DIFFERENTIALLY EXPRESSED GENE DETECTION WITH COVARIATE SELECTION UNDER SMALL SAMPLE SIZE GENOMIC SETTING

by

Kexin Guo

BS in Biological Science, Nankai University, China, 2016

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2018

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Kexin Guo

It was defended on

March 27, 2018

and approved by

Thesis Advisor: Tseng George, ScD, Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Robert T. Krafty, PhD, Associate Professor, Department of Biostatistics
Graduate School of Public Health, University of Pittsburgh

Hyun Jung Park, PhD, Assistant Professor, Department of Human Genetics
Graduate School of Public Health, University of Pittsburgh

George Tseng, ScD

DIFFERENTIALLY EXPRESSED GENE DETECTION WITH COVARIATE SELECTION UNDER SMALL SAMPLE SIZE GENOMIC SETTING

Kexin Guo

University of Pittsburgh, 2018

ABSTRACT

In the genomic setting, most data have relative small sample size (n) considering large number of covariates (p). For this type of data structure, it is not appropriate to fit simple linear regression models since the variance would be large and it could encounter over-fitting. Methods for restraining the number of variables contained in the model are necessary.

In this study, constrained best subset (CBS) and LASSO methods were performed to select covariates and detect differentially expressed (DE) genes. For comparison purpose, we set two different simulation settings for each method. Under univariate settings, all methods had type I error well controlled and CBS methods were more powerful than LASSO. However, LASSO had better prediction results compared to CBS methods even though it had more false positive covariates selected. Under genome-wide simulation settings, FDR only well controlled for larger sample size (n=50, 100). Other results have a similar trend as in the univariate setting.

Beyond simulations, eight transcriptomic studies from post-mortem brain tissues of major depressive disorder (MDD) patients were used as a real data application to further compare the CBS2 method and LASSO. As the result of meta-analysis combining all eight studies, CBS2 method generated more DE genes compared to LASSO. It also detected more significant pathways compared to LASSO. Our evaluations suggest that no method performs universally the best in the

small-n-large-p scenario and selection of the best method depends on sample size, dimensionality and the desired biological purpose. From the public health significance perspective, using CBS2 method under small sample size genomic setting could help us detect more DE genes as well as more meaningful pathways.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1.0 INTRODUCTION

Linear regression models are a simple linear approach modeling the relationship between a dependent variable and many independent variables. In real applications, there are usually a pool of candidate covariates among which only a few are predictive of the outcome. As a result, we need to apply model selection methods to find the important predictors. Common model selection techniques include forward selection, backward elimination, stepwise regression and criterion-based procedures using e.g. Akaike information criterion (AIC), Bayes Information Criterion (BIC) or Mallow's Cp statistics, etc. Tibshirani (1996) proposed a regularization method called "LASSO (least absolute shrinkage and selection operator)" to perform variable selection by putting a L1 penalty on the coefficients. As compared to other regularization methods such as ridge regression, the L1 penalty in LASSO can shrink coefficients to exact zeros, thus achieving the goal of variable selection (those variables with non-zero coefficients are retained). Due to its sparsity-inducing property and easiness of implementation, it became a popular and useful alternative to traditional variable selection methods. However, when the sample size is small, it may encounter over-fitting problem (James, Gareth, et al. 2013). The BIC-based best subset selection (i.e. choosing the model that minimizes BIC) method can be a good alternative. In real scenarios, the searching space can grow too large when there are many candidate covariates, thus it is reasonable to put a maximum number of covariates allowed to balance biological interpretation and computation feasibility. Such methods can be called Constrained Best Subset selection (CBS)

methods. Wang et al. (2012) proposed a CBS approach using either BIC or p-values to determine the best set of confounders in each gene while detecting disease-associated genes with a random intercept model. However, when the number of covariate (p) is large, computing will become a problem to get the best model among all possible models (2^p). As an alternative, we could use shrinkage method, also known as the regularization method. More details about LASSO regression will be explained in the method section.

In addition to prediction and variable selection, quantifying the uncertainty in the coefficient estimates is also important in linear regression models to allow for valid statistical inference. Lee et al. (2016) developed a general approach to valid inference for LASSO estimates after model selection. For the CBS based methods, to correct the potential bias of the variable selection procedure, people usually use permutation to get the p-values for inference (Wang et al., 2012). Till today, there is a lack of guidance for researchers on which methods to use for variable selection in real studies, especially when sample size is small. In this thesis, we will perform a thorough comparison between LASSO and CBS methods under small sample size genomic setting and evaluate their performance in type I error and false discovery rate (FDR) control, detection power, prediction, variable selection as well as valid inference. We will perform both simulations and real data analysis to assess the two methods.

As the technology of generating genomic data improves, there are various kinds of genomic data type in biomedical field. (Babu, M. Madan, 2004) In our study, the real data sets that we used is the micro-array data which is obtained by microarray technology. It is an indispensable tool that many biologists use to monitor genome wide expression levels of genes in a given organism (Babu, M. Madan, 2004). Specifically, we will investigate eight major depressive disorder (MDD) datasets from Wang et al. (2012). They come from three patient cohorts (MD1, MD2, MD3)

obtained from different sources at different times. Tissues from the dorsolateral prefrontal cortex (DLPFC), anterior cingulated cortex (ACC) and amygdala (AMY) brain regions were analyzed by microarray experiments to generate eight data sets. Since the tissue used for microarray experiment comes from brain, it is expensive and rare. Therefore, the sample size for each study is small. Beyond that, there are three additional clinical variables (alcohol dependence, evidence of taking anti-depressant drugs and death by suicide) and two technical variables (PH level of brain tissues and post-mortem interval PMI) available for each patient. Our interest of this study is to identify differentially expressed genes that are related to MDD and corrected for the confounding effects from the clinical and technical variables.

Consider the relative small sample size and the noise of the technology, it is of great importance to use the above variable selection methods to constrain the covariates that can be included in the model. Beyond that, we also performed meta-analysis using Fisher's method to combine the p-values since the signal in each study is relatively weak. It helps us to improve the results of DE gene detection and find enriched biological pathways.

2.0 METHODS

2.1 CONSTRAINED BEST SUBSET METHOD

Due to the relatively small sample size and large number of covariates in the data set, it is not appropriate to use simple linear regression for variable selection and identifying differentially expressed genes. In this study, improved linear model called Constrained best subset model (CBS) with gene-specific variable selection was performed. Denote gene by g ($1 \leqslant g \leqslant G$), sample by i ($1 \leqslant i \leqslant n$), covariate by j ($1 \leqslant j \leqslant p$). To adjust for potential cofactors while detecting disease associated genes, consider the following linear model for each gene:

$$Y_{gi} = \sum_{j=1}^{p} X_{ij}\beta_{gj} + \beta_{gd} * disease_i + \varepsilon_{gi}$$

Y_{gi}: gene expression level for gene g in ith sample.

X_{ij}: value of jth covariate in ith sample.

β_{gj}: effect of jth covariate on the expression level of gene g.

β_{gd}: disease effect on the expression level of gene g.

$disease_i$: disease indicator for ith sample.

ε_{gi}: error term of ith sample for gene g.

Among all the p covariates available, we assume only a small number of covariates (denoted as s, $s<<p$) are predictive of the gene expression thus β's of the other (p-s) covariates are zeros. The CBS method we introduce here test the disease effect while at the same time selecting for the best subset of covariates. It constrains the number of covariates that can be selected to be

at most K and searches for the best subset of variables which has the smallest Bayesian Information Criterion (BIC) among all possible pC1+pC2+…+pCk subsets. Three CBS methods with different K values (CBS1 (K=1), CBS2 (K=2) and CBS3 (K=3)) were compared for variable selection and differentially expressed gene detection in both simulation and real data application.

Taking CBS2 method as an example, all possible linear models that included the disease indicator variable and at most two covariates were fitted and the one with the smallest BIC was selected. After the best model was determined for each gene, we used likelihood ratio test to test whether there was disease effect on the gene expression for each gene (H0: $\beta_{gd}=0$). However, this p-value is biased due to the best subset selection procedure and type I error is not well controlled. To solve this problem, we treated the observed p-value as the test statistics and performed a permutation test by shuffling sample labels (i.e. disease status) of patients. In this study, Bp is used to denote the number of permutations. At each permutation, the p-value of disease effect is obtained for all genes and generates a large matrix (G x Bp) as the null distribution of the p-value for disease status variable. Denoted by $p_g^{(o)}$ the derived p-value for gene g in observed data and $p_{g'}^{(b)}$ the derived p-value for gene g' in the b-th permuted data. The permutation test p-value for gene g is obtained by:

$$p_g = \frac{\sum_{g'=1}^{G} \sum_{b=1}^{Bp} I(p_{g'}^{(b)} \leq p_g^{(o)})}{G * Bp}$$

Similar process was performed for CBS1 and CBS3 models.

2.2 LASSO METHOD

For comparison, we will compare CBS method to the commonly used regularization and variable selection method LASSO (R. Tibshirani 2016). Letting $\hat{\boldsymbol{\beta}} = \{\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_j, ..., \hat{\beta}_p\}$, the LASSO estimate:

$(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = argmin \sum_{i=1}^{n} (y_i - \alpha - \sum_{j=1}^{p} \beta_j * x_{ij})^2$ subject to $\sum_{j=1}^{p} |\beta j| \leq t$,

which is equivalent to: $(\hat{\alpha}, \hat{\boldsymbol{\beta}}) = argmin \sum_{i=1}^{n} (y_i - \alpha - \sum_{j=1}^{p} \beta_j * x_{ij})^2 + \lambda * \sum_{j=1}^{p} |\beta_j|$

It starts with the full model with all the covariates included in the model and later shrinks some coefficients directly to zero based on the tuning parameter λ. The non-zero coefficients returned by LASSO indicates the selected variables. We use the R package "glmnet" (Jerome Friedman et al. 2017) to implement the LASSO method. Different values of the tuning parameter λ will give different degrees of sparsity. Therefore, the selection of tuning parameter is very important. In this study, we use leave-one-out cross-validation to select the best λ that minimizes the mean square error (MSE) for each gene.

For LASSO, we use post-selective method to generate p-values of variables for inference. The post-selection method for LASSO was developed by Lee et al. (2016) to generate valid inference after model selection, so only those variables selected in the previous step will have meaningful p-values (Lee et al. Exact post-selection inference, with application to the lasso). If the disease indicator variable is not selected, we will enforce its p-value to be 1 since LASSO method does not treat it as an important predictor. We use the R package "selectiveInference" (Ryan Tibshirani et al. 2017) to perform the post-selection inference.

2.3 EVALUATION CRITERIA

In the simulation section, we will perform univariate and genome-wide settings. For univariate setting in simulation, we evaluated the power, Type I error as well as variable selection performance for constrained best subset methods and LASSO. For genome-wide setting, we evaluated the detection power of differentially expressed genes, False Discovery Rate (FDR) and the variable selection performance.

To evaluate variable selection performance of four methods (CBS1, CBS2, CBS3 and LASSO), we compared sensitivity, specificity and Youden index under both simulation settings. Sensitivity is calculated as the number of detected true features (true positives) divided by the total number of true features. Specificity is obtained by the number of negative features that are not detected (true negatives) divided by the total number of negative features. Negative features are those covariates not supposed to have relationship with the gene expression. To evaluate the overall performance, we use Youden index = sensitivity+ specificity -1. (All results are averaged over B=1000 replications for univariate setting and B=50 replications for genome-wide setting)

Under univariate simulation setting, both Power and Type I error for differentially expressed gene detection were evaluated at the significance level of 0.05. Power is defined as the average number of detected DE genes divided by the simulation times (B=1000 replications). Type I error is defined as the incorrect positive findings under null ("false positive"). Beyond that, we also calculated the root mean square error (RMSE) in an independent testing set with sample size n=10,000 under univariate simulation setting. RMSE is defined as below:

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{b=1}^{B}\sum_{j=1}^{J}(y_j^{test}-\hat{\beta}^{(b)}*x_j^{test})^2}{B*J}}$$

Under genomic simulation setting, we evaluated the false discovery rate (FDR) considering the multiple comparison issue. FDR in this study is defined as the number of false positive genes among all detected genes. We use Benjamini-Hochberg method to correct p-value for multiple comparison and get q-values. We set q-value cutoff of 0.05 (nominal FDR) to detect differentially expressed genes between case and control groups and assess the true FDR.

To evaluate the detection power of disease associated genes under genomic simulation setting, we plotted the number of true positive genes on the y-axis against top ranked genes (by p-values) on the x-axis for each of the four methods.

For real genomic data, since we do not know what is the true set of covariates that are related with gene expression level for each specific gene and which gene is significantly associated with the disease, we only compare CBS2 and LASSO model based on the number of detected genes given different FDR threshold. We also perform pathway enrichment analysis to examine biological annotation of detected genes using different methods.

3.0 SIMULATION

3.1 SIMULATION SETTINGS

We conducted both univariate and genomic simulations to evaluate the performance of constrained best subset method and LASSO method for differentially expressed gene detection and variable selection. For univariate simulation, we will assess the type I error, power, variable selection performance as well as the prediction error. For genomic simulation, we will assess both false discovery rate (FDR) and power.

3.1.1 UNIVARIATE SETTING

1. We simulated sample size $n \in \{20, 50, 100\}$. Each sample has 10 covariates ($X_1 – X_{10}$) and a disease indicator variable. Each of X_1-X_{10} was sampled from a standard normal distribution, i.e. $X_1 – X_{10} \sim N(0, 1)$. Disease indicator variable was randomly sampled from {0, 1} for each sample. In addition, we also simulated an independent test set with sample size n=10,000 to evaluate the prediction errors (RMSE).

2. We assume s=2 or 4 covariates that are predictive of the outcome. For s=2, we simulate each y_i from $N(\beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \beta_0 \cdot disease_i, 1)$. $\beta_0=\beta_1=\beta_2=\beta$ were set as {0.5, 1.3, 2} for each of the setting for a thorough comparison of the four methods.

For s=4, we simulate each y_i from $N(\beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \beta_3 \cdot X_{3i} + \beta_4 \cdot X_{4i} + \beta_0 \cdot disease_i, 1)$.

3. We similarly set $\beta_0=\beta_1=\beta_2=\beta_3=\beta_4=\beta \in \{0.5, 1.3, 2\}$. We run each of the univariate settings for B=1000 time and take the averaged results.

3.1.2 GENOMIC SETTING

To better mimic the real data scenarios, we simulate another genomic setting with 100 genes in total. Further, we also simulate gene-gene dependency structure.

1. We simulate 40 genes from 2 gene clusters (20 genes in each cluster) and another 60 genes

not belonging to any clusters. Each cluster has 5 differentially expressed genes and 15 noise genes. For the 60 independent genes, 10 out of 60 are generated as DE genes. In total, we have 20 DE genes out of 100.

2. For each of the two clusters (k=1,2), we sampled $\sum_{k\ 20*20} \sim W^{-1}(\psi,60)$ (k=1,2), where $\Psi = 0.5*I_{20\times 20} + 0.5*J_{20\times 20}$, W^{-1} denotes the inverse Wishart distribution, I is the identity matrix and J is the matrix with all elements equal to 1.

3. We assume s=2 or 4 covariates predictive of the outcome as in the univariate setting. For s=2, we simulate $(y_{1i}, y_{2i}, \ldots y_{20i})$ from N ($\boldsymbol{\mu}$, $\sum_{k\ 20*20}$) for genes from cluster k=1,2, where $\boldsymbol{\mu}= (\mu, \mu, \mu, \mu, \mu, 0, \ldots, 0)^T$, $\mu= \beta_{11}\bullet X_{1i}+ \beta_{12}\bullet X_{2i}+ \beta_{10}\bullet disease_i$. For s=4, $\mu= \beta_{11}\bullet X_{1i}+ \beta_{12}\bullet X_{2i}+ \beta_{13}\bullet X_{3i}+ \beta_{14}\bullet X_{4i}+ \beta_{10}\bullet disease_i$

The 60 independent genes are generated in a similar way as in the univariate setting. All effective coefficients βs are set as {0.5, 1.3, 2} for each of the two sub-settings.

4. We run each of the genomic settings for B=50 times and take the averaged results.

3.2 SIMULATION RESULTS

3.2.1 UNIVARIATE SETTING

1. Type I error rate

Table 1 shows the type I error rate of disease variable inference for four methods under different sample size situations. The simulation was repeated for B=1000 times for CBS and B=5000 for LASSO. To calculate permutation p-value for CBS, Bp=100 was used.

As we can see from Table 1, all of the four methods well controlled the Type I error even when sample size is very small.

Table 1. Type I error rate of four methods under different sample size situation

	n=20	**n=50**	**n=100**
CBS1	0.049(0.007)	0.039(0.006)	0.057(0.007)
CBS2	0.051(0.007)	0.046(0.007)	0.030(0.005)
CBS3	0.047(0.007)	0.053(0.007)	0.038(0.006)
LASSO	0.0528(0.003)	0.0514(0.003)	0.0506(0.003)

2. Power

Table 2 shows the results of power for disease inference under different sample size and different effect size settings for each of the four methods when s=2.

When the sample size is small (n=20), all the CBS methods are more powerful than LASSO method. As the sample size increases, LASSO becomes more powerful. In addition, the CBS methods outperform LASSO method when the signals are weaker (β=0.5, 1.3) and perform more

similarly as LASSO when the signal is strong and sample size is large. Since the true s=2, CBS2, which selects covariates up to 2, is the most powerful one among all three CBS methods.

Table 3 show the results of power for disease inference under different sample size and different effect size settings for each of the four methods when s=4.

When the sample size is small (n=20), all the CBS methods are more powerful than LASSO method. As the sample size increases, LASSO becomes more powerful. As effect size increases, the power of CBS methods increases in a much large scale than LASSO. In addition, the CBS methods outperform LASSO method when the signals are weaker (β=0.5, 1.3) and perform more similarly as LASSO when the signal is strong and sample size is large. Even though the truth is s=4, CBS methods still perform better at disease inference when sample size is small and CBS3 is close to the truth so that it is the most powerful one among all three methods.

Table 2. Power of four methods under different sample size and different effect size situation when s=2

	β=0.5			β=1.3			β=2		
	n=20	n=50	n=100	n=20	n=50	n=100	n=20	n=50	n=100
CBS1	0.132	0.312	0.604	0.339	0.766	0.975	0.429	0.872	0.995
CBS2	0.132	0.357	0.689	0.626	0.989	1	0.964	1	1
CBS3	0.122	0.344	0.682	0.563	0.984	1	0.927	1	1
LASSO	0.074	0.423	0.741	0.098	0.824	0.914	0.119	0.883	0.944

Table 3. Power of four methods under different sample size and different effect size situation when s=4

	β=0.5			β=1.3			β=2		
	n=20	n=50	n=100	n=20	n=50	n=100	n=20	n=50	n=100
CBS1	0.113	0.245	0.486	0.18	0.41	0.738	0.19	0.446	0.776
CBS2	0.105	0.269	0.528	0.189	0.534	0.856	0.225	0.603	0.901
CBS3	0.104	0.302	0.586	0.269	0.734	0.967	0.35	0.839	0.991
LASSO	0.056	0.455	0.778	0.067	0.855	0.931	0.088	0.909	0.953

3. Youden index of variable selection performance

Youden index is calculated as sensitivity+ specificity -1. Table 8-9 below show the Youden Index for each of the four methods.

All the three CBS methods perform better than LASSO in variable selection. Among them, since the true s=2, CBS2 performs the best. When sample size is fixed for each method, Youden Index increases as the effect size increases. In addition, when effect size is fixed for each method, Youden Index increases as the sample size increases.

When the true s=4, CBS methods performs better than LASSO when sample size is small (n=20) and similar as LASSO when sample size is large. LASSO method performs better than CBS methods when sample size is large (n=50, 100) and effect size is large (β=2).

Table 4. Youden Index of four methods under different sample size situation and different effect size for variable selection performance when s=2

	n=20			n=50			n=100		
	β=0.5	β=1.3	β=2	β=0.5	β=1.3	β=2	β=0.5	β=1.3	β=2
CBS1	0.307	0.485	0.494	0.476	0.5	0.5	0.498	0.5	0.5
CBS2	0.474	0.981	0.999	0.855	1	1	0.995	1	1
CBS3	0.483	0.900	0.913	0.842	0.952	0.949	0.963	0.965	0.968
LASSO	0.378	0.684	0.778	0.553	0.643	0.816	0.569	0.619	0.821

Table 5. Youden Index of four methods under different sample size situation and different effect size for variable selection performance when s=4

	n=20			**n=50**			**n=100**		
	β=0.5	β=1.3	β=2	β=0.5	β=1.3	β=2	β=0.5	β=1.3	β=2
CBS1	0.192	0.230	0.239	0.244	0.249	0.250	0.250	0.250	0.250
CBS2	0.316	0.458	0.468	0.480	0.499	0.500	0.499	0.500	0.500
CBS3	0.396	0.693	0.721	0.687	0.750	0.750	0.749	0.750	0.750
LASSO	0.345	0.604	0.695	0.425	0.565	0.758	0.441	0.528	0.755

4. Root mean square error (RMSE)

Tables 10-11 below show the results of RMSE for each of the four methods.

LASSO method has the smallest RMSE in all scenarios for both s=2 and s=4, even with small sample size (n=20). When sample size is fixed for each method under either setting, RMSE will increase if effect size increases. In addition, when effect size is fixed for each method, RMSE will decrease if sample size increases. Among the three CBS models, CBS2 and CBS3 have a better prediction performance (smaller RMSE) than CBS1.

Table 6. RMSE of four methods under different sample size situation and different effect size when s=2

	n=20			n=50			n=100		
	β=0.5	β=1.3	β=2	β=0.5	β=1.3	β=2	β=0.5	β=1.3	β=2
CBS1	1.282	1.804	2.424	1.161	1.689	2.300	1.136	1.665	2.269
CBS2	1.300	1.443	1.772	1.110	1.375	1.745	1.073	1.367	1.739
CBS3	1.278	1.438	1.783	1.101	1.376	1.746	1.072	1.367	1.739
LASSO	1.238	1.414	1.767	1.097	1.130	1.370	1.045	1.054	1.230

Table 7. RMSE of four methods under different sample size situation and different effect size when s=4

	n=20			**n=50**			**n=100**		
	β=0.5	β=1.3	β=2	β=0.5	β=1.3	β=2	β=0.5	β=1.3	β=2
CBS1	1.493	2.719	3.967	1.370	2.544	3.722	1.347	2.506	3.667
CBS2	1.461	2.463	3.542	1.308	2.322	3.366	1.287	2.305	3.342
CBS3	1.416	2.371	3.416	1.296	2.308	3.346	1.284	2.297	3.330
LASSO	1.333	1.702	2.293	1.119	1.204	1.539	1.055	1.083	1.317

3.2.2 GENOMIC SETTING

1. False discovery rate (FDR)

Table 12 show the true FDR at nominal FDR equal to 0.05 for the four methods when s=2.

When the sample size is small (n=20) and signal is weak (β=0.5), FDR of all the four methods is not well controlled at 0.05. However, as effect size increases (β=1.3 and β=2), CBS2 and CBS3 will have much smaller FDR compared to CBS1 and LASSO method. When n=50, all the methods except for CBS1 can control the FDR at 0.05. When sample size is large (n=100), all the four methods control the FDR well.

Similar results of FDR from s=4 sub-setting are shown in table 13.

Table 8. FDR of four methods under different effect size and sample size situation for disease inference when s=2

	n=20			**n=50**			**n=100**		
	β=0.5	β=1.3	β=2	β=0.5	β=1.3	β=2	β=0.5	β=1.3	β=2
CBS1	0.931	0.827	0.826	0.439	0.237	0.235	0.070	0.037	0.037
CBS2	0.838	0.048	0.059	0.057	0.036	0.037	0.046	0.046	0.049
CBS3	0.798	0.045	0.055	0.055	0.038	0.037	0.044	0.046	0.044
LASSO	0.930	0.897	0.908	0.070	0.037	0.041	0.034	0.033	0.033

Table 9. FDR of four methods under different effect size and sample size situation for disease inference when s=4

	n=20			**n=50**			**n=100**		
	β=0.5	β=1.3	β=2	β=0.5	β=1.3	β=2	β=0.5	β=1.3	β=2
CBS1	0.951	0.895	0.903	0.647	0.276	0.232	0.318	0.046	0.047
CBS2	0.939	0.818	0.160	0.596	0.034	0.035	0.085	0.046	0.047
CBS3	0.938	0.741	0.154	0.364	0.030	0.028	0.062	0.044	0.044
LASSO	0.930	0.900	0.905	0.060	0.038	0.040	0.035	0.034	0.033

2. Detection power of disease associated genes

Since the power plots have similar trend under s=2 and s=4 for fixed effect size and fixed sample size, we just show the results for s=4 to compare the four methods.

Figure 1-3 show the detection power comparison for different effect size when n=20. X axis is the top number of declared genes ranked by p-values of the disease effect, Y axis refers to the number of true disease associated genes. We can see that all the three CBS methods perform better than LASSO method and CBS3 method has the best power among the four methods. As the

effect size increases, the power of LASSO method does not increase much compared to the other three methods. Figure 4-6 show the power plot for different effect size when n=50. Here since the sample size increases, LASSO method performs almost equally well as the CBS2 and CBS3 methods.

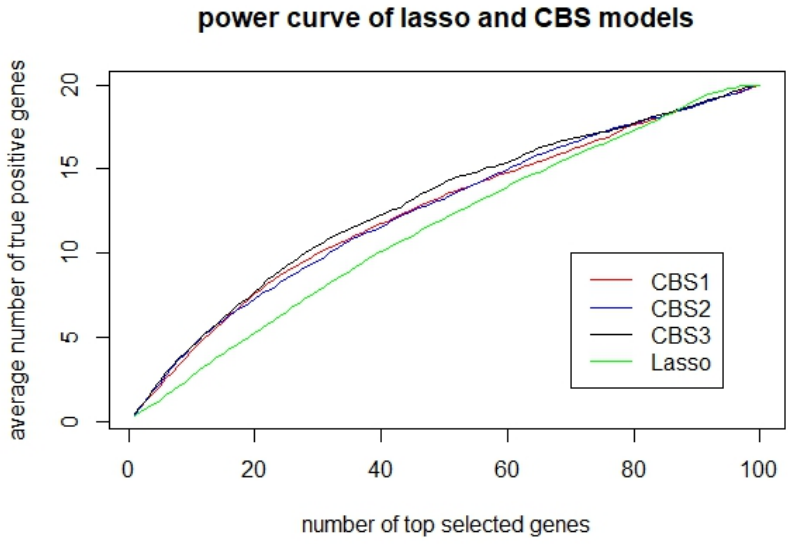

Figure 1. Rank-based power curve of four methods with β=0.5, n=20

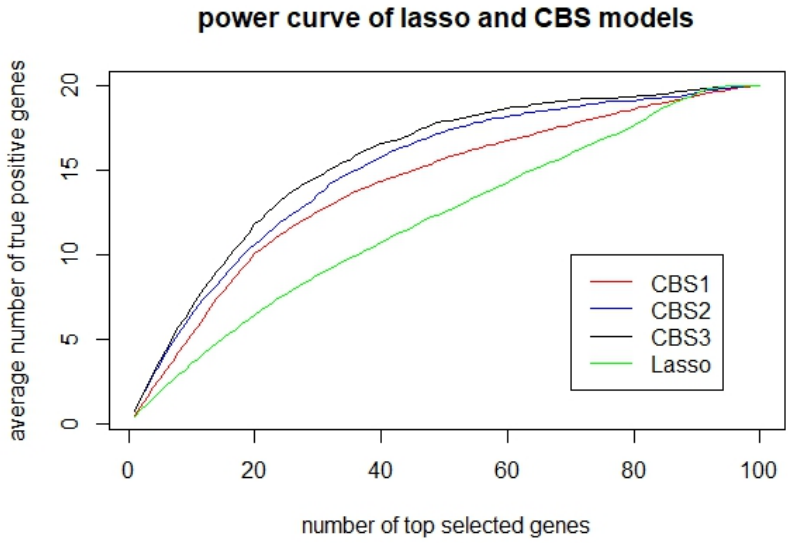

Figure 2. Rank-based power curve of four methods with β=1.3, n=20 when s=4

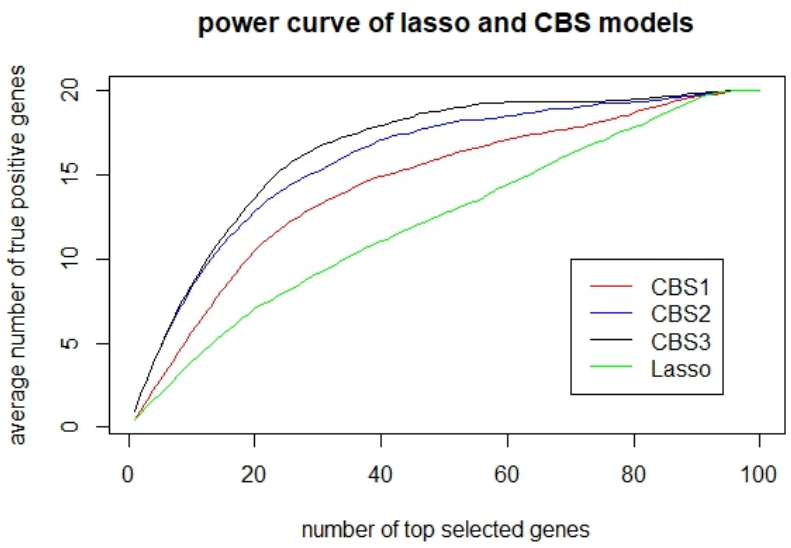

Figure 3. Rank-based power curve of four methods with β=2, n=20 when s=4

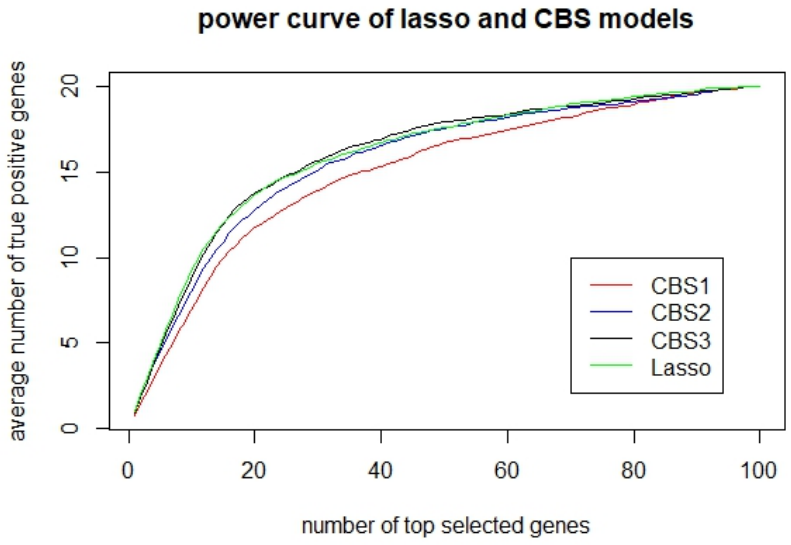

Figure 4. Rank-based power curve of four methods with β=0.5, n=50 when s=4

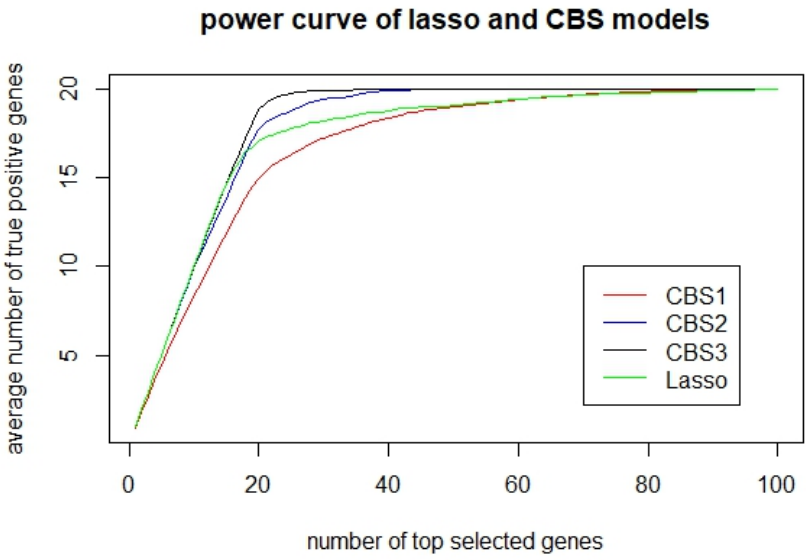

Figure 5. Rank-based power curve of four methods with β=1.3, n=50 when s=4

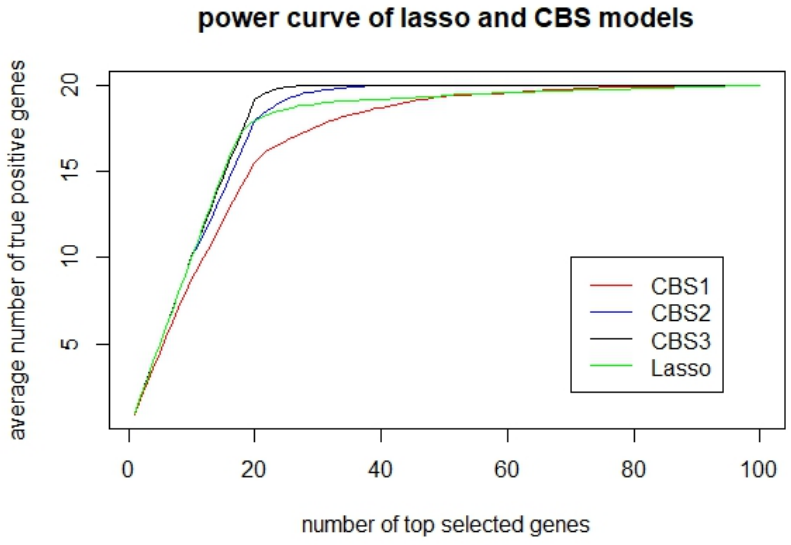

Figure 6. Rank-based power curve of four methods with β=0.5, n=50 when s=4

3. Variable selection performance

Tables 14-19 below show sensitivity, specificity and Youden Index of variable selection for the four methods under different settings.

When sample size is fixed, sensitivity and specificity will increase as effect size increases for each method. Overall, when s=2, CBS2 and CBS3 methods have larger Youden index than LASSO method. More specifically, LASSO method and CBS methods have similar sensitivity. However, CBS methods have higher specificity since LASSO tend to select more false positive features. When s=4, LASSO method has larger Youden index than all CBS methods under most settings. Like in s=2 scenario, LASSO method has higher sensitivity than CBS methods but lower specificity.

Table 10. Sensitivity, specificity and Youden Index of four methods under different effect size situation with sample size n=20 when s=2

	β=0.5			**β=1.3**			**β=2**		
n=20	sen	spe	youden	sen	spe	youden	sen	spe	youden
CBS1	0.401	0.938	0.339	0.487	0.941	0.428	0.494	0.941	0.435
CBS2	0.733	0.888	0.621	0.988	0.895	0.882	0.999	0.895	0.894
CBS3	0.769	0.844	0.612	0.992	0.848	0.84	1	0.849	0.849
LASSO	0.791	0.804	0.595	0.975	0.813	0.787	0.983	0.823	0.806

Table 11. Sensitivity, specificity and Youden Index of four methods under different effect size situation with sample size n=50 when s=2

	β=0.5			**β=1.3**			**β=2**		
n=50	sen	spe	youden	sen	spe	youden	sen	spe	youden
CBS1	0.486	0.964	0.449	0.5	0.964	0.464	0.5	0.964	0.464
CBS2	0.931	0.95	0.881	1	0.952	0.952	1	0.952	0.952
CBS3	0.949	0.939	0.888	1	0.94	0.94	1	0.94	0.94
LASSO	0.982	0.818	0.8	1	0.849	0.849	1	0.869	0.869

Table 12. Sensitivity, specificity and Youden Index of four methods under different effect size situation with sample size n=100 when s=2

	β=0.5			**β=1.3**			**β=2**		
n=100	sen	spe	youden	sen	spe	youden	sen	spe	youden
CBS1	0.5	0.977	0.477	0.5	0.977	0.477	0.5	0.977	0.977
CBS2	0.997	0.972	0.969	1	0.972	0.972	1	0.972	0.972
CBS3	0.998	0.966	0.964	1	0.966	0.966	1	0.966	0.966
LASSO	1	0.83	0.83	1	0.859	0.859	1	0.88	0.88

Table 13. Sensitivity, specificity and Youden Index of four methods under different effect size situation with sample size n=20 when=4

	β=0.5			**β=1.3**			**β=2**		
n=20	sen	spe	youden	sen	spe	youden	sen	spe	youden
CBS1	0.207	0.936	0.143	0.241	0.938	0.179	0.243	0.938	0.181
CBS2	0.381	0.884	0.265	0.486	0.89	0.376	0.492	0.89	0.382
CBS3	0.532	0.847	0.379	0.639	0.851	0.49	0.644	0.852	0.496
LASSO	0.681	0.8	0.481	0.787	0.81	0.596	0.775	0.817	0.592

Table 14. Sensitivity, specificity and Youden Index of four methods under different effect size situation with sample size n=50 when s=4

	β=0.5			**β=1.3**			**β=2**		
n=50	sen	spe	youden	sen	spe	youden	sen	spe	youden
CBS1	0.244	0.962	0.206	0.25	0.963	0.213	0.25	0.963	0.213
CBS2	0.472	0.948	0.421	0.5	0.95	0.45	0.5	0.95	0.45
CBS3	0.614	0.942	0.556	0.638	0.943	0.581	0.638	0.943	0.581
LASSO	0.841	0.823	0.665	0.837	0.847	0.684	0.788	0.864	0.652

Table 15. Sensitivity, specificity and Youden Index of four methods under different effect size situation with sample size n=100 when s=4

	β=0.5			**β=1.3**			**β=2**		
n=100	sen	spe	youden	sen	spe	youden	sen	spe	youden
CBS1	0.25	0.976	0.226	0.25	0.976	0.226	0.25	0.976	0.226
CBS2	0.499	0.971	0.47	0.5	0.971	0.471	0.5	0.971	0.471
CBS3	0.633	0.968	0.601	0.634	0.968	0.602	0.634	0.968	0.602
LASSO	0.857	0.836	0.692	0.844	0.86	0.704	0.787	0.878	0.665

3.3 CONCLUSION OF SIMULATIONS

From univariate and genomic simulation settings, we can see CBS2 and CBS3 have a better performance in both variable selection and power when sample size is small regardless of the effect size. The reason why CBS methods perform better is that they are very close to the truth and we constrain the number of covariates to be selected. This indicates that LASSO method will contain more false covariates in the model. However, LASSO method is preferred when sample size is large (50-100) since it has similar performance as CBS methods and it require less computing. In addition, LASSO method is better at making prediction for future patient compared to CBS methods. Although it does not always select the correct covariates, the estimate coefficient for the false covariates would be very close to 0 so that it could still generate the smallest RMSE.

4.0 REAL DATA APPLICATION

4.1 METHOD USED FOR REAL DATA APPLICATION

1. Data pre-processing method

We collected the eight datasets from Wang et al. (2012). The microarrays were scanned and summarized as by the manufacturers' default settings and the raw microarray data was preprocessed in a standard pipeline (Ding et al., 2015). The data was log2 transformed and the gene symbols were matched across 8 studies. In total, there are 16,689 common genes matched for eight studies before data pre-processing. Two sequential steps of gene filtering were implement according to Ding et al. (2015). We first calculated the mean expression level for each gene of eight studies and rank them from the smallest to largest. Then we sum the ranks for each gene across 8 studies and filters 30% genes that has the lowest rank sum. Similar process was done to filter the 40% genes that has the lowest rank sum of standard deviation. After gene filtering, 10680 common genes were left for eight studies.

2. Identification of disease associated genes.

There are a total of seven candidate confounders and we would like to select the ones most predictive of the gene expression. We applied both CBS2 method as well as LASSO method to identify MDD associated genes while adjusting for potential cofactors.

3. Meta-analysis method

Since the signal for each study is weak and individual study sample size is small, we use Fisher's method to combine p-values from eight studies to improve the DE gene detection power.

Fisher's test statistic is defined as $T_{gfisher}$-$2*\sum_{k=1}^{8} \log(p_{gk})$ $(1 \leq g \leq 10680)$. Under null, $T_{gfisher}$ ~ chi (16), and we can obtain the p-value for the test statistic. To control for multiple comparison, we use Benjamini-Hochberg method to obtain q-values.

4. Pathway Enrichment analysis

To annotate the DE genes detected, we further performed pathway enrichment analysis by using Fisher's exact test based on four pathway databases: Gene ontology Biological Process (GOBP), Biocarta, Kyoto Encyclopedia of Genes and Genomes (KEGG), and Reactome. To ensure a fair comparison, we selected the same number of top DE genes (top 500) from each method to perform pathway enrichment analysis.

4.2 RESULTS

1. DE genes detection under different q-value cutoff for each individual study

Table 20 shows the results of DE gene detection under different q-value cutoff of each individual study for CBS2 and LASSO method. From the table we can see that both methods do not detect much DE genes for most individual studies which indicates that the signal is very weak. Therefore, we need meta-analysis to improve the detection results.

Table 16. DE gene detection under different q-value cutoff of each individual study for CBS2 and LASSP

	CBS2 method			LASSO method		
	q<0.15	q<0.2	q<0.25	q<0.15	q<0.2	q<0.25
Study1	0	820	1463	254	393	632
Study2	5	9	23	NA	NA	NA
Study3	1	1	2	136	166	240
Study4	0	0	0	NA	NA	NA
Study5	545	968	1487	91	121	149
Study6	0	0	0	32	37	39
Study7	0	0	0	73	100	141
Study8	406	717	1094	12	19	72

2. DE genes detection under different q-value cutoff for meta-analysis

Table 21 shows the results of DE gene detection under different meta-analysis q-value cutoff for CBS2 and LASSO method. From the table we can see that CBS2 method detects many more DE genes than LASSO method at all q-value cutoffs. This indicates that the CBS2 method is more powerful than LASSO method when sample size is small and signal is weak (e.g. with MDD disease) in real studies.

Table 17. DE gene detection under different q-value cutoff of meta-analysis for CBS2 and LASSO method

q-value cutoff	Number of DE genes detected by CBS2 method	Number of DE genes detected by LASSO method
0.05	286	43
0.1	711	46
0.15	1053	48
0.2	1380	55

3. Pathway enrichment analysis

We picked top 500 DE genes from each method and perform pathway enrichment analysis. Table 22 shows the number of pathways detected by CBS2 and LASSO method under different q-value cutoff. We can see that LASSO method does not detect any pathways under q-value cutoff 0.5. In addition, only 11 pathways are detected as q-value<1. In the contrast, CBS2 method detects many more pathways under certain q-value cutoff. For instance, there are 10 pathways have q-value<0.2 and 26 pathways have q-value<0.3.

From table 23, we can see that for the top5 pathways detected by CBS2 method, q-value of LASSO method all equal to 1. However, for the top5 pathways detected by LASSO method, q-value of CBS2 method are smaller than that of LASSO method. In addition, we can see some of the detected pathways are related with brain disease. Therefore, CBS2 method can detect more meaningful pathways than LASSO method.

Table 18. Number of detected pathways under different q-value cutoff for CBS2 and LASSO method

q-value cutoff	CBS2 method	LASSO method
q-value<0.2	10	0
q-value<0.3	26	0
q-value<0.5	49	0
q-value<0.7	118	5
q-value<1	294	11

Table 19. p-value and q-value of top 10 pathways detected by CBS2 and LASSO method

	CBS2 method		LASSO method	
Detected Pathways	p-value	q-value	p-value	q-value
Reactome Formation of ATP by chemiosmotic coupling	0.000133	0.15	0.105889	1
KEGG Leukocyte transendothelial migration	0.000191	0.15	0.362605	1
BioCarta Thrombin signaling and protease-activated receptors	0.000237	0.15	0.141239	1
Reactome G alpha (12/13) signalling events	0.000318	0.151	0.134154	1
Reactome Axon guidance	0.00044	0.168	0.647771	1
KEGG Oxidative phosphorylation	0.003682	0.269	0.001373	0.662
KEGG Huntington's disease	0.047343	0.721	0.001416	0.662
KEGG Alzheimer's disease	0.01037	0.444	0.001461	0.662
Reactome Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins.	0.005136	0.334	0.001739	0.662
KEGG Steroid biosynthesis	0.016352	0.516	0.002403	0.711

5.0 DISCUSSION AND CONCLUSION

As we can see from the results of two simulation settings, CBS2 and CBS3 method perform well in both variable selection and power when sample size is small. LASSO method is better at prediction since it always has the smallest RMSE. When sample size is large, LASSO method would have similar results as two CBS methods. In conclusion, no method performs universally the best in the small-n-large-p scenario and selection of the best method depends on sample size, dimensionality and the desired biological purpose.

From real data application results, we conclude that CBS2 can detect more meaningful DE genes by using Fisher's method to implement meta-analysis in the motivation MDD data. Some of the pathways are related to brain disease which indicates that the DE genes detected is important referring to MDD disease even though each of the 8 studies has very weak signal. Although we use same top number of DE genes for both method, LASSO method did not perform well in pathway analysis since most of the top genes do not contain much information of disease and the number of DE genes is limited. Therefore, it is not worthwhile for LASSO to do pathway analysis in this setting.

One limitation in our study is that the comparison among CBS methods and LASSO method is not fair, since we force the disease variable to be always selected and constrain the maximum number of covariates that could be chosen for three CBS methods. The reason that we do not force disease variable to be always retained in the model for LASSO method is the lack of valid method to obtain the p-value for that specific variable. The method that we used in this study for LASSO inference is the post-selective inference method (reference). It will return the p-values for all the coefficients that do not shrink to 0. Although we could use "glmnet" package for LASSO

to fix a variable to always be in the model and obtain the coefficient estimate, right now there is no efficient package to obtain the corresponding p-value for inference purpose. We would like to develop a method that helps us obtain the valid p-value for the fixed variable in order to make it a fair comparison in the future.

BIBLIOGRAPHY

Tibshirani, Robert. "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* (1996): 267-288.

James, Gareth, et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.

Wang, Xingbin, et al. "Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: with application to major depressive disorder." *BMC bioinformatics* 13.1 (2012): 52.

Lee, J. D., Sun, D. L., Sun, Y., & Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, *44*(3), 907-927.

Babu, M. Madan. "Introduction to microarray data analysis." *Computational genomics: Theory and application* 17.6 (2004): 225-49.

Friedman, Jerome, et al. "Lasso and Elastic-Net Regularized Generalized Linear Models. R-package version 2.0-5. 2016." (2016).

Tibshirani, R., et al. "selectiveInference: Tools for Post-Selection Inference." *R package version* 1.3 (2016).

Ding, Ying, et al. "Molecular and genetic characterization of depression: overlap with other psychiatric disorders and aging." *Molecular neuropsychiatry* 1.1 (2015): 1-12.