An Investigation of Fully Relaxed Lasso and Second-Generation P-Values for High-Dimensional Feature Selection

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

Vanderbilt University

March 23, 2020

Contact: yi.zuo@vanderbilt.edu

Outline

Problem

- How to select good models for inference in high dimensional settings?
- Regularization can be used to reduce the feature space.
- Fully relaxed LASSO retains desirable prediction performance while yielding a model with coefficients on the original data scale.
- Second-generation p-values (SGPV) were proposed in large-scale multiple testing where an interval null hypothesis can be constructed to indicate when the data support only null, only alternative hypotheses or inconclusive.

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
- Sample size doesn't confound comparison.

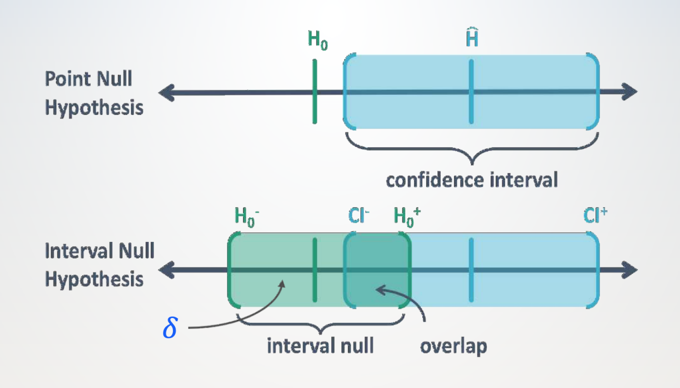


Figure 1: SGPV example 1

Second-Generation P-Values

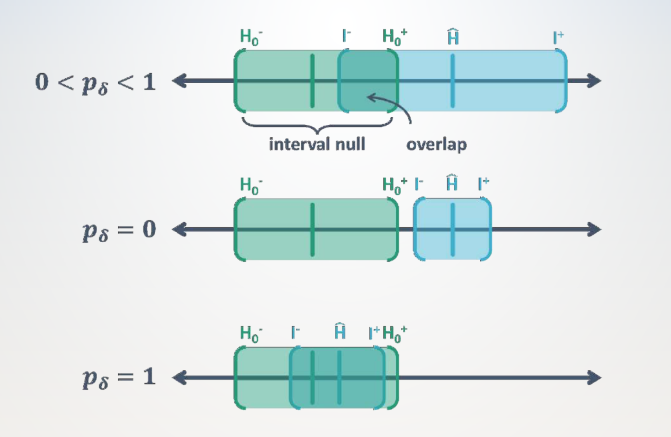


Figure 2: SGPV example 2

Second-Generation P-Values

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 $|I| > 2|H_0|$

Figure 3: Definition of SGPV

LASSO and Fully Relaxed LASSO

- The objective function of LASSO:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \frac{1}{N} ||\boldsymbol{y} - \boldsymbol{X}\beta||_2^2 + \lambda ||\boldsymbol{\beta}||_1 \right\}$$

- Implementation in `R`: `cv.glmnet`, `lambda.min`
- Fully relaxed LASSO: OLS on variables with non-zero shrunken coefficients

  Coefficients on the original data scale

Proposed Algorithm

- Use LASSO to first screen the independent variables (`cv.glmnet`, `lambda.min`)
- Fit OLS model on variables selected from the previous step
- Use the mean standard error of all β coefficients as the null bound
- Use SGPV to drop variables with trivial effects
- Model on variables that survive SGPV cutoff

Simulation

Simulation setup

- p=50, ten signals, $n = 100, 200, ..., 1000$, num.sim=1000
- Correlation structure of X
  - Independent: $\mathbf{\Sigma}_X = I$
  - Auto-regressive: $\mathbf{\Sigma}_X = \sigma_{i,j} = 0.5^{|i-j|}$ for $|i-j| \leq 10$, and $\sigma_{i,j} = 0$ otherwise.
  - Block diagonal: each block submatrix has dimension 5 x 5 and constant entries $\sigma_{i,j} = 0.5$, $i \neq j$ for $|i-j| \leq 5$, and $\sigma_{i,j} = 0$ otherwise.
- Signal to noise ratio, $\sigma_{noise} = 1$
  - High: $|\beta_k| \geq 0.4$, $k = 1, 2, ..., 10$
  - Low: $|\beta_k| \geq 0.2$, $k = 1, 2, ..., 10$
- Evaluate the rate of capturing the exactly true model using the proposed algorithm

Simulation

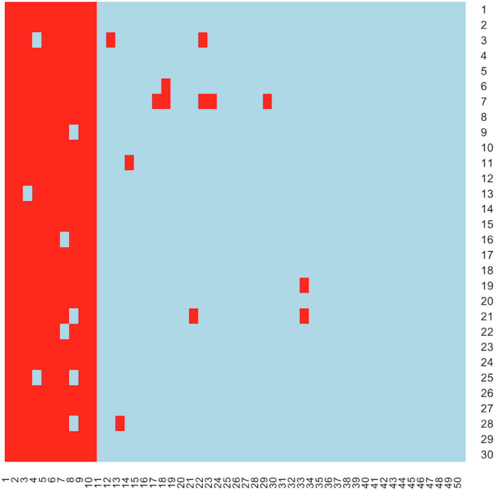


Figure 4: Example output: autoregressive, n=200, high SNR

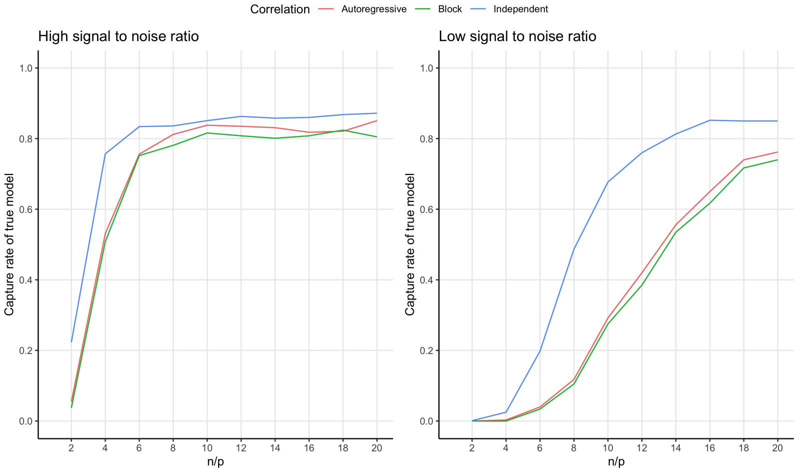


Figure 5: Simulation results

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, $n/p \approx 15$

5 project physical and financial features

- 1. Total floor area of the building
- 2. Lot area
- 3. Preliminary estimated construction cost based on the prices at the beginning of the project
- 4. Duration of construction
- 5. Price of the unit at the beginning of the project per m^2

19 economic variables and indices

- 6. The number of building permits issued
- 7. Building services index (BSI) for preselected base year
- 8. Wholesale price index (WPI) of building materials for the base year
- 9. Total floor areas of building permits issued by the city/municipality
- 10. Cumulative liquidity
- 11. Private sector investment in new buildings
- 12. Land price index for the base year
- 13. The number of loans extended by banks in a time resolution
- 14. The amount of loans extended by banks in a time resolution
- 15. The interest rate for loan in a time resolution
- 16. The average construction cost by private sector at the completion of construction
- 17. The average cost of buildings by private sector at the beginning of construction
- 18. Official exchange rate with respect to dollars
- 19. Nonofficial (street market) exchange rate with respect to dollars
- 20. Consumer price index (CPI) in the base year
- 21. CPI of housing, water, fuel & power in the base year
- 22. Stock market index
- 23. Population of the city
- 24. Gold price per ounce

Descriptive statistics - clusters

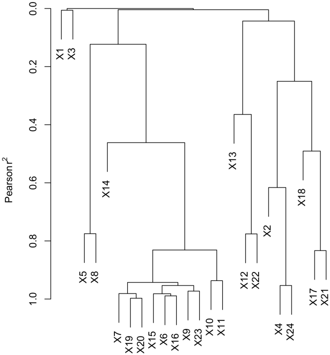


Figure 6: Varclus plot

Proposed algorithm

- Scale all inputs
- Cross-validated LASSO
  10-fold, $\lambda_{min} = 0.000129$, $\lambda_{1se} = 0.00174$
- Fully relaxed LASSO on all 24 Xs, 23 Xs
- Calculate the null bound
  Average of standard error of β coefficients is 0.0825
- Calculate the confidence interval of all $\hat{\beta}$
- Calculate SGPV using `R` package `sgpv`
- Pick the variables with SGPV of 0
  5. Price of the unit at the beginning of the project per m^2
  7. Building services index (BSI) for a preselected base year
  8. Wholesale price index (WPI) of building materials for the base year
  10. Cumulative liquidity
  20. Consumer price index (CPI) in the base year,
  21. CPI of housing, water, fuel & power in the base year

Descriptive statistics - clusters

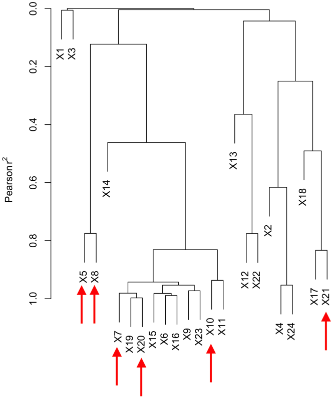


Figure 7: Varclus plot

Regression output

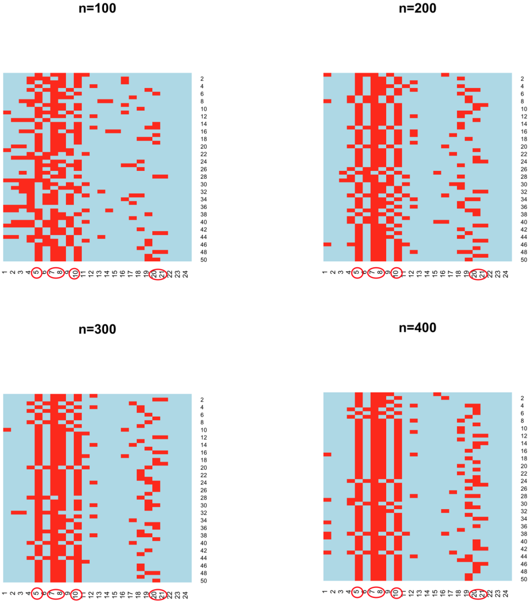
	Estimate	Standard error	t value	p-value
X5	0.999	0.011	89.941	$<$0.0001
X7	-1.591	0.120	-13.270	$<$0.0001
X8	0.668	0.101	6.603	$<$0.0001
X10	0.955	0.067	14.320	$<$0.0001
X20	0.508	0.216	2.347	0.019
X21	-0.559	0.223	-2.513	0.012

Table 1: Regression output, $R^2_{\text{adjusted}} = 0.973$

Validation

- Generate multivariate normal data from the observed data with different sample sizes
- Estimate the residual distribution from the fully relaxed LASSO model with selected variables
- Generate response Y using the mean function of the fully relaxed LASSO model
- Follow the proposed procedure and get a new model
- Repeat this N times to see whether the new model contains those six selected variables

Validation results



Takeaway

- Past: shrink β for model selection
- Now: incorporate confidence interval in model selection
- Fully relaxed LASSO + second generation p-values
  $\Rightarrow$ almost always capture the exact true model
  - When you have decent sample size n as compared to the number of variables p
  - When data are not highly correlated

References

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

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

3. 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.