Part III

Second-generation p-values: equivalence tests, statistical properties, and false discovery rates

Jeffrey D. Blume, PhD
School of Data Science
University of Virginia

and

Megan H. Murray, PhD
Department of Biostatistics
Vanderbilt University

Course Layout

- Slides Part I: Introduction and applications
  - Coding Part I
- Lunch (12:00-1:00pm)
- Slides Part II: Statistical Properties, Equivalence tests, false discovery rates, and study planning
  - Coding Part II
- Slides Part III: SGPV Variable Selection
  - Coding Part III
- Questions and Discussion

Outline

- Introduction
  - Tasks
  - Current approaches
  - Second-generation p-values
- Proposed algorithm
  - Steps
  - Implementation
- Numerical studies
  - Simulation studies
  - Real-world example
  - ProSGPV
- Coding Vignettes

Variable Selection

- SGPV variable selection (regression modeling) by Yi Zuo
  - Jeffrey's student
  - Yi's dissertation work January 2022
  - Currently working at Merck
  - R Package: ProSGPV

Variable Selection

- Traditional p-values do not perform well in variable selection
  - Tends to include trivial effects and results are sensitive to small modifications
- Second-generation p-values can improve variable selection using clinical significance
  - Superior support recovery, parameter estimation, and even prediction in certain scenarios, when compared to current standard procedures
  - Can accommodate continuous, binary, count, and time-to-event data
  - An R package made it easy to implement

Introduction

- Data are typically comprised of an outcome and features.
- A common scientific task is to separate the relevant features from the noise features.
- This task is called support recovery, which involves variable selection.

- We also want precise & unbiased parameter estimation, and good prediction

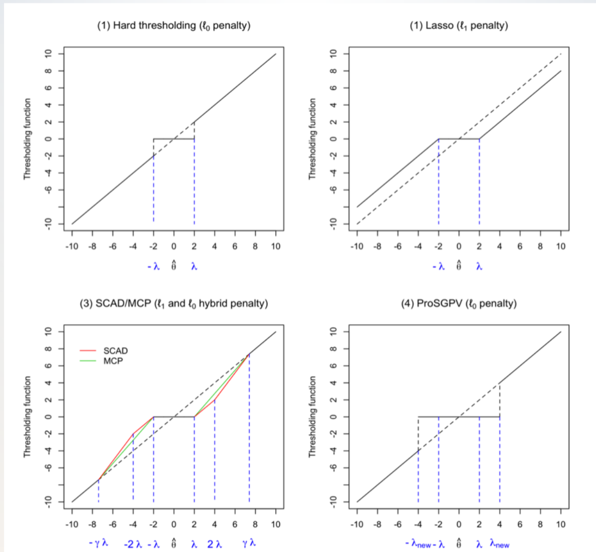
Current Approaches

- P-value based approaches
  - Forward, backward, stepwise selection - unstable
- l_0-based approach
  - Best subset selection (NP-hard, nonconvex optimization problem) very fast implementation: BeSS (Wen et al. 2020)
  - Select variables using AIC and BIC
- l_1-based approaches
  - Lasso, basis pursuit in compressed sensing
  - Adaptive lasso (Zou 2006)

Current Approaches cont.

- l_0 and l_1 hybrid approaches
  - SCAD (Fan and Li 2001) and MC+(Zhang 2010)
- l_1 and l_2 hybrid approach
  - Elastic net (lasso + ridge regression)
- Marginal correlation based approach
  - Iterative Sure Independence Screening (ISIS) (Fan and Lv 2008)

Thresholding functions of different algorithms



ASA SGPV Short Course Blume and Murray, 2022

Drawbacks of current approaches

- All procedures (particularly, adaptive lasso, ISIS, SCAD, and MC+) are great, in theory.
- The actual variable selection results depend on tuning parameters that are hard to specify in practice.
- A prediction-optimal lasso selects all signals + noise variables. It is a good place to start with.

ASA SGPV Short Course Blume and Murray, 2022

Second-generation p-values

- Second-generation p-values (SGPVs) were proposed in the high dimensional multiple testing context (Blume, D'Agostino McGowan, et al. 2018; Blume, Greevy, et al. 2019).
- SGPVs replace the point null hypothesis with a pre-specified interval null, which can be used to select effects that are clinically meaningful.

Second-generation p-values

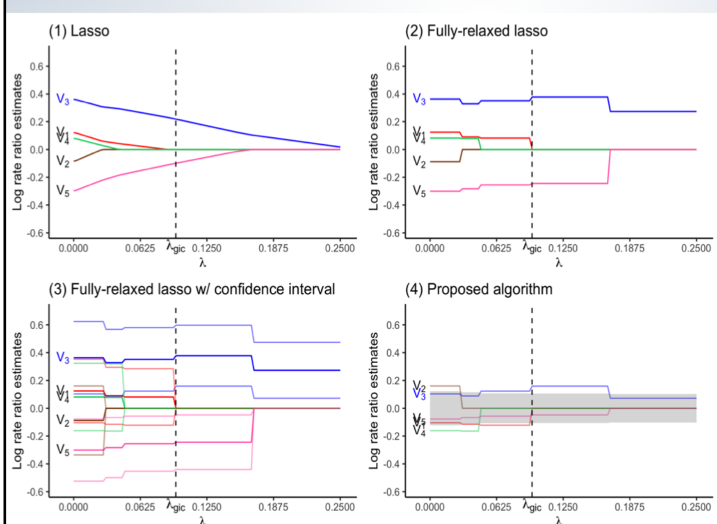
- SGPV enjoys several appealing properties:
  - SGPV close to 0 indicates support for the alternative hypothesis; close to 1 indicates support for the null hypothesis; and near 1/2 is inconclusive.
  - It doesn't need a threshold in the interpretation.
  - It values clinically meaningful effects over traditionally statistically significant effects, which could be valuable to support recovery.

Penalized regression with Second-Generation P-Values (ProSGPV)

Algorithm 1 The ProSGPV algorithm

1: **procedure** ProSGPV($\boldsymbol{X}$, $\boldsymbol{Y}$)
2: **Stage one**: Find a candidate set
3: Standardize explanatory variables
4: Fit a lasso and evaluate it at λ_{gic}
5: Fit OLS/GLM/Cox models on the lasso active set
6: **Stage two**: SGPV screening
7: Extract the confidence intervals of all variables from the previous step
8: Calculate the mean coefficient standard error $\overline{SE}$
9: Calculate the SGPV for each variable where $I_j = \hat{\beta}_j \pm 1.96 \times SE_j$ and $H_0 = [-\overline{SE}, \overline{SE}]$
10: Keep variables with SGPV of zero
11: Refit the OLS/GLM/Cox with selected variables
12: **end procedure**

ProSGPV Illustration



- Note that the null region in figure (4) is in grey. The value is 2 × SE and it converges to 0 at $\sqrt{n}$ rate.

Simulation

- Generate simulation data
- Run ProSGPV, lasso, BeSS, and ISIS on the data and record support recovery rate, parameter estimation mean absolute error (MAE), prediction root mean square error (RMSE) and area under the curve in a separate test set, and running time.
- Compare the performance of four algorithms over 1000 repetitions

Simulation Parameters

Table 1: Summary of parameters in simulation studies.

	Linear regression	Logistic regression	Poisson regression	Cox regression
n	100	32:320	40	80:800
p	100:1000	16	40:400	40
s	10	6	4	20
β_l	1	0.4	0.2	0.3
β_u	2	1.2	0.5	1
ρ	0.3	0.6	0.3	0.3
σ	2	2	2	2
ν	2			
Intercept t		0	2	0
Scale				2
Shape				1
Rate of censoring				0.2

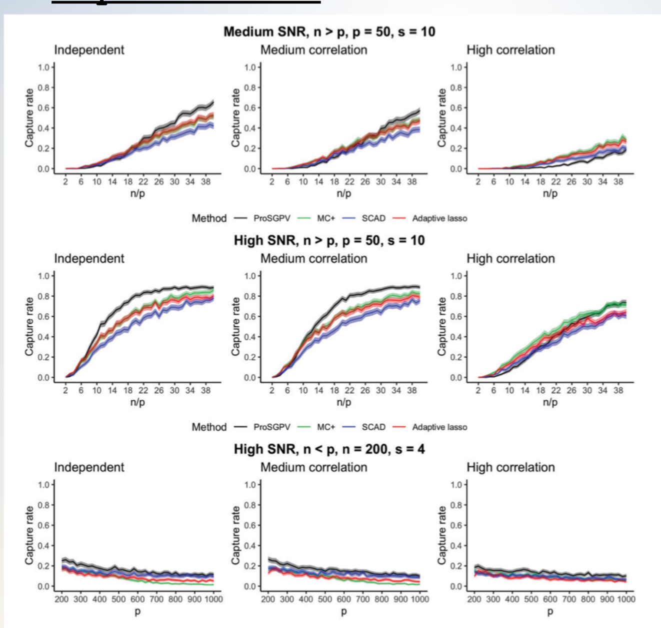
Metrics

- Support recovery is defined as capturing the exact true support, not containing it.
- An estimate of MAE is the following where $\beta_{0,j}$ is the jth true coefficient

$$\frac{1}{p}\sum_{j=1}^{p} \| \widehat{\beta}_j - \beta_{0,j} \|$$

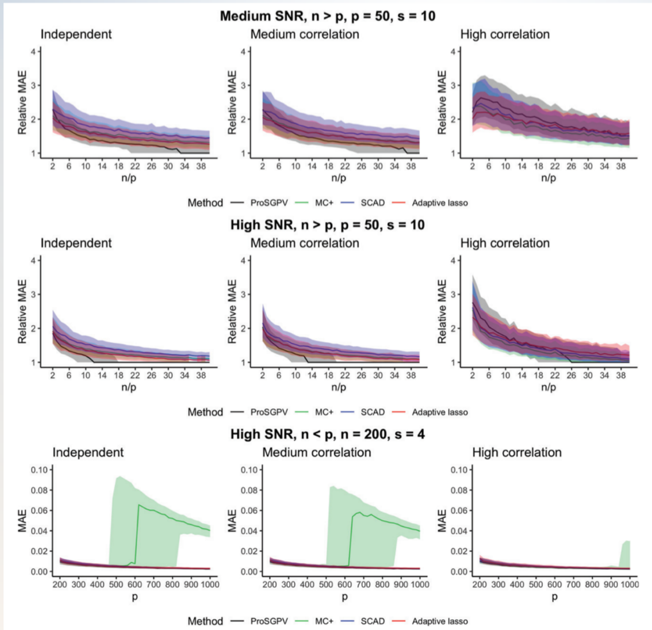
- MSE is the usual mean squared error

Capture Rate



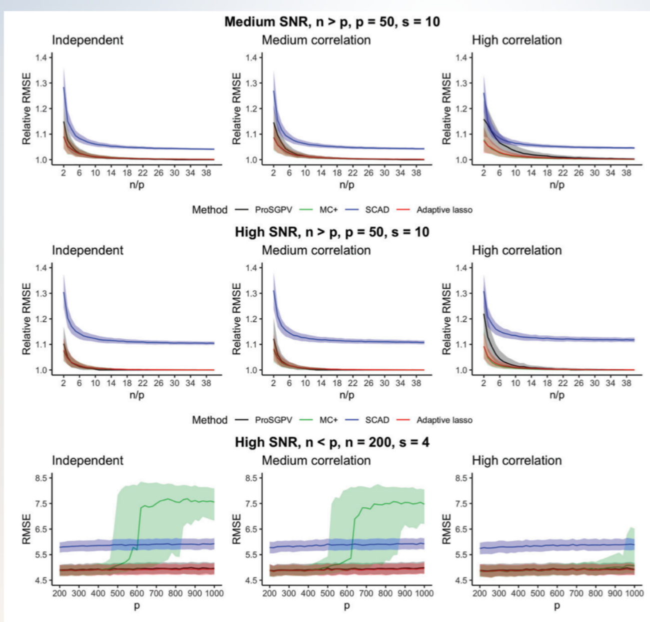
- Capture rate of the exact true model under combinations of autocorrelation level, signal-noise-ratios, and (n, p,s). In each panel, one algorithm has a colored solid line representing the average capture rate surrounded by the shaded 95% Wald interval over 1000 simulations.

Parameter Estimation Error



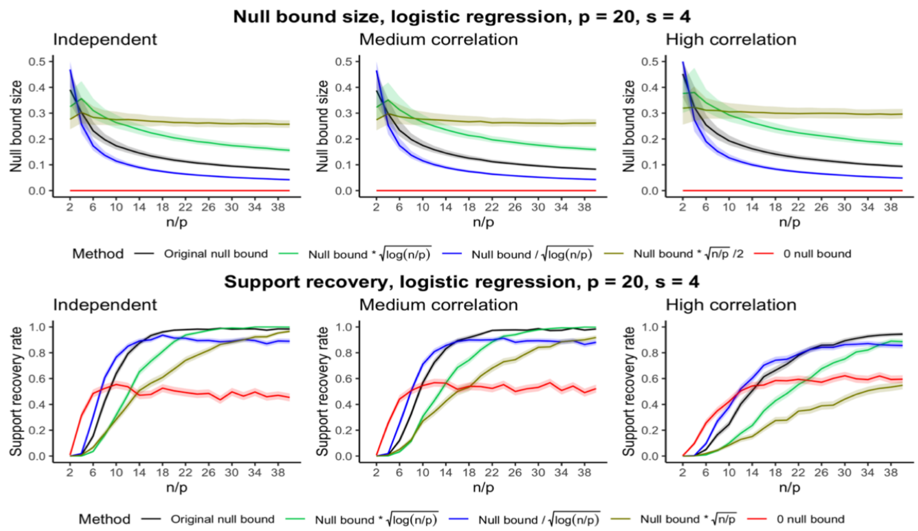
- Parameter estimation error of all algorithms under combinations of autocorrelation level, SNR, and (n, p,s). In each panel, one algorithm has a colored solid line representing the median (relative) MAEs surrounded by the shaded first and third quartiles over 1000 simulations.

Prediction Accuracy



- Comparison of prediction accuracy of all algorithms under combinations of autocorrelation level, SNR, and (n, p,s). Median (relative) root mean square errors are surrounded by their first and third quartiles over 1000 simulations.

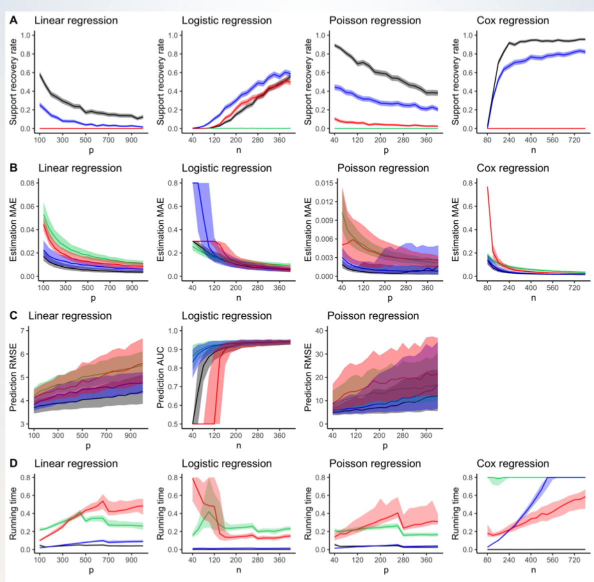
How null bound affects support recovery



ASA SGPV Short Course

Blume and Murray, 2022

Simulation Results



ASA SGPV Short Course

Blume and Murray, 2022

GVIF: improve performance with highly correlated data

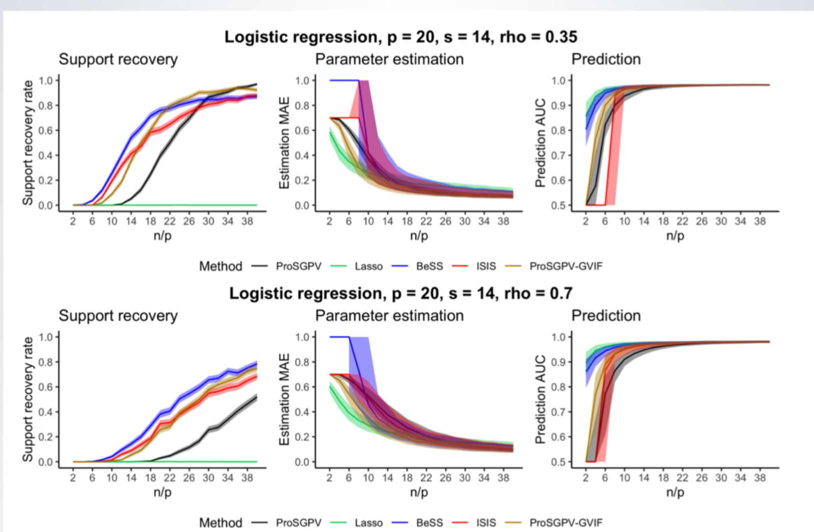
- Support recovery performance suffers when data are highly correlated. The original bound leads to too sparse a model.
- This can be fixed by replacing the constant null bound with an adjusted null bound based on the generalized variance inflation factor (GVIF) (Fox and Monette 1992).

$$GVIF = \frac{\det R_{11} \times \det R_{22}}{\det R} \tag{2}$$

- Each coefficient standard error is inversely weighted by GVIF and then summed to derive an adjusted null bound.

ASA SGPV Short Course | Blume and Murray, 2022

GVIF: improve performance with highly correlated data



ASA SGPV Short Course | Blume and Murray, 2022

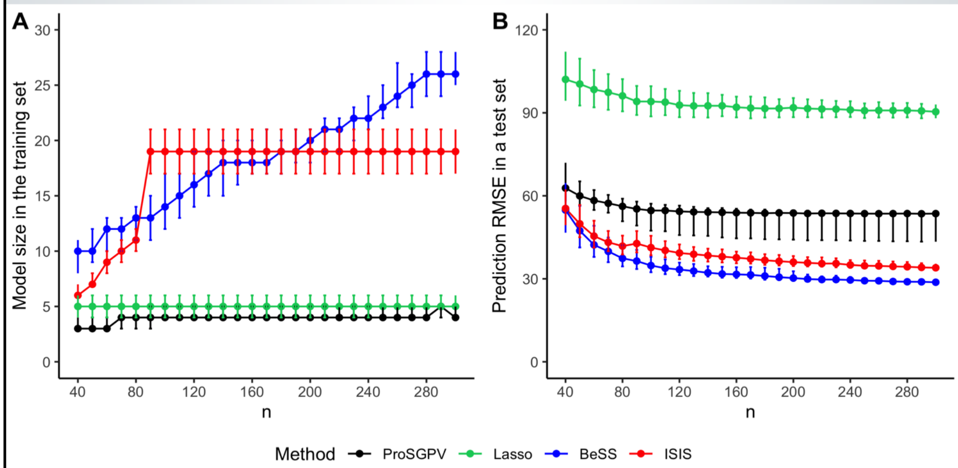
Real World Example

- The close price of Dow Jones Industrial Average (DJIA) was documented from Jan 1, 2010 to November 15, 2017
- Eight groups of primitive, technical indicators, big U.S. companies, commodities, exchange rate of currencies, future contracts and worlds stock indices, and other sources of information (Hoseinzade and Haratizadeh 2019) were collected to predict the DJIA close price.

Real World Example

- There are 1114 observations and 82 predictors. We randomly sampled 614 observations as a fixed test set. We allowed the training set size n to vary from 40 to 300. At each n, we recorded the distribution of the training model size for each algorithm as well as the distribution of the prediction RMSE over 1000 repetitions.
- Variables frequently selected by ProSGPV include 5-, 10-, and 15-day rate of change, and 10-day exponential moving average of (DJIA). Technical indicators seem more predictive than other world indices, commodity, exchange rate, futures, etc.

Real World Example



Takeaways

- Traditional p-values do not perform well in variable selection
  - Tends to include trivial effects and results are sensitive to small modifications
- Second-generation p-values can improve variable selection using clinical significance
  - Superior support recovery, parameter estimation, and even prediction in certain scenarios, when compared to current standard procedures
  - Can accommodate continuous, binary, count, and time-to-event data
  - An R package made it easy to implement

Links and Vignettes

- ProSGPV GitHub
  - https://github.com/zuoyi93/ProSGPV
- Paper
  - f1000research.com/articles/11-58
- Linear ProSGPV
  - https://cran.r-project.org/web/packages/ProSGPV/vignettes/linear-vignette.html
- GLM and Cox ProSGPV
  - https://cran.r-project.org/web/packages/ProSGPV/vignettes/glm-cox-vignette.html

Time for Code Part 3!

10 Minute Break!

Review of Topics Learned

- Traditional p-values are flawed
- Second-generation *p*-value framework and definition
- Outrageous claim: The SGPV achieves the inferential properties that many scientists hope, or believe, are attributes of the classic *p*-value.
- Statistical Properties of SGPVs
- Comparison to Equivalence Tests: Two One-Sided Tests (TOST)
- False Discovery Rates
- Study Planning
- SGPV Variable Selection
- Coding Examples

Acknowledgements

- Collaborators
  - William D. Dupont
  - Robert A. Greevy
  - Lucy D'Agostino McGowan
  - Jonathan Chipman
  - Valerie Welty
  - Lisa Lin
  - Jeffrey R. Smith
  - Yi Zuo
  - Thomas G. Stewart
  - Vanderbilt SEDS Lab

- Website
  - www.statisticalevidence.com

ASA SGPV Short Course   Blume and Murray, 2022

Questions?

ASA SGPV Short Course   Blume and Murray, 2022