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

Yi Zuo

Vanderbilt University

March 23, 2020

Contact: yi.zuo@vanderbilt.edu

Outline

1. Background
2. Second-Generation P-Values
3. LASSO and Fully Relaxed LASSO
4. Proposed Procedure
5. Simulation
6. Real-world data

Background

- In high-dimensional settings where inference is desirable, regularization can be used to reduce the feature space.
- Fully relaxed LASSO retains much of the desirable prediction performance from regularization while yielding a model with interpretable coefficients.
- 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

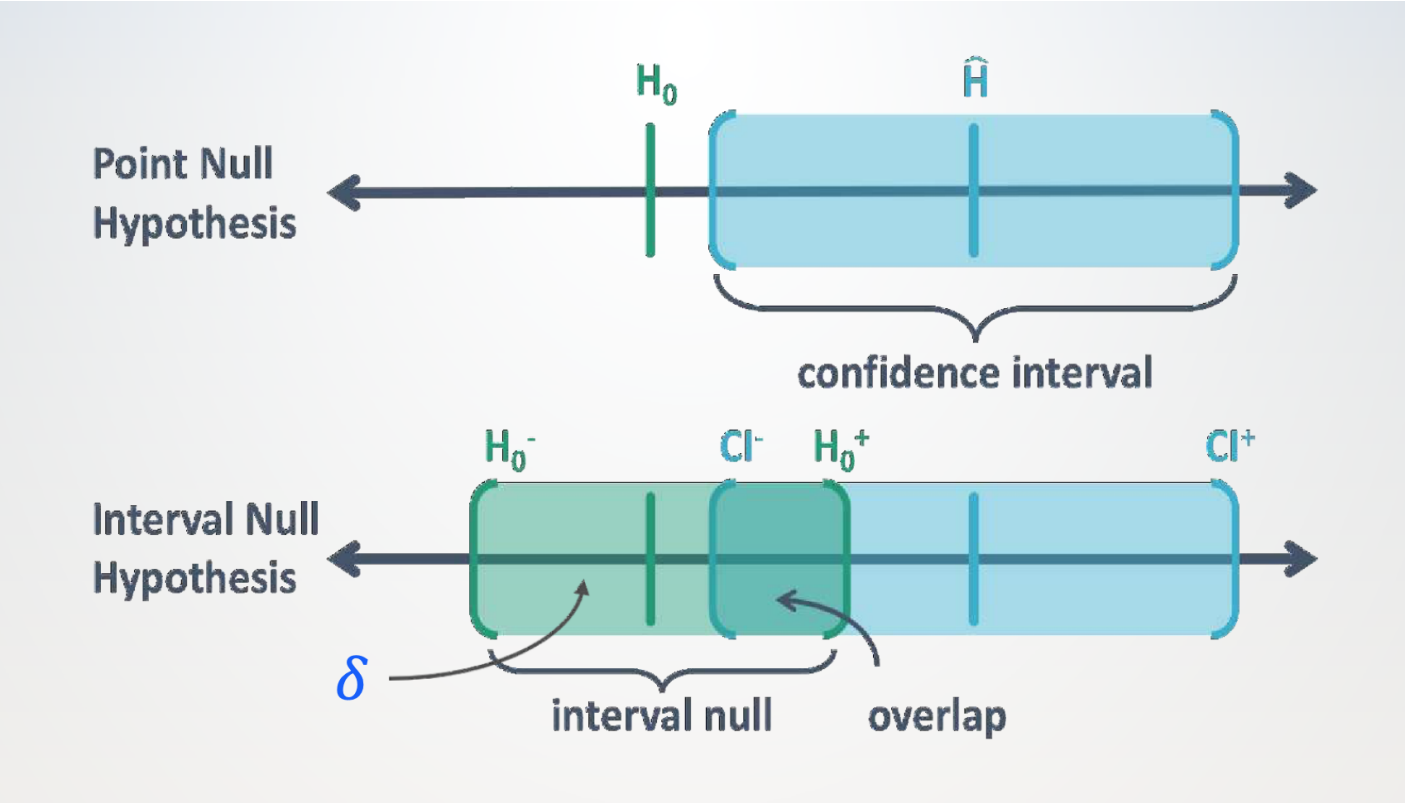
P-values

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

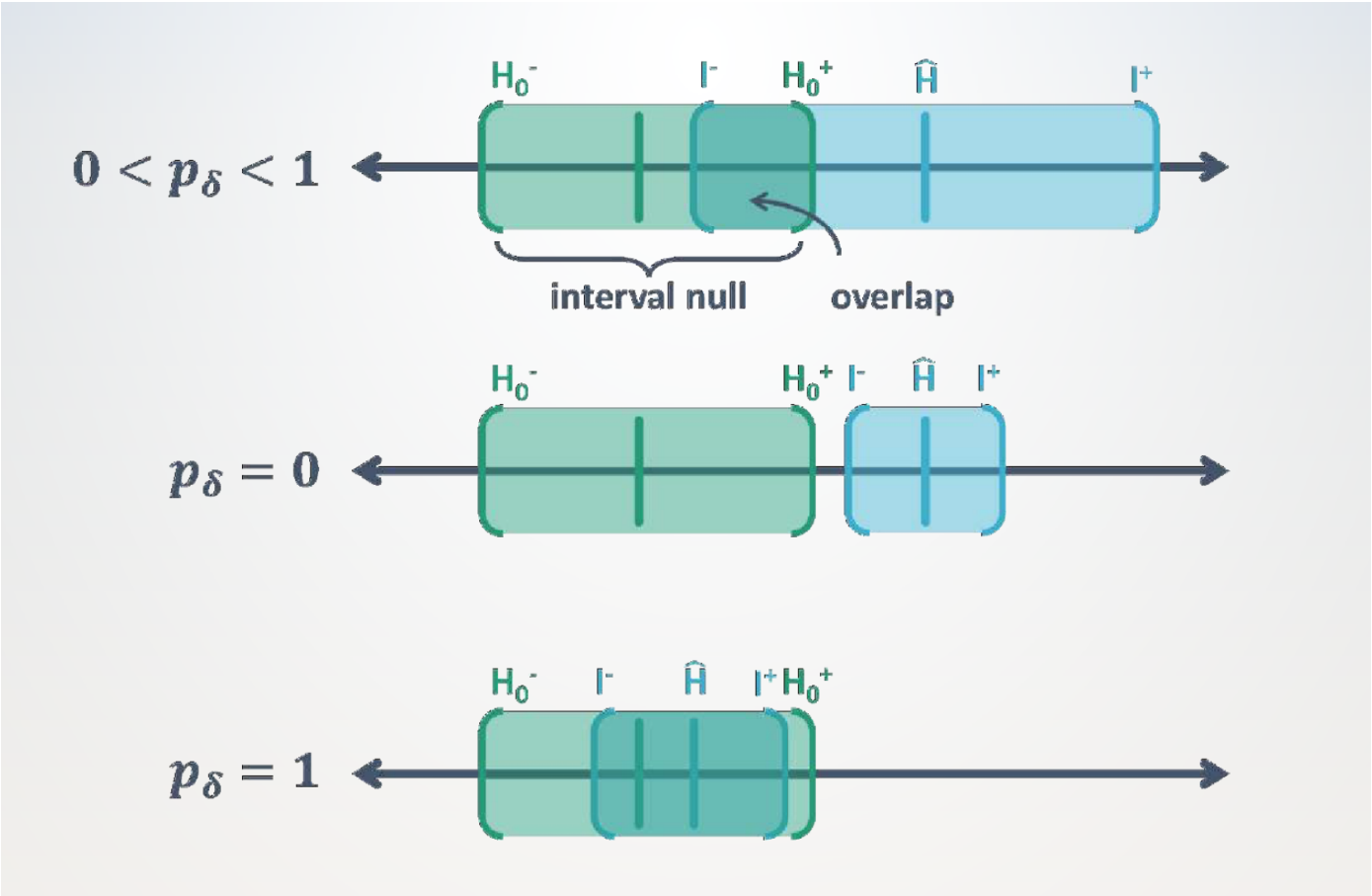
SGPV

- Small value $\Rightarrow$ support for the alternative hypothesis
- Large value $\Rightarrow$ support for the null hypothesis
- $1/2 \Rightarrow$ inconclusive
- Sample size doesn't confound comparison.

Second-Generation P-Values



Second-Generation P-Values



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 $|I| > 2|H_0|$

LASSO and Fully Relaxed LASSO

- The objective function of LASSO:

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

- Implementation in R: cv.glmnet, lambda.min

- Fully relaxed LASSO: OLS on selected variables
  Interpretable coefficients

Proposed Procedure

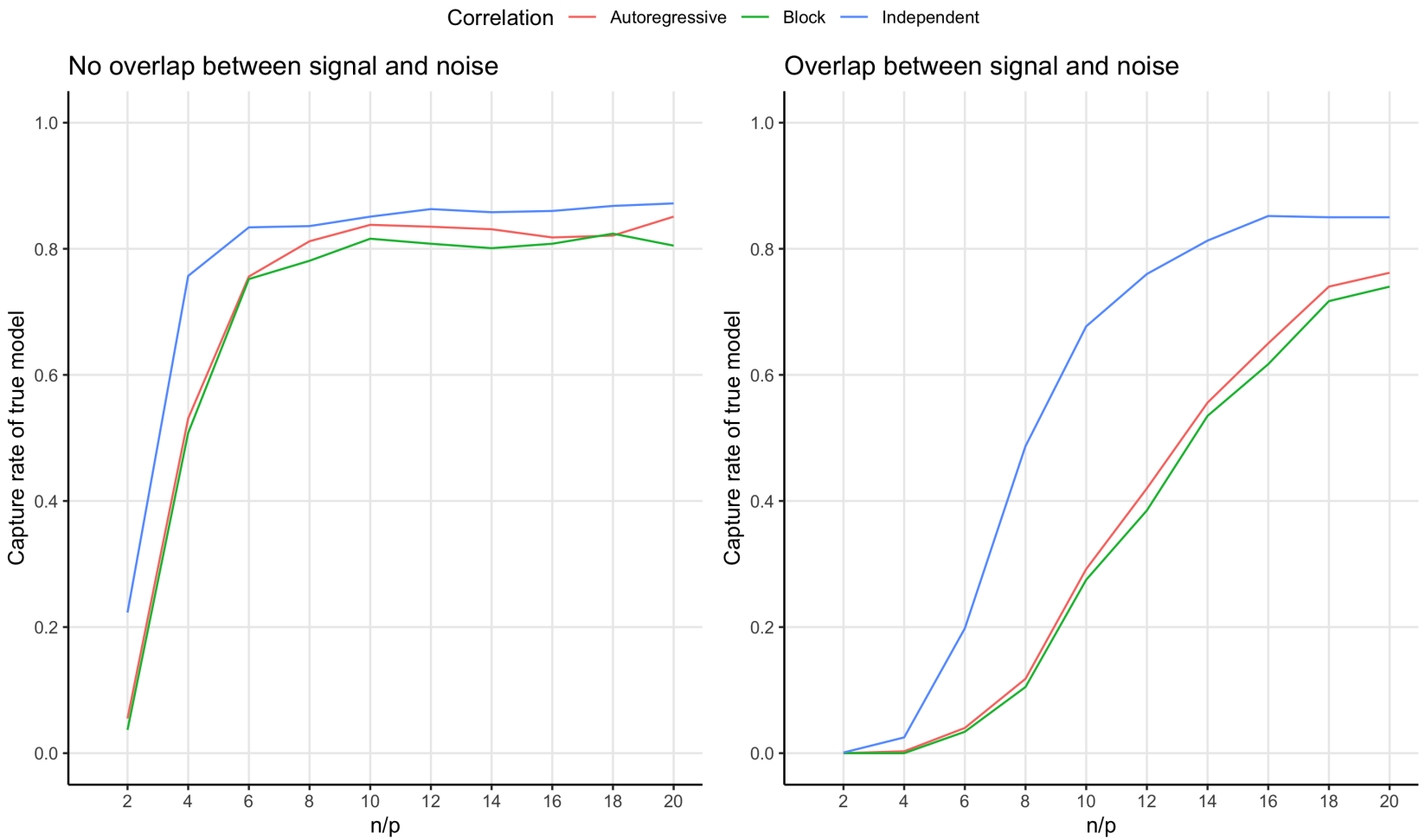
- Use LASSO to first screen the independent variables (cv.glmnet, lambda.min)
- Fit OLS model on variables selected from the previous step
- Use the mean standard error of all β coefficients as the null bound
- Use SGPV to drop variables with trivial effects
- Model on variables that survive SGPV

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, $\sigma_{noise} = 1$
  - No overlap: true $\beta = 0.4, -0.4, 0.4, 0.6, -0.6, -0.6, 1, 1, -1, -1$
  - Overlap: true $\beta = 0.2, -0.2, 0.2, 0.3, -0.3, -0.3, 1, 1, -1, -1$

Second-Generation P-Values



Real-world data

- Data description: Data set includes construction cost, sale prices, project variables, and economic variables corresponding to real estate single-family residential apartments in Tehran, Iran.
- Variables
  - 5 project physical and financial variables
  - 19 economic variables and indices
  - Outcome: Actual sales prices in 10,000 IRR
- 372 observations

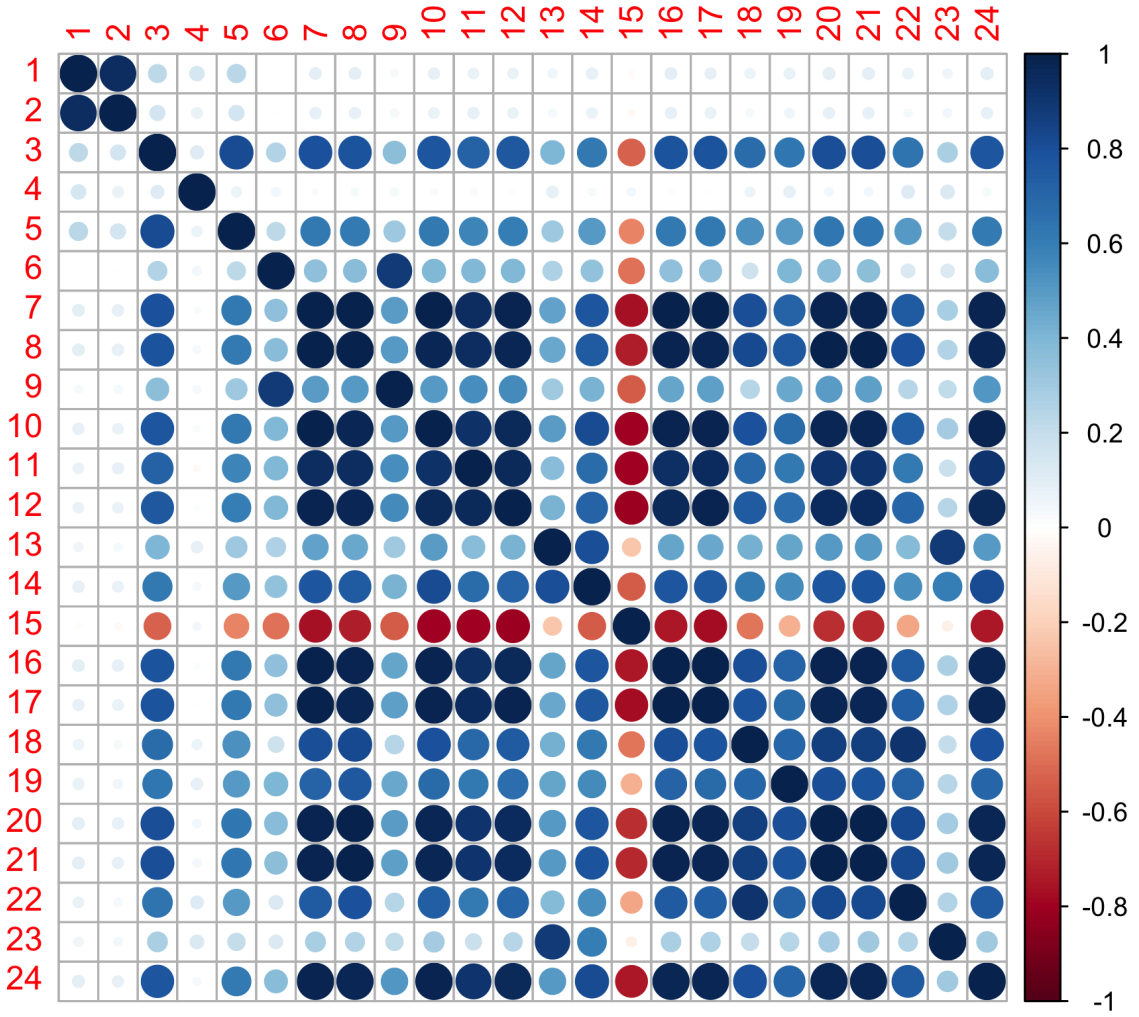
Variables

- Five project physical and financial variables
  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

Variables

- Nineteen 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 - correlation



Proposed procedure

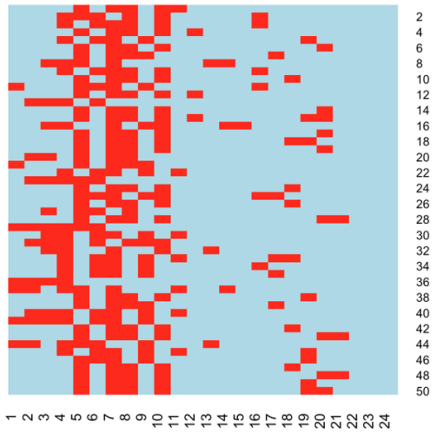
- Scale all inputs
- Cross-validated LASSO
  10-fold, $\lambda_{min} = 0.000297787$
- Fully relaxed LASSO on all 24 Xs
- Calculate the null bound
  Average of standard error of β coefficients is 0.08252418
- Calculate the confidence interval of all $\hat{\beta}$
- Calculate SGPV using R package *sgpv*
- Pick the variables with SGPV of 0
  Price of the unit at the beginning of the project per m^2
  Building services index (BSI) for a preselected base year
  Wholesale price index (WPI) of building materials for the base year
  Cumulative liquidity
  Consumer price index (CPI) in the base year,
  CPI of housing, water, fuel & power in the base year

Validation

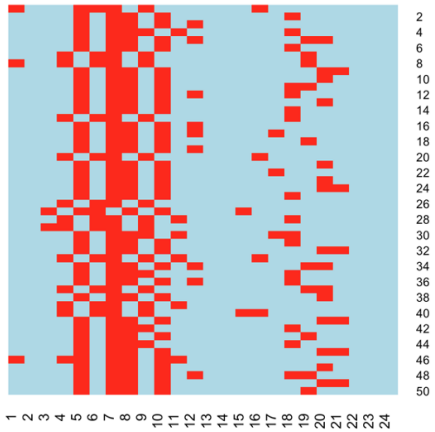
- Generate multivariate normal data from the observed data with different sample sizes
- Estimate the residual distribution from the fully relaxed LASSO model with selected variables
- Generate response Y using the mean function of the fully relaxed LASSO model
- Follow the proposed procedure and get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

Validation results

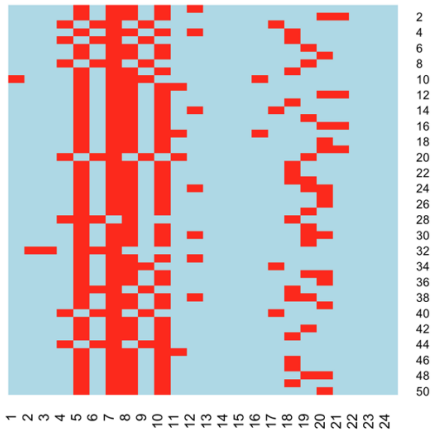
n=100



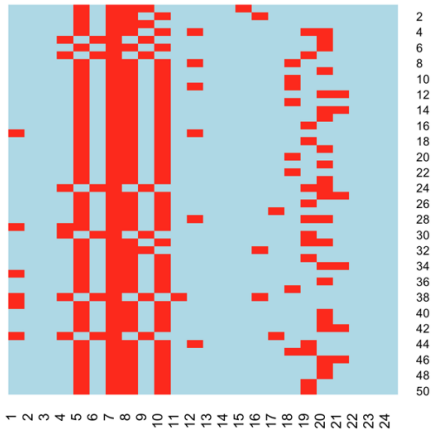
n=200



n=300



n=400



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.