Inference on Treatment Effects with High-dimensional Controls: Frequentist and Bayesian approaches

Duong Trinh

MRes Dissertation (ECON 5088P)

2020-2021

Supervisors:

Professor *Dimitris Korobilis* - Dr *Kenichi Shimizu*

University of Glasgow

Outline

Introduction

An observational study under *Conditional-on-Observables* assumption.
We focus on a *linear regression model*:

$$\underbrace{y_i}_{\text{outcome}} = \overbrace{\alpha}^{\text{effect}} \underbrace{T_i}_{\text{treatment}} + \underbrace{\sum_{j=1}^{p} x_{ij}\beta_j}_{\text{controls}} + \underbrace{\epsilon_i}_{noise}, \quad \mathbb{E}\left[\epsilon \mid T_i, x_i\right] = 0 \tag{1}$$

the number of *possible* controls is large, *specific* controls needed are unknown.

$X_{n\times p}$ - the dictionary of *possible* controls:

- can be richer as more features become available
- can contain transformation of "raw" controls in an effort to make models more flexible

$\Longrightarrow$ which *specific* controls do we use?

Regression with High-dimensional Controls

which *specific* controls do we use?

- If we use too few, or use the wrong ones, then OLS gives us a biased estimate of α because of omitted variable bias.
- If we use too many, the estimate is far less precise. When $p > n$, using them all is just impossible (OLS estimator is not identified).

$\Longrightarrow$ this forces us to consider **variable selection** to select controls that are "most relevant"?

Outline

Modern Variable Selection

1. The general model for variable selection in a high-dimensional setting:

$$y_i = \boldsymbol{z}_i'\boldsymbol{\theta} + \epsilon_i \tag{2}$$

2. Frequentist penalized likelihood methods, e.g. The LASSO

Least Absolute Shrinkage and Selection Operator [Tibshirani, 1996]

$$\hat{\boldsymbol{\theta}}_L = \arg\min \underbrace{\sum_{i=1}^{n} (y_i - \boldsymbol{z}_i'\boldsymbol{\theta})^2}_{\text{SSE}} \quad \text{s.t.} \quad \underbrace{\sum_{j=1}^{p} |\theta_j|}_{\text{penalty function } (\ell_1-\text{norm})} < \tau \tag{3}$$

The effect of the penalization is that LASSO sets the θ_j for some variables to zero, i.e. it does the **variable selection** for us.

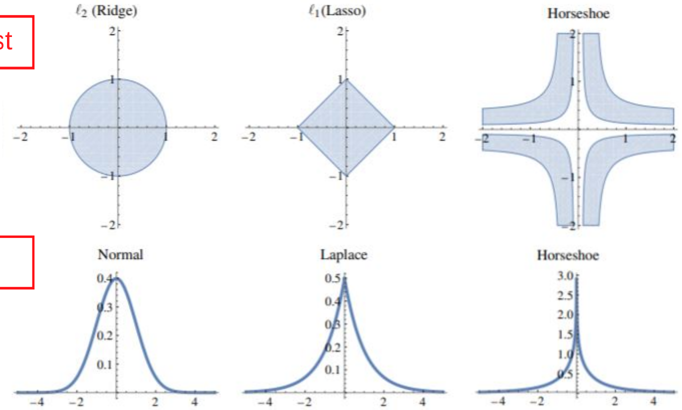
❸ Frequentist penalized likelihood methods vs. Bayesian shrinkage methods

Frequentist

Penalty function

Bayesian

Shrinkage prior



Shrink small coefficients towards zero and Leave substantially large coefficients large

Variable Selection when the Goal is Causal Inference

Our model is:

$$\text{Original Eq.: } y_i = \underbrace{\alpha T_i}_{\text{interest}} + \underbrace{\boldsymbol{x}_i' \boldsymbol{\beta}}_{\text{controls}} + \epsilon_i, \quad \forall i = 1, \dots, n \tag{4}$$

- Naive approach: apply directly a variable-selection technique (e.g. Lasso) directly to (4), and use the selected controls. $\longrightarrow$ Badly biased!
- How to make it right?

  (4) could be written as:

$$\text{Treatment Eq.: } T_i = \boldsymbol{x}_i' \boldsymbol{\beta_t} + \nu_i, \quad \forall i = 1. \dots, n \tag{5}$$

$$\text{Outcome Eq.: } y_i = \boldsymbol{x}_i' \boldsymbol{\beta_y} + \eta_i, \quad \forall i = 1, \dots, n \tag{6}$$

  We should avoid *Post-Single Selection*.

Frequentist Approach

Post-Double-Selection (PDS) Lasso [Belloni et al., 2014]:

- Step 1: Use Lasso to estimate Treatment equation (5),

$$T_i = \beta_{t1}x_{i,1} + \beta_{t2}x_{i,2} + \ldots + \beta_{tp}x_{i,p} + \nu_i$$

  Denote the set of Lasso-selected controls by S_1.

- Step 2: Use Lasso to estimate Outcome equation (6),

$$y_i = \beta_{y1}x_{i,1} + \beta_{y2}x_{i,2} + \ldots + \beta_{yp}x_{i,p} + \eta_i$$

  Denote the set of Lasso-selected controls by S_2.

- Step 3: Estimate original model (4) by OLS using the union of selected controls from steps 1 and 2, i.e. $\boldsymbol{w}_i' = S_1 \cup S_2$:

$$y_i = \underbrace{\alpha T_i}_{\text{interest}} + \underbrace{\boldsymbol{w}_i'\boldsymbol{\beta}}_{\text{controls}} + \epsilon_i$$

Finally, we can make inference on the treatment effect α of interest.

Bayesian Approach

We generalize High-dimensional Confounding Adjustment [Antonelli et al., 2019]

- Step 1: Use Lasso to estimate Treatment equation (5),

$$T_i = \beta_{t1} x_{i,1} + \beta_{t2} x_{i,2} + \ldots + \beta_{tp} x_{i,p} + \nu_i$$

  Denote the set of Lasso-selected controls by S_1.

- Step 2: Apply a Bayesian method to the original model (4), yet reduce the amount of shrinkage on coefficients of selected controls in step 1 (set $S1$).

$$y_i = \underbrace{\alpha T_i}_{\text{interest}} + \underbrace{\boldsymbol{x}_i' \boldsymbol{\beta}}_{\text{controls}} + \epsilon_i, \quad \forall i = \overline{1, n}$$

  $\leftrightarrow$ borrow information from the treatment model to guide the amount of shrinkage in the original model.

Bayesian Approach

- Step 2 (cont.):

For $j = \overline{1, p}$, a hierarchical model can be summarized as:

$$\boldsymbol{y}_i \mid \boldsymbol{T}_i, \boldsymbol{x}_i', \alpha, \boldsymbol{\beta}, \sigma^2 \sim \text{Normal}\ \left(\alpha \boldsymbol{T}_i + \boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2\right) \quad \forall i = \overline{1, n}$$

$$\alpha \mid \sigma^2 \sim \text{Normal}\ \left(0, \sigma^2 K\right)$$

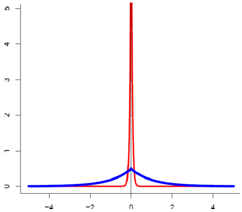
$$\beta_j \mid \gamma_j, \sigma^2, \tau_{0j}^2, \tau_{1j}^2 \sim (1 - \gamma_j) \underbrace{\text{Normal}\ \left(0, \sigma^2 \tau_{0j}^2\right)}_{\text{spike}} + \gamma_j \underbrace{\text{Normal}\ \left(0, \sigma^2 \tau_{1j}^2\right)}_{\text{slab}}$$

$$\tau_{0j}^2, \tau_{1j}^2 \sim \pi\left(\tau_{0j}^2, \tau_{1j}^2\right) \forall j = 1, \ldots, p$$

$$\sigma^2 \mid c, d \sim \text{Inv-Gamma}\ (c, d)$$

$$\gamma_j \mid \theta, \omega_j \sim \text{Bernoulli}\ \left(\theta^{\omega_j}\right)$$

$$\theta \mid a, b \sim \text{Beta}\ (a, b)$$



where $\pi\left(\tau_{0j}^2, \tau_{1j}^2\right)$ depends on the specified prior.

We consider `SSVS Normal`, `SSVS Lasso`, `SSVS Horseshoe`, `SSVS Student`

Monte-Carlo Study

1. Aim: Evaluating finite-sample performance of different variable selection methods (Frequentist and Bayesian) for inference on the treatment effect with high-dimensional controls:

$$y_i = \underbrace{\alpha T_i}_{\text{interest}} + \underbrace{\boldsymbol{x}_i'\boldsymbol{\beta}}_{\text{controls}} + \epsilon_i, \quad \forall i = \overline{1, n}$$

2. Methods:
   - Frequentist method: `PDS Lasso`;
   - Bayesian methods:

     `SSVS Normal`, `SSVS Lasso`, `SSVS Horseshoe`, `SSVS Student`.

Monte-Carlo Study

3. Data-Generating Processes: $(n, p) = (300, 400)$, $\alpha = 1$ (true TE)

$$\text{First Stage} \quad T_i = \boldsymbol{x}_i' \boldsymbol{\beta}_t + \nu_i$$
$$\text{Second Stage} \quad y_i = \alpha T_i + \boldsymbol{x}_i' \boldsymbol{\beta} + \epsilon_i$$

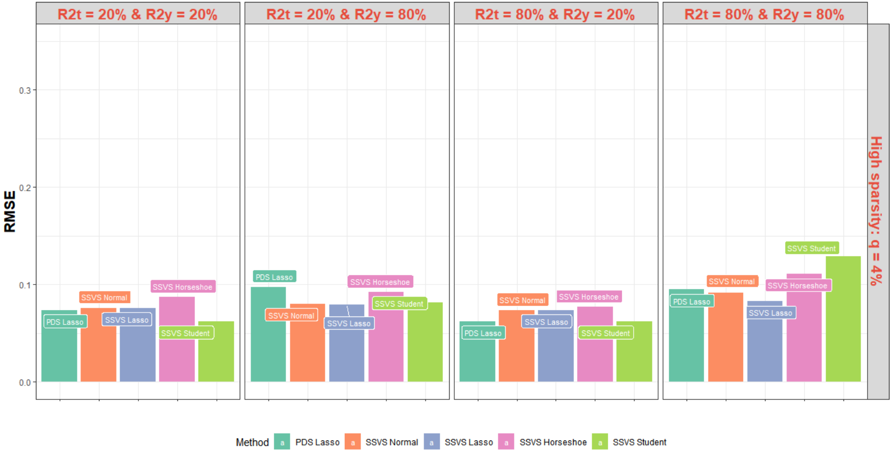
- Uncorrelated covariates: $x_{ij} \overset{iid}{\sim} N(0,1)$ for all j and i
- Homoskedastic errors: $\epsilon_i \sim$ i.i.d. $N(0,1)$ and $\nu_i \sim$ i.i.d. $N(0,1)$
- 5 types of covariates: strong confounders, weak confounders, instrumental variables, strong predictors, noise variables.
- Sparsity level: $q = 4\%$ (high sparsity) or $q = 40\%$ (low sparsity)
- Signal-to-noise ratios: $\left(R_t^2, R_y^2\right) \in \{(20\%, 20\%), (20\%, 80\%), (80\%, 20\%), (80\%, 80\%)\}$

$\longrightarrow$ 8 scenarios. We run $N_{sim} = 48$ simulations for each.

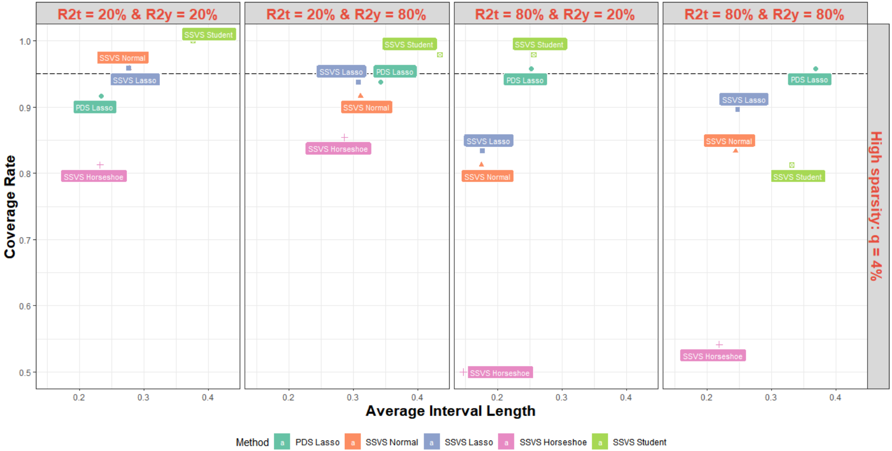
4. Performance Metrics: Mean-absolute-error (MAE), Root-mean-squared-error (RMSE), Empirical coverage, Inclusion probability.

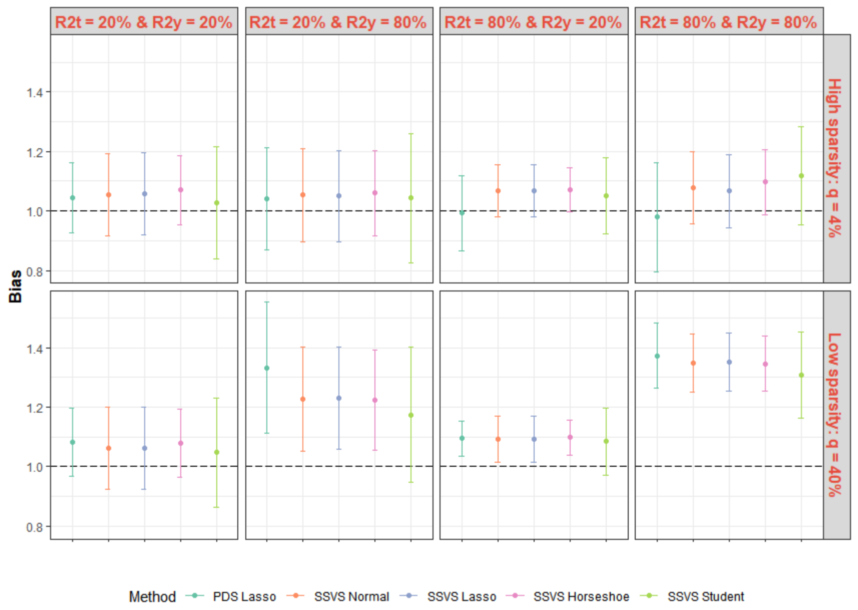
Simulation Results

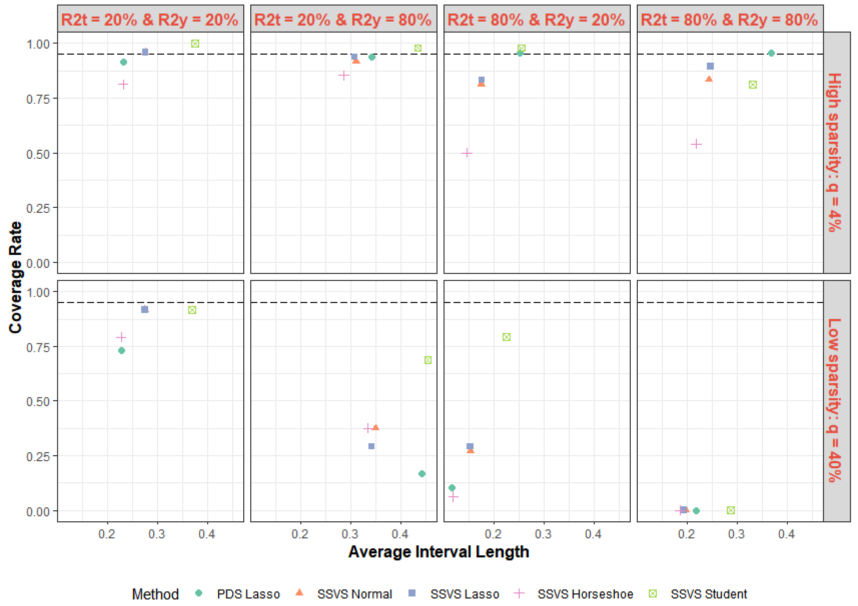
Simulation Results



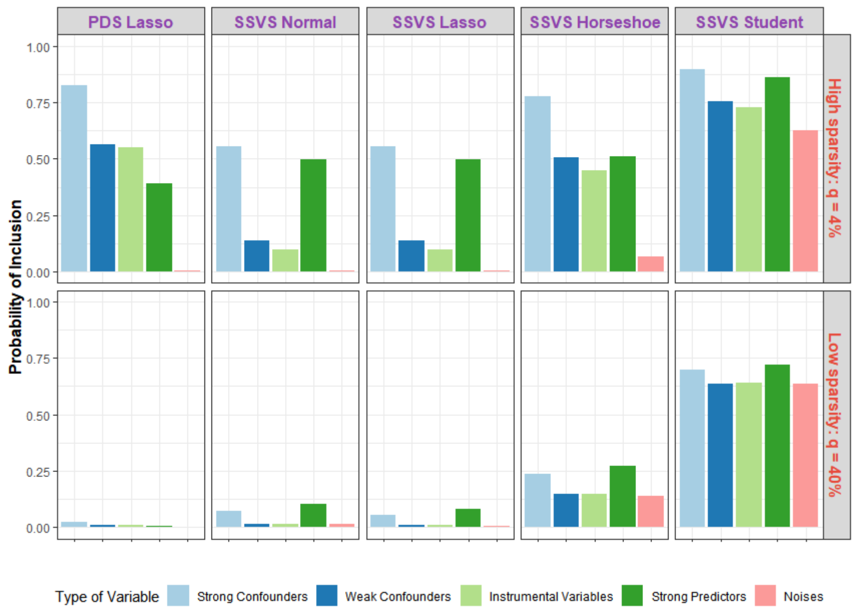
Simulation Results




R2t = 20% & R2y = 20%
R2t = 20% & R2y = 80%
R2t = 80% & R2y = 20%
R2t = 80% & R2y = 80%
High sparsity: q = 4%
Low sparsity: q = 40%
Bias
1.4
1.2
1.0
0.8
Method
PDS Lasso
SSVS Normal
SSVS Lasso
SSVS Horseshoe
SSVS Student


R2t = 20% & R2y = 20%
R2t = 20% & R2y = 80%
R2t = 80% & R2y = 20%
R2t = 80% & R2y = 80%
High sparsity: q = 4%
Low sparsity: q = 40%
Coverage Rate
1.00
0.75
0.50
0.25
0.00
0.2
0.3
0.4
Average Interval Length
Method
PDS Lasso
SSVS Normal
SSVS Lasso
SSVS Horseshoe
SSVS Student


R2t = 20% & R2y = 20%
R2t = 20% & R2y = 80%
R2t = 80% & R2y = 20%
R2t = 80% & R2y = 80%
High sparsity: q = 4%
Low sparsity: q = 40%
RMSE
0.0
0.1
0.2
0.3
Method
PDS Lasso
SSVS Normal
SSVS Lasso
SSVS Horseshoe
SSVS Student


PDS Lasso
SSVS Normal
SSVS Lasso
SSVS Horseshoe
SSVS Student
High sparsity: q = 4%
Low sparsity: q = 40%
Probability of Inclusion
1.00
0.75
0.50
0.25
0.00
Type of Variable
Strong Confounders
Weak Confounders
Instrumental Variables
Strong Predictors
Noises

Monte-Carlo Study (cont.)

5. Remarks:
   - PDSLasso dominates Bayesian methods in high-sparsity designs thanks to its stable performance. However, in low-sparsity designs, PDSLasso fails to select confounders and performs the worst.
   - SSVSStudent produces relatively large standard errors as a trade-off for the highest coverage rates. SSVSStudent is good at selecting confounders, but simultaneously includes the highest number of non-confounding factors.

     SSVSNormal and SSVSLasso perform quite similar. Although they cannot achieve the high coverage rates as SSVSStudent, their standard errors are smallest. SSVSLasso and SSVSNormal perform well in excluding non-confounding factors, but at the same time, they select less confounders into slab.

     SSVSHorseshoe often produces the highest bias and RMSE.
   - Regarding implementation, Bayesian methods require MCMC samplings $\longrightarrow$ more computational intensive compared to PDSLasso.
   - Low-sparsity or high SNR in the second stage degrades performance of all methods.

Empirical Illustration

We revist an observational study, under Conditional-on-Observables assumption: *Media and Political Persuasion: Evidence from Russia* [Enikolopov et al., 2011]

1. Aim: the causal effect of the only independent national TV channel, NTV, on the official voting results during the Russian 1999 election.
2. Hypothesis:
   - There is a *significant negative effect* of NTV availability on vote for pro-government **Unity**, which was criticized by NTV and praised by other national TV channels.
   - There is a *significant positive effect* of NTV availability on vote for all parties supported by NTV (centrist opposition **OVR** and liberal opposition **Yabloko** and **SPS**).
   - The effect of NTV availability on vote for parties which get similar coverage by NTV and state TVs (communist **KPRF** and nationalist **LDPR**) is *ambiguous*.
   - Prediction about the effect of NTV on **voter turnout** is *ambiguous*.

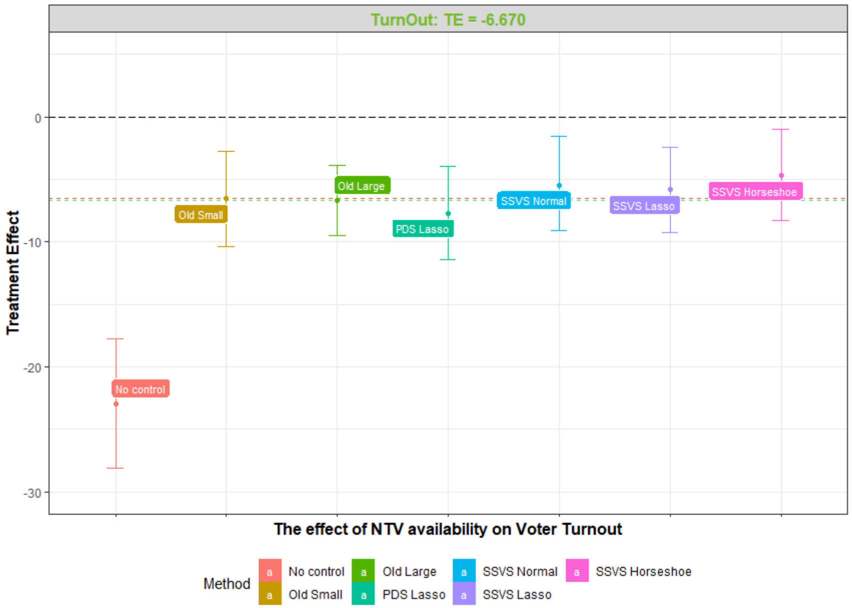
Empirical Illustration

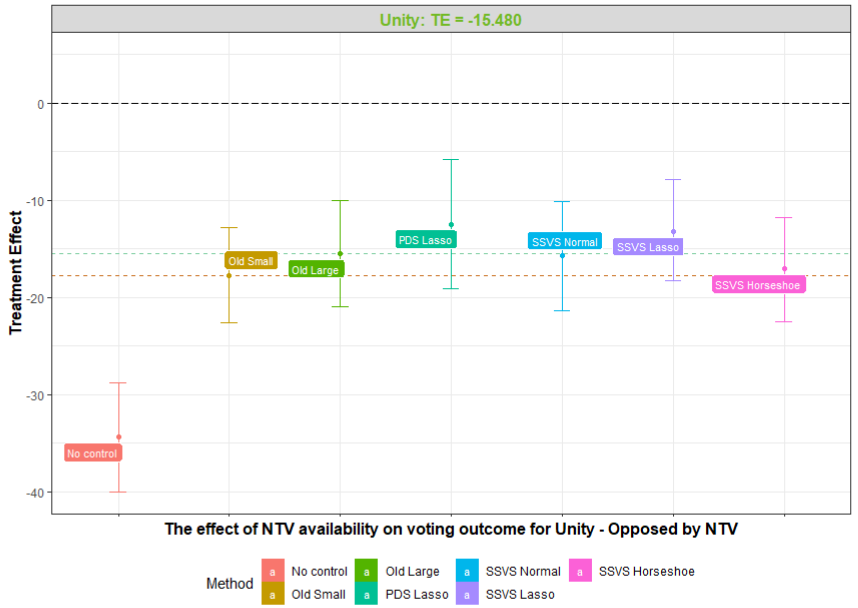
❸ Model Specification

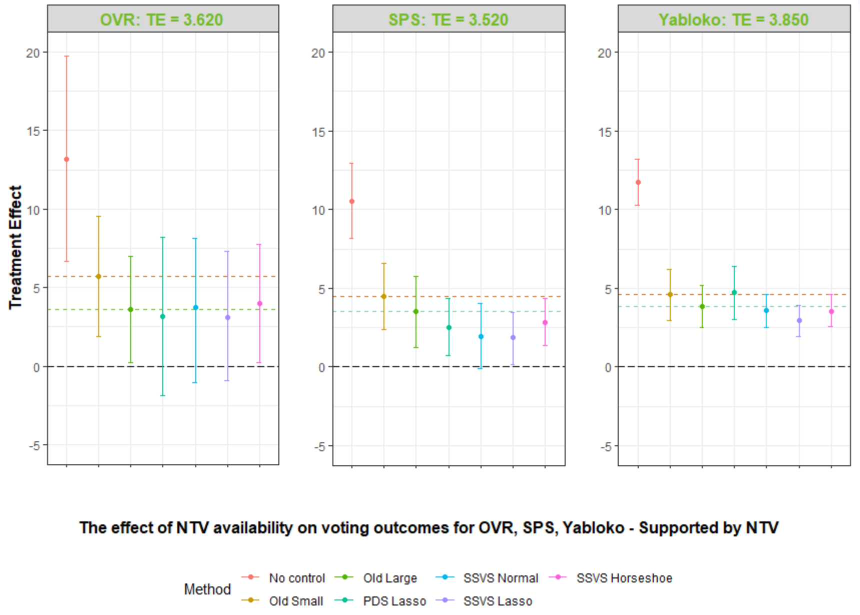
$$\underbrace{\text{vote}^{j}_{s,1999}}_{\text{voting outcome}} = \beta_0 + \overbrace{\alpha}^{\text{effect}} \underbrace{\text{NTV}_{s,1999}}_{\text{NTV availability}} + \underbrace{\overbrace{\beta_1' \boldsymbol{X}_{s,1995}}^{\text{results 1995}} + \overbrace{\beta_2' \boldsymbol{E}_{s,1998}}^{\text{socioeconomic}} + \overbrace{\delta_r}^{\text{regional}}}_{\text{controls}} + \underbrace{\varepsilon^{j}_{s}}_{\text{noise}} \quad (7)$$

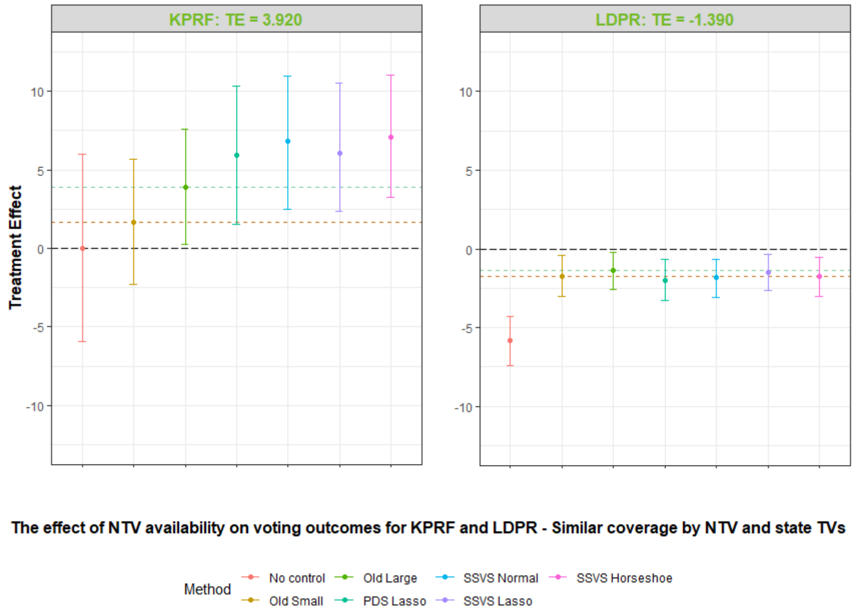
- Sample size: $n = 1568$ (sub-regions)
- Original analysis: Small set $= 91$ controls; Large set $= 99$ controls.
- New analysis: We add all *second-order polynomials* of the continuous covariates and all possible *first-order interactions* of non-region variables.
  $\longrightarrow$ New set $= 325$ *potential* controls to flexibly select among.

  OLS fails, can new techniques help?


TurnOut: TE = -6.670
Treatment Effect
0
-10
-20
-30
No control
Old Small
Old Large
PDS Lasso
SSVS Normal
SSVS Lasso
SSVS Horseshoe
The effect of NTV availability on Voter Turnout
Method
No control
Old Small
Old Large
PDS Lasso
SSVS Normal
SSVS Lasso
SSVS Horseshoe


Unity: TE = -15.480
Treatment Effect
0
-10
-20
-30
-40
No control
Old Small
Old Large
PDS Lasso
SSVS Normal
SSVS Lasso
SSVS Horseshoe
The effect of NTV availability on voting outcome for Unity - Opposed by NTV
Method
No control
Old Small
Old Large
PDS Lasso
SSVS Normal
SSVS Lasso
SSVS Horseshoe


OVR: TE = 3.620
SPS: TE = 3.520
Yabloko: TE = 3.850
Treatment Effect
20
15
10
5
0
-5
The effect of NTV availability on voting outcomes for OVR, SPS, Yabloko - Supported by NTV
Method
No control
Old Small
Old Large
PDS Lasso
SSVS Normal
SSVS Lasso
SSVS Horseshoe


KPRF: TE = 3.920
LDPR: TE = -1.390
Treatment Effect
10
5
0
-5
-10
The effect of NTV availability on voting outcomes for KPRF and LDPR - Similar coverage by NTV and state TVs
Method
No control
Old Small
Old Large
PDS Lasso
SSVS Normal
SSVS Lasso
SSVS Horseshoe

Empirical Illustration (cont.)

4. Remarks:

- Only `SSVSStudent` suffers from the issue which hinders the use of OLS so that nothing is estimated.
- `PDSLasso` tends to retain a more parsimonious final set of controls than the *ad hoc* approach (24 to 63 controls).
- Credible interval length of `SSVSLasso` is always smaller than `PDSLasso`, consistent with Antonelli et al. [2019].
- New methods complement the usual careful specification analysis by providing an efficient, data-driven way for regression sensitivity analysis.

Conclusion

We examined the application of Frequentist and Bayesian variable selection methods to inference on treatment effects with high-dimensional controls:

1. The advantages of modern methods in comparison with traditional counterparts in high-dimensional scenarios are undeniable.
2. These methods offer a coherent data-driven complement to ad hoc robustness checks, thus, support causal analysis in linear regression models.
3. The Bayesian approach offers a huge choice of shrinkage priors, thus being potential for new methods. However, a thoughtful implementation is required.

Conclusion

Limitations and implications for future studies:

1. A special case of inference on treatment effects:
   - Restrictions on the set of potential controls: Assumed to be "not bad", e.g. it at least does not contain pre-treatment variables. Distinguishing "good" from "bad" controls remains ambiguous.
     $\longrightarrow$ pay more systematic attention to integrate the idea of Directed Acyclic Graphs (DAGs).
   - Rely on conditional-on-observable assumption: A simplest form in the literature of causal inference.
     $\longrightarrow$ examine the merits of new methods in settings which allow for unmeasured confounding factors, e.g. selecting among IVs, fixed effects, etc.

2. Machine labor cannot replace brain power. It is evident that the methods introduced in this study are potential yet far from the cure-all. New methods should be applied with inference in mind, to quantify our degree of confidence, also importantly, our degree of uncertainty.

Reference

J. Antonelli, G. Parmigiani, and F. Dominici. High-dimensional confounding adjustment using continuous spike and slab priors. *Bayesian analysis*, 14(3):805, 2019.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies*, 81(2):608–650, 2014.

R. Enikolopov, M. Petrova, and E. Zhuravskaya. Media and political persuasion: Evidence from russia. *American Economic Review*, 101(7):3253–85, 2011.

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