A Comparative Study of Variable Screening Methods: Univariate versus Multivariate Screening

Cong Liu, Tao Shi and Yoonkyung Lee

Department of Statistics, The Ohio State University

THE OHIO STATE UNIVERSITY

Model and Goal

Model:

Consider a linear regression model:

$Y = \beta_1 X_1 + ... + \beta_p X_p + \varepsilon$

Without loss of generality, we assume only the first q (q<p) covariates $(X_1,...,X_q)$ are "relevant" with non-zero coefficients $(\beta_1,...,\beta_q)$

Goal:

Compare two types of variable screening methods that try to identify those q relevant covariates from all p candidates.

- q is expected to be very small and n is small, too.
- p can be very large, "curse of dimensionality".

Why screening?

- Better prediction accuracy
- Better model interpretation
- Necessary when p>n

Screening Methods: Univariate vs. Multivariate

Screening methods:

- Univariate screening: (Correlation screening)
  - Screening variables only based on the absolute value of marginal Pearson's correlation. Simply Pick X_i's with large $|corr(Y, X_i)|$.
  - Computationally efficient, but may overlook some important predictors due to multicollinearity.
- Multivariate screening:
  - Subset selection: Best subset selection; forward and backward stepwise selection.
  - Penalized least square methods: LASSO, LARS, Dantzig selector, Elastic net, etc.
- Our study: Correlation screening vs. LASSO
  - LASSO is a well-studied method which automatically select variables by shrinking some coefficients to 0.

Simulation Setup

Simulation study:

Correlation screening and LASSO are compared via a simulation study. We consider different data structures by controlling four factors as described below.

Factors:

1. Total number of covariates p, and relevant covariates q

   Evolving from the classical small p, small q setting to large p, small q (p>n) setting.
2. The sign of non-zero regression coefficients $(\beta_1,...,\beta_q)$
3. The off-diagonal elements r in the covariance matrix $\sum_p$

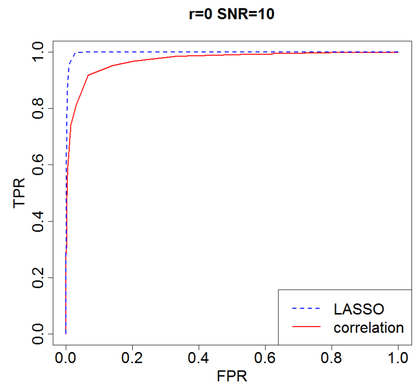
   Assume $(X_1,...,X_p)$ follow a multinormal distribution $N(0,\sum_p)$, and $\sum_p$ has common diagonal elements 1 and off-diagonal elements r.
4. Signal-to-noise ratio (SNR)

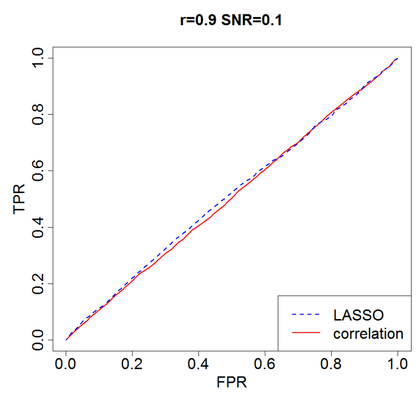
   SNR is defined by $Var(\beta_1 X_1 + ... + \beta_p X_p) / Var(\varepsilon)$

Model Assessment: ROC curve

Model assessment: ROC curve

Rather than picking up the "best" model, we compare the two methods on an universal scale via Receiving Operating Characteristic (ROC) curve.





- X-axis: False positive rate (FPR)

$$\text{FPR}=\frac{\text{\# of selected irrelavent variables}}{\text{total \# of irrelavent variables}}$$

- Y-axis: True positive rate (TPR)

$$\text{TPR}=\frac{\text{\# of selected relavent variables}}{\text{total \# of relavent variables}}$$

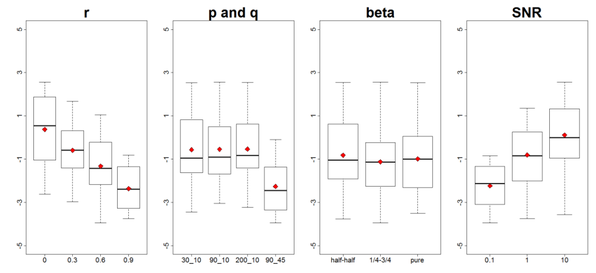
- By varying the threshold or the penalty parameter in correlation screening and LASSO respectively, the size of selected model increases from 0 (Null model, FPR=TPR=0) to p (Full model, FPR=TPR=1)
- Therefore, the ROC curve starts at (0,0) and ends at (1,1), and the Area Under the Curve (AUC) is desired to be large for better overall selection.

ANOVA Study

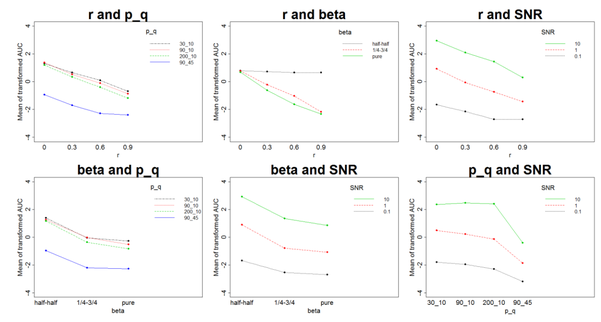
ANOVA study

We perform ANOVA studies to investigate the importance of the four factors in terms of affecting AUC for both methods and the difference in AUC between them. This may provide a guideline for choosing screening method under different data structures.

The box plots showing main effects (for correlation screening)



The interaction plots (for LASSO)



- The response variable is AUC (with a proper transformation), and the model includes the main effects and all two-way interaction.
- Box plots and interaction plots are used to visualize the main effects and interactions.
- The importance order of factors for correlation screening and LASSO are quite different, which is related to the "Irrepresentable Condition". (Zhao and Yu (2006))

Discussion and References

Discussion

1. In most cases in our study, LASSO performs better than correlation screening when SNR is moderate (1) or high (10). When SNR is low (0.1), both methods perform poorly.
2. We also perform ROC analysis on the correlation-LASSO "hybrid" method and found no significant advantage.
3. Our simulation setting doesn't cover the situation where p>>n and it is part of the future work.

References:

1. Tibshirani,R. (1996) Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* **58** 267-288

2. Zhao.P and Yu, B. (2006) On Model selection consistency of LASSO. *J. Machine Learning Res.* **7**, 2541-2567