Permutation Variable Selection for High-Dimensional Regression and Balancing Estimators of Causal Effects

Apoorva Lal[1]

[1] *Stanford University ; apoorval@stanford.edu*

Abstract

Researchers are often interested in estimating average treatment effects in observational settings based on treatment unconfoundedness, which is often only plausible conditional on a large number of pretreatment variables. We provide methods to combine recent advances in model-agnostic variable selection using 'knockoffs' with high-dimensional regularized regression and entropy-balancing to perform covariate adjustment. These methods allow for the inclusion of complex interaction terms in both the outcome and treatment-assignment regressions, which makes the conditional unconfoundedness assumption more plausible than the conventional practice of controlling for them linearly. Knockoff-based variable selection methods improve the precision of estimates and substantially alleviate the curse of dimensionality in balancing methods with many controls and interactions. We provide simulation evidence that finds these selection methods often outperform other standard treatment effect estimators, and ease computational constraints for balancing. Finally, we provide an illustration of these methods using two empirical examples.

Word Count: 6132

Preliminary draft

1 Introduction

Observational causal inference in the social sciences often involves the incorporation of a large set of covariates, conditional on which the treatment is argued to be as good as random. Researchers typically perform covariate adjustment by including controls on the right hand side of a linear regression to justify the treatment unconfoundedness assumption that permits a causal interpretation of the regression coefficient as the average treatment effect (ATE). While this practice of 'kitchen-sink' regressions has been studied and critiqued extensively in methodological work related to the 'credibility revolution' in the social sciences (Angrist and Pischke 2010; Samii 2016), it remains widespread in applied work, especially in settings where experiments and more credible research designs are unavailable for the question at hand. Researchers' substantive knowledge typically informs a set of variables that might be important to adjust for, but seldom identifies exactly which ones or the functional form with which they ought to enter the model. The choice over functional form and what covariates to adjust provides for ample 'researcher degrees of freedom', which gives rise to concerns over p-hacking and 'the garden of forking paths' wherein researchers report subgroup-effects that are likely mined statistical noise (Gelman and Loken 2013).

To address this problem, this paper proposes a method that combines recent advances in variable selection tools flexibly adjust for covariates in both regression and balancing estimators for treatment effect estimation, and an `R` implementation[1]. The approach involves a variable selection followed by estimation and inference. In the first step, covariates (and their basis expansion and interactions) that are important predictors of *either the treatment or the outcome* are selected following the double-LASSO approach (Belloni, Chernozhukov, and Hansen 2014b) (henceforth BCH), where variable selection is performed using permutation-based 'knockoffs' methodology (Barber and Candès 2015). Next, treatment effects are estimated using OLS (post-LASSO) or weighting, for example using entropy-balancing weights (Hainmueller 2012) with standard errors that account for the high-dimensional covariates and heteroskedasticity (Cattaneo, Jansson, and Newey 2018). The variable selection step selects the variables and their corresponding higher moments and interactions for covariate adjustment in the estimation and inference step. This is particularly important in light of recent work by Wuthrich and Zhu (2021) shows that the BCH approach of performing variable selection using the standard LASSO produces non-negligible omitted-variables bias in finite samples, and as such a better variable selection approach is needed. We find that the combination of covariate selection using knockoffs and estimation using entropy and residual balancing improves upon popular estimators in mean squared error (MSE) terms.

The use of $\ell 1$-based knockoff based variable selection before balancing permits researchers to consider balancing on many possible covariates and their interactions. Researchers seldom know particular interaction terms that might be important to adjust for, or the exact functional form with which the variables of interest enter the model. It is plausible that unbeknownst to the researcher, particular combinations of covariate values are particularly predictive of treatment takeup, or are correlated with an unobserved confounder, and therefore must be adjusted. With even a moderate number of controls, the total number of regressors from including pairwise interactions is $p + \frac{p(p-1)}{2}$ may exceed the number of observations in the data, which renders classical estimation infeasible, and also yields an infeasible number of constraints for balancing estimators.

We contribute to a growing literature on the use of high-dimensional regression for causal inference (Belloni *et al.* 2012; Belloni, Chernozhukov, and Hansen 2014b, 2014a; Chernozhukov, Hansen, and Spindler 2015; Chernozhukov, Hansen, and Spindler 2016a; Imai and Ratkovic 2013; Farrell 2015; Bloniarz *et al.* 2016; Ratkovic 2020). Much of this literature begins with the notion of *regularization bias*: naive applications of regularized regression wherein a low-dimensional parameter of interest (the treatment effect) is left un-regularized while control coefficients are selected using the LASSO's variable selection property induces a form of omitted-variables bias (Leeb and Pötscher 2005) because of imperfect model selection. Subsequent work (Chernozhukov *et al.* 2017; Chernozhukov *et al.* 2018) on

1. R programs to implement the procedure is available at https://github.com/apoorvalal/KnockoffEntropyBalancing

Double Machine Learning generalises this approach to estimating any treatment effect or structural parameter that can be represented using a Neyman orthogonality condition[2] and permits the use of arbitrary machine learning algorithms such as random forests and neural nets. However, these methods tend to be more opaque, demand tuning several hyperparameters, and generally rely on algorithms that fewer social scientists are familiar with, and as such is beyond the scope of our paper. Our approach is also related to hierarchically regularized entropy balancing, as proposed by Yang and Xu (2021), who impose $\ell 2$ (ridge) regularization on the variables that enter into entropy balancing. Our approaches differ in the variable selection methods: hierarchical regularization does not impose sparsity and instead imposes larger penalties on higher order terms, while we remain agnostic on the extent of regularization and instead identify variables to match-on using the knockoff selection method. Our work also relates to recent work using double-LASSO approaches to model specification for interaction terms (Blackwell and Olson 2021) and the estimation of survey weights (Ben-Michael, Feller, and Hartman 2021).

The rest of the paper is organised as follows: we provide methodological background to the problem and the proposed solution in section 2, provide results from a simulation exercise that compares the performance of LASSO-selection estimators with that of other standard estimators in section 3, illustrate the application of these estimators to two examples in section 4, and conclude in section 5.

2 Covariate adjustment and balancing in high dimensions

Consider a setting with n units with data $(y_i, d_i, \mathbf{z}_i)_{i=1}^n$, where d_i is a binary treatment, y_i is the outcome, and $\mathbf{z}_i$ is a k-vector covariates. $y(d), d \in \{0, 1\}$ is the potential outcome unit i under each treatment status, only one of which is observed for any given unit. The individual treatment effect $\tau_i := y_i(1) - y_i(0)$ is never estimable, and most researchers are typically interested in the Sample Average Treatment effect SATE $:= \frac{1}{n}\sum_{i=1}^n \tau_i$ or Sample Average Treatment effect on the treated SATT $:= \frac{1}{n}\sum_{i:d_i=1} \tau_i$. To proceed, we need the standard identification assumptions

1. *No interference:* $y_i = y(d_i)\ \forall\ \mathbf{d}_{-i}$ where $\mathbf{d}_{-i}$ is the vector of treatment assignments for all other units.
2. *Unconfoundedness*: $y(1)\ y(0) \perp\!\!\!\perp d|\mathbf{x}$
3. *Overlap*: $0 < e(\mathbf{x}) < 1$ where $e(\mathbf{x})$ is the propensity score.

The inverse-probability-of-treatment weighting (IPTW) estimator for the ATT involves imputing average missing potential outcome for the treated units in the absence of treatment

$$\widehat{\mathbb{E}}[y(0)|d = 1] = \frac{\sum_{i:d_i=0} y_i w_i}{\sum_{i:d_i=0} w_i}$$

where w_i are weights that yield balance between treated and control units along covariates. These weights may be computed as $\widehat{\pi}(\mathbf{x}_i)/(1 - \widehat{\pi}(\mathbf{x}_i))$ using a parametric model for the propensity score $\pi(\mathbf{x}_i)$, at the risk of model misspecification.

We focus on Entropy balancing (Hainmueller 2012), which is a balancing method that directly incorporates covariate balance into the weight function in the form of moment conditions, and uses an entropy loss that is more robust under model misspecification (Imbens, Johnson, and Spady 1998). It is doubly-robust : it is consistent as long as either the assignment model or regression model is correctly specified (Zhao and Percival 2016), and typically achieves excellent balance and lower MSE relative to other matching and propensity score methods. The entropy-balancing weight w_i for each control unit is chosen by a reweighting scheme that solves the following optimization problem:

$$\max_{w_i} H(w) = -\sum_{i:D=0} w_i \log(w_i)$$

2. which sets up the problem such that the parameter of interest solves the equation $\mathbb{E}\partial_\eta \psi(Z_i, \tau, \eta_0) = 0$, where Z_i is the data, η is the nuisance function, and the Gateaux derivative of the score ψ with respect to the nuisance function is set to zero. Chernozhukov et al call the approach DML – *double machine learning* – because of their connection to the notion of double-robust estimation under unconfoundedness Robins, Rotnitzky, and Zhao (1994)

subject to balance/moment-condition $\sum_{i:D=0} w_i c_{ri}(\mathbf{X}_i) = m_r \quad r \in 1, \ldots, R$ and normalising ('proper weights') $\sum_{i:D=0} w_i = 1$ and $w_i \geq 0 \ \forall \{i : d_i = 0\}$ constraints. This problem is convex but has dimensionality of n_0 (non-negativity) + p (moment conditions) + 1 (normalisation). The dual, on the other hand, only has dimensionality $p + 1$ and unconstrained, which is considerably easier to solve using Newton-Raphson-like techniques. EB has the additional benefit of being achieving one-shot balancing, unlike other propensity-score approaches, which involve iteratively fitting the pscore model and assessing balance. It relates closely to Empirical Likelihood (EL) approaches that directly incorporate covariate balance into the objective function for propensity score fitting (Imai and Ratkovic 2014; Ben-Michael *et al.* 2021).

Nevertheless, in complex models with many interactions, $p+1$ can be quite large, and NR-like algorithms frequently run into numerical instability errors[3] with large sets of constraints if balance on all covariates is specified in the entropy balancing problem. Furthermore, the correctly specifying the outcome and/or assignment model remains a challenge with only linear lower-order terms [as is common in most applications of entropy balancing], and can be greatly aided by adding polynomial terms and interactions, which increases the dimension of the optimisation problem. Data-driven methods are a promising avenue for reducing the dimension of the problem. We propose selecting variables to balance on motivated by the double-LASSO, which discuss next.

Double-LASSO for covariate adjustment Belloni, Chernozhukov, and Hansen (2014b) motivate the use of LASSO regression (Tibshirani 1996) for covariate adjustment writing by linear approximations of the propensity score and outcome models as follows:

$$d_i = \underbrace{e(\mathbf{z}_i)}_{\mathbf{x}_i'\boldsymbol{\beta}_d + r_{ei}} + \nu_i \ , \mathbb{E}(\nu_i|\mathbf{x}_i) = 0 \tag{1}$$

$$y_i = \underbrace{m(\mathbf{z}_i)}_{\mathbf{x}_i'\boldsymbol{\beta}_y + r_{mi}} + \upsilon_i \ , \mathbb{E}(\upsilon_i|\mathbf{x}_i) = 0 \tag{2}$$

where the unknown flexible functions $e(\cdot)$ and $m(\cdot)$ are approximated using a dictionary of flexible polynomials and basis expansions $\mathbf{x}_i$ of the high dimensional covariates $\mathbf{z}_i$[4]. The linear-expansion approach can be thought of as a Taylor expansion of the true conditional expectation function with some approximation error $r_{\cdot i}$.[5]

Omitted Variables Bias (OVB) arises when one fails to adjust for x's that are correlated with *both* d and y, i.e. have nonzero true coefficients in both $\boldsymbol{\beta}_y$ and $\boldsymbol{\beta}_d$, thereby resulting in a failure of the unconfoundedness assumption. To guard against resultant OVB from such omissions, BCH suggest using *both* equations for selection, which immunizes the resulting procedure against model-selection mistakes in one of the two steps, thereby giving it the double-robustness property. This is a special case of the general Neyman Orthogonalization approach described in Chernozhukov *et al.* (2018)[6].

The double selection approach involves estimating both reduced form equations using LASSO, and using the union of the selected variables for the final estimation step. Let $\mathcal{S}_1$ and $\mathcal{S}_2$ denote the controls selected by the LASSO in the

3. particularly because of the Hessian in the denominator

4. the existence of such basis representations is guaranteed under standard regularity conditions for e and m: that they have bounded variation on a compact interval. This guarantees the existence of a Fourier expansion

5. a key assumption underlying the BCH approach is (approximate) *sparsity*, wherein $s := ||\beta_y||_0 = \sum_{j=1}^{p} 1\{\beta_{0j} \neq 0\} << n$, which is that the number of *relevant* regressors is much smaller than sample size. Equivalently, we need it to be the case that $n \to \infty$, $s \log p / n \to 0$. This formulation also allows for small estimation error r_{gi} in the approximation of the nonlinear function $g(\cdot)$

6. BCH also propose a Frisch-Waugh-Lovell style-procedure of using predictive tools on both outcome and treatment regressions, and using the residuals to estimate treatment effects, as proposed by Robinson (1988). This is typically implemented with regressions of the form $y - \widehat{\mathbb{E}}[y|\mathbf{x}] = \tau(d - \widehat{\mathbb{E}}[d|\mathbf{x}]) + \nu$ where $\widehat{\mathbb{E}}[\cdot]$ are estimates from LASSO regressions. This procedure is an early version of double-machine learning (Chernozhukov *et al.* 2018), which proves that the residuals-on-residuals approach is $\sqrt{n}$ consistent for treatment effects even when the CEFs are fitted using nonparametric regressions with slower rates.

regression of y_i on $\mathbf{x}_i$ and d_i on $\mathbf{x}_i$ respectively. The 'double-selection' estimator can be written as

$$(\check{\tau}, \check{\boldsymbol{\beta}}) = \text{argmin}_{\tau \in \mathbb{R}, \boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{N} [y_i - \tau d_i - \mathbf{x}_i' \boldsymbol{\beta}]^2 : \beta_j = 0 \; \forall j \notin \mathcal{S} := \mathcal{S}_1 \cup \mathcal{S}_2$$

We prove that these estimators satisfy the Neyman orthogonality condition in appendix section A.1. The double-LASSO is robust to 'small' model misspecification. It is are implemented in the `hdm` package (Chernozhukov, Hansen, and Spindler 2016b) in `R` and `pdslasso` package in `STATA`.

It is worth noting that the specific value of LASSO coefficients isn't important to us, since we are interested in the treatment effect τ instead of the entire (potentially very large) coefficient vector β, and as such we can treat the latter as a nuisance parameter. Double-LASSO therefore uses 'post-LASSO' OLS as the final estimation step, where the LASSO is used purely as a model-selection device and un-regularized regression is fit with the subset of variables selected by the LASSO.

2.1 Variable Selection Step

The double-LASSO approach relies on the LASSO estimator

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij} \beta_j \right) + \underbrace{\lambda \sum_{j=1}^{p} \|\beta_j\|_1}_{\text{regularisation term}}$$

where λ is a penalty term that penalises model complexity, and the use of the $\ell 1$ norm $\|.\|_1$ induces sparsity in the estimated coefficient vector $\widehat{\boldsymbol{\beta}}$. The kink in the penalty term produced by the L1 norm induces $\hat{\beta}$ to have lots of zeros, which has resulted in the LASSO's popularity as a model-selection device The LASSO is well suited to $p > n$ problems and estimates $\widehat{\boldsymbol{\beta}}$ with many zeroes in it, depending on the penalty term λ. The penalty is typically chosen by cross-validation, but other procedures have been proposed, including the rate-optimal procedure[7], perfect selection procedure [8], and the iterative procedure [9] ,which is recommended for the double-LASSO approach by BCH.

Choosing λ for model selection is a challenging problem because a single parameter needs to perform both shrinkage and selection, it trades off shrinking relevant variables' coefficients with the inclusion of irrelevant variables. Wuthrich and Zhu (2021) show that in finite samples, the BCH regularization parameter typically under-selects (i.e. zeroes-out too many variables), resulting in severe omitted-variables bias, especially in settings where the R^2 of the generative model is low. Yang (2005) and Meinshausen, Bühlmann, *et al.* (2006) show that a single value for λ cannot simultaneously be used for model-selection as well as regression function estimation, and therefore various other regularisation approaches have been proposed[10]. This is a particular problem when Xs are correlated, as is typical in social-scientific data with highly correlated, since the LASSO's 'oracle' model selection property relies on the *irrepresentable condition*, which requires that the sum of signed regression coefficients of unimportant variables (i.e. xs absent from the true model) on the important variables (xs present in the true model) cannot exceed 1 (Zou 2006; Fan *et al.* 2020), which is highly restrictive.

7. $\lambda = 2\sigma\sqrt{(2(1+\upsilon)\log(p))/n}, \quad \upsilon > 0$, Bickel, Ritov, and Tsybakov (2009)
8. $\sigma\psi^{-1}\sqrt{\log p/n}, \quad \psi \in (0, 1]$, Wainwright (2009)
9. This involves rewriting the penalty term as $\frac{\lambda}{n}\|\hat{\Psi}\beta\|_1$, where $\widehat{\boldsymbol{\Psi}} := \text{diag}(\widehat{\psi}, \ldots, \widehat{\psi}_p)$ is a diagonal matrix of penalty loadings that can be chosen for the problem at hand, such as 0 for parameters one wants to avoid shrinking, such as treatment effects or group-fixed effects. These penalty loadings can be chosen to address heteroskedasticity in errors, in which case the loadings are set to $\widehat{\psi}_j = \sqrt{\mathbb{E}_n[\mathbf{x}_{ij}^2 \hat{\varepsilon}_i^2]}$, where $\hat{\varepsilon}$ are preliminary estimates of residuals, Belloni *et al.* (2012)
10. There have been attempts to improve the LASSO by moving beyond $l1$ penalties and introducing a mixture of $l1$ and $l2$ using the elastic net (Zou and Hastie 2005), introducing a non-convexity in the penalty term (Zhang 2010; Mazumder, Friedman, and Hastie 2011) introducing an additional 'relaxing' parameter (Meinshausen 2007).

Variable selection using permutation knockoffs Since the LASSO is far from the best variable selection procedure available in finite samples, we contend that improving upon the LASSO's variable selection properties is likely to help us estimate treatment effects better. In finite samples, the variable selection step is paramount: Under-inclusion of variables relevant for both outcome and treatment models results in omitted variables bias (Wuthrich and Zhu 2021), while over-inclusion of variables in the treatment equation (that have no direct effect on the outcome) can hurt precision by reducing variation in the residualised treatment $\tilde{d}$, thereby increasing the SE of treatment effect (by simple Frisch-Waugh-Lovell reasoning). [11]

To improve upon the LASSO's model selection, we propose adopting a permutation-based version of the knockoffs methodology proposed by Barber and Candès (2015) and Barber, Candes, and Samworth (2020) for LASSO. The basic intuition for variable-selection using knockoffs is that if a variable X predicts the outcome well, it ought to do better than a knockoff $\tilde{X}$, which mimics the covariance structure similar to the data matrix but is independent of Y. This then motivates the use of regularized regression on the augmented data matrix $[\mathbf{X}\ \tilde{\mathbf{X}}]$; this increases problem dimension to $2p$. This approach is also related to the popular `boruta` algorithm for variable selection and feature importance for random forests (Kursa, Rudnicki, *et al.* 2010) and is widely used in model selection procedures with various other machine learning algorithms.

Knockoff construction is a burgeoning literature with different procedures relying on different assumptions about the DGP. The exact finite-sample FDR control property of the original Barber and Candès (2015) procedure relies on a fixed design matrix $\mathbf{X}$, which is suitable for its motivating use case in genetics studies but is far too strong an assumption for typical observational causal inference settings. Candès *et al.* (2018) settings with known $\mathbb{F}_X$, which is weaker but remains too stringent for our case. We adopt the approach with no distributional assumptions on $\mathbf{X}$ and choose to construct knockoffs by permuting the rows of the regressor matrix $\mathbf{X}$ to construct $\widetilde{\mathbf{X}}$. This approach that does not provide finite sample FDR control guarantees[12] (Barber and Candès 2015, sec 3.1), but relies on substantially weaker assumptions and works well in simulation studies (Gégout-Petit, Gueudin-Muller, and Karmann 2020).

Knockoff construction procedure We propose constructing knockoffs simply by repeatedly permuting the rows of the design matrix $\mathbf{X}$ to construct $\tilde{\mathbf{X}}$ (Barber and Candès 2015; Gégout-Petit, Gueudin-Muller, and Karmann 2020). In each iteration k,

1. A permutation matrix $\mathbf{P}_k$ is constructed by permuting the rows of a $n \times n$ identity matrix. $\mathbf{P}_k$ contains precisely a single 1 in each column and row and 0 everywhere else [13] (Banerjee and Roy 2014).
2. A knockoff matrix $\widetilde{\mathbf{X}}_k$ with reshuffled rows is constructed by pre-multiplying the data matrix $\mathbf{X}$ with the permutation matrix $\mathbf{P}$
$$\tilde{\mathbf{X}}_k = \mathbf{P}_k \mathbf{X}$$
3. The knockoff matrix $\widetilde{\mathbf{X}}_k$ is (column-) concatenated with the original data matrix $\mathbf{X}$ to construct the $n \times 2p$ augmented data matrix $[\mathbf{X}, \widetilde{\mathbf{X}}_k]$.
4. The outcome $\mathbf{y}$ is regressed on the augmented data-matrix using LASSO regression. For each variable j,
   (a) Compute a LASSO test statistic $T_j := \sup\{\lambda > 0, \hat{\beta}_j(\lambda) \neq 0\}\ \ j \in \{1, \ldots, 2p\}$, which is the largest penalty λ at which a variable j has a nonzero coefficient in the LASSO.

11. Reducing the variance in treatment (by over-controlling for variables that predict the treatment but are uncorrelated with the outcome) reduces the denominator of the standard error estimator, which in turn increases the standard error of the treatment effect estimate.

12. this occurs because although the permuted matrix $\widetilde{\mathbf{X}}$ has the same covariance structure as $\mathbf{X}$, it is unable to mimic the correlation of $\mathbf{X}$ with $\boldsymbol{y}$

13. For example, a typical $\mathbf{P}$ may look like
$$\mathbf{P} = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

(b) Compute the following measure of how often j enters into the model before its knockoff:

$$W_j := T_j \vee \tilde{T}_j \times \begin{cases} +1 & T_j \geq \tilde{T}_j \\ -1 & T_j \leq \tilde{T}_j \end{cases}$$

We repeat this procedure a large K number of times.
We then keep a variable j if $W_j \geq q$, some threshold value of times the true variable enters the model before the knockoff across K iterations. This selection procedure is implemented in the `kosel` package [14].
This approach is based on the heuristic that If $\mathbf{x}_j$ is truly predictive of y, it should enter the model earlier (for larger penalty λ) than over many possible knockoffs $\tilde{\mathbf{x}}_j$. Gégout-Petit, Gueudin-Muller, and Karmann (2020) propose 0.1 as a rule of thumb value. These permutation knockoffs constitute an improvement over the LASSO and outperform both the original LASSO and Candes and Barber knockoffs in simulations (sec 3.2, Ibid).

2.2 Estimation Step

Entropy balancing delivers one-shot balancing weights, which constitute a major improvement over the aforementioned loop, but runs into convergence and numerical stability issues when balancing on many covariates and their interactions. Our proposed combination of the double-LASSO principle of balancing on selected variables both reduced form equations, combined with variable selection using knockoffs, promises to provide a list of variables and their higher moments[15] ease the computational burden of entropy balancing substantially. Alternatively, if the researcher believes the linear approximation of the outcome model is correct, they may use an amended version of double-LASSO.
In summary, the method we propose involves the following steps:

1. **Variable selection**:
   - regress outcome on covariates using LASSO, select predictive variables using the knockoff selector, call them $\mathcal{S}_1$
   - regress treatment on covariates using LASSO, select predictive variables using knockoff selector, call them $\mathcal{S}_2$
2. **Estimation**:
   - **Knockoff Entropy Balancing (KOEB)**: Perform entropy balancing on the set of moment conditions ($\mathcal{S}_1 \cup \mathcal{S}_2$ instead of the full set of predictors and polynomials and interactions)
   - **Knockoff Selection (KOSEL)**: perform 'post-LASSO' linear regression, as in Belloni, Chernozhukov, and Hansen (2014b), wherein the researcher regresses the outcome on treatment, controlling for variables that are predictive of either the outcome of the treatment (i.e. $\mathcal{S}_1 \cup \mathcal{S}_2$).

3 Monte-Carlo Simulations

We perform a monte-carlo exercise to evaluate the performance of LASSO-based methods with other popular estimators. To this end, we adapt experiment 2 in Hainmueller (2012), which involves using covariates from the Dehejia and Wahba (1999)) experimental sample of LaLonde (1986)) data, and specifying a highly nonlinear data-generating process for the propensity score and outcome with a known treatment effect in order to evaluate performance.
We focus on benchmarking our proposed knockoff-selection entropy balancing estimator (KOEB) against conventional regression adjustment using OLS, Double selection (DS), propensity score matching (PS), Mahalanobis distance matching (MD), entropy balancing (EB) on the entire design matrix. We provide a detailed description of the data-generating process in B.

14. https://cran.r-project.org/web/packages/kosel/
15. this is because balancing on a polynomial, say x^2, is equivalent to entropy-balancing on the variance of x since the columns in $\mathbf{X}$ are all centered

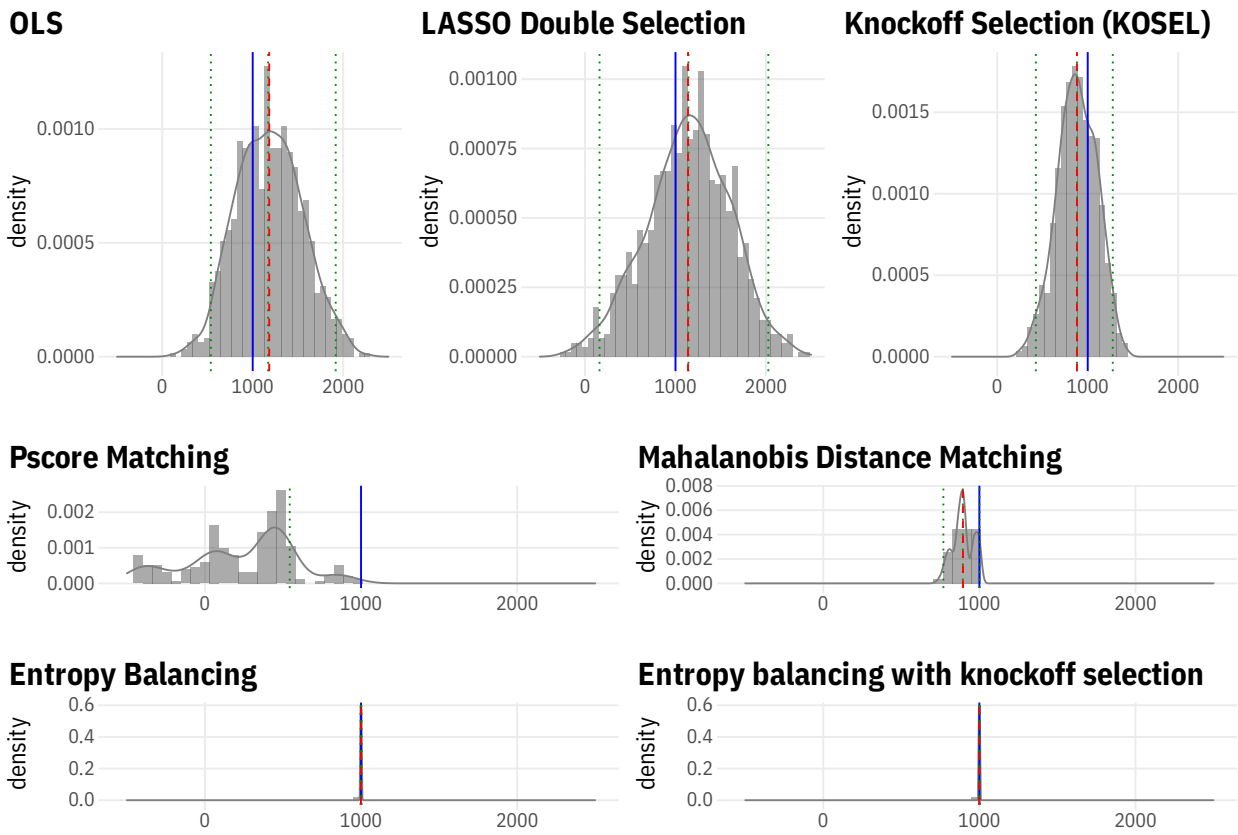

Figure 1. Distribution of effect estimates from simulations. The Blue vertical line is the truth, the red line is the mean estimate, and green lines indicate 2.5th, 50th, and 97.5th percentile

We plot the distribution of estimates from 10,000 replications of the simulation procedure in figure 1, and summary statistics in table 1. Among regression-based methods, we find that DS and KOSEL improve substantially upon OLS, and that KOSEL has the lowest variance. Knockoff selection is more precise (i.e. it has lower RMSE) than the LASSO based alternatives. Among weighting methods, we find the LASSO-selected entropy balancing method is best-performing in terms of bias, median absolute deviation (MAD), and root-means-squared-error (RMSE), with Propensity score Matching (PS) performing worst[16]. We find that knockoff-augmented entropy-balancing (KOEB) performs comparably to 'full' entropy balancing on the covariate matrix (EB), and is easier to compute. Furthermore, the 'reduced' entropy-balancing weights are computable the large matrix of covariates with many interactions, while the 'full' entropy-balancing weights computation on the wide covariate matrix very frequently ran into numerical errors, which is not reflected in the figure because the resultant effect estimates are `NA`s.

In summary, we find that KOSEL and KOEB are the best performing (in MSE terms) in their respective classes of regression and re-weighting estimators. KOEB is unbiased in our simulations with a highly nonlinear treatment and outcome model, and therefore should be preferred when its implementation is feasible. However, if researchers prefer to use regressions for other (computational, longitudinal data, or interpretation) reasons, KOSEL clearly outperforms naive covariate adjustment in both bias and variance terms, and is weakly superior to double-selection LASSO thanks to higher precision.

16. Indeed, much of the mass of the PS estimate distribution is to the left of the x-scale in the figure, hence the suspiciously thin histogram

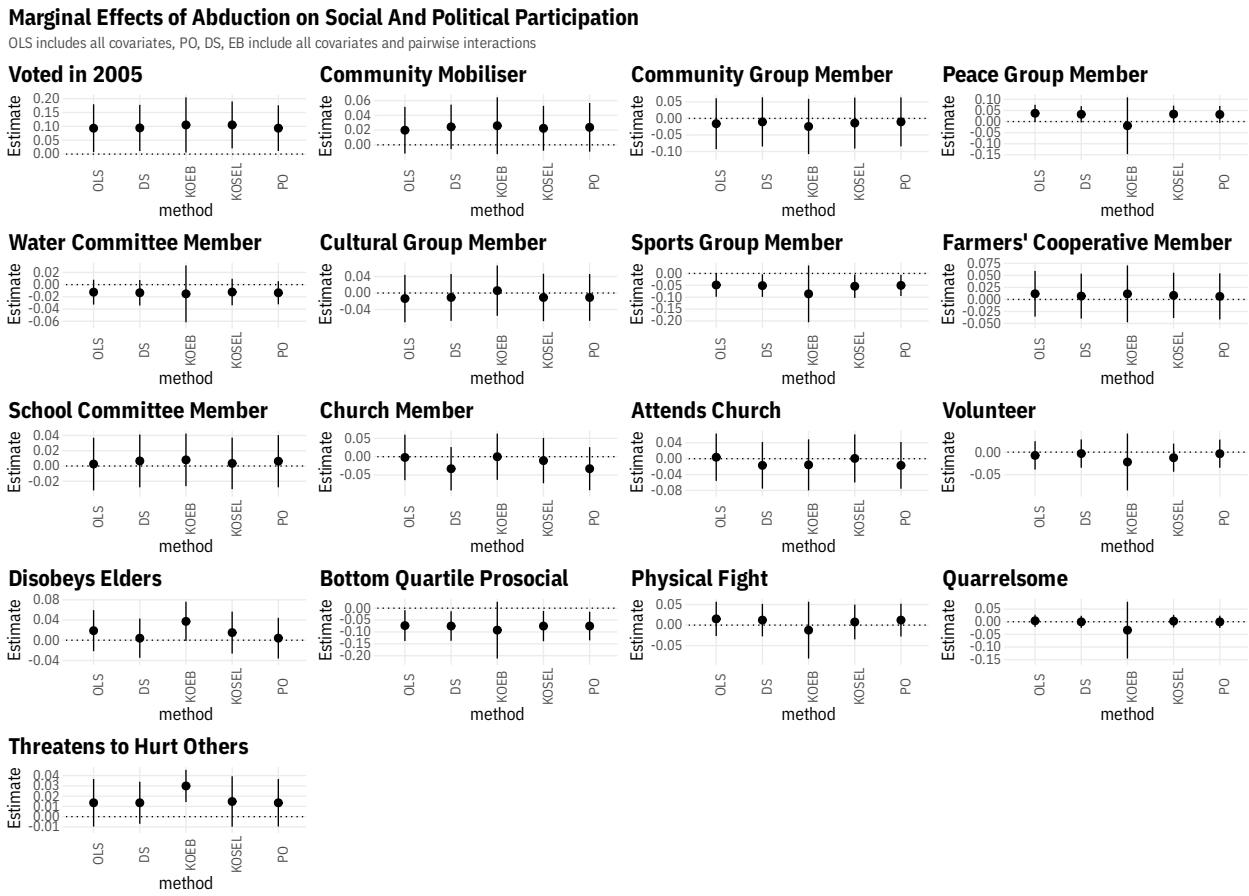

Figure 2. OLS, Double-Selection, Knockoff Entropy Balancing, Knockoff LASSO, and Partialling Out for all outcomes in Blattman (2009) table 3

4 Empirical Examples

4.1 Blattman (2009)

Blattman (2009) attempts to provide causal evidence of the political legacy of violent conflict by studying the effects of involuntary rebel recruitment on postwar political engagement and socio-political behaviour of ex-combatants using an individual level dataset. Blattman argues that patterns in rebel abduction during Uganda's civil war generated 'nearly exogenous variation' in recruitment, and as such, causal estimates of its impact on later-life outcome such as political participation can be identified. Blattman uses data from the Survey of War Affected Youth (SWAY) and estimates the effects of being abducted on downstream outcomes conditional on a list of controls using logistic regression (table 3).

We replicate Blattman's findings using OLS,[17] controlling for possible every pairwise interaction between the 36 controls that Blattman includes his regressions, and report the results in figure 2. This produces a matrix of 666 covariates.[18] For several of the outcomes (notably `vote05`, the titular voting indicator that is the focus of much of the paper), the number of nonmissing observations is lower than 666 – in the case of voting, the number of observations is 542. Thus, $p >> n$ in this setting even with pairwise interactions.

Double-selection, partialling out, selected-entropy balancing, and Knockoff-selection generally produce estimates that agree with the vanilla OLS estimates on sign and magnitude, with slightly higher precision. This, combined with

17. the marginal effects estimates are very close to the estimates from logistic regression
18. this is equal to $36 + (36 \times 35)/2$

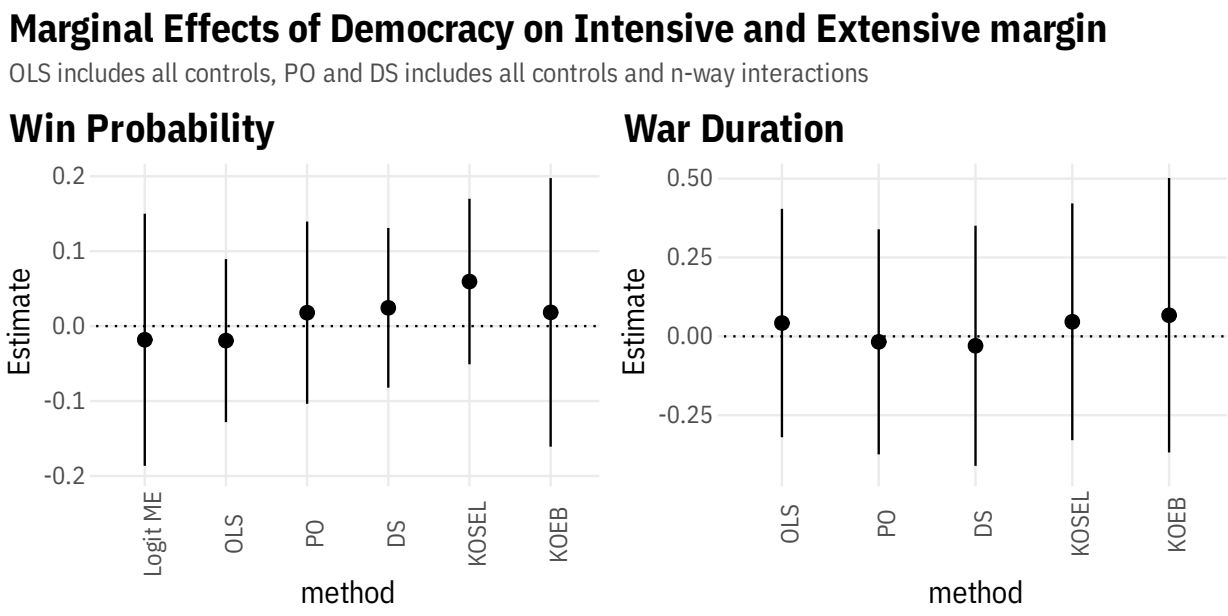

Figure 3. Logit, OLS, Partialling Out, Double-Selection, Knockoff, and Entropy Balancing Estimates for Lyall (2010) table 2 (model 3), table 5 (model 19)

Blattman's sensitivity analysis to evaluate how large the effects of unobserved confounders needs to be in order to reduce the voting effect by half, increases our confidence in the findings.

4.2 Lyall (2010)

Lyall (2010) studies the effect of democracy's impact on counterinsurgency (COIN) war outcomes and duration. He collects data on war outcomes and covariates for internal conflicts from 286 insurgencies from 1800-2005, and estimates the 'effect' of a democratic regime on COIN outcomes such as war outcomes (win/draw/loss, truncated to win/loss for a subset of the analysis) and war duration. The paper begins by noting that the conventional wisdom that democracies are bad at COIN operations is rendered questionable by omitted variables bias, and proceeds to estimate logit regressions on the full and matched samples and finds that the effect of democracy on outcomes is generally small and statistically indistinguishable from zero.

We replicate Lyall's analysis using all possible pairwise interactions and report the results in figure 3. We find that although the confidence intervals still cover zero for both outcomes, the sign of the estimated coefficients changes when using double-selection and knockoff-selection with more flexible sets of controls for the dichotomous victory outcome. The confidence intervals for the estimated coefficient for the war duration outcome is substantially narrower than the vanilla OLS (or Weibull, as reported in the paper).

5 Conclusion

This paper has outlined a method of combining recent advances in variable selection with balancing and regularized regression (double-LASSO) estimators to estimate causal effects. This approach promises to be particularly useful in settings where researchers use a selection-on-observables approach to estimate treatment effects. The high-dimensional regression-based approach permits researchers to adjust for many interactions of control variables, possibly more than the number of observations, in order to render the conditional unconfoundedness assumption more credible. As such, it presents a simple and interpretable step in credibly estimating treatment effects in observational settings. In addition to improving the plausibility of the unconfoundedness assumption, these methods

also reduce researcher degrees of freedom in specifying the regression used to estimate effects of interest. We see from the two applications replicated in the paper that these methods tend to find smaller and sometimes precise effects than the original studies, and as such may reduce the number of published false-positives.

References

Angrist, J. D., and J.-S. Pischke. 2010. “The credibility revolution in empirical economics: How better research design is taking the con out of econometrics.” *Journal of economic perspectives* 24 (2): 3–30.

Banerjee, S., and A. Roy. 2014. *Linear algebra and matrix analysis for statistics.* Vol. 181. Crc Press Boca Raton, FL, USA:

Barber, R. F., E. J. Candes, and R. J. Samworth. 2020. “Robust inference with knockoffs.” *Annals of Statistics* 48 (3): 1409–1431.

Barber, R. F., and E. J. Candès. 2015. “Controlling the false discovery rate via knockoffs.” *The Annals of Statistics* 43 (5): 2055–2085.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen. 2012. “Sparse models and methods for optimal instruments with an application to eminent domain.” *Econometrica* 80 (6): 2369–2429.

Belloni, A., V. Chernozhukov, and C. Hansen. 2014a. “High-dimensional methods and inference on structural and treatment effects.” *Journal of Economic Perspectives* 28 (2): 29–50.

———. 2014b. “Inference on treatment effects after selection among high-dimensional controls.” *The Review of Economic Studies* 81 (2): 608–650.

Ben-Michael, E., A. Feller, and E. Hartman. 2021. “Multilevel calibration weighting for survey data” (February). arXiv: 2102.09052 [stat.ME]. http://arxiv.org/abs/2102.09052.

Ben-Michael, E., A. Feller, D. A. Hirshberg, and J. R. Zubizarreta. 2021. “The Balancing Act in Causal Inference” (October). arXiv: 2110.14831 [stat.ME]. http://arxiv.org/abs/2110.14831.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov. 2009. “Simultaneous analysis of Lasso and Dantzig selector.” *The Annals of statistics* 37 (4): 1705–1732.

Blackwell, M., and M. P. Olson. 2021. “Reducing model misspecification and bias in the estimation of interactions.” *Political Analysis,* 1–20.

Blattman, C. 2009. “From violence to voting: War and political participation in Uganda.” *American political Science review,* 231–247.

Bloniarz, A., H. Liu, C.-H. Zhang, J. S. Sekhon, and B. Yu. 2016. “Lasso adjustments of treatment effect estimates in randomized experiments.” *Proceedings of the National Academy of Sciences* 113 (27): 7383–7390.

Candès, E., Y. Fan, L. Janson, and J. Lv. 2018. “Panning for gold: ‘model-X’ knockoffs for high dimensional controlled variable selection” [in en]. *Journal of the Royal Statistical Society. Series B, Statistical methodology* 80, no. 3 (June): 551–577. http://doi.wiley.com/10.1111/rssb.12265.

Cattaneo, M. D., M. Jansson, and W. K. Newey. 2018. “Inference in Linear Regression Models with Many Covariates and Heteroscedasticity.” *Journal of the American Statistical Association* 113, no. 523 (July): 1350–1361. https://doi.org/10.1080/01621459.2017.1328360.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, and W. Newey. 2017. “Double/Debiased/Neyman Machine Learning of Treatment Effects.” *The American economic review Papers and Proceedings* 107, no. 5 (May): 261–265. https://www.aeaweb.org/articles?id=10.1257/aer.p20171038.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. 2018. “Double/debiased machine learning for treatment and structural parameters.” *The econometrics journal* 21, no. 1 (February): C1–C68. http://doi.wiley.com/10.1111/ectj.12097.

Chernozhukov, V., C. Hansen, and M. Spindler. 2016a. “hdm: High-dimensional metrics.” *arXiv preprint arXiv:1608.00354.*

———. 2016b. “High-dimensional metrics in R.” *arXiv preprint arXiv:1603.01700.*

Chernozhukov, V., C. Hansen, and M. Spindler. 2015. “Valid post-selection and post-regularization inference: An elementary, general approach.” *Annual Review of Economics.*

Dehejia, R. H., and S. Wahba. 1999. “Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs.” *Journal of the American statistical Association* 94 (448): 1053–1062.

Fan, J., R. Li, C.-H. Zhang, and H. Zou. 2020. *Statistical foundations of data science.* Chapman / Hall/CRC.

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

Gégout-Petit, A., A. Gueudin-Muller, and C. Karmann. 2020. “The revisited knockoffs method for variable selection in L 1-penalized regressions.” *Communications in Statistics-Simulation and Computation,* 1–14.

Gelman, A., and E. Loken. 2013. “The garden of forking paths: Why multiple comparisons can be a problem, even when there is no “fishing expedition” or “p-hacking” and the research hypothesis was posited ahead of time.” *Department of Statistics, Columbia University.*

Hainmueller, J. 2012. “Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies.” *Political analysis,* 25–46.

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

———. 2014. “Covariate balancing propensity score.” *Journal of the Royal Statistical Society: Series B: Statistical Methodology,* 243–263.

Imbens, G., P. Johnson, and R. Spady. 1998. “BInformation theoretic approaches to inference in moment condition models.” *Econometrica* 66.

Kursa, M. B., W. R. Rudnicki, *et al.* 2010. “Feature selection with the Boruta package.” *J Stat Softw* 36 (11): 1–13.

LaLonde, R. J. 1986. “Evaluating the econometric evaluations of training programs with experimental data.” *The American economic review,* 604–620.

Leeb, H., and B. M. Pötscher. 2005. “Model selection and inference: Facts and fiction.” *Econometric Theory,* 21–59.

Lyall, J. 2010. “Do Democracies Make Inferior Counterinsurgents? Reassessing Democracy’s Impact on War Outcomes and Duration.” *International Organization* 64 (1): 167–192.

Mazumder, R., J. H. Friedman, and T. Hastie. 2011. “Sparsenet: Coordinate descent with nonconvex penalties.” *Journal of the American Statistical Association* 106 (495): 1125–1138.

Meinshausen, N. 2007. “Relaxed lasso.” *Computational Statistics & Data Analysis* 52 (1): 374–393.

Meinshausen, N., P. Bühlmann, *et al.* 2006. “High-dimensional graphs and variable selection with the lasso.” *Annals of statistics* 34 (3): 1436–1462.

Ratkovic, M. 2020. “Rehabilitating the Regression: Honest and Valid Causal Inference through Machine Learning.” *WP,* https://scholar.princeton.edu/sites/default/files/ratkovic/files/plce_public.pdf.

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

Robinson, P. M. 1988. “Root-N-consistent semiparametric regression.” *Econometrica: Journal of the Econometric Society,* 931–954.

Samii, C. 2016. “Causal empiricism in quantitative research.” *The Journal of Politics* 78 (3): 941–955.

Tibshirani, R. 1996. “Regression shrinkage and selection via the lasso.” *Journal of the Royal Statistical Society: Series B (Methodological)* 58 (1): 267–288.

Wainwright, M. J. 2009. “Sharp thresholds for High-Dimensional and noisy sparsity recovery using l1-Constrained Quadratic Programming (Lasso).” *IEEE transactions on information theory* 55 (5): 2183–2202.

Wuthrich, K., and Y. Zhu. 2021. “Omitted variable bias of Lasso-based inference methods: A finite sample analysis.” *The review of economics and statistics,* http://arxiv.org/abs/1903.08704.

Yang, E., and Y. Xu. 2021. “Hierarchically Regularized Entropy Balancing.” *Working Paper,* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3807620.

Yang, Y. 2005. “Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation.” *Biometrika* 92 (4): 937–950.

Zhang, C.-H. 2010. “Nearly unbiased variable selection under minimax concave penalty.” *The Annals of statistics* 38 (2): 894–942.

Zhao, Q., and D. Percival. 2016. “Entropy balancing is doubly robust.” *Journal of Causal Inference* 5 (1).

Zou, H. 2006. “The adaptive lasso and its oracle properties.” *Journal of the American statistical association* 101 (476): 1418–1429.

Zou, H., and T. Hastie. 2005. “Regularization and variable selection via the elastic net.” *Journal of the royal statistical society: series B (statistical methodology)* 67 (2): 301–320.

A Proofs

A.1 Proof of Neyman Orthogonalization for Double Selection

A parameter of interest τ_0 is the solution to an orthogonalized moment condition of the form

$$\mathbb{E}\psi(\mathbf{Z}_i, \tau_0, \boldsymbol{\eta_0}) = 0$$

where $\psi(\cdot)$ is a real-valued function satisfying the Orthogonality condition $\mathbb{E}\partial_{\boldsymbol{\eta}}\psi(\mathbf{Z}_i, \tau_0, \boldsymbol{\eta}_0) = 0$, a vector of observables $\mathbf{Z}_i := \{y_i, d_i, \mathbf{x}_i\}$, and a high-dimensional nuisance parameter η_0.
For our particular setting, recall the two equations

$$\begin{aligned} y_i &= \tau d_i + \mathbf{x}_i'\boldsymbol{\beta}_0 + \varepsilon_i \;\;,\; \mathbb{E}\left[\varepsilon_i|\mathbf{x}_i, d_i\right] = 0 \\ d_i &= \mathbf{x}_i'\boldsymbol{\delta}_0 + \nu_i \;\;,\; \mathbb{E}(\nu_i|\mathbf{x}_i) = 0 \end{aligned}$$

We denote the nuisance parameter $\boldsymbol{\eta} := (\boldsymbol{\beta}_0, \boldsymbol{\delta}_0)'$. We can now write the corresponding moment condition for our linear setup using the two residuals

$$\mathbb{E}\psi(\mathbf{Z}_i, \tau_0, \eta_0) = \mathbb{E}\Big(\overbrace{(y_i - d_i\tau_0 - \mathbf{x}_i\boldsymbol{\beta}_0)}^{\text{Residual from outcome regression}} \underbrace{(d_i - \mathbf{x}_i'\boldsymbol{\delta}_0)}_{\text{Residual from treatment regression}} \Big) = 0 \tag{3}$$

We now prove that this moment condition satisfies the orthogonality condition $\mathbb{E}\partial_\eta\psi(\cdot) = 0$. To see this, write

$$\partial_\eta\psi(\mathbf{z}_i, \tau, \eta) = \begin{bmatrix} \partial_\beta\psi(\mathbf{z}_i, \tau, \eta) \\ \partial_\eta\psi(\mathbf{z}_i, \tau, \eta) \end{bmatrix} = \begin{bmatrix} -(d_i - \mathbf{x}_i'\boldsymbol{\delta})\mathbf{x}_i \\ -(y_i - d_i\tau - \mathbf{x}_i'\boldsymbol{\beta})\mathbf{x}_i \end{bmatrix}$$

The two pieces of the above expression are normal equations for the corresponding treatment and outcome equations. Taking expectations of these normal equations

$$\mathbb{E}\left[(d_i - \mathbf{x}_i'\boldsymbol{\delta}_0)\mathbf{x}_i\right] = \mathbb{E}\left[\nu_i\mathbf{x}_i\right] = 0$$

and

$$\mathbb{E}\left[(y_i - \tau_0 d_i - \mathbf{x}_i'\boldsymbol{\beta}_0)\mathbf{x}_i\right] = \mathbb{E}\left[\varepsilon_i\mathbf{x}_i\right] = 0$$

Therefore, we have proved that equation 3 satisfies the Neyman Orthogonality condition.

B Simulation Setup and additional results

Here, we describe the simulation study described in section 3. We use covariates from the LaLonde (1986) experimental sample (445 observations from the NSW experiment). We assume a constant treatment effect of $\tau = \$1000$, and specify a non-linear DGP for the outcome

$$\begin{aligned} Y = {} & 1000D + 0.1\exp[0.7(\log(\texttt{re74} + 1))] + 0.7\log(\texttt{re75 +1}) + \\ & 0.6\exp(\log(\texttt{re74}) \times \texttt{hispanic}) - 0.01\texttt{black} \times \log(\texttt{age} + 1) + \epsilon \end{aligned}$$

where $\epsilon \sim \mathcal{N}(0, 10)$.
The true propensity score DGP is the following function

$$\begin{aligned} \pi_i = {} & \text{logit}^{-1}(1 + .4\mu + .1\texttt{age} - .3\texttt{educ} - .09\texttt{re74} - .05\texttt{re75} \\ & + .2\texttt{u74} \times \texttt{u75} + .3\texttt{married} \times \texttt{u75} - .2\log(\texttt{re75}) \times \log(\texttt{age})^2 \\ & - .1\texttt{black} \times \log(\texttt{age}) + .05\texttt{hispanic} \times \log \texttt{education} \\ & + .1\texttt{hispanic} \times \texttt{nodegree} \times \texttt{u74} - .05\texttt{black} \times \texttt{u74} \times \texttt{u75} \\ & - .05\texttt{married} \times \texttt{nodegree} \times \log(\texttt{re74}) + \eta_i) \end{aligned}$$

Estimator	BIAS	MAD	RMSE	Runtime
Difference in Means	-8688.830	6701.07	8693.297	0.002
OLS	182.809	170.63	406.217	0.008
double-LASSO (Double Selection)	138.579	143.19	483.177	5.656
double-LASSO (Knockoff Selector)	-118.660	-119.02	248.369	5.241
double-LASSO (Partial Out)	-209.952	-192.02	449.809	5.589
Entropy Balancing	-0.017	-0.02	1.627	0.098
Entropy Balancing (Knockoff selection)	-0.017	-0.02	1.627	0.012
Mahalanobis Distance Matching	-105.078	-106.04	124.366	0.096
Propensity Score + Mahalanobis Distance Matching	-493.439	-496.68	523.188	0.102
Propensity Score Matching	-2681.354	747.97	3060.527	0.055
Propensity Score Weighting	-191.590	-182.32	293.271	0.007

Table 1. Simulation study results. Bias is computed by averaging $\widehat{\tau} - \tau$ estimate minus the true treatment effect of 1000, averaged over simulations, Mean Absolute Deviation (MAD) by averaging $|\widehat{\tau} - \tau|$, and RMSE as $\sqrt{\overline{(\widehat{\tau} - \tau)^2}}$. Runtime (in seconds) is averaged over all runs. Genetic matching (`rgenoud` is omitted because of prohibitively long runtimes.)

where μ is obtained from regressing the treatment indicator on $\texttt{age}^2, \texttt{educ}^2, \texttt{re74}^2, \texttt{re75}^2, \texttt{u74}, \texttt{u75}, \texttt{black}, \texttt{hispanic}, \texttt{married}, \texttt{nodegree}$. In the monte carlo replications, we use an (incorrect) functional form to estimate the propensity score by regressing the treatment indicator on all the covariates linearly.

For variable-selection methods (Double LASSO and Knockoff-selection), we construct a large set of controls using the following steps:

- Construct a data matrix with the $\log$, linear, and quadratic terms of all continuous variables, and all binary indicators.
- Construct all pairwise interactions
- Construct Box-Cox polynomials ($\log x$, linear, quadratic, and cubic polynomials)
- Drop 0 or near-0 variables (this addresses mechanically impossible interaction terms like those between mutually exclusive categories)
- Drop highly correlated variables (those with pairwise correlation coefficient above 0.95) to avoid multicollinearity

We then use the resultant matrix $\mathbf{X}$ to estimate treatment effects using LASSO partialling out, double selection, and entropy balancing.

For each iteration of the simulation we + simulate treatment status using a Bernoulli draw with individual-treatment probabilities π_i + generate the outcome using the outcome DGP + Estimate treatment effects using all available methods either using no controls (for RAW), a 'narrow' covariate matrix with 10 columns (for MD, PS, PSMD, PSW, EB), or 'wide' covariate matrix (for LPO, LDS, LEB).