Improved Variable Selection for Regression and Balancing Estimators

Apoorva Lal

Stanford University

October 2, 2021

Overview

Introduction

Methodology

Simulation Study

Empirical Applications

Introduction

- Despite its questionable reputation, a selection-on-observables (SOO) research design is extremely common in the social sciences
  - The empirical strategy in such papers is ‘control for stuff and hope for the best’
- This may only be adequate if the set of controls is sufficiently large
- However, we need regularization and variable selection with high-dimensional $\mathbf{x}$
- High-dimensional regression performs Shrinkage and Selection with a single tuning parameter, and can only do well at both under *implausibly strong* assumptions
- **This paper**: apply simulation-based variable selection methods to improve covariate adjustment using balancing and regression estimators
  - Focus on *estimation*, no new results on *identification*

The Contribution

Estimation of treatment effects using two steps.

- (0) Basis expansion and interactions of covariates
- **Variable selection** using double-LASSO approach
    - covariates that predict treatment
    - covariates that predict output
- **Estimation**
    - Balancing estimator
    - Regression

Setup

- Data $\mathcal{D} = (Y_i, D_i, \mathbf{X}_i)_{i=1}^{N} \ \in \ \mathbb{R} \times \{0,1\} \times \mathbb{R}^k$
  - Conventional settings: $k << n$
  - Modern settings: $k \approx n$
- Potential outcomes $Y(D)$, $\tau_i = Y_i(1) - Y_i(0)$. Interested in SATT $:= \frac{1}{n}\sum_{i:D_i=1} \tau_i$
- SOO assumptions for ATT
  - **Unconfoundedness**: $Y(0) \perp\!\!\!\perp D|\ \mathbf{X}$ - need $\mathbf{X}$ to be potentially quite large for this to be true
  - **(weak) Overlap**: $p(\mathbf{X}) < 1$ where $p(\mathbf{X})$ is the propensity score . Higher-dimensional $\mathbf{X}$ might lead to violations of overlap, since in ever smaller cells of $\mathbf{x}$, we cannot find both treatment and control units.
- Need to impute

$$\widehat{\mathbb{E}}[Y(0)|D=1] = \frac{\sum_{i:D=0} Y_i w_i}{\sum_{i:D=0} w_i}$$

Entropy Balancing

- IPW estimators typically rely on inverse propensity score weights $w_i := \widehat{p}(\mathbf{X}_i)/(1 - \widehat{p}(\mathbf{X}_i)$ to make the re-weighted control group look like the treatment group
- Need well-specified $\widehat{p}$: fit, assess-balance loop
- Hainmueller (2012): Choose balancing weights w_i

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

Balance constraints: $\sum_{i:D_i=0} w_i c_{ri}(\mathbf{X}_i) = m_r$ with $r \in 1, \ldots, R$

'Proper' weights: $\sum_{i:D_i=0} w_i = 1$ and $w_i \geq 0 \; \forall \; \{i : D = 0\}$

- Choice of balance condition? Kernel (Wong and Chan 2018; Hazlett 2018), Hierarchical shrinkage (Yang and Xu 2021)

Double-LASSO Covariate Adjustment

- Stipulate a partially linear outcome model $Y_i = \tau D_i + g(\mathbf{X}_i) + U_i$, and approximate the nuisance $g(\cdot)$ using a basis expansion
  - a naive solution would be to simply estimate this regression without shrinking τ
  - this produces 'regularization bias' resulting in biased estimates of $\widehat{\tau}$: form of OVB reg-bias
- Need propensity score model too $D_i = m(\mathbf{X}_i) + V_i$
- **Double-LASSO** (Belloni, Chernozhukov, and Hansen 2014): Fit both outcome and pscore using LASSO regressions, adjust for $\mathcal{S}_1 \cup \mathcal{S}_2$
- **Double ML** (Chernozhukov et al. 2018): use FWL-logic (Robinson 1988) to partial out $\mathbf{X}$s and regress residuals on residuals

Variable Selection using LASSO Regression

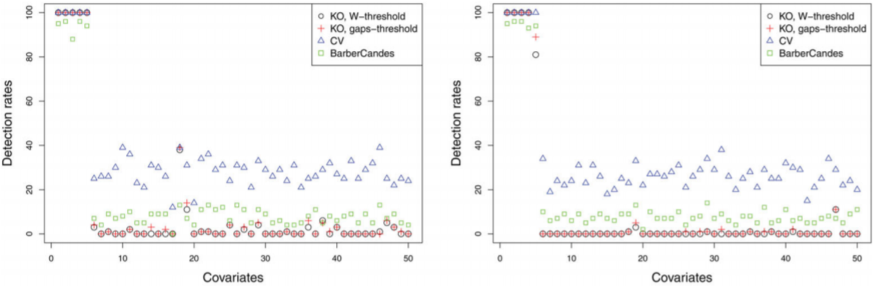
$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j \right) + \underbrace{\lambda \sum_{j=1}^{p} \|\beta_j\|_1}_{\text{regularisation term}}$$

- λ penalises model complexity, $\|.\|_1$ induces sparsity in the estimated coefficient vector $\widehat{\boldsymbol{\beta}}$
- Choosing λ is a challenging problem because a single parameter needs to perform both shrinkage and selection
- 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.

Better Variable Selection: Knockoffs

- Approach proposed by Barber and Candes Barber and Candès (2015)
- Basic idea - if 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
  - Under some implausibly strong assumptions, this controls the False Discovery Rate (FDR) in finite samples
- 'cheap' knockoff: permute rows of design matrix $\mathbf{X}$ to construct $\tilde{\mathbf{X}}$ Gégout-Petit, Gueudin-Muller, and Karmann (2020)
- for LASSO regression, define test statistic $T_j := \sup\{\lambda > 0, \hat{\beta}_j(\lambda) \neq 0\} \;\; j \in \{1, \ldots, 2p\}$

$$W_j := T_j \vee \tilde{T}_j \times \begin{cases} +1 & T_i \geq \tilde{T}_i \\ -1 & T_i \leq \tilde{T}_i \end{cases} \quad \text{keep if } W_j \geq q$$



(a) $\boldsymbol{\beta} = (1, 1, 1, 1, 1, 0, \ldots, 0)$.

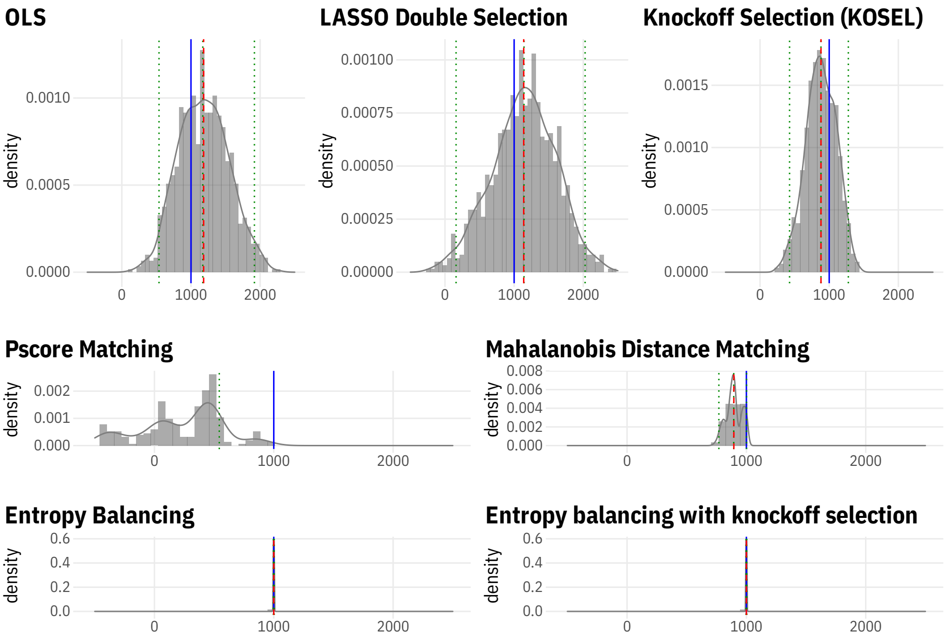
(b) $\boldsymbol{\beta} = (2.5, 2, 1.5, 1, 0.5, 0, \ldots, 0)$.

The Contribution

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 (2014), wherein the researcher regresses the outcome on treatment, controlling for variables that are predictive of either the outcome of the treatment (i.e. $\mathcal{S}_1 \cup \mathcal{S}_2$).

Simulation Study on Lalonde (1986) sample

- Take 445 observations from the original NSW study
- Specify highly nonlinear outcome and pscore model DGP
- Study bias and variance of standard estimators to benchmark the knockoff-based proposals over many replications


OLS
LASSO Double Selection
Knockoff Selection (KOSEL)
Pscore Matching
Mahalanobis Distance Matching
Entropy Balancing
Entropy balancing with knockoff selection
density
0.0000
0.0005
0.0010
0.00000
0.00025
0.00050
0.00075
0.00100
0.0015
0.000
0.001
0.002
0.002
0.004
0.006
0.008
0.0
0.2
0.4
0.6
0
1000
2000

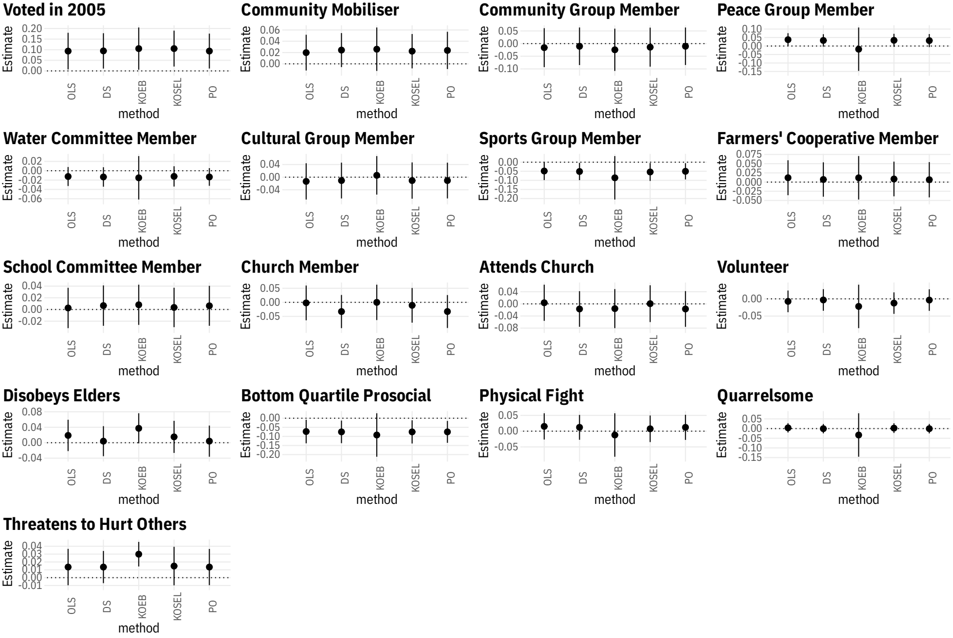
MSE Comparison

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
PScore + MD 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

Entropy Balancing also outright fails frequently, while knockoff entropy balancing does not.

Blattman (2009)

- studies the effects of involuntary rebel recruitment on postwar political engagement and socio-political behaviour of ex-combatants using an individual level dataset.
- 36 controls, 542 observations
- All pairwise interactions : $36 + (36 \times 35)/2 = 666$



Marginal Effects of Abduction on Social And Political Participation

OLS includes all covariates, PO, DS, EB include all covariates and pairwise interactions

Voted in 2005

Community Mobiliser

Community Group Member

Peace Group Member

Water Committee Member

Cultural Group Member

Sports Group Member

Farmers' Cooperative Member

School Committee Member

Church Member

Attends Church

Volunteer

Disobeys Elders

Bottom Quartile Prosocial

Physical Fight

Quarrelsome

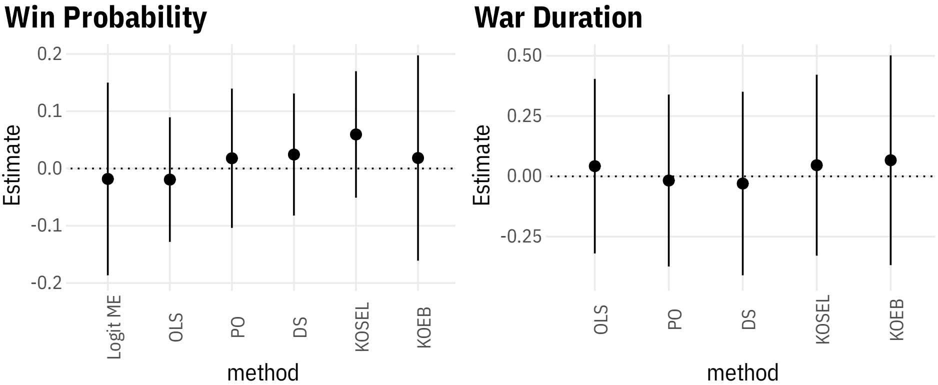
Threatens to Hurt Others

Lyall (2010)

- effect of democracy’s impact on counterinsurgency (COIN) war outcomes (extensive margin) and duration (intensive margin).
- data on war outcomes and covariates for internal conflicts from 286 insurgencies from 1800-2005

Marginal Effects of Democracy on Intensive and Extensive margin

OLS includes all controls, PO and DS includes all controls and n-way interactions



Win Probability

War Duration

Conclusion

- Choice of variables to adjust for is a major ‘researcher degree-of-freedom’
- Recent advances in using high-dimensional regression have appealing theoretical properties, but variable selection step has suboptimal finite sample properties
- We propose a method of combining recent advances in variable selection with balancing and regularized regression (double-LASSO) estimators to estimate causal effects
- **caution:** Conditioning on Post-treatment variables is still bad (Hünermund, Louw, and Caspi 2021)

[1] Rina Foygel Barber and Emmanuel J Candès. “Controlling the false discovery rate via knockoffs”. In: *The Annals of Statistics* 43.5 (2015), pp. 2055–2085.

[2] Alexandre Belloni, Victor Chernozhukov, and Christian Hansen. “Inference on treatment effects after selection among high-dimensional controls”. In: *The Review of Economic Studies* 81.2 (2014), pp. 608–650.

[3] Christopher Blattman. “From violence to voting: War and political participation in Uganda”. In: *American political Science review* (2009), pp. 231–247.

[4] Victor Chernozhukov et al. “Double/debiased machine learning for treatment and structural parameters”. In: *The econometrics journal* 21.1 (Feb. 2018), pp. C1–C68. URL: http://doi.wiley.com/10.1111/ectj.12097.

[5] Anne Gégout-Petit, Aurélie Gueudin-Muller, and Clémence Karmann. “The revisited knockoffs method for variable selection in L 1-penalized regressions”. In: *Communications in Statistics-Simulation and Computation* (2020), pp. 1–14.

[6] Jens Hainmueller. “Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies”. In: *Political analysis* (2012), pp. 25–46.

[7] Chad Hazlett. “Kernel balancing: A flexible non-parametric weighting procedure for estimating causal effects”. In: *Available at SSRN 2746753* (2018).

[8] Paul Hünermund, Beyers Louw, and Itamar Caspi. “Double Machine Learning and Bad Controls–A Cautionary Tale”. In: *arXiv preprint arXiv:2108.11294* (2021).

[9] Jason Lyall. “Do Democracies Make Inferior Counterinsurgents? Reassessing Democracy’s Impact on War Outcomes and Duration”. In: *International Organization* 64.1 (2010), pp. 167–192.

[10] Peter M Robinson. “Root-N-consistent semiparametric regression”. In: *Econometrica: Journal of the Econometric Society* (1988), pp. 931–954.

[11] Raymond KW Wong and Kwun Chuen Gary Chan. “Kernel-based covariate functional balancing for observational studies”. In: *Biometrika* 105.1 (2018), pp. 199–213.

[12] Kaspar Wuthrich and Ying Zhu. “Omitted variable bias of Lasso-based inference methods: A finite sample analysis”. In: *The review of economics and statistics* (2021). URL: http://arxiv.org/abs/1903.08704.

[13] Eddie Yang and Yiqing Xu. “Hierarchically Regularized Entropy Balancing”. In: *Working Paper* (2021). URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3807620.

Regularization Bias

Assume we fit

$$\widehat{\tau} = \left(\frac{1}{n}\sum_i D_i^2\right)^{-1}\left(\frac{1}{n}\sum_i D_i(Y_i - \widehat{g}(\mathbf{X}_i))\right)$$

If $D_i = m(\mathbf{X}_i) + V_i$,

$$\sqrt{n}(\widehat{\tau} - \tau) = \underbrace{\left(\frac{1}{n}\sum_i D_i^2\right)^{-1}\left(1/\sqrt{n}\sum_i D_i U_i\right)}_{\rightsquigarrow \mathcal{N}(0,\cdot)}$$

$$+ \underbrace{\left(\frac{1}{n}\sum_i D_i^2\right)^{-1}\left(\sqrt{n}\sum_i D_i(g - \widehat{g})\right)}_{\text{Since } \widehat{g} \text{ is biased, } \not\to 0}$$

Back

DGP for Simulation Study

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

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

Back