Feature Selection Bias in Assessing the Predictivity of SNPs for Alzheimer's Disease

Longhai Li

Department of Mathematics and Statistics
University of Saskatchewan
Saskatoon, SK, CANADA

3 June 2019
Presented at the University of Manitoba

Acknowledgements

- Thanks to my co-authors: Mei Dong and Lloyd Balbuena.
- Thanks to the fundings from NSERC and CFI of Canada.
- Thanks to Pingzhao for inviting me and hosting my visit.

Outline

1. Introduction

2. Methodologies
   - GWAS and False Discovery Rate
   - Regularized Logistic Regression Methods
   - Internal CV and External CV
   - Predictive Metrics

3. Data Analysis Results
   - Results on a Real Dataset
   - Results on two Synthetic Datasets

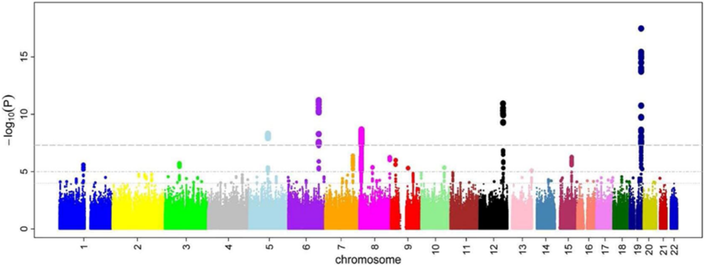
4. Conclusions and Future Work

Section 1

Introduction

GWAS (Feature Selection)

- Genome-wide association study (GWAS) has been widely applied for identifying the genetic variations (Single-nucleotide polymorphisms, SNP) that affect a phenotype. Typically, the sample size *n* is small but the number of features (SNPs), *p*, is huge.
- GWAS measures the association between a phenotype and a single SNP using traditional statistical methods, e.g. logistic regression, Fisher's exact test. Results are expressed with statistical significance measures, such as p-values/q-values.



Predictive Analysis

- Features (eg. SNPs) that are statistically significant are not necessarily good predictors of a phenotype.
- We are interested in applying statistical learning methods to measure the predictivity of **selected features** (eg. selected SNPs) for a phenotype.
- There are two methods to implement feature selection with cross-validation (CV):
  - External cross-validation (ECV): features are re-selected in each fold of CV *based on only training samples*.
  - Internal cross-validation (ICV): features are pre-selected before applying CV based on *all samples*.

Feature Selection Bias of ICV (objective of this talk)

- In ECV, the test samples are external to feature selection, which is a component of training process. ECV is a correct method. However, it is computationally intensive because the GWAS needs to be repeated in each fold of CV.
- In ICV, the test samples are not external (internal) to feature selection process. This is a wrong method. There is a feature selection bias in the predictivity given by ICV. However, people often use it unconsciously for convenience and other reasons.
- In this talk, we will use a real SNP data related to late onset Alzheimer's disease (LOAD) and two synthetic datasets to demonstrate the severity of the feature selection bias of ICV.

Section 2

Methodologies

Subsection 1

GWAS and False Discovery Rate

Notations

- Response (Phenotype): y_i is the binary indicator for the phenotype of individual i.

$$y_i = \begin{cases} 0 \; for \; controls \\ 1 \; for \; cases \end{cases}$$

- Feature (SNPs): For autosome, $x_i^{(j)}$ is a row vector with two dummy variables to represent SNP$_j$ for individual i.

$$x_i^{(j)} = \begin{cases} (0,0), & if \; genotype = AA \\ (0,1), & if \; genotype = Aa \\ (1,0), & if \; genotype = aa \end{cases}$$

For SNPs on chromosome X, $x_i^{(j)}$ is a row vector with four dummy variables because there are five categories considering sex.

GWAS with Logistic Regression with a Single SNP

The logistic regression of SNP$_j$ can be written as:

$$\log\left(\frac{Pr(y_i=1)}{1-Pr(y_i=1)}\right)=\beta_0+x_i^{(j)}\beta_j, \tag{1}$$

where β_j denotes the column vector of regression coefficients associated with SNP$_j$ and β_0 denotes the intercept. We can estimate (β_0, β_j) by minimizing the negative log likelihood:

$$\ell(\beta_0,\beta_j)=-\sum_{i=1}^{n}\left(y_i(\beta_0+x_i^{(j)}\beta_j)-\log(1+e^{(\beta_0+x_i^{(j)}\beta_j)}\right). \tag{2}$$

Likelihood Ratio Test (LRT)

We use likelihood ratio test (LRT) to compare the goodness of fit between a null model and an alternative model:

$$\begin{aligned} H_0^{(j)} &: y_i \sim f(y_i | x_i^{(COV)}), \text{in words, SNP}_j \text{ is unrelated to } y \\ H_1^{(j)} &: y_i \sim f(y_i | x_i^{(COV)}, x_i^{(j)}), \text{in words, SNP}_j \text{ is related to } y \end{aligned} \tag{3}$$

where f represents logistic regression model. Denote the maximized likelihoods of the null model (H_0) and the alternative model (H_1) by $L_0^{(j)}$ and $L_1^{(j)}$, respectively. Then the log likelihood ratio is defined as

$$\Lambda^{(j)} = -2(\log L_0^{(j)} - \log L_1^{(j)}). \tag{4}$$

Then $\Lambda^{(j)} \sim \chi^2_{C-1}$, where C is the number of variants of SNP$_j$.

Multiple Comparison Problem

- When performing m independent tests, family wise error rate (FWER) is the probability that at least one unrelated SNP is selected by thresholding p-values with α_0:

$$\text{FWER} = 1 - (1 - \alpha_0)^m \approx \min(1, m\alpha_0), \tag{5}$$

- Specifying α_0 as α/m can make sure that FWER is less than or equal to a given α (say 0.05).
- However, this criteria is often too stringent in practice.
- Hypothesis testing results

	Accept null	**Reject null**	**Total**
Null true	U	V	m_0
Alternative true	T	S	m_1
Total	W	R	m

False Discovery Rate

- Benjamini and Hochberg (1995) proposed the false discovery rate (FDR) as a measure of test error in testing multiple hypotheses. Story (2003) proposed to define FDR as follows:

$$\mathrm{FDR} = E\left[\frac{V}{R}\middle| R > 0\right] Pr(R > 0) \approx \frac{E(V)}{E(R)}. \tag{6}$$

- Given m ordered p-values $p_1, p_2, ..., p_m$, the false discovery rate with p_j as the cutoff in SNP selection can be estimated by:

$$\widehat{\mathrm{FDR}}(p_j) = Pr(H_0^{(j)} \text{ is true} | p \leq p_j) = \min_{t > p_j} \frac{\pi_0 t}{\hat{F}(t)}. \tag{7}$$

where π_0 is the proportion of null hypotheses, F_0 is the and $\hat{F}$ is an estimated CDF of all p-values.

Limitations of GWAS

- The p-value/q-value from GWAS only measures statistical significance but not practical significance. Strong statistical significance do not necessarily imply strong predictivity. For example, many SNPs selected by GWAS are not good predictors [1, 2].
- Joint effects of features on a phenotype cannot be measured in univariate GWAS. Many phenotypes are believed to be polygenic.
- The validity of q-values relies on the correctness of assumed models, which may not hold for real datasets.

Subsection 2

Regularized Logistic Regression Methods

LASSO Logistic Regression

Given a selected subset of SNPs with size k, denoted by $\{s_1, ..., s_k\}$, LASSO minimizes a negative likelihood function of β penalized by L_1 norm to find a sparse estimate of coefficients:

$$\hat{\beta}_{LASSO}(\lambda_1) = \underset{\beta_0, \beta}{\operatorname{argmin}} \; \ell(\beta_0, \beta) + \lambda_1 ||\beta||_1, \tag{8}$$

where,

$$\boldsymbol{x_i} = (x_i^{(s_1)}, ..., x_i^{(s_k)})^T, \text{a row vector representing selected SNPs,}$$

$$\beta = (\beta_{s_1}^T, ..., \beta_{s_k}^T)^T, \text{a column vector of coefficients,}$$

$$\ell(\beta_0, \beta) = -\sum_{i=1}^{n} \left(y_i(\beta_0 + \boldsymbol{x}_i^T \beta) - \log(1 + e^{(\beta_0 + \boldsymbol{x}_i^T \beta)}) \right),$$

$$||\beta||_1 = \sum_{j=1}^{p} |\beta_j|,$$

and λ_1 is a tuning parameter.

Elastic-net Logistic Regression

- Elastic-net combines L_1 and L_2 penalties to enforce the sparse solutions:

$$\hat{\beta}_{EN}(\alpha, \lambda_2) = \underset{\beta_0, \beta}{\operatorname{argmin}} \; \ell(\beta_0, \beta) + \lambda_2 \left[(1-\alpha)||\beta||_1 + \frac{1}{2}\alpha||\beta||_2 \right], \quad (9)$$

  where $||\beta||_2 = \sum_{j=1}^{p} \beta_j^2$, $\alpha \in (0,1)$ controls the weight between L_1 and L_2, λ_2 is a tuning parameter.
- LASSO tends to select one feature and ignores the rest when there are several features correlated. While elastic-net combines LASSO and ridge regression, it can select a group of correlated features.

Bayesian Hyper-LASSO Logistic Regression

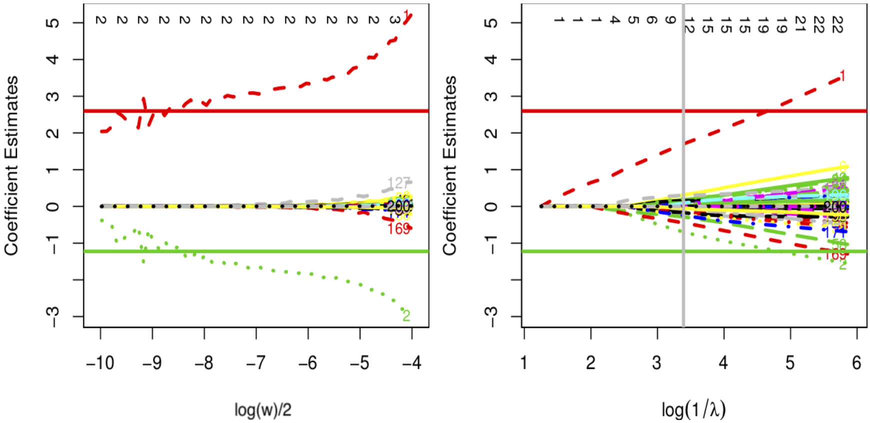
Consider a heavy-tailed distribution such as t-distribution as a prior for $\boldsymbol{\beta}$. The hierarchical Bayesian regression model can be described as follows:

$$
\begin{aligned}
Pr(y_i = 1 | \boldsymbol{x}_i, \beta_0, \beta) &= \frac{e^{\beta_0 + \boldsymbol{x}_i^T \beta}}{1 + e^{\beta_0 + \boldsymbol{x}_i^T \beta}}, \\
\beta_j | \sigma_j^2 &\sim N(0, \sigma_j^2), \text{ for } j = 0, 1, ..., p, \\
\sigma_j^2 &\sim IG(a/2, wa/2), \text{ for } j = 1, 2, ..., p.
\end{aligned}
$$

With σ_j^2 marginalized with respect to Inverse-Gamma prior, β_j a t prior with a degrees of freedom and scale $\sqrt{w}$.
Hyper-LASSO can shrink small coefficients to 0 more aggressively while keeping the large coefficients.

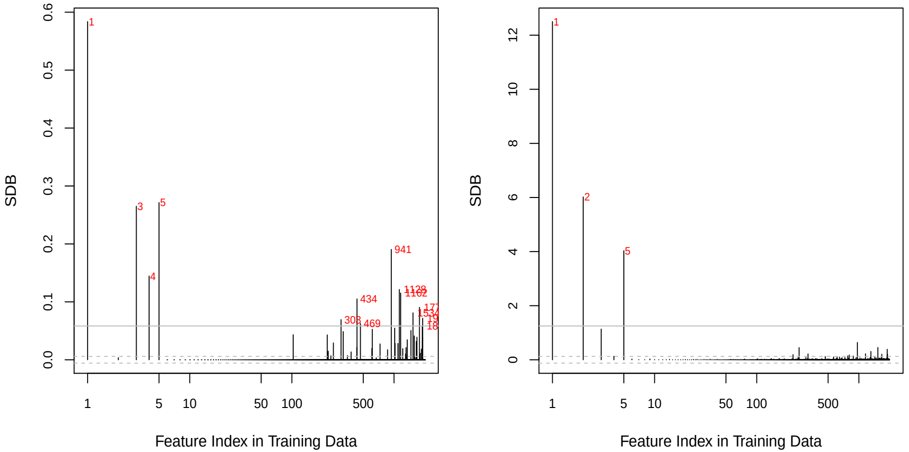
Comparison of LASSO and hyper-LASSO: I

Solution paths with LASSO and hyper-LASSO penalties:



Comparison of LASSO and hyper-LASSO: II

Coefficient estimates at a given choice of scale:



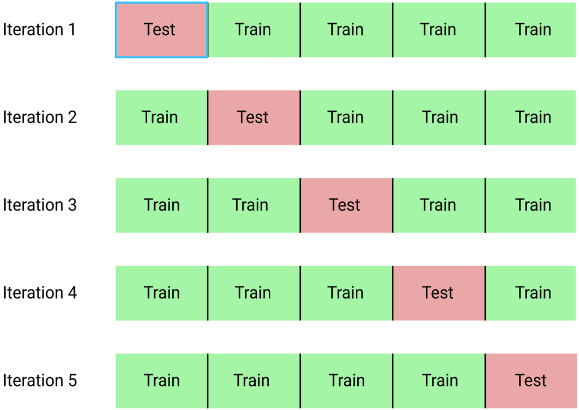
(a) LASSO

(b) Cauchy (t with one degree freedom)

Subsection 3

Internal CV and External CV

Cross-validation without Feature Selection



Iteration 1	Test	Train	Train	Train	Train
Iteration 2	Train	Test	Train	Train	Train
Iteration 3	Train	Train	Test	Train	Train
Iteration 4	Train	Train	Train	Test	Train
Iteration 5	Train	Train	Train	Train	Test

Cross-validation with Feature Selection

- ICV
    1. Pre-select a subset of features, say S with a GWAS method,based on all the samples
    2. Apply CV using only the data of the pre-selected subset S
- ECV
    1. Split the samples into K parts
    2. For $k = 1, \ldots, K$
        1. Select a subset of features, say S_k with a GWAS method, based on only the training samples in fold k
        2. Train a model using the data of S_k, based on only the training samples in fold k
        3. Predict the phenotype of test samples in fold k

A Picture for Explaining ICV and ECV

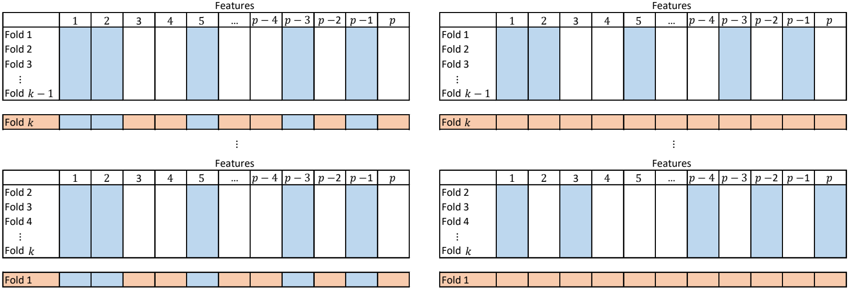


Figure 1: A picture of internal CV and external CV. Blue shadow represents the selected features. Orange shadow represents the test set.

(a) Internal Cross-validation

(b) External Cross-validation

Subsection 4

Predictive Metrics

Criteria for Measuring Predictive Performance

We can measure the goodness of predictive probabilities $\hat{P}_i(y_i|x_i)$ with the true observations of y_i using several criteria:

- ER(Error Rate): Estimate y_i by $\hat{y}_i = \operatorname{argmax} \hat{P}_i(y_i|\boldsymbol{x}_i)$.

$$ER = \frac{1}{n}\sum_{i=1}^{n} I(\hat{y}_i \neq y_i). \tag{10}$$

- AMLP(Average Minus Log-probability on observed y_i) (information criterion):

$$AMLP = -\frac{1}{n}\sum_{i=1}^{n} \log(\hat{P}_i(y_i)). \tag{11}$$

- AUC (Area under the curve)

Section 3

Data Analysis Results

Subsection 1

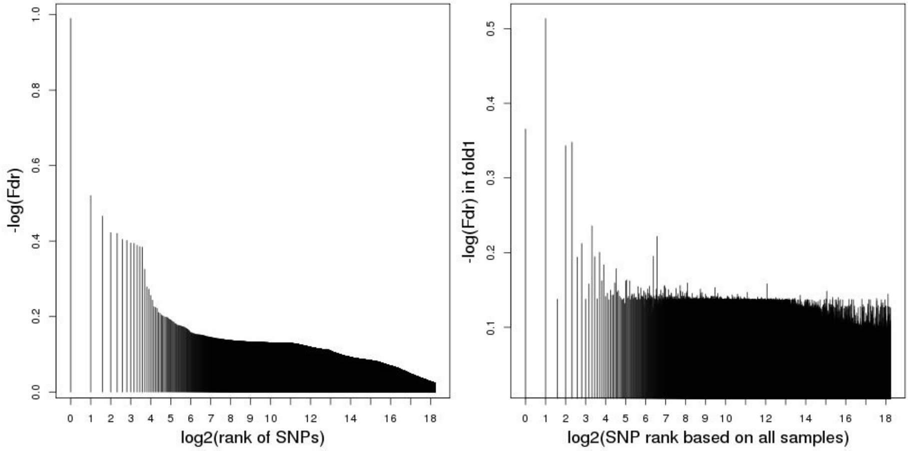
Results on a Real Dataset

Real Dataset

- Genotype was collected from Mayo Clinic LOAD GWAS using Illumina Human-Hap 3000 BeadChips, which includes 313,504 SNPs. (Carrasquillo et. al., Nature Genetics, 2009). We downloaded it from Synapse (syn5591675).
- Response of interest
  - 844 cases with Alzheimer's Disease
  - 1,255 controls without Alzheimer's Disease
- Using Plink, 3,955 SNPs failed to pass the quality control for missing rate > 0.05 and minor allele frequency < 0.01.
- 2,099 samples with 309,549 SNPs and APOE $\varepsilon 4$ are used for real data analysis.

False Discovery Rates

Figure 2: The $\widehat{\text{FDR}}$ based on all samples (left) and the $\widehat{\text{FDR}}$ based on the training samples in fold 1 of ECV given the SNP ranking based on all samples (right).



(a) All samples

(b) Fold 1 in ECV, re-ordered

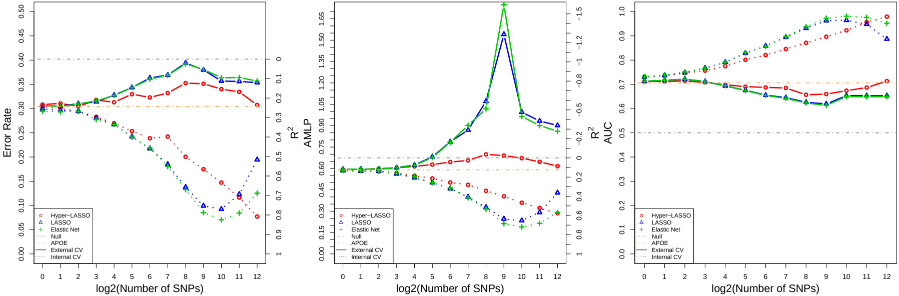
Comparisons of the Ordering of SNPs

Table 1: Top 10 SNPs selected based on all samples and based on the training samples in fold 1 of ECV. The number in the bracket indicates the rank of SNP ordered using the other method to select SNPs.

SNP rank	Based on all samples	Based on fold 1 in ECV
1	rs1279795 (2)	rs8039031 (2)
2	rs8039031 (1)	rs1279795 (1)
3	rs6649176 (378)	rs7318037 (5)
4	rs1552820 (4)	rs1552820 (4)
5	rs7318037 (3)	rs17103033 (10)
6	rs9788079 (11)	rs10753514 (95)
7	rs5915434 (7)	rs5915434 (7)
8	rs6671507 (375)	rs10519111 (13)
9	rs4435421 (21)	rs13237949 (84)
10	rs17103033 (5)	rs1552828 (11)

Predictive Performance

Figure 3: The plots of predictivity of selected SNPs averaged over 10-fold CV for the real dataset.



(a) Error Rate

(b) AMLP

(c) AUC

Subsection 2

Results on two Synthetic Datasets

Synthetic Datasets: I

We use the SNPs from real dataset and generate coefficients and phenotype by the following scheme:

- Select APOE $\varepsilon 4$ and 10 SNPs, denoted by $x_i^{(f)}$, for $f = 0, 1, ..., 10$, which are used as covariates for generating y_i.
- Generate two coefficient vectors β for $x_i^{(f)}$, for $f = 1, ..., 10$, from normal distribution $N(0, \sigma^2)$, where $\sigma = 0.1$ for dataset1 (weak signals) and $\sigma = 2$ for dataset2 (strong signals).
- For $i = 1, 2, ..., 2099$, the phenotype is generated from:

$$y_i \sim \text{Bernoulli}(p_i), \tag{12}$$

where

$$\begin{aligned} p_i &= Pr(y_i = 1 | x_i^{(0)}, ... x_i^{(10)}) \\ &= \frac{1}{1 + \exp(-(\beta_0 + \sum_{f=0}^{10} x_i^{(f)} \beta^{(f)}))}. \end{aligned} \tag{13}$$

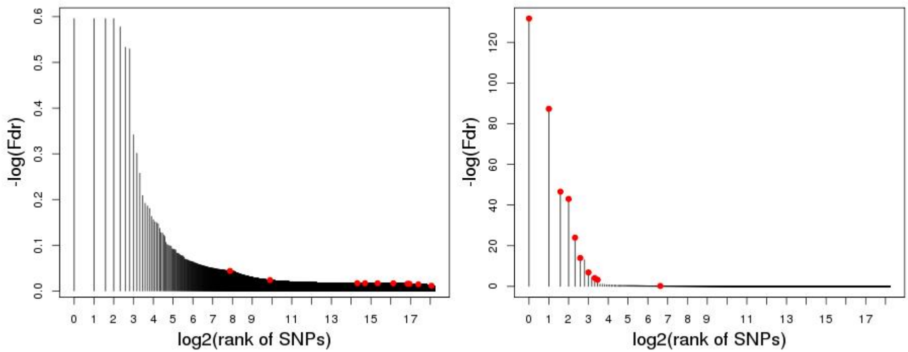
Synthetic Datasets: II

Table 2: Synthetic parameters for the 10 SNPs.

genetic loci	β (weak)	β (strong)	genetic loci	β (weak)	β (strong)
apoe-1	1.00	1.00	rs8106922-2	-0.18	1.84
apoe-2	2.00	2.00	rs405509-1	-0.05	3.44
rs2075650-1	-0.05	-2.45	rs405509-2	0.08	1.65
rs2075650-2	-0.04	0.35	rs8039031-1	-0.06	0.77
rs157580-1	0.05	-1.18	rs8039031-2	-0.03	-3.29
rs157580-2	-0.14	-3.53	rs7318037-1	0.03	1.31
rs439401-1	-0.13	2.19	rs7318037-2	0.09	-0.02
rs439401-2	-0.00	-1.35	rs1420566-1	-0.07	0.73
rs6859-1	-0.03	-1.77	rs1420566-2	0.11	-0.67
rs6859-2	-0.12	-1.80	rs10402271-1	-0.25	-2.98
rs8106922-1	-0.02	2.43	rs10402271-2	-0.04	-3.34

False Discovery Rates Based on All the Samples

Figure 4: The $\widehat{\text{FDR}}$ based on all samples for dataset1 (left) and dataset2 (right). The red dots indicate true signals.

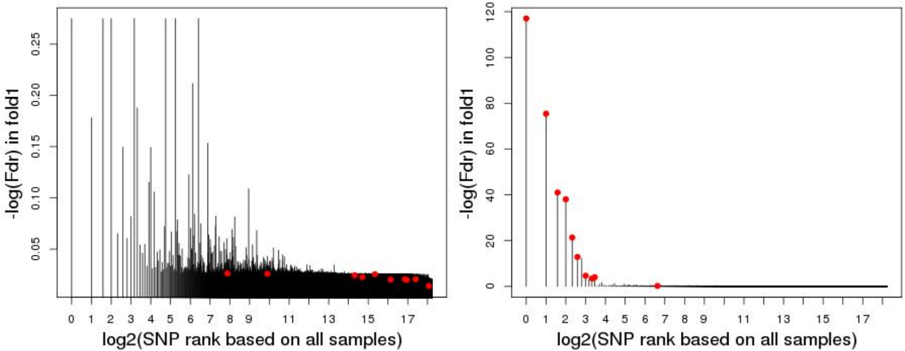


(a) Dataset1, all Samples

(b) Dataset2, all Samples

False Discovery Rates Based on Training Samples in Fold 1

Figure 5: $\widehat{\text{FDR}}$ of fold 1 in ECV given the SNP ranking based on all samples for dataset1 (left) and dataset2 (right). The red dots indicate true signals.



(a) Dataset1, fold 1 in ECV, re-ordered

(b) Dataset2, fold 1 in ECV, re-ordered

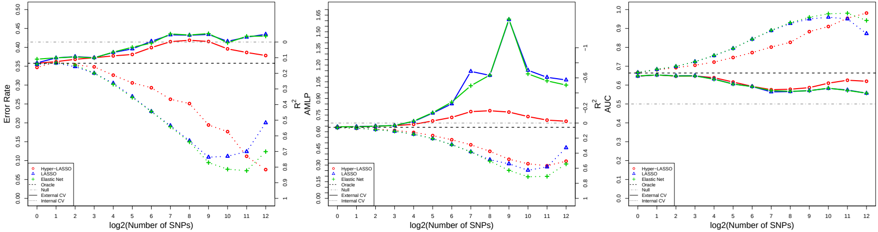
Comparisons of the Ordering of SNPs

Table 3: Top 10 SNPs selected based on all samples and based on the training samples in fold 1 of ECV for dataset1 and dataset2. SNPs with * represent the true signals. The number in the bracket indicates the rank of the SNP ordered using the other method to select SNPs.

Rank	Dataset1		Dataset2	
	All samples	Fold 1	All samples	Fold 1
1	rs1808380 (7)	rs13190617 (4)	rs405509* (1)	rs405509* (1)
2	rs2280201 (10)	rs1321981 (3)	rs8106922* (2)	rs8106922* (2)
3	rs1321981 (2)	rs1010196 (27)	rs157580* (3)	rs157580* (3)
4	rs13190617 (1)	rs10519980 (85)	rs439401* (4)	rs439401* (4)
5	rs11664142 (31)	rs10008892 (38)	rs2075650* (5)	rs2075650* (5)
6	rs2425483 (12)	rs2255994 (9)	rs10402271* (6)	rs10402271* (6)
7	rs2894111 (36)	rs1808380 (1)	rs460527 (7)	rs460527 (7)
8	rs2868574 (20)	rs11084445 (70)	rs8039031* (8)	rs8039031* (8)
9	rs2255994 (6)	rs1432679 (10)	rs2597504 (11)	rs7318037* (11)
10	rs1432679 (9)	rs2280201 (2)	rs1420566* (10)	rs1420566* (10)

Predictive Performance for Dataset 1

Figure 6: The plots of predictive metrics of selected SNPs for synthetic dataset 1.



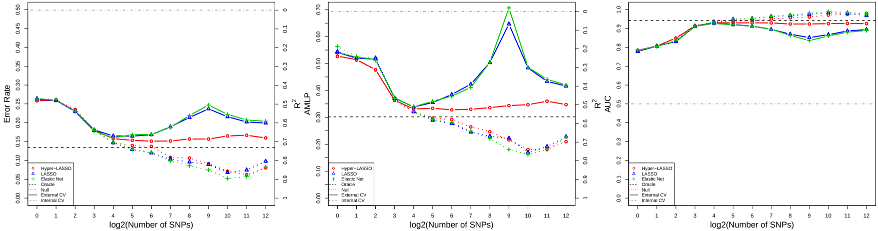
(a) Error Rate

(b) AMLP

(c) AUC

Predictive Performance for Dataset 2

Figure 7: The plots of predictive metrics of selected SNPs averaged over 10-fold CV for synthetic dataset 2.



(a) Error Rate (b) AMLP (c) AUC

Section 4

Conclusions and Future Work

Conclusions and Future Work

- Internal CV will introduce large bias leading severe false discovery, especially when the signal is very weak. External CV should be performed to obtain honest predictivity of selected SNPs.
- Hyper-LASSO with heavy tail performs better than LASSO and Elastic-net in GWAS data.
- Alternatives of ECV?

Selected References

[1] Lo A, Chernoff H, Zheng T, Lo SH. Framework for making better predictions by directly estimating variables' predictivity. *PNAS* Nov 2016; :201616 647doi:10.1073/pnas.1616647113.

[2] Gränsbo K, Almgren P, Sjögren M, Smith JG, Engström G, Hedblad B, Melander O. Chromosome 9p21 genetic variation explains 13% of cardiovascular disease incidence but does not improve risk prediction. *Journal of internal medicine* 2013; **274**(3):233–240.