$, and estimate bandwidth parameters for the interference components.

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

$$bases_f = screenmean(\mathbf{y}, \{\phi_{k,k'}(\mathbf{X}_{\cdot j}, \mathbf{X}_{\cdot j'})\}, \mathcal{S}_0) \tag{37}$$

$$bases_{\phi_y} = constructinterference(y - \widehat{\mathbb{E}}(y|bases_f), \{\phi_{k,k'}(\mathbf{X}_{\cdot j}, \mathbf{X}_{\cdot j'})\}, \mathcal{S}_0) \tag{38}$$

$$bases_{g_1} = screenmean(\mathbf{t}, \{\phi_{k,k'}(\mathbf{X}_{\cdot j}, \mathbf{X}_{\cdot j'})\}, \mathcal{S}_0) \tag{39}$$

$$\widehat{\mathbf{v}} = \mathbf{t} - \widehat{E}(\mathbf{t}|bases_{g_1}) \tag{40}$$

$$bases_{g_2} = screenmean(|\widehat{\mathbf{v}}|, \{\phi_{k,k'}(\mathbf{X}_{\cdot j}, \mathbf{X}_{\cdot j'})\}, \mathcal{S}_0) \tag{41}$$

$$bases_{\phi_t} = constructinterference(\mathbf{t} - \widehat{\mathbb{E}}(\mathbf{t}|bases_{g_1}, \widehat{\mathbf{v}} \odot bases_{g_2}), \{\phi_{k,k'}(\mathbf{X}_{\cdot j}, \mathbf{X}_{\cdot j'})\}, \mathcal{S}_0) \tag{42}$$

$$\widehat{\mathbf{v}}_2 = \widehat{\mathbf{v}} - \widehat{\mathbb{E}}(\widehat{\mathbf{v}}|bases_{\phi_t}) \tag{43}$$

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, I model the treatment heteroskedasticity and interference in the treatment. To do so, I want to look for any systematic trends in $|\mathbf{v}|$, the absolute value of the

error residual, which gives the bases in g_2. Here, $\widehat{\mathbb{E}}$ denotes the high-dimensional regression given above. Then I 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 $\mathcal{S}_0$. I then update $\widehat{\mathbf{v}}$ by using instead the residuals after regressing using the high-dimensional regression on $bases_{\phi_t}$, giving $\mathbf{v}_2$. At this point, I have what is needed 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 Double Machine Learning Here, all estimation is done only using data in $\mathcal{S}_1$. I regress $\mathbf{y}$ on $\{bases_f, bases_{\phi_y}\}$, retaining the point estimate, selected bases, and principal components of $\widehat{\mathrm{Var}}(\widehat{\mathbf{y}}|\cdot)$. I select the number of principal comoponents so as to include 90% of the variance in $\widehat{\mathrm{Var}}(\widehat{\mathbf{y}}|\cdot)$. Specifically, if I denote as $\widehat{\beta}$ the estimated coefficients from this model and B the full bases set, I take the matrix

$$\widehat{\mathrm{Var}}(\widehat{\mathbf{y}}) = B\widehat{\mathrm{Var}}(\widehat{\beta}|B)B^\top \tag{44}$$

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

I combine the point estimates, selected bases, and principal components into the matrix $\widehat{U}_{\widehat{u}}$. This matrix may not be full rank, with all of the elements coming in, so I 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 that sample can be used for inference.

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

I 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, with $HC3$ standard errors at default.[5] The diagnostics of Cinelli and Hazlett (2020) are run on this subsample.

E.4 Cross-fitting and Repeated Cross-fitting

I then cross-fit once, swapping the roles of $\mathcal{S}_1$ and $\mathcal{S}_2$, as most of the computational time occurs in subsample $\mathcal{S}_0$. I 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 $n = 1000$. The proposed method performs well across sample sizes. The settings with interference carry the same qualitative results as that in the body, where the proposed method performs well across settings and sample size. The same holds for the setting without interference. The proposed method still maintains some bias with the effect heterogeneity at the largest sample size, but outperforms the other methods in terms of bias, and offers estimates that are most robust to the inclusion or exclusion of random effects.

[5] I do so due to the uncertainty over the covariate set. Note that I implemented $HC0$, or the standard robust standard errors, in the experimental replication in the main body, to match the original authors' specification.

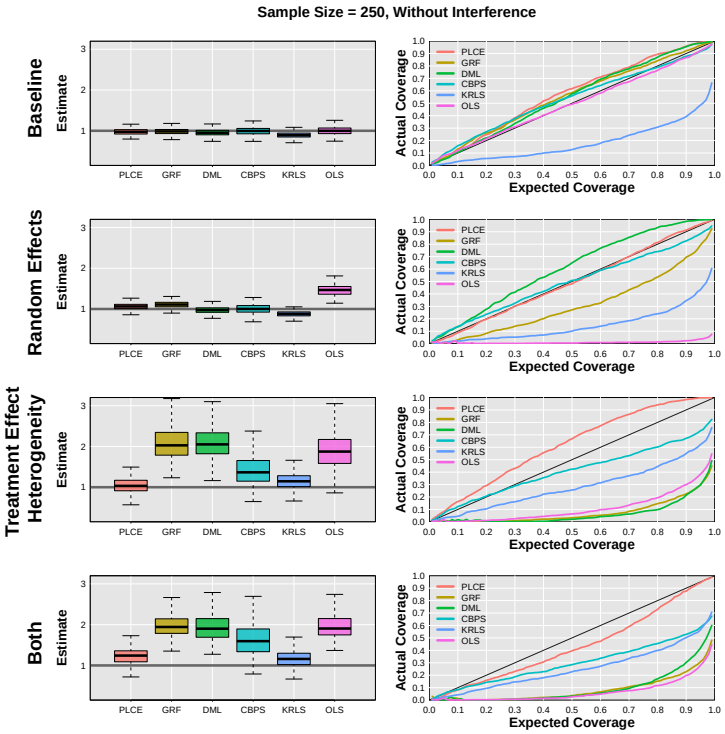

Figure 3: **Simulation Results Without Interference,** $n = 250$

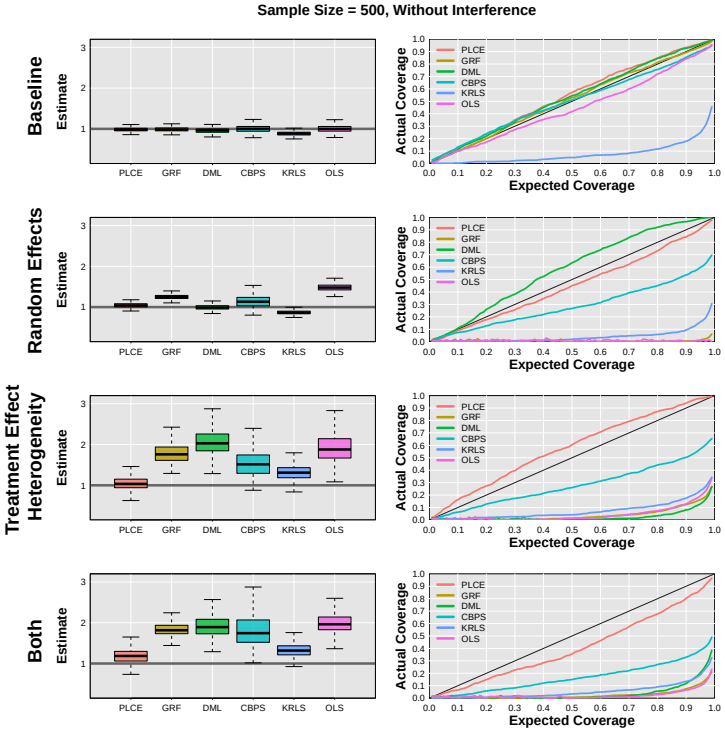

Figure 4: **Simulation Results Without Interference,** $n = 500$

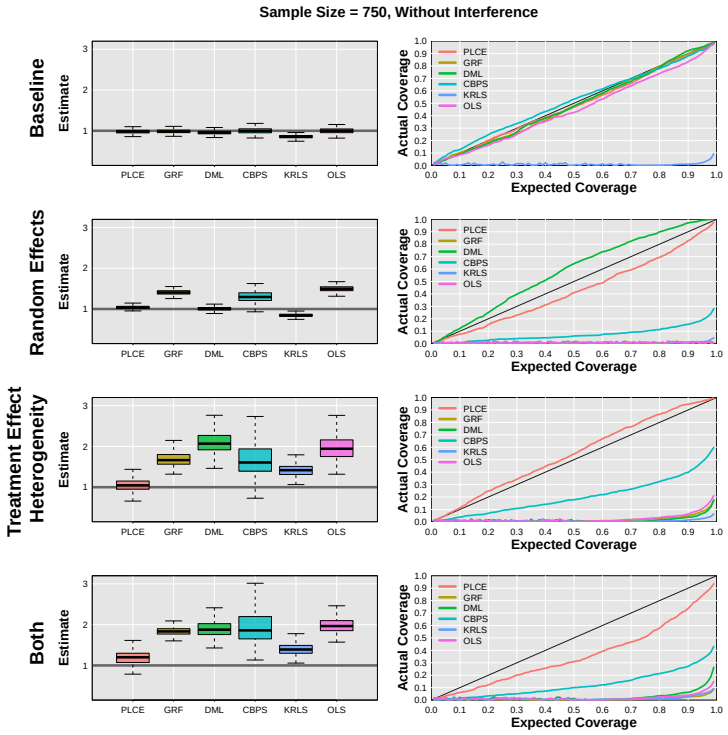

Figure 5: **Simulation Results Without Interference,** $n = 750$

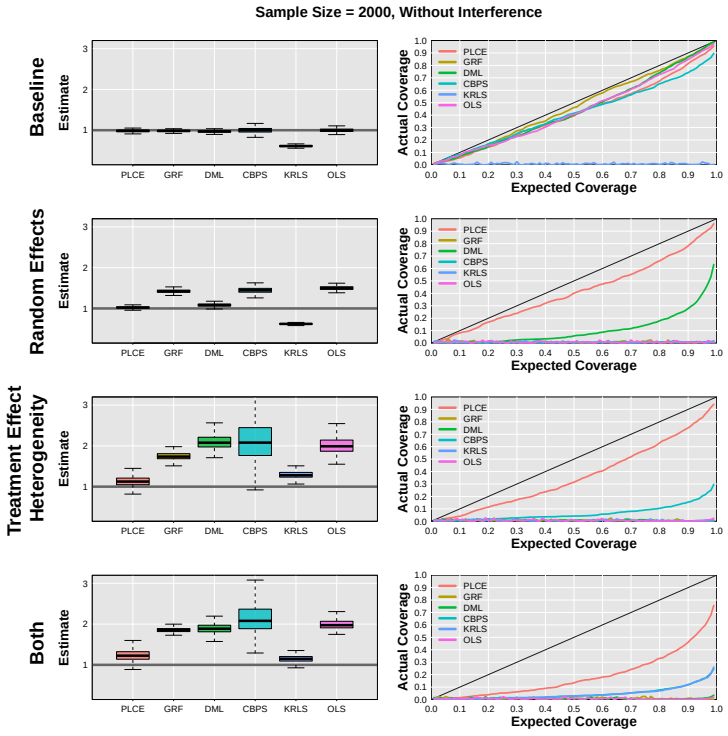

Figure 6: **Simulation Results Without Interference,** $n = 2000$

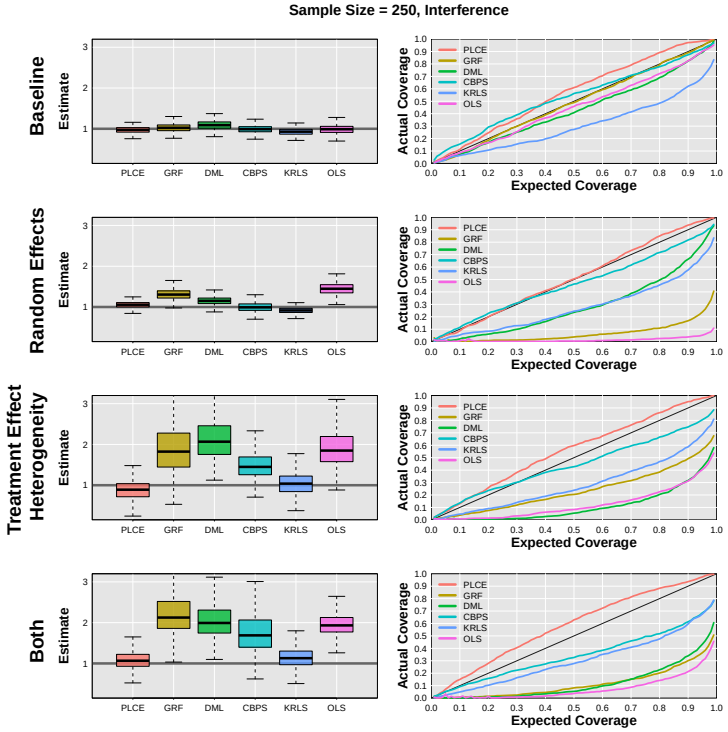

Figure 7: **Simulation Results With Interference,** $n = 250$

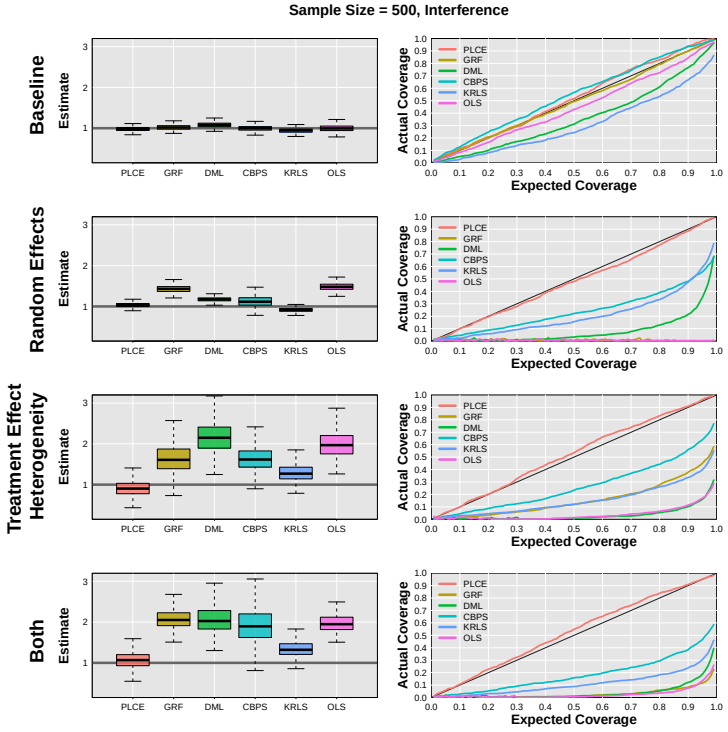

Figure 8: **Simulation Results With Interference,** $n = 500$

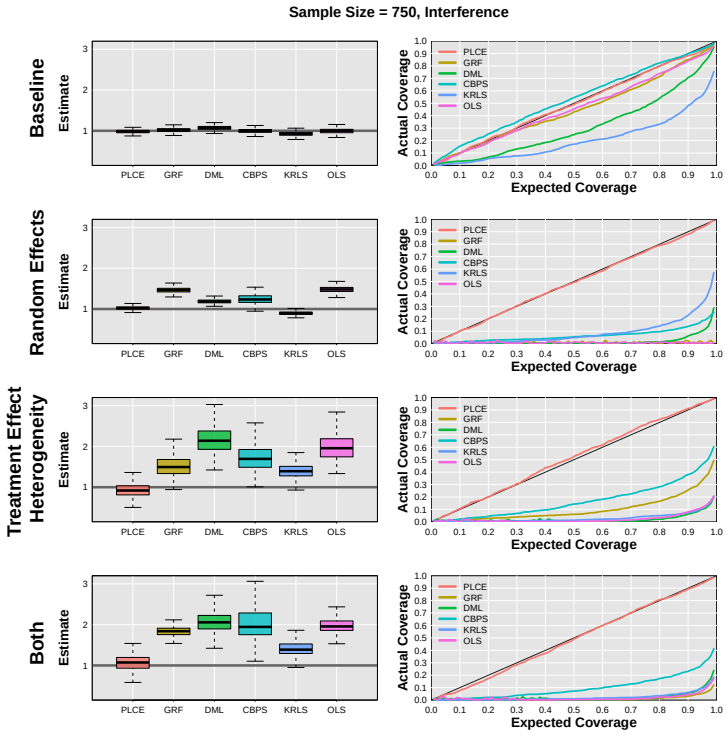

Figure 9: **Simulation Results With Interference,** $n = 750$

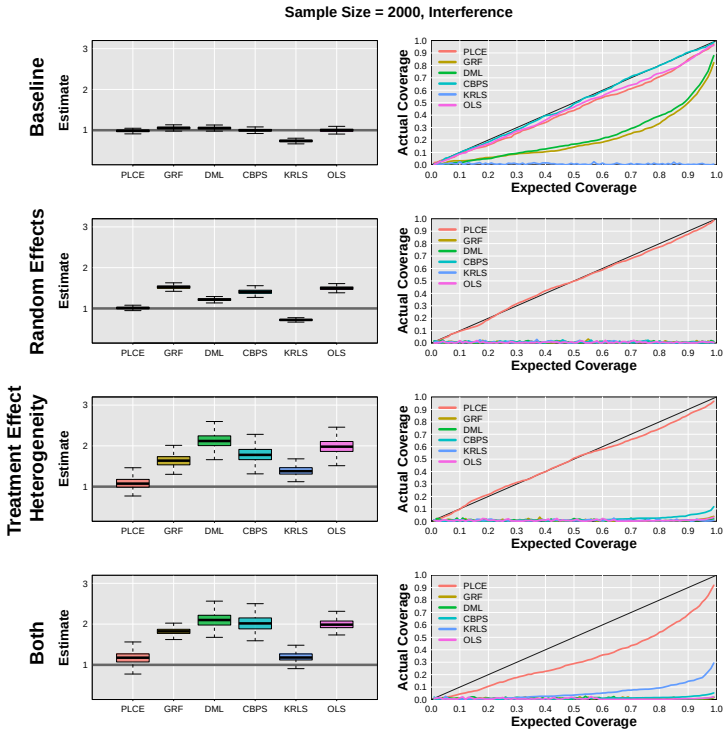

Figure 10: **Simulation Results With Interference,** $n = 2000$

References

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.

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

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

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

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