Feature Selection by Second-Generation P-Values Can Outperform Oracle Adaptive Lasso

Yi Zuo*, MPH, Jeffrey D. Blume, PhD

Vanderbilt University

March 17, 2021

Contact: yi.zuo@vanderbilt.edu

Outline

Main message *If you have to use p-values for variable selection, use second-generation p-values.*

1. Background
2. Second-Generation P-Values
3. Proposed Algorithm
4. Simulation
5. Real-world example

Background

- What is feature/variable selection?
  - Identify the right support (support recovery)
  - Derive valid parameter estimation

Background

- What is feature/variable selection?
  - Identify the right support (support recovery)
  - Derive valid parameter estimation
- What are current standard procedures?
  - Lasso, ridge, elastic net, adaptive lasso
  - SCAD, MC+
  - ...

Background

- What is feature/variable selection?
  - Identify the right support (support recovery)
  - Derive valid parameter estimation
- What are current standard procedures?
  - Lasso, ridge, elastic net, adaptive lasso
  - SCAD, MC+
  - ...
- Why use second generation p-values (SGPV)?
  - Those procedures don't balance support recovery and parameter estimation well in finite sample sizes.

Second-Generation P-Values

P-values $\in (0, 1)$

- Small value $\Rightarrow$ support for the alternative hypothesis
- Large value $\Rightarrow$ inconclusive
- Big sample size $\Rightarrow$ likely to reject the null even for "tiny" effects

Second-Generation P-Values

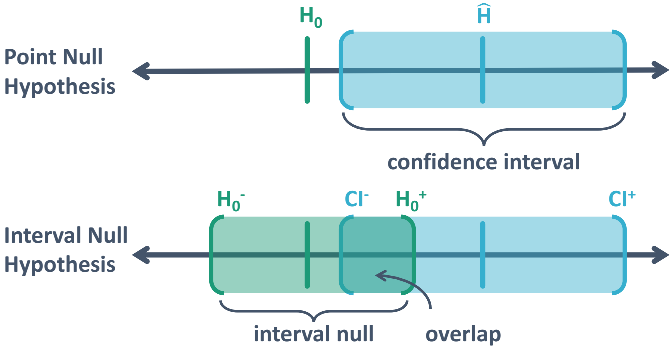
P-values $\in (0, 1)$

- Small value $\Rightarrow$ support for the alternative hypothesis
- Large value $\Rightarrow$ inconclusive
- Big sample size $\Rightarrow$ likely to reject the null even for "tiny" effects

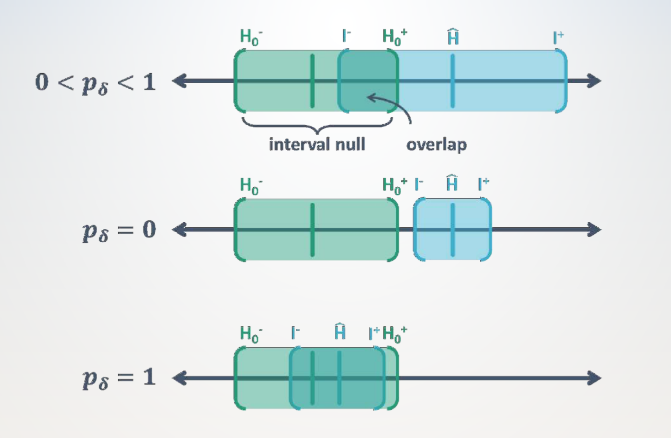
SGPV $\in [0, 1]$

- Small value $\Rightarrow$ support for the alternative hypothesis
- Large value $\Rightarrow$ support for the null hypothesis
- $\sim 1/2 \Rightarrow$ inconclusive

Second-Generation P-Values - Example 1



Second-Generation P-Values - Example 2



Second-Generation P-Values - Definition

Second-generation *p*-value (SGPV)

$$p_\delta = \frac{|I \cap H_0|}{|I|} \times \max\left\{\frac{|I|}{2|H_0|}, 1\right\}$$

Proportion of data-supported hypotheses that are also null hypotheses

Small-sample correction factor

shrinks proportion to ½ when $|I|$ wide

when $|I| > 2|H_0|$

Second-Generation P-Values

So, just use SGPVs to screen variables!

- You are correct... Well, that's the naive one-stage ProSGPV!
- It's extremely fast to implement and works well when $n > p$

So, just use SGPVs to screen variables!

- You are correct... Well, that's the naive one-stage ProSGPV!
- It's extremely fast to implement and works well when $n > p$

What about $p > n$?

- You need a first stage screening to reduce the feature space
- Then you can apply your new favorite SGPVs to screen variables!

Proposed Algorithm - Penalized Regression with SGPVs

```
procedure ProSGPV(X, Y)
    Stage one: fully relaxed lasso
        Standardize all inputs (X, Y)
        Fit cross-validated lasso on the data
        Fit OLS on the lasso active set evaluated at λ_1se
    Stage two: SGPV screening
        Extract confidence intervals of all variables from the last OLS
        Calculate mean standard error SE‾ from all coefficient estimates
        Calculate the SGPV for each variable
        Keep variables with SGPV of zero where the null bound is SE‾
        Re-run the OLS with selected variables on the original scale
end procedure
```

Notes on ProSGPV

The first stage is quite flexible.

- Other feature selection methods can also be used in the first stage. (Elastic net, sure independence screening, etc.)

Notes on ProSGPV

The first stage is quite flexible.

- Other feature selection methods can also be used in the first stage. (Elastic net, sure independence screening, etc.)

Lasso can be evaluated at λ other than λ_{1se}

- As long as it falls into a reasonable range, e.g., $(0.8 * \lambda_{min}, 1.2 * \lambda_{1se})$

Notes on ProSGPV

The first stage is quite flexible.

- Other feature selection methods can also be used in the first stage. (Elastic net, sure independence screening, etc.)

Lasso can be evaluated at λ other than λ_{1se}

- As long as it falls into a reasonable range, e.g., $(0.8 * \lambda_{min}, 1.2 * \lambda_{1se})$

An R package is created to compute the solution to the ProSGPV algorithm.

- ProSGPV: https://CRAN.R-project.org/package=ProSGPV

Simulation - design

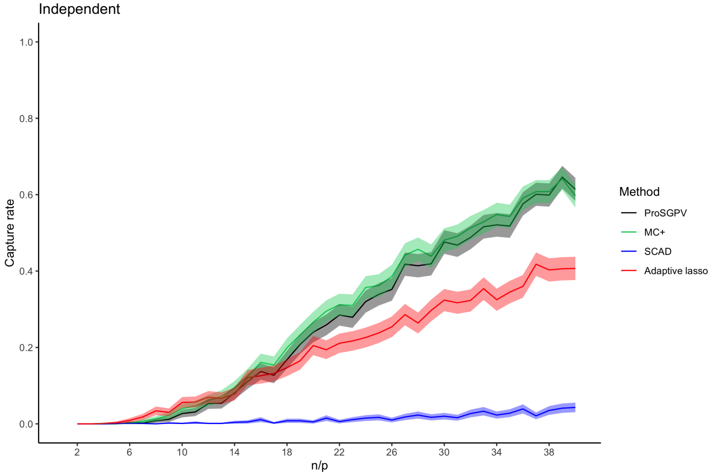
- Step 1. Draw n rows of the matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ i.i.d. from $N_p(0, \Sigma)$ where $\Sigma_{i,j} = \rho^{|i-j|}$
- Step 2. Generate $\boldsymbol{Y} \in \mathbb{R}^{n \times 1}$ from $N_n(\boldsymbol{X}\boldsymbol{\beta}_0, \sigma^2 \boldsymbol{I})$ with σ^2 defined to meet the desired SNR level ν, i.e. $\sigma^2 = \boldsymbol{\beta}_0^T \Sigma \boldsymbol{\beta}_0$
- Step 3. Run SCAD, MC+, adaptive lasso, and ProSGPV
- Step 4. Repeat the simulation 1000 times and aggregate the results

Simulation - setup

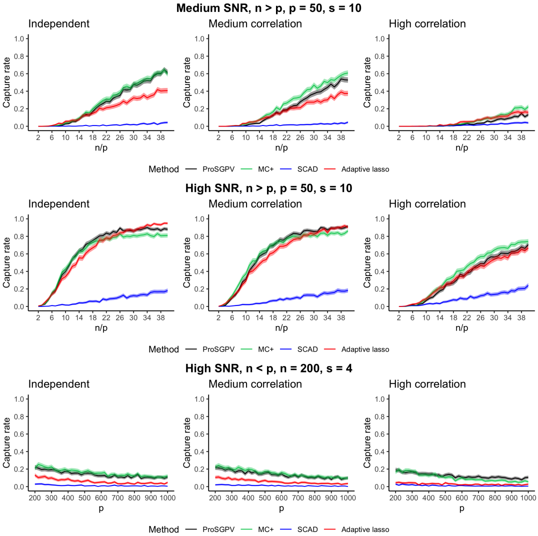
- $\boldsymbol{\beta}_0 \in \mathbb{R}^{p\times 1}$ has s non-zero elements from 1 to 5 with random signs
- $\rho = 0$ (independent), 0.35 (medium), 0.7 (high)
- SNR $= \mathrm{Var}(f(x))/\mathrm{Var}(\epsilon)$, 0.7 (medium), 2 (high)
- Low-dimensional setting $n > p$
  $p = 50$, $s = 10$, n ranges from 100 to 2000 by 50
- High-dimensional setting $p > n$
  $n = 200$, $s = 4$, p ranges from 200 to 2000 by 20

Simulation results - support recovery - example

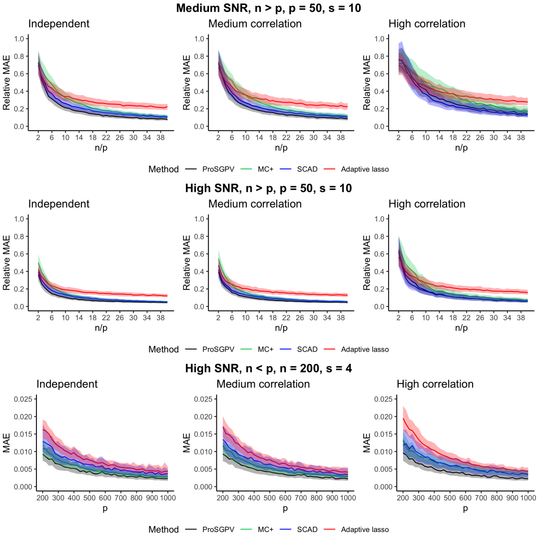
Medium SNR, $n > p$



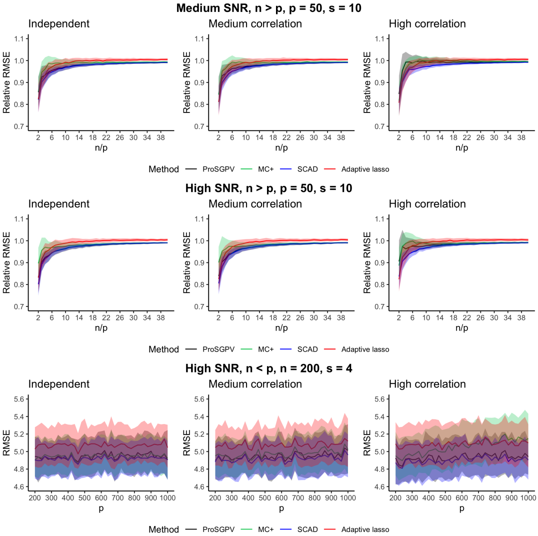
Simulation results - support recovery



Simulation results - parameter estimation bias



Simulation results - prediction accuracy



Real-world example

- Tehran single-family residential apartments data
- Features
  - 5 project physical and financial variables
  - 19 economic variables and indices
  - Outcome: Actual sales prices in 10,000 IRR
- 372 observations, $n/p \approx 14$

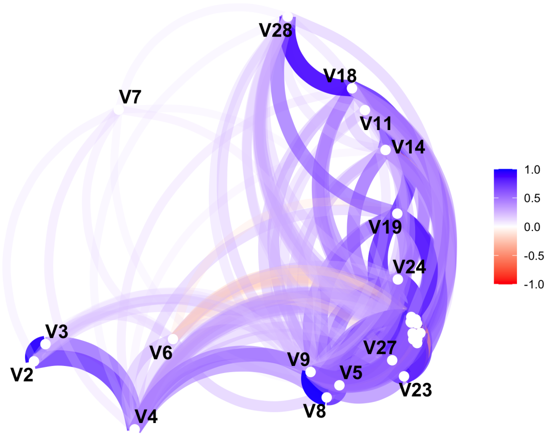
7 project physical and financial features

- V2. Total floor area of the building
- V3. Lot area
- V4. Total Preliminary estimated construction cost based on the prices at the beginning of the project
- V5. Preliminary estimated construction cost based on the prices at the beginning of the project
- V6. Equivalent preliminary estimated construction cost based on the prices at the beginning of the project in a selected base year
- V7. Duration of construction
- V8. Price of the unit at the beginning of the project per square meter

19 economic variables and indices

- V11. The number of building permits issued
- V12. Building services index (BSI) for preselected base year
- V13. Wholesale price index (WPI) of building materials for the base year
- V14. Total floor areas of building permits issued by the city/municipality
- V15. Cumulative liquidity
- V16. Private sector investment in new buildings
- V17. Land price index for the base year
- V18. The number of loans extended by banks in a time resolution
- V19. The amount of loans extended by banks in a time resolution
- V20. The interest rate for loan in a time resolution
- V21. The average construction cost by private sector at the completion of construction
- V22. The average cost of buildings by private sector at the beginning of construction
- V23. Official exchange rate with respect to dollars
- V24. Nonofficial (street market) exchange rate with respect to dollars
- V25. Consumer price index (CPI) in the base year
- V26. CPI of housing, water, fuel & power in the base year
- V27. Stock market index
- V28. Population of the city
- V29. Gold price per ounce

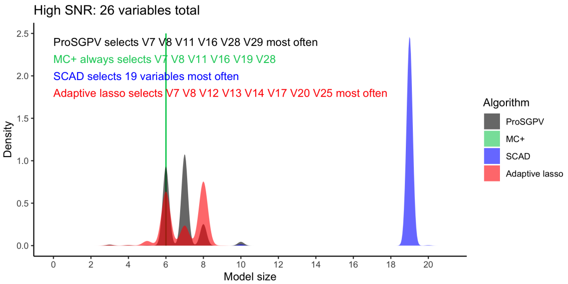
Descriptive statistics - clusters and correlation

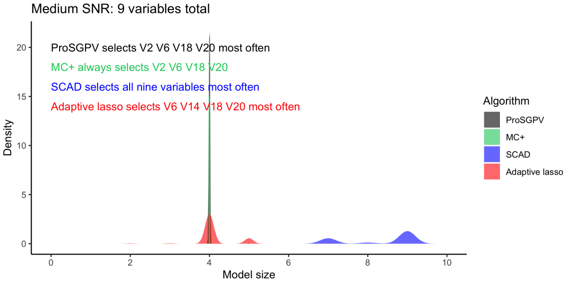


Preliminary analysis

- High SNR: 26 variables total, $R^2 = 0.98$
- Medium SNR: 9 variables total, $R^2 = 0.41$
  - By removing variables whose absolute correlation with the outcome is higher than 0.45.
- Randomly split data into a training set (n=260) and a test set (n=112), record the size of the selected model from each algorithm, and calculate prediction RMSE in the test set
- Repeat the process 1000 times

Feature selection results - model size





Feature selection results - prediction performance

- High SNR: 26 variables total
  - SCAD (294.01)
  - ProSGPV (330.70)
  - AL(348.73)
  - MC+ (458.74)
- Medium SNR: 9 variables total
  - SCAD (1110.46)
  - ProSGPV (1149.51)
  - AL(1189.97)
  - MC+ (4701.58): could be a scaling issue

Takeaway

- Past: shrink point estimates for feature selection
- Now: incorporate confidence intervals (uncertainty) in feature selection

Takeaway

- Past: shrink point estimates for feature selection
- Now: incorporate confidence intervals (uncertainty) in feature selection
- Fully relaxed lasso (first stage) +
  second generation p-values (second stage)
  - Capture the exact true model with high probability
  - Produce parameter estimates with lowest bias in general
  - Yield decent prediction

Takeaway

- Past: shrink point estimates for feature selection
- Now: incorporate confidence intervals (uncertainty) in feature selection
- Fully relaxed lasso (first stage) +
  second generation p-values (second stage)
  - Capture the exact true model with high probability
  - Produce parameter estimates with lowest bias in general
  - Yield decent prediction
- Check out the ProSGPV package on CRAN
  - `https://CRAN.R-project.org/package=ProSGPV`
  - `https://github.com/zuoyi93/ProSGPV`

References

1. Blume, Jeffrey D., et al. "Second-generation p-values: Improved rigor, reproducibility, & transparency in statistical analyses." *PLoS One* 13.3 (2018): e0188299.

2. Blume, Jeffrey D., et al. "An introduction to second-generation p-values." *The American Statistician* 73.sup1 (2019): 157-167.

3. Zuo, Y., Stewart Thomas G., and Blume, Jeffrey D.. "Variable Selection with Second-Generation P-Values." *arXiv preprint arXiv:2012.07941* (2020).

4. Zuo, Y., Stewart Thomas G., and Blume, Jeffrey D.. "ProSGPV: Penalized Regression with Second-Generation P-Values". R package version 0.1.0, URL: `https://CRAN.R-project.org/package=ProSGPV` (2021).

5. Rafiei, Mohammad Hossein, and Hojjat Adeli. "A novel machine learning model for estimation of sale prices of real estate units." *Journal of Construction Engineering and Management* 142.2 (2016): 04015066.