UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Sequential Multiple Testing for Variable Selection in High Dimensional Linear Model

Permalink

https://escholarship.org/uc/item/9gv015p9

Author

Chen, Hailu

Publication Date

2016

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Sequential Multiple Testing for Variable Selection in High Dimensional Linear Model

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Applied Statistics

by

Hailu Chen

August 2016

Dissertation Committee:

Dr. Xinping Cui, Chairperson
Dr. Subir Ghosh
Dr. Thomas Girke

Copyright by
Hailu Chen
2016

The Dissertation of Hailu Chen is approved:

Committee Chairperson

University of California, Riverside

Acknowledgments

I could not have got this far without help from many different people, and with them I have been had a wonderful graduate time during the five years at University of California, Riverside.

First, I would like to express my sincere gratitude to my graduate advisor, Prof. Xinping Cui, who provided the opportunity, guidance and support in my Ph.D. study. During these five years, she gave me many constructive ideas both for my research and future career plan. Without her help and support, it would not be possible for me to complete my doctoral study successfully. Thank you.

I would like to thank all the members of my dissertation committee, Prof. Subir Ghosh, Prof. Thomas Girke, Prof. Daniel Jeske and Prof. Shizhong Xu. Their review and evaluation of this work as well as their valuable suggestions and generous help during my Ph.D. study are deeply appreciated.

Additionally, I would like to sincerely thank all professors who instructed me during my graduate study at UCR and all the friends in the statistical department.

I greatly acknowledge the funding sources that supported in the Ph.D. study, the Dean's Distinguished Fellowship from the Department of Statistics and the Dissertation Year Fellowship from the Graduate Division.

Also, I would like to thank my parents, Xinxiang Chen and Yuling Ge, my sister Dongxiang Ge, my brother Haitao Chen, and my grandparents for their unconditional support and generosity. I am especially grateful to my husband, Haimao Zhan, for his friendship, encouragement, continuous love and support.

To my family for all the support.

ABSTRACT OF THE DISSERTATION

Sequential Multiple Testing for Variable Selection in High Dimensional Linear Model

by

Hailu Chen

Doctor of Philosophy, Graduate Program in Applied Statistics
University of California, Riverside, August 2016
Dr. Xinping Cui, Chairperson

Covariance test is proposed for testing the significance of the predictor variable that enters the current lasso model along the lasso solution path. In this paper, we propose the sequential multiple testing structure using covariance test p-values, which has good power properties with error rate controlled at a desired level. Specifically, we consider the full underlying hypotheses and the error rate control within each step as well as across all steps along the lasso solution path.

Our sequential multiple hypothesis structure becomes valid because of the asymptotic distribution of the covariance test. And we prove that the minimum composite p-values under null hypothesis get larger along the steps, which is desirable when we apply step-down procedure for error rate control. The benefit of making use of the multivariate structure is, for some scenarios, to increase power while other procedures stop the selection too early. Also, our Hybrid procedures show higher power through simulations for weak signals and for high-dimensional data.

To control FWER, we propose Hybrid Bonferroni-Holm Step-Down procedure

along with Hybrid Hochberg-Holm Step-Down and Hybrid Simes-Holm Step-Down procedures and compare them with StrongStop. To control FDR, we propose Hybrid Bonferroni-Benjamini-Liu Step-Down procedure and compare it with the ForwardStop, StrongStop, TailStop procedures and SLOPE . Simulation studies show that our proposed procedures have higher power with both FWER and FDR controlled at the desired level, especially for large scale and high-dimensional data, and very stable to use as well as in correlated design matrix.

In our work, we first review the variable selection in statistics and most popular variable selection methods. One of the most widely used and well developed method lasso is introduced in Chapter 2. The covariance test proposed for the significance test for lasso will be given in Chapter 3 , as well as its properties for orthogonal matrix X, and the asymptotic distribution. We then propose our sequential hypotheses multiple testing structure built on the lasso covariance test in Chapter 4. In Chapter 5, we develop the Hybrid Bonferroni-Holm Step-Down procedure to control FWER at α, and compare with StrongStop. In Chapter 6, we propose Hybrid Bonferroni-Benjamini-Liu Step-Down method and prove the FDR can be controlled at q. Comparisons are conducted with ForwardStop, StrongStop, TailStop and SLOPE through simulation studies. In Chapter 7, we show two applications of a diabetes data and a framingham heart study data for FDR control and compare it with the other procedures. The conclusion and discussion are given in the last. The proofs for the main theorems and lemmas are included in the appendix.

Contents

List of Figures **x**

List of Tables **xii**

1 Linear Regression **1**
1.1 Introduction 1
1.2 Variable selection 4
1.2.1 Best-subset selection, AIC, BIC, C_p criterion 4
1.2.2 Forward selection, Backward elimination, Stepwise selection, Forward Stagewise, LARS 6
1.2.3 Some common penalized variable selection methods 8
1.3 Selective Inference 14

2 Lasso **20**
2.1 Lasso 20
2.2 Lasso path 23
2.2.1 Orthogonal predictor matrix 24
2.2.2 General predictor matrix 24
2.2.3 Relationship with LARS 26

3 Covariance test **28**
3.1 The covariance test statistic 28
3.1.1 Orthogonal predictor matrix X 30
3.1.2 General predictor matrix X 30
3.1.3 Exact distribution of covariance test statistic 31
3.1.4 Asymptotic distribution of covariance test 32
3.2 Covariance statistics for other penalized methods 37
3.2.1 Covariance test statistic for SCAD 38
3.2.2 Covariance test statistic for MCP 38
3.3 Some new significance tests for variable selection 39
3.3.1 Spacing test for LARS 40
3.3.2 Penalized score test for lasso 41

4 Sequential multiple hypothesis testing **43**
4.1 Multiple testing . . . 43
4.1.1 FWER . . . 44
4.1.2 FDR . . . 49
4.2 Sequential multiple testing . . . 51
4.2.1 Single hypothesis test . . . 53
4.2.2 Multiple hypotheses test . . . 53

5 Hybrid sequential multiple testing to control FWER **56**
5.1 FWER control procedure . . . 57
5.2 Simulation results for FWER . . . 59

6 Hybrid sequential multiple testing to control FDR **65**
6.1 FDR control procedure . . . 66
6.2 Other methods . . . 68
6.2.1 ForwardStop . . . 68
6.2.2 StrongStop . . . 69
6.2.3 TailStop . . . 69
6.2.4 SLOPE . . . 73
6.3 Simulation results for FDR control . . . 76
6.4 Simulation results for FDR control in large-scale and high-dimensional data 79

7 Real data example **98**
7.1 Diabetes study . . . 98
7.2 Framingham heart study: . . . 101

8 Conclusions **105**

Bibliography **107**

A Derive the exact distribution in Chapter 3 **113**

B Proof of Theorems in Chapter 5 **117**

C Proof of Theorems and Lemmas in Chapter 6 **120**

List of Figures

3.1 Simulation results to verify the asymptotic distributions of covariance test statistics for $d > 1$ in general design matrix. 36

5.1 Average power and observed FWER for Hybrid Bonferroni-Holm Step-Down, Hybrid Hochberg-Holm Step-Down, Hybrid Simes-Holm Step-Down and StrongStop procedures for $\alpha \in \{0.05,\ 0.1,\ 0.15,\ 0.2,\ 0.25\}$. 61
5.2 High-dimensional: Average power and observed FWER for the sequential multiple testing at α=0.05. Hybrid Bonferroni-Holm Step-Down, Hybrid Hochberg-Holm Step-Down, Hybrid Simes-Holm Step-Down and StrongStop methods are evaluated. 64

6.1 Density functions of pivotal statistics for TailStop and Hybrid sequential under null distribution at $k_0 = 0, 1, 2, 5$ for $m = 100$. 72
6.2 Average power and observed FDR for Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH, TailStop-BL and SLOPE procedures for $q \in \{0.05,\ 0.1,\ 0.15,\ 0.2,\ 0.25\}$. 77
6.3 Setup 1: Average power and observed FDR for the sequential multiple testing at q=0.01. Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop, TailStop-BH, TailStop-BL and SLOPE procedures are evaluated. 82
6.4 Setup 1: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH, TailStop-BL and SLOPE procedures are evaluated. 83
6.5 Setup 2: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH, TailStop-BL and SLOPE procedures are evaluated. 85
6.6 Setup 3: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH, TailStop-BL and SLOPE procedures are evaluated. 87
6.7 Setup 4: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH, TailStop-BL and SLOPE procedures are evaluated. 89
6.8 Setup corr1: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH and TailStop-BL procedures are evaluated. 92
6.9 Setup corr2: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH and TailStop-BL procedures are evaluated. 93

6.10 Setup corr3: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH and TailStop-BL procedures are evaluated. 94
6.11 Setup corr4: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH and TailStop-BL procedures are evaluated. 95

7.1 Lasso path for Diabetes study. 99
7.2 Histogram of HDL for Framingham study. 102

List of Tables

4.1 Multiple testing notation. 44
4.2 Compare FWER procedures in subspaces for m=3. 48

5.1 Comparison among model a - d for FWER control. 60
5.2 Comparison among model e - h for FWER control. 62

6.1 Comparison among model a - d for FDR control. 76
6.2 Comparison among model e - h for FDR control. 78
6.3 The calculated FDR and average selected number of variables for SLOPE in Setup corr4. 91

7.1 Lasso path for Diabetes study. 98
7.2 Coefficients at each steps for Diabetes study. 100
7.3 Covariance drop in Diabetes study. 101
7.4 Framingham heart study results. 103

Chapter 1

Linear Regression

1.1 Introduction

We consider the usual linear regression model,

$$\mathbf{y} = \mathbf{X}\beta + \epsilon, \quad \epsilon \sim N(0, \sigma^2\mathbf{I}) \tag{1.1}$$

where $\mathbf{y} \in \mathbb{R}^n$ is an outcome vector and $\mathbf{X} \in \mathbb{R}^{n\times m}$ is the matrix of predictor variables, $\beta \in \mathbb{R}^m$ is a vector of unknown coefficients which need to be estimated.

One of the most popular estimation method for linear model is least squares, in which the estimates of the coefficients β are chosen to minimize the residual sum of squares

$$\text{RSS}(\beta) = \sum_{i=1}^{n}(y_i - x_i^T\beta)^2,$$

where x_i records the observations on the m predictor variables of the ith subject. Equivalently we can write it as

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta).$$

RSS is a quadratic function of the parameters. We can differentiate RSS with respect to β to obtain the unique solution

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{1.2}$$

that minimizes RSS. The variance-covariance matrix of the least squares estimates $\hat{\beta}$ is

$$\text{Var}(\hat{\beta}) = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2.$$

RSS is a squared error loss function which is one of the most widely used loss functions in statistics. In general, a squared error loss function can be written as $L(\mathbf{y}, \hat{\mathbf{y}}) = (\mathbf{y} - \hat{\mathbf{y}})^2$, and the expected squared error loss (also called expected generalization error) is

$$\begin{aligned} E(L(y, \hat{y})) &= Var(\epsilon) + Var(\hat{f}(x)) + [E(\hat{f}(x)) - f(x)]^2 \\ &= Var(\epsilon) + MSE. \end{aligned}$$

Here $y = f(x) + \epsilon$ and $\hat{y} = \hat{f}(x)$. $Var(\epsilon)$ is called the irreducible error, which comes from the underlying true model. $Var(\hat{f}(x))$ is the variance of the prediction model, which measures how sensitive the prediction model is to changes in the data set, regardless to the underlying true model. $[E(\hat{f}(x)) - f(x)]^2$ is the bias of the prediction model relative to the underlying true model. Usually, there is bias-variance trade-off, i.e., the increase of variance is accompanied by the decrease of bias, and vice versa. The Gauss-Markov theorem implies that the least squares estimator has the smallest mean squared error (MSE) among all unbiased linear estimators. However, sometimes a biased estimator with a smaller mean squared error might be desirable over the least square method. For example, shrinkage methods, which set some of the least squares coefficients to zero, are preferred when the

prediction accuracy and parsimony of the model are of greater interest. Moreover, when there are more predictors than the number of sample observations, the least square method would not work since the inverse of the matrix $X^T X$ is not unique.

Shrinkage method based regression is also called penalized regression, which chooses the estimates of β that minimize

$$||\mathbf{y} - \mathbf{X}\beta||^2 + \sum_{j=1}^{m} p_\lambda(|\beta_j|). \tag{1.3}$$

The first term in (1.3) is the residual sum of square, and the second term is the penalty function. There are a wide range of penalty functions p_λ. Frank and Friedman (1993) introduced the bridge penalty. As will be shown below, the Ridge penalty (Hoerl and Kennard 1970) and the lasso penalty (Tibshirani 1996) are special cases of bridge penalty,

$$p_\lambda(\beta_j) = \lambda|\beta_j|^\gamma \quad \gamma \neq 0,$$

where $\lambda \geq 0$ is a nonnegative regularization parameter, controlling the level of sparsity in $\hat{\beta}$.

For $\gamma = 0$, it can be thought of no penalty and therefore is equivalent to the least square method.

For $\gamma = 1$, it is lasso penalty and the lasso estimators of β in orthogonal matrix are

$$\hat{\beta}^{lasso} = sign(\hat{\beta}^{ols})(\hat{\beta}^{ols} - \lambda)_+,$$

where $\hat{\beta}^{ols}$ is the ordinary least squares (OLS) estimator of β and $x_+ = \max(x, 0)$, which is called soft thresholding. It is obvious that the lasso estimator is a biased estimator.

In Chapter 2, we will introduce lasso regression, solution path and its properties in more details.

For $\gamma = 2$, it is the Ridge regression. The Ridge estimator related to OLS is

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} = \frac{1}{2\lambda + 1}\hat{\beta}^{ols}.$$

Ridge regression performs proportional shrinkage, the larger the variable, the more it shrinks. Ridge regression was originally proposed by Hoerl and Kennard (1970). When there are many correlated variables in the linear regression model, their coefficients can become poorly determined and exhibit high variance. A wildly large positive coefficient on one variable can be cancelled by a similarly large negative coefficient on its correlated variable. The ridge penalty can alleviate this problem, which makes it popular to deal with multicollinearity.

1.2 Variable selection

In this section, we review a number of least square method based and shrinkage method based variable selection approaches for linear regression model.

1.2.1 Best-subset selection, AIC, BIC, C_p criterion

There are several old-fashioned variable selection criterions which are based on RSS combined with the number of the variables considered in the model.

Best subset selection intents to find the best model with subset of size k, $k \in \{1, \cdots, m\}$, according to a given selection criterion, such as AIC, BIC and Mallow's C_p.

There are a total of $2^m - 1$ potential subsets to choose, so the computational cost of the best subset selection increases exponentially with m.

The Akaike Information Criterion (AIC) (Akaike 1974) for variable selection in linear regression models is defined as

$$AIC = nlog\left(\frac{RSS}{n}\right) + 2k.$$

The best model is the one that gives the smallest AIC over all the possible models considered.

Similarly, the Bayesian Information Criterion (BIC) (Schwarz 1978) is another criterion, which is defined as

$$BIC = nlog\left(\frac{RSS}{n}\right) + \log(n)k.$$

The best model is the one that gives the smallest BIC over all the possible models considered. When $n > e^2 \approx 7.4$, $log(n) > 2$, BIC tends to penalize complex models more heavily, giving preference to simpler models. Larger models will fit better and therefor it has smaller RSS but uses more parameters. Thus the best choice of model is the balance between the model fit and the model size. For model selection, there is no clear preference between AIC and BIC. BIC is asymptotically consistent as a selection criterion.

Mallow's C_p criterion (Mallows 1973) is similar as AIC and BIC,

$$C_p = \frac{RSS_k}{\hat{\sigma}^2} + 2k - n,$$

where $\hat{\sigma}^2$ is the estimation of the variance component from the model with all predictors and RSS_k indicates the RSS from a model with k parameters. For the Gaussian distribution, the AIC criterion is equivalent to C_p. We choose the best model with the lowest C_p values.

1.2.2 Forward selection, Backward elimination, Stepwise selection, Forward Stagewise, LARS

In this section, we review some sequential selection procedures including forward selection, backward selection, stepwise selection, forward stagewise and LARS.

In forward selection, one starts with no variables in the model. At each step, the most significant variable is added if its p-value is less than some pre-set level. This procedure is repeated until all the remaining variables are statistically non-significant. This procedure is also called "p-to-enter".

In backward elimination, one starts with all candidate variables in the model. At each step, the least significant variable is dropped if its p-value is larger than some pre-set level. This procedure is repeated until all remaining variables are statistically significant. This procedure is also called "p-to-remove".

The stepwise regression is a combination of backward elimination and forward selection. This avoids the situation where some variables are added or removed early in the process and one finds them important later. At each stage a variable may be added or removed.

There are some drawbacks of these procedures. First, because of the "one-at-a-time" nature of adding or dropping variables, it is possible to miss the optimal model. Second, stepwise selection tends to pick models that are smaller than optimal for prediction purpose. In addition, backward elimination and stepwise selection are not working for high dimensional data since all the p-values or F test are based on the least square method.

The forward stagewise regression starts like forward regression, with an intercept equal to $\bar{y}$, and centered predictors with coefficients initially all 0. At each step the algorithm identifies the variable most correlated with the current residual. It then computes the simple linear regression coefficient of the residual on this chosen variable, and then adds it to the current coefficient for that variable. This is continued till none of the variables have correlation with the residuals.

Least Angle Regression (LARS) (Efron, Hastie and Johnstone 2004) uses the idea of the forward Stagewise procedure to accelerate the computation. In the LARS procedure, one starts with all coefficients equal to zero, and finds the predictor most correlated with the response, say x_1. One then takes the largest step possible in the direction of this predictor until some other predictor, say x_2, has as much correlation with the current residual. Instead of continuing along x_1, LARS proceeds in a direction equiangular between the two predictors until a third variable x_3 earns its way into the "most correlated" set. LARS then proceeds equiangularly among x_1, x_2 and x_3, that is, along the "least angle direction" until a fourth variable enters, and so on. This is also where the name came from.

With any variable selection methods, it is important to keep in mind that model selection cannot be deviate from the underlying purpose of the investigation. The variables been dropped can still be correlated with the response. It would be wrong to say these variables are unrelated to the response, while it's just that they provide no additional explanatory effect beyond those variables already included in the model.

1.2.3 Some common penalized variable selection methods

Fan and Li (2001) suggested that any penalty function p_λ should result in an estimator $\hat{\beta}$ that fulfills the following properties (Oracle property):

Unbiasedness: penalized least squares and similar methods introduce bias into the estimator of β in the hope of reducing variance. We would want that any p_λ we choose will result in an estimator that is approximately unbiased for large components of β.

Sparsity: we have prior reasons to believe that some of the components of β should be 0. We want to choose the p_λ which results in a method that sets some components of β to 0.

Continuity: we would like our predictors to be continuous to prevent instabilities in prediction.

Next, some popular penalized variable selection methods will be introduced.

SCAD

Most of the bridge penalties $p_{\lambda(|\beta|)} = \lambda \sum_j |\beta_j|^\gamma$ don't satisfy all the three conditions (sparsity, unbiasedness and continuity) for the estimation. As a result, Fan and Li (2001) suggested the Smoothly Clipped Absolute Deviation Penalty (SCAD), which does fulfill all these three conditions. SCAD was defined by

$$p_{\lambda,\gamma}(\beta) = \begin{cases} \lambda\beta, & \text{if} \quad \beta \leq \lambda \\ \frac{\gamma\lambda\beta - 0.5(\beta^2+\lambda^2)}{\gamma-1}, & \text{if} \quad \lambda < \beta \leq \gamma\lambda \\ \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)}, & \text{if} \quad \beta > \gamma\lambda \end{cases}$$

$$p'_{\lambda,\gamma}(\beta) = \begin{cases} \lambda, & \text{if} \quad \beta \leq \lambda \\ \frac{\gamma\lambda - \beta}{\gamma - 1}, & \text{if} \quad \lambda < \beta \leq \gamma\lambda \\ 0, & \text{if} \quad \beta > \gamma\lambda \end{cases}$$

for $\lambda \geq 0$ and $\gamma > 2$. The penalty begins by applying the same rate of penalization as the lasso, and reduces that rate to 0 as β gets further away from zero;

MCP

Zhang (2010) proposed the Minimax Concave Penalty (MCP), which also has the oracle properties. MCP was defined as

$$p_{\lambda,\gamma}(\beta) = \begin{cases} \lambda\beta - \frac{\beta^2}{2\gamma}, & \text{if} \quad \beta \leq \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & \text{if} \quad \beta > \gamma\lambda \end{cases}$$

$$p'_{\lambda,\gamma}(\beta) = \begin{cases} \lambda - \frac{\beta}{\gamma}, & \text{if} \quad \beta \leq \gamma\lambda \\ 0, & \text{if} \quad \beta > \gamma\lambda \end{cases}$$

for $\lambda \geq 0$ and $\gamma > 1$. The rationale behind the penalty can be understood by considering its derivative: MCP begins by applying the same rate of penalization as the lasso, but continuously relaxes that penalization until, when $\beta > \gamma\lambda$, the rate of penalization drops to 0. The solution of the MCP is employed by the coordinate descent algorithm to obtain the coordinate-wise minimizer of the objective function.

Adaptive lasso

Lasso cannot be an oracle procedure, since the lasso estimates of the regression coefficients are biased due to the shrinkage imposed by the l_1 penalty. To increase the accuracy of the estimation of large signals and eliminate some false discoveries, Zou (2006) proposed the Adaptive lasso, considering the weighted penalty:

$$\hat{\beta} = \arg\min ||\mathbf{y} - \sum_{j=1}^{m} \mathbf{x}_j \beta_j||^2 + \lambda \sum_{j=1}^{m} w_j |\beta_j|,$$

where $\mathbf{w}$ is a known weight vector. They have shown that the weights are data-dependent and wisely chosen, then the weighted lasso can have all the three properties.

In these procedures the smoothing parameter $\lambda_1, \cdots, \lambda_m$ are adjusted to the unknown signal magnitudes based on some estimates of regression coefficients, perhaps obtained through previous iterations of lasso. The idea is then to consider a weighted penalty $\sum_i w_i |\beta_i|$, where w_i is inversely proportional to the estimated magnitudes so that large regression coefficients are shrunk less than smaller ones. In some circumstances, such adaptive versions of lasso outperform the regular lasso for selection.

Elastic Net

The elastic net was proposed by Zou and Hastie (2005) to solve two problems. The first is because the lasso doesn't have a better performance when the dimension of the variables m is much greater than n. The Ridge regression is able to select larger models but does not perform variable selection. Secondly, if a model contains a number of correlated variables, the lasso is most likely to choose one of the correlated variables. In

many applications, this is desirable. For example, in ultra-high dimensional model, it is likely that every significant variable will have a number of 'noise' variables correlated to it. In that case, we prefer to only select the most significant variable. The Ridge tends to pick the correlated variables together in a model.

Therefore, the Elastic Net uses a linear combination of the lasso and the Ridge

$$p_\lambda(|\beta|) = \lambda_1 \sum_{j=1}^{m} |\beta_j| + \lambda_2 \sum_{j=1}^{m} \beta_j^2.$$

This also can be rewritten as

$$p_\lambda(|\beta|) = \sum_{j=1}^{m} (\alpha|\beta_j| + (1-\alpha)\beta_j^2).$$

We can consider this as a compromise between lasso and Ridge regression. The second term encourages highly correlated features to be averaged, while the first term encourages a sparse solution in the coefficients of these averaged features. The Elastic Net can yield more than n non-zero coefficients when $m > n$, a potential advantage over the lasso. The Elastic Net penalty can be used with any linear model, in particular for regression or classification.

SLOPE

Sorted L-One Penalized Estimation (SLOPE) proposed by Bogdan et al. (2013) is to use the sorted l_1 norm for variable selection and the estimation of SLOPE is the solution of

$$min_\beta \frac{1}{2}||y - X\beta||_{l_2}^2 + \lambda_1|\beta|_{(1)} + \lambda_2|\beta|_{(2)} + \cdots + \lambda_m|\beta|_{(m)}.$$

Here, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ and $|\beta|_1 \geq |\beta|_2 \geq \cdots \geq |\beta|_m$ are ordered by the magnitudes of λ and β. In short, the regularizer is a sorted l_1 norm which penalizes the regression coefficients according to their rank: the higher the rank-the closer to the top, the larger the penalty.

They proposed the proximal gradient algorithms and fast prox algorithm to compute the estimation of SLOPE.

Similarly as adaptive lasso, SLOPE uses the adaptive penalty for different signals to eliminate the false discoveries. However, the idea behind SLOPE for the penalty is entirely different from adaptive lasso. In the adaptive lasso, the penalty tends to decrease as the magnitude of coefficients increases, while SLOPE is the opposite. SLOPE is to adapt to the unknown signal sparsity while controlling the FDR. In practice, since the SLOPE sequence $\lambda_1 \geq \cdots \geq \lambda_m$ leading to FDR control is typically rather large, they don't recommend using SLOPE directly for the estimation of regression coefficients. Instead, SLOPE uses the two-stage procedure: in the first step, SLOPE is used to identify significant predictors; in the second step, the corresponding regression coefficients are estimated using the least squares method within the identified sparse regression model.

Tuning parameters

Every penalized regression considered above involves a regularization parameter λ (sequence of λ_i for SLOPE). This parameter regulates the severity of the penalty p_λ and also it indicates how far we are along the bias-variance spectrum. Small values of λ reduce the penalty and therefore allow more complex models, which reduces the bias but increases the variance; Large values of λ increase the penalty and therefore allow a

more parsimonious models, which increases the bias but reduces the variance. There are some common approaches to choose the general parameter λ, such as cross validation and generalized cross validation.

Cross validation (Stone 1974) solves the problem by splitting the data into K separate segments, by fitting the model using $K-1$ of these segments and assessing the goodness of the fit using the last segment. We use the cross validation score as an estimate to the expected generalization error. The cross validation score is defined as:

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}^{-\kappa(i)}(X_i)),$$

where $L(\cdot)$ is the loss function and $\hat{f}^{\kappa(i)}$ is the linear model fitted to the set without the segment containing data point i.

To fit our model using cross validation, we will select the model with the lowest cross validation score. There are several options to divide the data set, such as K fold, leave-one-out, leave-p-out cross validation. The computational cost could be expensive, especially when K is large. However, it doesn't cost lots of computation for linear model. For the fitted $\hat{\mathbf{Y}} = H\mathbf{y}$, it can be shown (Hastie and Tibshirani 1990) that the cross validation score takes the form

$$CV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{f}(X_i)}{1-H_{ii}}\right)^2.$$

The generalized cross validation score is defined as:

$$GCV(\hat{f}) = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{f}(X_i)}{1-trace(H)/N}\right)^2.$$

The generalized cross validation has computational advantage compared to cross validation since the trace of H is easier to calculate than the diagonal element.

The most common method to choose tuning parameter is cross validation, and it tends to obtain the prediction model with higher power but offers no FDR control. Cross validation chooses a λ that minimizes an estimate of prediction error: this would increase if irrelevant variables are selected, but also if important ones are omitted, or if their coefficients are excessively shrunk. As a result, the λ_{CV} is quite smaller than a penalization parameter chosen with FDR control in mind.

1.3 Selective Inference

Selective inference, mentioned by Taylor and Tibshirani (2015), addresses the following challenges: having a set of data to find potential associations, how do we properly assess the strength of these associations. This challenge becomes more important in the era of big data and complex statistical modeling. There are increasing number of references trying to address this problem in most recent years.

The selection process is based on variable selection, and the methods of selecting potential features include forward selection, stepwise selection, marginal selection, lasso, ridge regression, and other penalized regression. The statistical investigation can be thought of as two stages (Fithian, Sun and Taylor 2015):

1. Selection stage: the analyst chooses a probabilistic model for the data, and formulates testing, estimation, or other problems in terms of unknown aspects of the model;

2. Inference stage: The analyst attempts the chosen problems using the data and the selected model.

The inference, confidence interval and hypothesis testing, made after selection process is referred as post-selection inference (Berk et al. 2013, Lee et al. 2015, Lee and Taylor 2014, Fithian, Sun and Taylor 2015, Tibshirani et al. 2015, Reid, Taylor and Tibshirani 2015 and etc.).

For penalized regression, such as lasso, the estimator of lasso is biased and is based on sparse model. Some of the articles were aiming at de-biasing or de-sparsing the estimator and then making statistical inference. Van De Geer et al. (2014) proposed a general method for constructing confidence intervals and statistical tests by using desparsifying lasso, and applied lasso with model-wise regression. They also extended their methods into generalized linear model. Javamard and Monranari (2013) proposed an efficient algorithm for constructing confidence intervals and p-values. The confidence intervals had nearly optimal size and the hypothesis testing had nearly optimal power based on constructing a 'de-biased' version of regularized M-estimators. Furthermore, Javamard and Monranari (2014) developed an approach based on debiasing of the lasso estimator, and developed a de-biased testing procedure of lasso for type I error control. For Gaussian random design models, they developed a general upper bound on the minimax power of the test with a given significance level. Lee et al. (2015) proposed the framework for post selection inference that conditions on which model was selected. They provided post-selection intervals for regression coefficients for False Coverage Rate (FCR) and positive False Coverage Rate (pFCR). Reid, Taylor and Tibshirani (2015) adapted post selection inference framework by

Lee et al. (2015) for the Lasso under the orthogonal setting, where sample elements have different underlying signal sizes. Ahrens and Bhattacharjee (2015) considered the two-step estimation strategy for spatial autoregressive panel model, and derived the convergence rates for the two-step lasso estimator. Dezeure et al. (2015) reviewed recent high dimensional inference methods for constructing p-values and confidence intervals in both linear and generalized linear models by implementing sample-splitting, de-sparsifying lasso, and ridge projection and bias correction. Su, Bogdan and Candes (2015) demonstrated the false discoveries occur early on the lasso path no matter how strong the signals are, and derived an asymptotic trade-off between false and true positive rates along the lasso path.

For forward stepwise model selection, Loftus and Taylor (2014) described a truncated χ test conditional on the outcome of the model selection procedure. The truncated χ test statistic followed unif (0,1) distribution under the global null. The spacing test proposed by Tibshirani et al. (2015) for forward selection, LARS and lasso, and the spacing test is asymptotically equivalent to the covariance test(Lockhart et al. 2014). Lee and Taylor (2014) proposed a framework for post model selection inference for marginal screening in linear regression. The marginal screening is to choose the largest k components marginally and fit them with the least square method. The confidence interval and hypothesis test were made conditional on selection. Voorman, Shojaie and Witten (2014) applied score test based on penalized regression. To test one individual parameter while treating others as nuisance parameters, the score test was applied in lasso and ridge regression. Bühlmann (2013) applied a ridge estimator correction to construct p-values for individual and general hypotheses in a high-dimensional linear model with fixed design and Gaussian errors.

There are other techniques used for statistical selection inference. Zhang and Zhang (2014) proposed a low dimensional projection (LDP) approach to construct confidence intervals for regression coefficients. They proved the asymptotic unbiasedness and normality of finite dimensional distribution of the estimators and the consistency of their estimated covariance. The methods were applied to make inference about a preconceived low dimensional parameters and also applied to all regression coefficients via simultaneous interval estimation. Taylor, Loftus and Tibshirani (2014) derived the exact p-value for testing a global null hypothesis in a general adaptive regression setting via the Kac-Rice formula assuming Gaussian errors, and their approach can be used for lasso, group lasso, principal components and matrix completion problems. Bogdan et al. (2013) introduced a new sorted l_1 norm penalty, called SLOPE, and provided efficient algorithm for variable selection. Bogdan et al. (2014) and Su and Candes (2015) improved algorithm to choose the λ_i sequence and control FDR and extend the result to designs with a wide range of sparsity.

For correlated covariates in the high dimensional model, Mandozzi et al. (2015) borrowed the idea of hierarchical cluster tree and proceeded the process in a sequential manner until a cluster is not significant. Ning and Liu (2015), on the other hand, used a decorrelated score function, which is uncorrelated with the nuisance score function, to construct both hypothesis tests and confidence regions.

Most of the selection inferences are made conditional on the selected models. But these methods of inference ignore the fact that the inferences are preceded by selected

processes that produce models that are themselves stochastic. If we choose the wrong model the inference will break down. Berk et al. (2013) proposed the structure of the simultaneous inference and give some asymptotic results, which does not depend on the truth of the selected model and produces valid inference even with the wrong model.

We will focus our work on error rate control for variable selection. So far it is extremely difficult for researchers to find a way to estimate the standard error of the estimators for shrinkage variable selection (Tibshirani 1996, Osborne et al. 2000, Pötscher and Leeb 2009 and etc.). A variety of ways have been developed to control the Type I error rate. Some methods combined the lasso with splitting or resampling to bypass estimating the standard error directly, and some used specifically developed selection penalties to perform multiple testing under Familywise Error Rate (FWER) or False Discovery Rate (FDR) control. Wasserman and Roeder (2009) splitted the data into two parts. The number of variables was reduced to a manageable size by applying a variable selection procedure on the first split, such as lasso, and the traditional hypothesis testing was then applied to the second split. However, an onetime random split of the data was sensitive and irreproducible. Meinshausen et al. (2009) extended the inference through multiple random splits and used quantile functions to obtain robust cutoff values. On the other hand, Meinshausen and Bühlmann (2010) introduced stability selection, which was based on subsampling, and provided finite sample control for some error rates of false discoveries. Shah and Samworth (2013) improved the stability selection by introducing complementary pairs stability selection. Other methods used two-step procedures, by using penalty function to choose important variables first and then fitting them with least square to control FDR

or FWER. But sometimes it may not be appropriate to split the sample or the sample size is not large enough. Under those circumstances, the accuracy and power will be compromised.

To control the error rate for variable selection, G'Sell et al. (2015) considered a multiple hypothesis testing where the hypotheses are ordered. The rejection rule is to choose the stopping point k, and they proposed two testing procedures, ForwardStop and StrongStop, to control the false discovery rate in the ordered setting. However, they haven't fully used the joint asymptotic distribution of covariance test and suffer power loss in some cases. As an extension of StrongStop, TailStop incorporats lasso covariance test and controls the FDR nearly exactly.

Our research work is motivated by multiple testing for lasso variable selection. Therefore, in the next chapter, we will introduce lasso regression more comprehensively, and our future work will mainly focus on error rate control for lasso regression.

Chapter 2

Lasso

2.1 Lasso

The lasso estimator is defined as

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^m} \frac{1}{2}||\mathbf{y} - \mathbf{X}\beta||_2^2 + \lambda||\beta||_1 \tag{2.1}$$

where $\lambda \geq 0$ is a nonnegative regularization parameter, controlling the level of sparsity in $\hat{\beta}$. The second term in (2.1) is l_1 penalty, which is crucial for lasso. Setting $\lambda = 0$ reverses the lasso problem to OLS which minimizes the unregularized empirical loss. The l_1 penalization approach is also called basis pursuit in signal processing (Chen, Donoho, and Saunders 2001).

The lasso continuously shrinks the coefficients towards 0 as λ increases and some coefficients are shrunk to exact 0 if λ is sufficiently large. Donoho et. al. (1995) proved the near-minimax optimality of soft thresholding (the lasso shrinkage with orthogonal predictors). It also has been shown that the l_1 penalty is able to discover the "right" sparse

representation of the model under certain conditions (Donoho and Huo 2002, Donoho and Elad 2002, Donoho 2004).

It is well known that the estimator in (2.1) fulfills the Karush-Kuhn-Tucker (KKT) condition:

$$-\mathbf{X}^T(Y - \mathbf{X}\hat{\beta})/n + \lambda\hat{\kappa} = 0$$

$$||\hat{\kappa}||_\infty \leq 1 \text{ and } \hat{\kappa}_j = \text{sign}(\hat{\beta}_j) \quad \text{if } \hat{\beta}_j \neq 0$$

Under some regularity conditions on the design, Knight and Fu (2006) have shown estimation consistency of lasso for fixed p and fixed β as $n \to \infty$. In particular, they have shown that $\hat{\beta}(\lambda_n) \to \beta$ in probability and asymptotic normality of the estimates provided that $\lambda_n = o(n)$. In addition, it is shown that for $\lambda_n \propto n^{1/2}$ (on the same order of $n^{1/2}$), and as $n \to \infty$ there is a non-vanishing positive probability for lasso to select the true model.

On the model selection consistency front, Meinshausen and Bühlmann (2006) have shown that under a set of conditions, lasso is consistent in estimating the dependency between Gaussian variables even when the number of variables m grows faster than n. Addressing a slightly different but closely related problem, Leng et al. (2004) have shown that for a fixed m and orthogonal designs, the lasso estimate that is optimal in terms of parameter estimation does not give consistent model selection.

Zhao and Yu (2006) formalized the neighborhood stability condition in the context of linear regression models as a strong irrepresentable condition. They showed that under this crucial condition and certain other regularity conditions, lasso is consistent for variable selection, even when the number of variables m is as large as $\exp(n^a)$ for some $0 < a < 1$.

Thus, their results are applicable to high-dimensional regression problems, provided that the conditions, in particular the strong irrepresentable condition, are reasonable for the data. Under certain conditions, the sparsity pattern of the lasso coefficient $\hat{\beta}$ is able to identify the nonzero elements.

Also pointed out by Tibshirani (2013), when the columns of $\mathbf{X}$ are in general position, there is a unique lasso solution at each $\lambda > 0$, and hence a unique path. Points $\mathbf{x}_1, \cdots, \mathbf{x}_m \in \mathbb{R}^n$ are said to be in general position provided that no k dimensional affine subspace $L \subseteq \mathbb{R}^n$, $k < \min\{n, m\}$, contains more than $k + 1$ elements of $\{\pm\mathbf{x}_1, \cdots, \pm\mathbf{x}_m\}$, excluding antipodal pairs. The assumption that $\mathbf{X}$ has columns in general position is a very weak one. For example, if the entries of $\mathbf{X}$ are drawn from a continuous probability distribution on $\mathbb{R}^{nm}$, then the column of $\mathbf{X}$ are almost surely in general position, and this is true regardless of the sizes of n and m.

The lasso method using convex surrogates for sparsity, such as the l_1 or the sorted l_1 norms, can have major difficulties in controlling the FDR (Bogdan et al. 2013, 2015). This phenomenon has to do with the shrinkage of regression coefficients.

We consider the lasso for simplicity and assume that the columns of X have unit norm. The optimality conditions for the lasso solution $\hat{\beta}$ take the form

$$\hat{\beta} = \eta_\lambda(\hat{\beta} - X'(X\hat{\beta} - y)),$$

where η_λ is the soft-thresholding operator, $\eta_\lambda(t) = \text{sign}(t)(|t| - \lambda)_+$, applied componentwise. Set

$$v_i =< X_i, \sum_{j \neq i} X_j(\beta_j - \hat{\beta}_j) >,$$

we can rewrite the optimality conditions as

$$\hat{\beta}_i = \eta_\lambda(\beta_i + X_i'z + v_i).$$

Observe that conditional on X, $X_i'z \sim N(0,1)$. In an orthogonal design, $v_i = 0$ for all i, and taking λ at the appropriate level controls the FDR and even the FWER. Imagine now that X has correlated columns and that k regression coefficients are large. Then the lasso estimates of these coefficients will be shrunk towards zero by an amount roughly proportional to λ. The effect of this is that v_i now looks like a noise term whose size is roughly proportional to λ times the square root of k. In other words, there is an inflation of the 'noise variance' so that we should not be thresholding at λ but at a higher value. The consequence is that the lasso procedure is too liberal and selects too many variables for which $\beta_i = 0$; hence, the FDR becomes just too large.

2.2 Lasso path

The lasso path $\hat{\beta}(\lambda)$ is a continuous and piecewise linear function of λ, with knots at values $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$ (those knots are dependent on $\mathbf{y}, \mathbf{X}$). At $\lambda = \infty$, the solution has no active variables (i.e. all variables have zero coefficients). As λ decreases along the lasso path, each knot λ_k indicates the entry and removal of a variable from the current active set (i.e. its coefficient becomes nonzero and zero respectively).

2.2.1 Orthogonal predictor matrix

We examine the special case of an orthogonal predictor matrix $\mathbf{X}$, that is, one that satisfies $\mathbf{X}^T\mathbf{X} = \mathbf{I}$. The result can be seen as special cases of the general $\mathbf{X}$.

The lasso solution (2.1) at any given value of λ has closed-form

$$\hat{\beta}_j(\lambda) = S_\lambda(X_j^T y), \quad j = 1, \cdots, m,$$

where $X_1, \cdots, X_m$ are columns of X, and S_λ is the soft-thresholding function

$$S_\lambda(x) = \begin{cases} x - \lambda, & \text{if } x > \lambda \\ 0, & \text{if } -\lambda \leq x \leq \lambda \\ x + \lambda, & \text{if } x < -\lambda \end{cases}$$

Letting $U_j = X_j^T y,\ j = 1, \cdots, m$, the knots in the lasso path are simply the values of λ at which the coefficients become nonzero. In linear regression, U_j is the solution of $\hat{\beta}_j$

$$\lambda_1 = |U_{(1)}|, \lambda_2 = |U_{(2)}|, \cdots, \lambda_m = |U_{(m)}|, \tag{2.2}$$

where $|U_{(1)}| \geq |U_{(2)}| \geq \cdots \geq |U_{(m)}|$ are the order statistics of $|U_1|, \cdots, |U_m|$.

For an orthogonal design matrix which satisfies the general position requirement, the lasso solution path will be unique and is the same as LARS procedure.

2.2.2 General predictor matrix

In this part, we consider a general predictor matrix $\mathbf{X}$, with columns normalized. If A denotes the active set and s_A denotes the signs of active coefficients at knot λ_k,

$$A = \text{supp}(\hat{\beta}(\lambda)), \qquad s_A = \text{sign}(\hat{\beta}_A(\lambda_k)).$$

And the knots λ_{k+1} in the lasso path is given by

$$\lambda_{k+1} = \max\{\lambda_{k+1}^{\text{join}}, \lambda_{k+1}^{\text{leave}}\}, \tag{2.3}$$

where $\lambda_{k+1}^{\text{join}}$ is the value at which a variable would join the active set A and $\lambda_{k+1}^{\text{leave}}$ is where a variable would leave from the current trajectory for the lasso coefficients at λ_k. These values are

$$\lambda_{k+1}^{\text{join}} = \max_{j \notin A, s \in \{-1,1\}} \frac{X_j^T(I - P_X)y}{s - X_j^T(X_A^T)^+ s_A} \cdot \mathbf{1}\left\{\frac{X_j^T(I - P_X)y}{s - X_j^T(X_A^T)^+ s_A} < \lambda_k\right\}, \tag{2.4}$$

where $P_A = X_A(X_A^T X_A)^{-1} X_A^T$ and $(X_A^T)^+ = X_A(X_A^T X_A)^{-1}$; and

$$\lambda_{k+1}^{\text{leave}} = \max_{j \in A} \frac{[(X_A)^+ y]_j}{[(X_A^T X_A)^{-1} s_A]_j} \cdot \mathbf{1}\left\{\frac{[(X_A)^+ y]_j}{[(X_A^T X_A)^{-1} s_A]_j} < \lambda_k\right\}. \tag{2.5}$$

Therefore, the active set and the signs of the active coefficients remain constant in between knots. At any point in the path, the corresponding active set $A = \text{supp}(\hat{\beta}(\lambda))$ of the lasso solution indexes a linearly independent set of predictor variables, that is, rank $R(X_A) = |A|$, where we use X_A to denote the columns of X in A.

For a matrix X satisfying the positive cone condition (a restrictive condition that covers, e.g., orthogonal matrices), there are no variables removed from the active set as λ decreases, and then the number of knots is m.

2.2.3 Relationship with LARS

Lasso is related to LARS procedure. Under some lasso modification and assuming the "one at a time" condition, the LARS algorithm yields all Lasso solutions. Therefore, we can use the LARS algorithm to produce lasso path. Generally, assume equicorrelation set A and equicorrelation signs s. For the path solution at λ_k, the non-equicorrelation coefficients equal to zero, $\hat{\beta}_{-A}^{LARS}(\lambda_k) = 0$, and the equicorrelation coefficients as

$$\hat{\beta}_{A}^{LARS}(\lambda_k) = (X_A)^{+}(y - (X_A^T)^{+}\lambda_k s). \tag{2.6}$$

The LARS algorithm for the lasso path is given by

Algorithm 1 *Given y and $\mathbf{X}$. Start with the iteration counter $k = 0$, regularization parameter $\lambda_0 = \infty$, equicorrelation set $A = \emptyset$ and equicorrelation signs $s = \emptyset$.*

While $\lambda_k > 0$:

1. Compute the LARS lasso solution at λ_k by least squares, as in (2.6). Continue in a linear direction from the solution for $\lambda \leq \lambda_k$.

2. Compute the next joining time λ_{k+1}^{join}, when a variable outside the equicorrelation set achieves the maximal absolute inner product with the residual, as in (2.4).

3. Compute the next crossing time λ_{k+1}^{leave}, when the coefficient path of an equicorrelation variable crosses through zero, as in (2.5).

4. Set $\lambda_{k+1} = \max\{\lambda_{k+1}^{join}, \lambda_{k+1}^{leave}\}$. If $\lambda_{k+1}^{join} > \lambda_{k+1}^{leave}$, then add the joining variable to A and its sign to s; otherwise, remove the crossing variable from A and its sign from s. Update $k = k + 1$.

Tibshirani (2013) showed that computed path satisfies the KKT conditions at each λ in this algorithm, and is hence indeed a lasso solution path. It also shows that the computed path is continuous at each knot in the path λ_k, and hence is globally continuous in λ.

The lasso estimator and lasso path can be generated properly as stated above to choose the significant variables. However, as mentioned by Bogdan et al. (2013, 2015), lasso can not control FDR because of the inflation of the 'noise variance' and too many variables selected. To use lasso as a tool for variable selection with error rate control, we need to bypass the traditional way and apply new ideas for the significant test.

Recent work illustrated the distribution or asymptotic distribution of the test statistics developed for lasso and lasso related variable selection methods. Lockhart (2014) proposed a significance test based on the covariance test statistic for lasso model. They have shown that the proposed covariance test statistics have asymptotic exponential distributions under the null hypothesis. The covariance test statistic measures the covariance drop for two adjacent variables in the lasso path and hold good properties for significance test. We will illustrate it in more details in the next chapter.

Chapter 3

Covariance test

Lockhart et al. (2014) proposed a significance test for lasso, which was based on the lasso solution path, and proved the asymptotic exponential null distribution of the proposed covariance test statistic. These results were derived under lasso critical assumptions such as normality, sure screening or model selection consistency of lasso. We will introduce the covariance test statistic and its asymptotic distribution here.

3.1 The covariance test statistic

The covariance test statistic is proposed to test the significance of the variable that enters the active set (i.e. the variables have nonzero coefficients) along the lasso path. Therefore, the test is defined at a given step (kth step) along the solution path.

First, let A be the active set just before λ_k and suppose the predictor j enters at λ_k. Then denote $\hat{\beta}(\lambda_{k+1})$ as the solution at the next knot λ_{k+1} in the path , using predictors $A \cup \{j\}$. Finally, let $\hat{\beta}_{\mathcal{A}}(\lambda_{k+1})$ be the solution at $\lambda = \lambda_{k+1}$ using only the active predictors

$X_{\mathcal{A}}$. To be explicit,

$$\hat{\beta}_A(\lambda_{k+1}) = \arg\min_{\beta_A \in \mathbb{R}^{|A|}} \frac{1}{2}||y - X_A\beta_A||_2^2 + \lambda_{k+1}||\beta_A||_1.$$

The covariance test statistic is defined as

$$T_k = (< \mathbf{y}, \mathbf{X}\hat{\beta}(\lambda_{k+1}) > - < \mathbf{y}, \mathbf{X}_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}(\lambda_{k+1}) >)/\sigma^2, \tag{3.1}$$

where

$$\hat{\beta}(\lambda_{k+1}) = (\mathbf{X}^T_{\mathcal{A}\cup\{j\}}\mathbf{X}_{\mathcal{A}\cup\{j\}})^{-1}\mathbf{X}^T_{\mathcal{A}\cup\{j\}}\mathbf{y} - \lambda_{k+1}(\mathbf{X}^T_{\mathcal{A}\cup\{j\}}\mathbf{X}_{\mathcal{A}\cup\{j\}})^{-1}s_{\mathcal{A}\cup\{j\}},$$

$$\hat{\beta}_A(\lambda_{k+1}) = (\mathbf{X}^T_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}^T_{\mathcal{A}}\mathbf{y} - \lambda_{k+1}(\mathbf{X}^T_{\mathcal{A}}\mathbf{X}_{\mathcal{A}})^{-1}s_{\mathcal{A}}.$$

The covariance statistic uses the difference of two inner products with y, which can be roughly thought of as an (uncentered) covariance, hence, the name of the test. The larger the covariance of y with $X\hat{\beta}$ compared to that with $X_{\mathcal{A}}\hat{\beta}_{\mathcal{A}}$, the more important the role of variable j in the proposed model $\mathcal{A} \cup \{j\}$.

The test statistic can be simplified as $T_k = C(\mathcal{A}, s_{\mathcal{A}}, j, s)\lambda_k(\lambda_k - \lambda_{k+1})$, where $C(\mathcal{A}, s_{\mathcal{A}}, j, s) = ||(X_{\mathcal{A}\cup\{j\}})^+ - (X^T_{\mathcal{A}})^+ s_{\mathcal{A}}||_2^2$, and s is the sign upon entry (sign of variable j upon entry). The assumption is that the signs of s_A at λ_k is the same as s_A at λ_{k+1}, which means that no variables are deleted from the active set in the solution path $\tilde{\beta}_A(\lambda)$ for $\lambda \in [\lambda_{k+1}, \lambda_k]$. When X satisfies the positive cone condition (which includes X orthogonal), no variables ever leave the active set, and then the assumption holds. For general case, both the lasso and least squares on X_A recover the signs of the true coefficients.

3.1.1 Orthogonal predictor matrix X

When the design predictor matrix X is orthogonal, which satisfies $X^TX = I$. In fact, for orthogonal X, it is straightforward to check that $C(\mathcal{A}, s_{\mathcal{A}}, j, s) = 1$, therefore, the covariance test statistic for testing the entry of the variable at step k is

$$T_k = \lambda_k(\lambda_k - \lambda_{k+1})/\sigma^2. \tag{3.2}$$

When $k = 1$, $T_1 = \lambda_1(\lambda_1 - \lambda_2)/\sigma^2$, and it tests the first variable entering the model.

3.1.2 General predictor matrix X

For general predictor matrix X with columns in general position, the proposed covariance test statistic is related with the knots $\lambda_1 \geq \cdots \geq \lambda_m$ in the lasso path. The knots of λ_k is the maximum of the knots joining the model or leaving the model: $\lambda_{k+1} = \max\{\lambda_{k+1}^{\text{join}}, \lambda_{k+1}^{\text{leave}}\}$, see in (2.3)-(2.5) for details. The covariance test statistics at step k is

$$T_k = C(\mathcal{A}, s_{\mathcal{A}}, j, s)\lambda_k(\lambda_k - \lambda_{k+1}),$$

where $C(\mathcal{A}, s_{\mathcal{A}}, j, s) = ||(X_{\mathcal{A}\cup\{j\}})^+ - (X_{\mathcal{A}}^T)^+ s_{\mathcal{A}}||_2^2$. The test statistic is to test the entry of the kth variable in the lasso path.

There is an interesting connection between the covariance test statistic and the degrees of freedom of a fitting procedure. The degrees of freedom is defined by Efron (1986) as

$$\text{df}(\hat{y}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}\text{Cov}(y_i, \hat{y}_i).$$

For the lasso regression, if a model containing k nonzero coefficients, the degrees of freedom

of the lasso fit is equal to k. For details, please refer Efron (2004), Zou, Hastie and Tibshirani (2007).

Next the exact and asymptotic distribution of the covariance test statistic for case $k = 1$ and $k > 1$ and for both orthogonal and general design matrix X will be presented.

3.1.3 Exact distribution of covariance test statistic

As we see from the covariance test statistics, it is complicated to get the exact distribution of T_k and there is no closed form for the distribution which will be illustrated next. We will derive the exact distribution of the first two statistics for covariance test for orthogonal matrix here. In orthogonal design, $T_k = \lambda_k(\lambda_k - \lambda_{k+1})$. According to equation (2.2), let $V_1 \geq V_2 \geq \cdots \geq V_m$ be the order statistics of an independent sample of χ_1 variates. Let $T_k = V_k(V_k - V_{k+1})$. By our calculation, the exact distribution of T_1 and T_2 are as follows.

The distribution of $T_1 = V_1(V_1 - V_2)$ is

$$f(T_1) = \int_0^\infty \frac{1}{T} 4m(m-1)(2\Phi(T_1/T - T) - 1)^{m-2}\phi(T_1/T - T)\phi(T_1/T)dT,$$

and distribution of $T_2 = V_2(V_2 - V_3)$ is

$$f(T_2) = \int_0^\infty \frac{1}{T} 4m(m-1)(2\Phi(\frac{T_2}{T}) - 1)^{m-3}(2\Phi(\frac{T_2}{T}) - 1)\phi(\frac{T_2}{T} - T)\phi(\frac{T_2}{T})dT.$$

And the joint distribution of $T_1 = V_1(V_1 - V_2)$ and $T_2 = V_2(V_2 - V_3)$ is

$$f(T_1, T_2) = \int_0^\infty \frac{1}{T^2 - T_1} 8m(m-1)(m-2)(2\Phi(T - \frac{T_1}{T} - \frac{TT_2}{T^2 - T_1}) - 1)^{m-3} \times$$
$$\phi(T)\phi(T - \frac{T_1}{T})\phi(T - \frac{T_1}{T} - \frac{TT_2}{T^2 - T_1})dT.$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of normal distribution. The detailed calculations are presented in Appendix A.

The distributions of T_1 and T_2 are already quite complicated and don't have closed forms and so does their joint distribution. Also due to the m possible covariance test statistics, it is impossible to calculate all the exact distributions. Luckily, Lockhart et al. (2014) found their asymptotical distributions which have simpler format.

3.1.4 Asymptotic distribution of covariance test

The following three results show the limiting distributions of the covariance test statistics when considering both orthogonal and general predictor matrix.

Result 1: Assume that $X \in \mathbb{R}^{n\times m}$ is orthogonal, and the true coefficient has k_0 nonzero components. Assume the smallest non-zero true coefficient is large compared to $\sigma\sqrt{2\log(m)}$, then for fixed $d \geq 0$,

$$(T_{k_0+1}, T_{k_0+2}, \cdots, T_{k_0+d}) \xrightarrow{d} (\text{Exp}(1), \text{Exp}(1/2), \cdots, \text{Exp}(1/d)) \text{ as } m \to \infty \tag{3.3}$$

Result 1 suggests that the covariance test statistics are asymptotically independent and converge in distribution to an exponential family when the design matrix is orthogonal. Under the global null hypothesis $y \sim N(0, \sigma^2)$, the covariance test statistic T_k is increasingly conservative at the kth step of the lasso solution path. At the first step, T_1 follows EXP(1) with mean 1 asymptotically, and at the second step, T_2 follows the EXP(1/2) with mean 1/2 asymptotically, and the covariance test statistics are decreasing with mean. This progress is desirable as we are less and less likely to reject the null hypothesis, which is the same intuition with the lasso path where the variables are less and less important in order. In

addition, the covariance test statistics for $T_1, T_2, \cdots$ are asymptotically independent, then the corresponding p-values can also be considered as independent. This property is also desirable for multiplicity adjustment, where some multiple testing procedure may require independency or positive dependency for test statistics.

Moreover, if there is k_0 nonzero coefficients in the true model, starting from the $k_0 + 1$ variable in the lasso path, we still have the result with probability approaching to one. The assumption for the strong signal, with coefficients larger compared to $\sigma\sqrt{2\log(m)}$, ensures the nonzero entries of β are large in absolute value, also it is the common constraint to separate the nonzero variables in lasso and in other variable selection methods as well.

Result 2: (For general X, $k = 1$) Assume that $\mathbf{X} \in \mathbb{R}^{n\times m}$ has unit normed columns in general position, and let $R = X^T X$. Assume also that there is some $\delta > 0$ such that for each $j = 1, \cdots, p$, there exists a subset of indices $S \subseteq \{1, \cdots, m\} \setminus \{j\}$ with

$$1 - R_{i,S\setminus\{i\}}(R_{S\setminus\{i\},S\setminus\{i\}})^{-1}R_{S\setminus\{i\},i} \geq \delta^2 \quad \text{for all } i \in S,$$

and the size of S growing faster than $\log m$,

$$|S| \geq d_m \qquad \text{where}\frac{d_m}{\log m} \to \infty \qquad \text{as } m \to \infty$$

Under the global null distribution, we have $P(T_1 > t) \to e^{-t}$ for all $t \geq 0$.

Result 3: (For general X, $k > 1$) Assume that $X \in R^{n\times m}$ has columns in general position, and $y \in R^n$ is drawn from the normal regression model. Assume that with probability approaching one: the lasso estimate at step k_0 in the path has support A_0 and signs s_{A_0} and the least squares estimate at step k_0 has the same signs as this lasso estimate.

Assume that there exists a constant $0 < \eta \leq 1$ such that

$$||(X_{A_0})^+ X_j||_1 \leq 1 - \eta \qquad \text{for all } j \notin A_0.$$

Define the matrix R by

$$R_{ij} = X_i^T (I - P_{A_0}) X_j \qquad \text{for } \ i, j \notin A_0.$$

Assume that the diagonal elements in R are all of the same order, that is $R_{ii}/R_{jj} \leq C$ for all i, j and some constant $C > 0$. Finally assume that for each fixed $j \notin A_0$, there is a set $S \subseteq \{1, \cdots, m\} \backslash (A_0 \cup \{j\})$ such that for all $i \in S$,

$$[R_i i - R_{i,S\backslash\{i\}} (R_{S\backslash\{i\},S\backslash\{i\}})^{-1} R_{S\backslash\{i\},i}]/R_i i \geq \delta^2$$

$$|R_i j|/R_j j < \eta/(2 - \eta)$$

$$||(X_{A_0 \cup \{j\}})^+ X_i||_1 < 1$$

where $\delta > 0$ is a constant (not depending on j), and the size of S grows faster than $\log m$

$$|S| \geq d_m \qquad \text{where} \frac{d_m}{\log m} \to \infty \qquad \text{as } m \to \infty$$

Then at step $k = k_0 + 1$, we have $\lim_{m \to \infty} P(T_k > t) \leq e^{-t}$ for all $t \geq 0$.

Simulations in Lockhart et al. (2014) demonstrated that these results hold for both orthogonal designs and general design. For result 2, the authors give a upper bound, which requires a sure screening property and other conditions (Fan and Ke 2014). For result 2 and 3, they proved that the covariance test statistic has an asymptotically exponential distribution under both global null and null model for k_0 nonzero coefficients.

From these results, the theoretical asymptotical distributions get more and more complicated when the predictor matrix is general and for $k > 1$. And they don't have the asymptotic distribution of joint covariance test statistics for general matrix as Result 1.

We conjecture that under general design matrix, the covariance test statistics also have the joint asymptotic distributions, even though Lockhart et al. (2014) haven't proved it. We test the asymptotic distributions of covariance test statistics in the following simulations. Let $n = 1000, m = 2000$, consider general gaussian X, with $x_i \sim i.i.d.\ N(0, 1/n)$, $\sigma = 1$. We calculate the covariance test statistics based on the lasso path, and draw the qq-plot with the corresponding exponential distribution. Each simulation is repeated 500 times.

Figure 3.1 displays the simulation results for $T_{k_0+d} \sim \text{Exp}(1/d)$ under the global null distribution, where fitting the dth covariance test statistics with the random generated exponential with mean $1/d$, and the diagonal line is with slope 1. The results generally support that these test statistics fit their asymptotical distributions very well. In addition, the performance gets more deviation when d gets larger, which also share the same conclusion with Fan and Ke (2014) for orthogonal case. Also, under some other structures of the general predictor matrix, the test statistics perform well. Therefore, we will incorporate these results into our proposed procedures later.

The covariance test statistic T_k is based on the known variance σ^2. In practice, the error variance is typically unknown. If $n > m$, we are able to estimate it in linear models. We can estimate σ^2 by the mean squared residual error $\hat{\sigma}^2 = ||y - X\hat{\beta}^{ols}||_2^2/(n-m)$, where $\hat{\beta}^{ols}$ is the least square regression. Applying the estimation of σ^2 in the covariance test

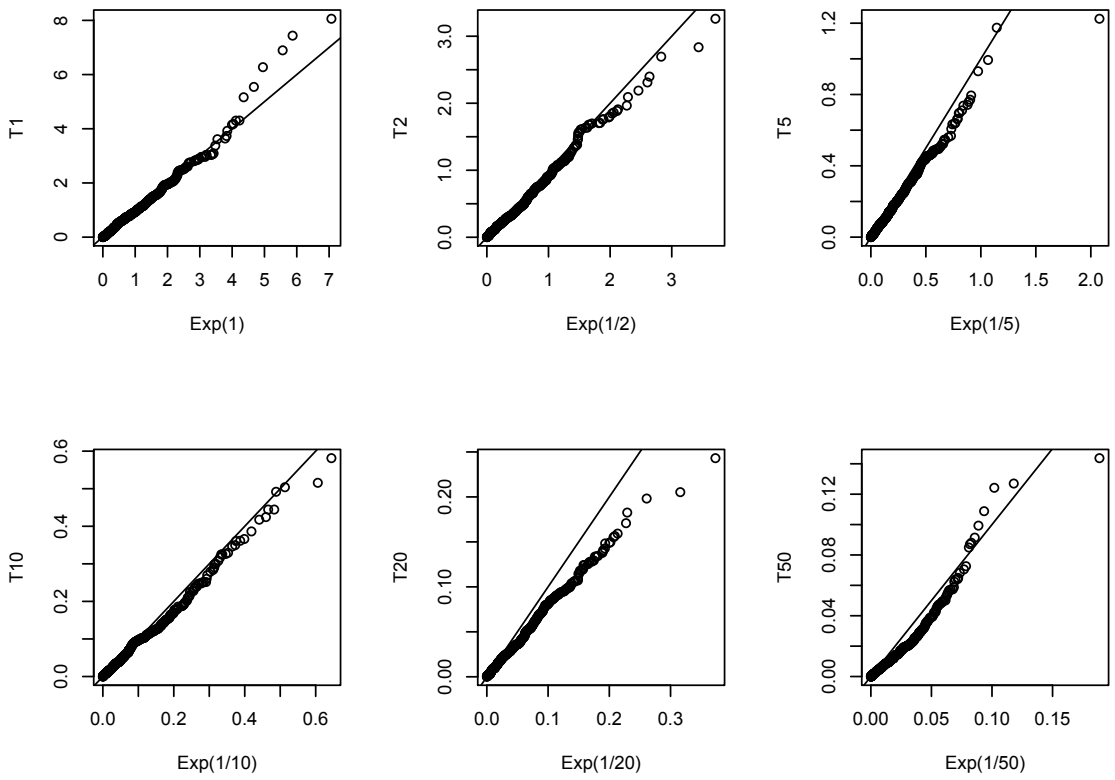


Figure 3.1: Simulation results to verify the asymptotic distributions of covariance test statistics for $d > 1$ in general design matrix.

statistic yields a F test statistic which has an asymptotic F distribution under the null:

$$F_k = \frac{< y, X\hat{\beta}(\lambda_{k+1}) > - < y, X_A\tilde{\beta}_A(\lambda_{k+1}) >}{\hat{\sigma}^2} \xrightarrow{d} F_{2,n-m}.$$

This follows because $F_k = T_k/(\hat{\sigma}^2/\sigma^2)$, where the numerator T_k being asymptotically Exp(1) $= \chi^2_2/2$, the denominator $\hat{\sigma}^2/\sigma^2$ being asymptotically $\chi^2_{n-m}/(n-m)$ and these two are independent.

When $m > n$, we are not able to estimate the σ^2 by ordinary least square method. One way is to estimate σ^2 from the least squares fit on the support of the model selected

by cross validation. This estimation is sometimes inflated. Then the test statistic follows the asymptotic $F_{2,n-r}$ under the null, where r is the size of the model selected by cross validation. Also we may use the exponential distribution by assuming the variance is the estimation, while it is less accurate.

3.2 Covariance statistics for other penalized methods

Pointed out by Fan and Ke (2014), there are similar results of the solution paths created by other penalty functions. For a general penalty function $p_\lambda(\cdot)$, the covariance test statistics at the knot λ_k can be conducted as lasso, except that $\hat{\beta}(\lambda_{k+1})$ and $\tilde{\beta}_A(\lambda_{k+1})$ are computed by different $\sum_j p_\lambda(\beta_j)$.

For the orthogonal design matrix, the penalized least square with folded concave penalty is unique (Fan and Li 2001). The covariance test statistic for penalized estimator is

$$T_k = V_k \cdot h_{V_{k+1}}(V_k)/\sigma^2,$$

where $h_\lambda(\cdot)$ is a thresholding function defined by $h_\lambda = \arg\min_u\{\frac{1}{2}(u-x)^2 + p_\lambda(u)\}$. And V_k is the kth largest elements of $\{|X_j^T y| : 1 \le j \le m\}$ for the orthogonal predictor, which follows the independent χ_1 distribution. They only have covariance test statistics results for orthogonal design matrix for SCAD and MCP now, which will be illustrated next.

3.2.1 Covariance test statistic for SCAD

For the SCAD penalty (Fan and Li 2001) with a parameter $\gamma > 2$, the covariance test statistics can be written as

$$T_k^{\text{scad}} = \begin{cases} V_k(V_k - V_{k+1})/\sigma^2, & \text{if} \quad V_k \leq 2V_{k+1} \\ \frac{\gamma-1}{\gamma-2} V_k(V_k - \frac{\gamma}{\gamma-2} V_{k+1}), & \text{if} \quad 2V_{k+1} < V_k < \gamma V_{k+1} \\ V_k^2/\sigma^2, & \text{if} \quad V_k \geq \gamma V_{k+1}. \end{cases}$$

They similarly show that for any fixed $k \geq 1$ the joint distribution of covariance test statistics have

$$(T_1^{\text{scad}}, T_2^{\text{scad}}, \cdots, T_k^{\text{scad}}) \xrightarrow{d} (\text{Exp}(1), \text{Exp}(1/2), \cdots, \text{Exp}(1/k)) \ \text{ as } \ m \to \infty$$

under the global null true model.

Similarly, the covariance test statistics are used to test the entry of kth variable in the solution for SCAD method.

3.2.2 Covariance test statistic for MCP

For the MCP penalty (Zhang 2010) with a parameter $\gamma > 1$, the covariance test statistics can be shown similarly as

$$(T_1^{\text{mcp}}, T_2^{\text{mcp}}, \cdots, T_k^{\text{mcp}}) \xrightarrow{d} \frac{\gamma}{\gamma - 1}(\text{Exp}(1), \text{Exp}(1/2), \cdots, \text{Exp}(1/k)) \ \text{ as } \ m \to \infty$$

under the global null true model.

To utilize the asymptotic null distribution without conservatism, we need to have the model selection consistency assumption: the first k_0 variables contain all important

variables. In addition, we can use these asymptotic distribution for statistical inference, such as hypothesis testing, confidence interval estimation. These results need to be confirmed for orthogonal design matrix X and the assumptions sometimes are difficult to meet in practice. Also they haven't have similar result for general design matrix which will be more complicated.

So far, we have introduced the penalized variable selection methods and the covariance test statistics. Benefit from the significance test, we can utilize the asymptotic distribution of the covariance test statistics, and get the p-values for testing the variable that entering the current active model to control error rate with p-value corrections.

3.3 Some new significance tests for variable selection

There are some new proposed significance test for variable selection related to post selection. Some of them are proposed for lasso and LARS. Taylor et al. (2014) proposed a sequence of spacing test p-values corresponding to each step in the LARS path. On the other hand, the penalized score test (Voorman, Shojaie and Witten 2014) was introduced for lasso path based on penalized regression. The sparsity pattern of the lasso corresponded exactly to a decision based on the proposed test. The test statistics were shown to be independent and uniformly distributed for spacing test and asymptotic normal for penalized score test under the null hypothesis.

3.3.1 Spacing test for LARS

Taylor et al. (2014) proposed the spacing test for Least-angle path. Same as lasso, we assume that X has columns in general position, which ensures the uniqueness of the selected variables and the sign at each step. Similarly, suppose that the LARS algorithm produces the list of active variables A_k and signs s_{A_k} after k steps. The null hypothesis being considered is

$$H_0 : e_k^T X_{A_k}^+ \beta = 0$$

i.e., the spacing test is a test for the kth coefficient in the multiple regression of β on X_{A_k}. Letting

$$\omega_k = ||(X_{A_k}^+)^T s_{A_k} - (X_{A_{k-1}}^+)^T s_{A_{k-1}}||_2,$$

the one-sided spacing test statistic is defined by

$$R_k = \frac{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(\lambda_k \frac{\omega_k}{\sigma})}{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(M_k^+ \frac{\omega_k}{\sigma})},$$

and the two-sided spacing test statistic is defined by

$$T_k = 2 \cdot \min\{R_k, 1 - R_k\}.$$

where λ_{k-1} and λ_k are the knots at steps $k-1$ and k in the LAR path, and M_k^+ is the random variable for the maximum of linear functions of y.

Testing the null hypothesis $H_0 : e_k^T X_{A_k}^+ \beta = 0$, the one-sided spacing statistic R_k and the two-sided statistic T_k are, both conditional on $\Gamma y \geq U$:

$$\Pr_{e_k^T X_{A_k}^+ \beta = 0} (R_k \leq \alpha | \Gamma y \geq U) = \alpha$$

$$\Pr_{e_k^T X_{A_k}^+ \beta = 0}(T_k \leq \alpha | \Gamma y \geq U) = \alpha$$

for any $0 \leq \alpha \leq 1$. The event $\Gamma y \geq U$ is approximate to LAR selection event. Then, the one-sided and two-sided spacing test follow approximate uniform distribution.

There is an interesting connection between the LAR spacing test and the covariance test for lasso. After a fixed number of k steps of LAR, the one-sided spacing p-values is asymptotically equivalent to the covariance test. Review Taylor et al. (2014) for details.

3.3.2 Penalized score test for lasso

Voorman, Shojiaie and Witten (2014) pointed out that the covariance test statistic can be used to test whether all relevant variables are non-zero, however, it can not provide confidence intervals or other statistical inference for lasso, which also means the analysis based on covariance test is conditional on λ_k. And the active set of the lasso path at each λ_k will vary across samples, which makes the interpretation difficult. They proposed a more universal way to test any interested variables and further analysis.

To simplify the notation, consider a single variable of interest $x = x_j$, and denote $Z = [x_k, k \neq j] \in \mathbb{R}^{n \times (m-1)}$ as the matrix containing all other variables. The linear model can be re-written as

$$y = \alpha x + Z\beta + \epsilon$$

where $\alpha \in \mathbb{R}$ and $\beta \in \mathbb{R}^{m-1}$.

The goal is to test $H_0 : \alpha = 0$. The score test, which is based on the derivative of the log-likelihood, also known as the score, is evaluated under the null hypothesis. Also the parameters are estimated under the constraints imposed by the null hypothesis.

Assume the coefficient estimation for lasso at knot λ and form the test statistic $T_\lambda = (y - X\beta_\lambda)/\sqrt{n}$. Under some conditions (which depend on λ),

$$\frac{T_\lambda}{\sigma_\epsilon \sqrt{x^T(I_n - P_A)x/n}} \xrightarrow{d} N(0,1)$$

where $x^T(I_n - P_A)x/n$ having the upper bound 1.

The idea of penalized score test can be used in other penalty functions, such as Ridge, SCAD, elastic net and etc.

Those two recently proposed significant test will be very useful for post selection inference. However, we are not interested in any particular variable or other selection procedures now. I will focus on the lasso solution path and sequentially test the entering variable in the model for their significance. Therefore, in next chapter, our testing structure will be based on lasso covariance test and two type I errors for multiple testing and procedures to control error rate will be introduced.

Chapter 4

Sequential multiple hypothesis testing

4.1 Multiple testing

First, we will introduce the error rates in general multiple testing before we connect them with variable selection. And then, we will review some multiple testing procedures to control the corresponding error rate.

Suppose we have m hypotheses, denoted by $H_1, H_2, \cdots, H_m$ and among those there are m_0 null hypothesis. Using a statistical test, we reject the null hypothesis if the test is declared significant. And we fail to reject the null hypothesis if the test is non-significant. It is possible to make wrong decisions, i.e. errors, based on our knowledge. The true hypothesis for each individual hypothesis can be true or false, and on the other hand for each decision we make, there will be chance to claim each hypothesis as significant and

non-significant. Therefore, these decisions can be summed up in Table 4.1 as follows:

Table 4.1: Multiple testing notation.

	Null hypothesis is True	Null hypothesis is False	Total
Declared significant	V	S	R
Declared non-significant	U	T	$m - R$
Total	m_0	$m - m_0$	m

Parameters	Notation
m_0	number of true null hypotheses
$m - m_0$	number of true alternative hypotheses
V	number of false positives
S	number of true positives
T	number of false negatives
U	number of true negatives
R	number of rejected null hypotheses

There are two types of Type I error rates associated with the decision making for multiple testing. One is Familywise error rate (FWER) and the other one is False discovery rate (FDR).

4.1.1 FWER

FWER is the probability of making one or more false discoveries among all the hypotheses, which can be expressed as:

$$\text{FWER} = \Pr(V \geq 1) = 1 - \Pr(V = 0). \tag{4.1}$$

From the definition of FWER, we can tell that this type I error is a very stringent one, not allowing any wrong claims for all the null hypothesis. There are number of ways available now to control the FWER at α, for example, 0.05. Since the Bonferroni inequality leads to a more conservative test procedure, there have been several attempts to improve the method. Holm's (1979) step-down procedure and Hochberg's (1988) step-up procedure for

multiple testing are both developed to control FWER, where Holm's procedure is thought of as a step-down version of the Bonferroni test, and Hochberg's procedure is thought of as a step-up version of the Bonferroni test. The Holm's step-down procedure is compared from the larger test statistics to smaller test statistics, and similarly, the Hochberg's step-up procedure is to move from the smaller test statistics to the larger ones. For the p-values, the probability of rejecting the hypothesis under the null distribution, the Holm's step-down procedure is compared from the smallest p-value to the larger ones and the Hochberg's step-up is compared from the larger ones to the smaller p-values.

Let $p_i, i = 1, \cdots, m$ denote the sample p-values of tests for $H_i, i = 1, \cdots, m$, computed without multiplicity adjustment. Let $[1], \cdots, [m]$ denote the random indices such that

$$p_{[1]} \leq \cdots \leq p_{[m]},$$

where $[i]$ is the anti-rank of p_i among $p_1, \cdots, p_m$.

Bonferroni procedure proceeds as follows:

Reject the hypotheses H_i if $p_i \leq \alpha/m$.

Holm's step-down procedure proceeds as follows:

Step 1: If $p_{[1]} < \alpha/m$, reject $H_{[1]}$ and go to Step 2; otherwise stop.

Step 2: If $p_{[2]} < \alpha/(m-1)$, reject $H_{[2]}$ and go to Step 3; otherwise stop.

...

Step m: If $p_{[m]} < \alpha$, reject $H_{[m]}$ and stop.

Hochberg's step-up procedure proceeds as follows:

Step 1: If $p_{[m]} < \alpha$, reject $H_{[i]}, i = 1, \cdots, m$ and stop; otherwise go to Step 2.

Step 2: If $p_{[m-1]} < \alpha/2$, reject $H_{[i]}, i = 1, \cdots, m-1$ and stop; otherwise go to Step 3.

...

Step k: If $p_{[1]} < \alpha/m$, reject $H_{[i]}, i = 1$ and stop.

Bonferroni procedure is the same as the Holm's step-down procedure when we consider there is at least one hypothesis is significant. Holm's step-down procedure will only move forward unless the $H_{[1]}$ is rejected.

The Bonferroni procedure tends to be rather conservative if the number of hypotheses is larger or the test statistics are strongly positively correlated. The Holm procedure is, in fact, a closed testing procedure in which each intersection hypothesis is tested using a global test based on the Bonferroni procedure. Since the Bonferroni procedure cannot reject more hypotheses than the Holm procedure, the latter serves as an example of how the power of a multiple testing procedure can be enhanced in a uniform manner by applying the closure principle. The procedures of Bonferroni and Holm control the FWER under any dependence structure of the p-values (or equivalently the individual test statistics). Essentially, this is achieved by assuming a 'worst-case' dependence structure (which is close to an assumption of independence for all practical purposes). But such an approach generally leads to power loss.

The Simes test (Simes 1986) focused on testing the global hypothesis of no effect, i.e. $H_0 = \cap_{i=1}^{m} H_i$. It rejects H_0 if $p_{[i]} \leq i\alpha/m$ for at least one $i = 1, \cdots, m$.

Note that the Simes test makes use of all the ordered p-values to test the global hypothesis and thus it is more powerful than a global test based on the Bonferroni proce-

dures, and the Simes test cannot be used directly to test the individual hypotheses. The Hochberg procedure is based on the Simes global test while it can be used to test the individual hypothesis. The Hochberg procedure controls the FWER under the same conditions as the Simes test, so it holds only under non-negative dependency. Further, it is uniformly more powerful than the Holm procedure, but on the other hand , it is uniformly less power than the Hommel procedure. Of course, there are some other extended versions of these procedures (Wiens 2003, Dmitrienko and Tamhane 2007, Hommel 1989, Dunnett and Tamhane 1991 and etc.) which also can be used to control FWER.

We are interested in testing the intersection of each individual hypothesis, $H_0 = \cap_{i=1}^{m} H_i$. Most of the methods for FWER control procedures are testing individual hypothesis except for Simes test. Test the joint hypotheses can be thought of as long as all the individual hypothesis is not being rejected, then the joint hypothesis is not being rejected.

For FWER control, the cutoff values for Holm's procedure and the Hochberg procedure are the same but different from Simes. We take an example by comparing their performance along with the Bonferroni procedure in the joint hypotheses testing to express this point. Suppose there are 3 hypotheses been tested at one step, then the intersection hypotheses becomes

$$H_0 : \beta_1 = 0; \beta_2 = 0; \beta_3 = 0.$$

And $p_i, i = 1, 2, 3$ denote the p-values of tests for individual $H_{0i} : \beta_i = 0$. The final decision of the intersection hypothesis is that if reject at least one of the individual hypothesis, then H_0 is rejected. If we apply the multiple testing procedures for FWER control for the intersection hypothesis, then we can claim that the Holm's step-down procedure is the

same as Bonferroni method, different from Hochberg procedure. Like the partitioning test (Huang and Hsu 2007), we can partition the sample space into 8 non-overlapped subspaces. We get the following result in Table 4.2.

Table 4.2: Compare FWER procedures in subspaces for m=3.

	8 combinations	Hochberg's step-up	Holm's step-down	Bonferroni
1	$p_{[1]} < \frac{\alpha}{3}; p_{[2]} < \frac{\alpha}{2}; p_{[3]} < \alpha$	Reject $H_{0[1]}$, $H_{0[2]}$, $H_{0[3]}$	Reject $H_{0[1]}$, $H_{0[2]}$, $H_{0[3]}$	At least Reject $H_{0[1]}$
2	$p_{[1]} < \frac{\alpha}{3}; p_{[2]} < \frac{\alpha}{2}; p_{[3]} > \alpha$	Reject $H_{0[1]}$, $H_{0[2]}$	Reject $H_{0[1]}$, $H_{0[2]}$	At least Reject $H_{0[1]}$
3	$p_{[1]} < \frac{\alpha}{3}; p_{[2]} > \frac{\alpha}{2}; p_{[3]} < \alpha$	Reject $H_{0[1]}$, $H_{0[2]}$, $H_{0[3]}$	Reject $H_{0[1]}$	At least Reject $H_{0[1]}$
4	$p_{[1]} < \frac{\alpha}{3}; p_{[2]} > \frac{\alpha}{2}; p_{[3]} > \alpha$	Reject $H_{0[1]}$	Reject $H_{0[1]}$	At least Reject $H_{0[1]}$
5	$p_{[1]} > \frac{\alpha}{3}; p_{[2]} < \frac{\alpha}{2}; p_{[3]} < \alpha$	Reject $H_{0[1]}$, $H_{0[2]}$, $H_{0[3]}$	Fail to Reject all	Fail to Reject all
6	$p_{[1]} > \frac{\alpha}{3}; p_{[2]} < \frac{\alpha}{2}; p_{[3]} > \alpha$	Reject $H_{0[1]}$, $H_{0[2]}$	Fail to Reject all	Fail to Reject all
7	$p_{[1]} > \frac{\alpha}{3}; p_{[2]} > \frac{\alpha}{2}; p_{[3]} < \alpha$	Reject $H_{0[1]}$, $H_{0[2]}$, $H_{0[3]}$	Fail to Reject all	Fail to Reject all
8	$p_{[1]} > \frac{\alpha}{3}; p_{[2]} > \frac{\alpha}{2}; p_{[3]} > \alpha$	Fail to Reject all	Fail to Reject all	Fail to Reject all

The reason we use the same cutoff value for all the 8 combination is that the Holm and Hochberg procedures share the same cutoff values. Holm's step-down and Bonferroni reach the same decisions: reject H_0 for subspaces 1 - 4 and fail to reject H_0 for subspaces 5 - 8, while Hochberg's step-up procedure reject H_0 for subspaces 1 - 7 and fail to reject H_0 only in subspace 8. Even though the cutoff values for the individual hypothesis for Holm's step-down and Hochberg's setup-up is the same, they don't share the same conclusion. Since the cutoff value of Simes' procedure is not the same as Hochberg's step-up and Holm's step-down procedure, they won't reach the same decisions for the intersection hypotheses. In Chapter 5 for our proposed FWER control procedures, we will incorporate the intersection hypotheses and will use Bonferroni, Hochberg and Simes methods within each step.

4.1.2 FDR

FDR (Benjamini and Hochberg 1995) is the expected proportion of rejected null hypotheses that are incorrect rejections, defined as:

$$\text{FDR} = E\left(\frac{V}{V+S}\right) = E\left(\frac{V}{R}\right). \tag{4.2}$$

FDR provides less stringent control compared to FWER for type I errors. Therefore, FDR procedures usually reach high power at the cost of more false positives. If all the null hypotheses are true, the FDR is equivalent to FWER, otherwise FDR is much smaller than FWER. The FDR control is more often used for high dimensional hypothesis, and sometimes the FWER is not very practical in reality.

There are several popular procedures to control FDR at q. Similarly as the procedures as in FWER control, the Benjamini-Hochberg procedure (Benjamini and Hochberg 1995) examines the p-value from the the smaller test statistics to the larger test statistics and Benjamini-Liu procedure (Benjamini and Liu 1999) moves from the more significant to the less significant hypotheses. Benjamini-Hochberg-Yekutieli procedure (Benjamini and Yekutieli 2001) relaxes the independent requirement for test statistics.

Benjamini-Hochberg procedure (BH step-up) proceeds as follows:

Step 1: For a given q, find the largest k such that $p_{[k]} \leq \frac{k}{m}q$;

Step 2: Reject the null hypothesis for all $H_{[1]}, \cdots, H_{[k]}$.

Benjamini-Hochberg-Yekutieli procedure proceeds as follows:

Step 1: For a given q, find the largest k such that $p_{[k]} \leq \frac{k}{m \cdot c(m)}q$;

Step 2: Reject the null hypothesis for all $H_{[i]}$ for $i = 1, \cdots, m$.

where $c(m) = 1$ if the tests are independent or positively correlated; $c(m) = \sum_{i=1}^{m} \frac{1}{i}$ for any arbitrary dependence.

Benjamini-Liu procedure (BL step-down) proceeds as follows:

Step 1: Find the smallest k such that $p_{[k]} > \delta_i$;

Step 2: Reject the null hypothesis for all $H_{[1]}, \cdots, H_{[k-1]}$.

where $\delta_i = 1 - \left[1 - \min(1, \frac{m}{m-i+1}q)\right]^{1/(m-i+1)}$.

The BH procedure is valid when the m tests are independent and also in some dependence case. In addition, the Benjamini-Hochberg-Yekutieli procedure is proposed to improve the BH procedure, and it can control the false discovery rate under positive dependence assumptions. The BH step-up procedure is compared the p-values from the larger ones to the smaller p-values, and find the maximum index and reject the corresponding hypotheses with the small p-values. The BL step-down procedure proceeds the opposite way, considering from the smaller p-values to the larger p-values. All these procedures can control the FDR at q.

For FDR control, since all the cutoff values are not the same for each individual hypothesis testing procedures, then the decisions will be vary by different procedures. The discussion for the joint hypothesis testing will be dismissed.

There are some extended versions of FDR, referring pFDR (Storey 2001) for positive FDR, local FDR (Efron 2004), mdFDR (Guo et al. 2010) for mixed directional FDR and etc. for details.

4.2 Sequential multiple testing

Since our work will be focused on the error rate control for variable selection, now we have to connect the multiple testing with variable selection. First of all, we will introduce some definitions related error rate in terms of selection. Denote $\hat{D}_n$ as the selected variables, and D as the set of true variables in the model. Suggested by Lin, Foster and Ungar (2011), Meinshausen and Bühlmann (2010), Shah and Samworth (2012), Meinshausen et al. (2009), Wu, Boos and Stefanski (2007), Wasserman and Roeder (2009), G'Sell, Hastie and Tibshirani (2013), and the type I error rate definition (4.1) and (4.2) in multiple testing, we define Family-wise Error Rate (FWER), False Discovery Proportion (FDP) and False Discovery Rate (FDR) in terms of variable selection as:

$$FWER(\hat{D}_n) = Pr\left(|\{j \in \hat{D}_n : \beta_j = 0\}| \geq 1\right) \tag{4.3}$$

$$FDP(\hat{D}_n) = |\{j \in \hat{D}_n : \beta_j = 0\}|/|\hat{D}_n| \tag{4.4}$$

$$FDR(\hat{D}_n) = E(FDP) \tag{4.5}$$

FWER is defined as the probability that there is at least one rejected variable whose true coefficient is 0. FDP is the proportion of falsely discovered variable, which is defined as the proportion of falsely rejected variables. And the FDR is the expectation of the FDP.

Now back to our variable selection process. Our variable selection methods are based on the penalized methods, such as lasso. The lasso path is changed when the tuning parameter decrease from ∞. When λ decrease, there will be some variable enters into or leaves the model. We can consider this process as sequentially examining the entered

variable whether this variable is significant or not, and we refer this process as sequential multiple testing.

The lasso variable selection consistency assumes that the first k_0 variables contain all important variables, which means the significant variables precede all the non-significant ones. For the fixed number of steps $k \geq 1$ along the solution path, the covariance test examines the set of variables selected by the lasso before the kth step sequentially. Then the null hypothesis being tested at the kth step is

$$H_{0k} : A \supseteq supp(\beta) \tag{4.6}$$

where β is the true underlying coefficient vector in the model. For $k = 1$, $A = \emptyset; H_{01} : \beta = 0; H_{\text{a}1}$: at least one $\beta_i \neq 0$. For $k > 1$, the set A (i.e. current active set not including the variable to be added at step k) is random and depends on the previous steps and y, which makes the covariance test a conditional hypothesis test. Then the null distributions depend conditionally on the event $A \supseteq supp(\beta)$. At the kth step, the hypotheses in terms of the coefficients can also be written as:

$$\begin{aligned} A \supseteq supp(\beta) &\Leftrightarrow \{1, \cdots, m\} \backslash A \subseteq \{j : j \in \{1, \cdots, m\}, \beta_j = 0\} \\ &\Leftrightarrow \beta_{\{1,\cdots,m\}\backslash A} = 0 \end{aligned}$$

Since the current active set A is depend on y, X and previous lasso path, then at the kth step the hypotheses becomes

$$H_{0k} : \beta_{\{1,\cdots,m\}\backslash A|_{y,X,k}} = 0. \tag{4.7}$$

We can clearly see the hypotheses depend on the previous steps along the path and in each

step the hypotheses can be treated as the multiple hypotheses setting. When the lasso path or other regression path is unique, we can test the kth entered variable sequentially. Some authors also considered a similar setup (Zhang 2011, Buhlmann 2012), while they consider the hypothesis as fixed.

4.2.1 Single hypothesis test

The single hypothesis proposed in Lockhart (2014) is to test one coefficient at each step sequentially along the lasso path, without considering all the underlying multiple structure. This results in an ordered hypothesis testing problem that is conditional on the order in which the variables are added along the lasso path.

By the result of the covariance test, if we only test one variable at a time, the sequential hypothesis test at the kth step becomes $H_{0k} : \beta_k = 0$. If H_{0k} is rejected, continue to the $(k+1)$th step. Here the β_k is the coefficient of the kth variable in the lasso path. In terms of the distribution of the covariance statistics, it is equivalent to check if $T_k \sim \text{Exp}(1)$. The tests are preformed in the sequential manner, which means that, for instance, we will move forward to the second step if and only if the first hypothesis is rejected.

4.2.2 Multiple hypotheses test

Consider the full underlying hypotheses in (4.7), which formulated in multivariate setting at each step. Stated by the asymptotic theory as Result 1, starting from the first non-significant variable (i.e. the (k_0+1)th variable in the path since the true coefficient has k_0 nonzero components), the covariance test statistics $T_{k_0+1}, T_{k_0+2}, \cdots$ follow exponential family with mean 1, 1/2, $\cdots$. By our sequential rule, before we move on to the kth step, all

the first $(k-1)$ variables in the solution path are claimed as significant variables. At step k, test if $H_{0k}: \beta_{\{1,\cdots,m\}\setminus A|_{y,X,k}} = 0$, which is equivalent to test if $H_{0k}: T_k \sim \text{Exp}(1); T_{k+1} \sim \text{Exp}(1/2); \cdots$. If fail to reject H_{0k}, it means that starting from the kth variable, the followed variables in the lasso path are non-significant. If reject H_{0k}, we move on to test the $(k+1)$th variable and all the first k variables are concluded as important variables.

Therefore, we propose the sequential multiple hypotheses testing procedure structure, which works as follows:

Step 1: test $H_{01}: \beta_{\{1,\cdots,m\}\setminus A|_{y,X,k=1}} = 0$ at significance level α_1;

If fail to reject all the hypotheses, stop; variable selected $\hat{D}_n = \{\emptyset\}$;

If reject at least one of the hypotheses, continue to Step 2;

Step 2: test $H_{02}: \beta_{\{1,\cdots,m\}\setminus A|_{y,X,k=2}} = 0$ at significance level α_2;

If fail to reject all the hypotheses, stop; $\hat{D}_n = \{1^{st}\}$;

If reject at least one of the hypotheses, continue to Step 3;

...

Step k: test $H_{0k}: \beta_{\{1,\cdots,m\}\setminus A|_{y,X,k}} = 0$ at significance level α_k;

If fail to reject all the hypotheses, stop; $\hat{D}_n = \{1^{st}, \cdots, (k-1)^{th}\}$;

If reject at least one of the hypotheses, continue to Step $k+1$;

...

Step m: test $H_{0m}: \beta_{\{1,\cdots,m\}\setminus A|_{y,X,k}} = 0$ at significance level α_k;

If fail to reject all the hypotheses, stop; $\hat{D}_n = \{1^{st}, \cdots, (m-1)^{th}\}$;

If reject at least one of the hypotheses, stop; $\hat{D}_n = \{1, \cdots, m\}$.

As an added benefit, the sequential multiple testing structure increases power for the selection procedure when the true coefficient is large while the drop of the covariance test is relatively small. For instance, when the design matrix X is orthogonal, $T_k = \lambda_k(\lambda_k - \lambda_{k+1})$, and λ_k is also the estimation of absolute value of the kth largest true coefficient. When the true coefficients are equal or close but with large signal, it does not guarantee a larger drop of covariance. As a result of the single hypothesis structure, the procedure will fail to reject the hypothesis and end too early to approach higher power.

In Chapter 5 and Chapter 6, the procedures proposed for FWER and FDR control are based on our sequential multiple testing structure and different cutoff values of α_i and q_i will be chosen accordingly.

Chapter 5

Hybrid sequential multiple testing to control FWER

FWER is the probability of making one or more false discoveries among all the hypotheses when performing multiple hypotheses tests. We've introduced some popular FWER control procedures in Chapter 4. In this chapter, we will proposed new procedures for FWER control at α for lasso variable selection by incorporating the well-known procedures.

In this chapter, we first propose a hybrid multiple testing procedure, which is based on the covariance test statistics and Bonferroni adjusted p-values, for testing H_{0k} : $\beta_{\{1,\cdots,m\}\setminus A|_{y,X,k}} = 0$ at the kth step. Test statistics at each step are following asymptotic exponential family and our proposed procedures will be demonstrated with p-values in the rest of this paper. The performance of the proposed procedure is then evaluated through simulation studies.

5.1 FWER control procedure

To calculate p-values using asymptotic exponential distribution, we assume model selection consistency of lasso, i.e., the first k_0 variables contain all the important variables and the variables are ordered from the most significant to least significant in the lasso path. In our multivariate structure, we borrow the idea of Holm's step-down procedure for FWER control at α. Assuming the associated p-values $P_1, \cdots, P_m \in [0,1]$ for a multiple testing, Holm's step-down procedure starts by ordering the p-values from the lowest to the highest $P_{(1)}, \cdots, P_{(m)}$. Let R be the smallest k such that $P_{(k)} > \frac{\alpha}{m-k+1}$, and then the null hypotheses $H_{(1)}, \cdots, H_{(R-1)}$ will be rejected. Unlike one group of p-values above, our p-values from our structure are two dimensional. Consider the p-values in our multivariate structure at each step as follows:

Step 1 : $P_1^1 \quad P_2^1 \quad P_3^1 \quad \cdots \quad P_m^1$; Let $P^1 = \min\{P_i^1\}, i = 1, \cdots, m$;

Step 2 : $P_1^2 \quad P_2^2 \quad P_3^2 \quad \cdots \quad P_{m-1}^2$; Let $P^2 = \min\{P_i^2\}, i = 1, \cdots, m-1$;

$\cdots$

Step k : $P_1^k \quad P_2^k \quad P_3^k \quad \cdots \quad P_{m-k+1}^k$; Let $P^k = \min\{P_i^k\}, i = 1, \cdots, m-k+1$;

$\cdots$

Step m : P_1^m; Let $P^m = P_1^m$.

Our strategy is to use the composite p-value, the minimum p-value, at each step. The desired property of the composite minimum p-values is that they are increasing across steps, given by Lemma 2. And more importantly, the order of the p-values is consistent with the order of hypotheses. Therefore, we propose the Hybrid Bonferroni-Holm Step-Down procedure

as described in Procedure 1 for sequential multiple testing in variable selection of lasso to control FWER.

Lemma 2 *Suppose* $\hat{D}_n \neq \emptyset$ *and* $|\hat{D}_n| = S$. *Let* P^k *be the minimum p-value for the kth step, i.e.* $P^k = \min\{P_1^k, P_2^k, \cdots, P_{m-k+1}^k\}$, *where* P_i^k *is the p-value for* $H_{0k,i} : \beta_{k+i-1} = 0$ *at kth step. Then* $P^1 < P^2 < \cdots < P^S$.

Procedure 1 (Hybrid Bonferroni-Holm Step-Down) *Let* P^k *and* P_i^k *be defined as Lemma 2, Hybrid Bonferroni-Holm Step-Down procedure can be described as:*

Step 1: *test* $H_{01} : \beta_{\{1,\cdots,m\}\setminus A|_{y,X,k=1}} = 0$;

If $P^1 \leq \frac{\alpha}{m(m-1+1)}$, *reject* H_{01}, *continue to Step 2;*

Otherwise, fail to reject, stop; variable selected $\hat{D}_n = \{\emptyset\}$;

Step 2: *test* $H_{02} : \beta_{\{1,\cdots,m\}\setminus A|_{y,X,k=2}} = 0$;

If $P^2 \leq \frac{\alpha}{m(m-2+1)}$, *reject* H_{02}, *continue to Step 3;*

Otherwise, fail to reject, stop; $\hat{D}_n = \{1^{st}\}$;

...

Step k***:*** *test* $H_{0k} : \beta_{\{1,\cdots,m\}\setminus A|_{y,X,k}} = 0$;

If $P^k \leq \frac{\alpha}{m(m-k+1)}$, *reject* H_{0k}, *continue to Step* $k+1$;

Otherwise, fail to reject, stop; $\hat{D}_n = \{1^{st}, \cdots, (k-1)^{th}\}$;

...

Step m***:*** *test* $H_{0m} : \beta_{mth} = 0$;

If $P^m \leq \frac{\alpha}{m}$, *reject* H_{0m}, *stop;* $\hat{D}_n = \{1, \cdots, m\}$;

Otherwise, fail to reject, stop; $\hat{D}_n = \{1^{st}, \cdots, (m-1)^{th}\}$.

Theorem 3 *Hybrid Bonferroni-Holm Step-Down (Hybrid BHSD) procedure controls FWER at level α.*

The proof of Lemma 2 and Theorem 3 show in Appendix. We note that this procedure is applied in the covariance test because they follow the exponential distribution family across the steps, while it should be taken extra calibration for other test statistics.

5.2 Simulation results for FWER

Next the simulations are based on our proposed procedure and we compare its result with other methods.

1. **Hybrid Hochberg-Holm Step-Down (Hybrid HHSD):** Hochberg procedure is incorporated under our sequential multiple testing structure within each step since the asymptotic distributions are independent. Within the kth step, the ith ordered p-values is compared with $\frac{\alpha}{(m-i+1)(m-k+1)}$;

2. **Hybrid Simes-Holm Step-Down (Hybrid SHSD):** Similar as Hochberg, within the kth step, the ith ordered p-value is compared with $\frac{i\alpha}{m(m-k+1)}$; And we expect those two methods have higher power than Procedure 1; However, we haven't proved that the FWER control for Hybrid Hochberg-Holm Step-Down and Hybrid Simes-Holm Step-Down procedures.

3. **StrongStop:** StrongStop (G'Sell et al. 2014) is proposed to control FWER and FDR. Let $\tilde{Y}_k = -log(p_k)$, $\tilde{Z}_k = \sum_{j=k}^{m} Y_j/j$, $\tilde{q}_k = e^{-\tilde{Z}_k}$. Here the p_k is the p-value for the lasso single hypothesis test at the kth step. Reject hypotheses $1, \cdots, \hat{k}$ (select the

first $\hat{k}$ variables), where

$$\hat{k} = \max\{k \in \{1, \cdots, m\} : \tilde{q}_k < \frac{\alpha k}{m}\}$$

The weakness of StrongStop is that the decision to reject at kth step depends on the p-values after this variable. From intuition, it is not how we make decisions. Also if the last p-values are larger than the uniform or deviate from the uniform distribution, it will suffer power loss and be unstable.

The first group of simulations are set for orthogonal design matrix. Let $n = 200,\ m = 20,\ \sigma = 1,\ \alpha = 0.05$. Among the 20 variables, first 4 variables are significant and the following 16 are non-significant. We simulate 1000 times for different coefficients of true β with result showing in Table 5.1. Four different model coefficients are: a) $\beta = (8,\ 6,\ 4,\ 2,\ 0,\ \cdots,\ 0)$; b) $\beta = (10,\ 8,\ 6,\ 4,\ 0,\ \cdots,\ 0)$; c) $\beta = (6,\ 6,\ 6,\ 6,\ 0,\ \cdots,\ 0)$; d) $\beta = (4,\ 4,\ 4,\ 4,\ 0,\ \cdots,\ 0)$.

Table 5.1: Comparison among model a - d for FWER control.

	a		b		c		d	
	Power	FWER	Power	FWER	Power	FWER	Power	FWER
Hybrid BHSD	0.5893	0.0080	0.8452	0.0050	0.9607	0.0040	0.4973	0.0110
Hybrid HHSD	0.5915	0.0080	0.8488	0.0050	0.9625	0.0040	0.5008	0.0120
Hybrid SHSD	0.5923	0.0080	0.8495	0.0050	0.9627	0.0040	0.5035	0.0120
StrongStop	0.4980	0.0000	0.7083	0.0000	0.7125	0.0000	0.1573	0.0000

From Table 5.1 , the first three methods using our multiple structure do not have a big difference, while generally Simes has the highest power, followed by Hochberg and then Bonferroni. And more importantly, all the methods can control FWER at $\alpha = 0.05$. With larger effect size (i.e. model b and c), the power is higher as expected. Our structure

gains more power when the effect sizes are equal, and the reason for this is because of the covariance test. When the design matrix are orthogonal and the coefficients are equal (i.e. model c and d), the estimates of β are also very close, therefore the hypotheses are easily failed to be rejected for other methods even though this variable is significant. However, our structure can avoid this because there would be a drop in covariance between the group of significant and the non-significant ones. Clearly, the hybrid sequential multiple testing procedure has more advantage than other testing procedures.

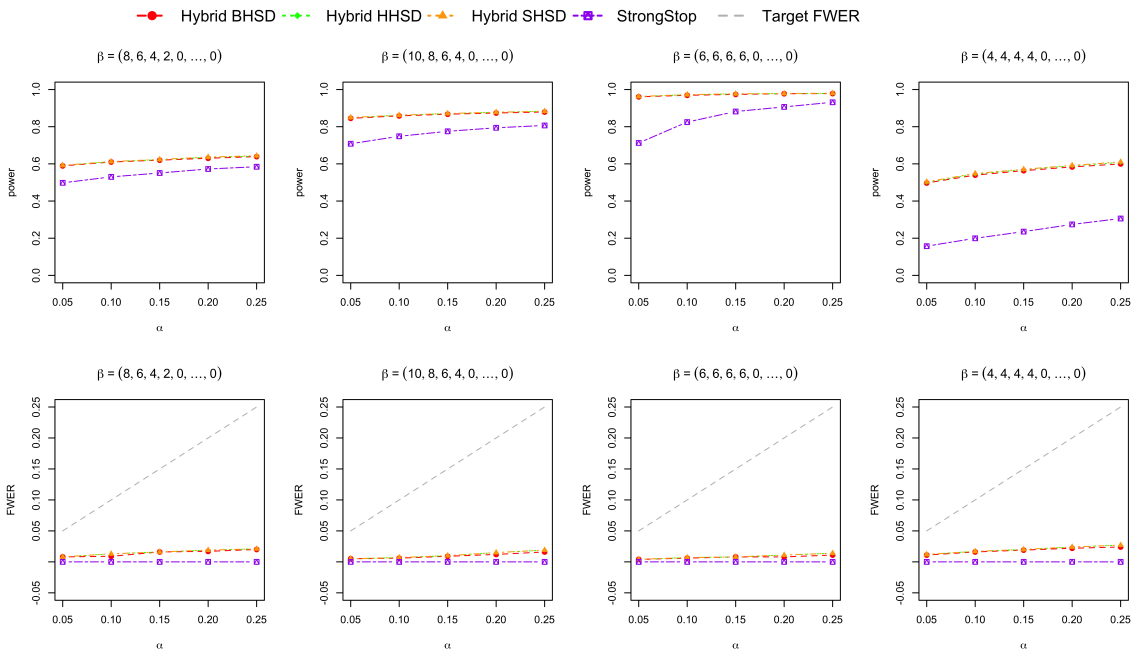


Figure 5.1: Average power and observed FWER for Hybrid Bonferroni-Holm Step-Down, Hybrid Hochberg-Holm Step-Down, Hybrid Simes-Holm Step-Down and StrongStop procedures for $\alpha \in \{0.05,\ 0.1,\ 0.15,\ 0.2,\ 0.25\}$.

Figure 5.1 summaries the performance of Hybrid Bonferroni (Hochberg, Simes)-Holm Step-Down, and StrongStop methods for sequential multiple testing for same coefficients over 1000 simulations. The target FWER is $\alpha \in \{0.05,\ 0.1,\ 0.15,\ 0.2,\ 0.25\}$.

All these methods are able to control FWER at their target level, with our structure having the highest power than others. The Hybrid Bonferroni-Holm Step-Down, Hybrid Hochberg-Holm Step-Down and Hybrid Simes-Holm Step-Down have very similar power, and it is difficult to clearly separate them from the plot. For equally weak signals, the performance of StrongStop is very poor and could not identify the significant variables efficiently. Therefore, our procedures are very competitive for the orthogonal design.

Moreover, we implement our procedures for a general predictor matrix. The simulations show performance for different coefficients and different correlation structures. The covariates from model e) and f) are generated from $X \sim i.i.d.N(0, 1/n),\ \sigma = 1$. The covariates from model g) and h) are generated from the multivariate normal distribution with correlation $cor(x_i, x_j) = 0.5^{|i-j|},\ \sigma = 1$. All setups have $n = 200,\ m = 20$ and $\alpha = 0.05$. Among the 20 variables, 4 variables are significant and the other 16 variables are non-significant. There are two kinds of different model coefficients, for model e) and g), non-zero variables with $\beta = (10,\ 8,\ 6,\ 4,\ 0,\ \cdots,\ 0)$; For model f) and h), non-zero variables with $\beta = (6,\ 6,\ 6,\ 6,\ 0,\ \cdots,\ 0)$. We simulate 1000 times.

Table 5.2: Comparison among model e - h for FWER control.

	e		f		g		h	
	Power	FWER	Power	FWER	Power	FWER	Power	FWER
Hybrid BHSD	0.8343	0.0090	0.9443	0.0110	1.0000	0.0580	1.0000	0.0540
Hybrid HHSD	0.8380	0.0120	0.9457	0.0110	1.0000	0.0610	1.0000	0.0600
Hybrid SHSD	0.8383	0.0120	0.9457	0.0110	1.0000	0.0610	1.0000	0.0600
StrongStop	0.7817	0.0000	0.8600	0.0010	1.0000	0.0000	1.0000	0.0000

The performance of the highly correlated design matrix models (i.e. model g and h) are generally better and all methods are able to detect the important variables with

power approaching to 1. Sometimes the FWER is slightly greater than the desired rate, and this may be due to the simulation error and also because of the more hypotheses we test, the worse the empirical distributions get as compared to the asymptotic distribution. For the independent generated design matrix (i.e. model e and f), our procedure also has higher power than StrongStop.

We also apply the procedure in high-dimensional data, with very sparse structure. The simulation is set as follows: let $n = 5000,\ m = 5000,\ \alpha = 0.05$. Consider general gaussian $\mathbf{X}$, with $\mathbf{x_i} \sim i.i.d.\ N(0, 1/n)$, $\sigma = 1$. The true significant number of variables $k_0 =$ 0, 10, 20, 30, 40, 50. Non-zero variables for weak signal with coefficient $\beta_i = \sqrt{2\log(m)} \approx$ 4.127; Non-zero variables for medium signal with coefficient $\beta_i = 2\sqrt{2\log(m)} \approx 8.255$; Non-zero variables for strong signal with coefficient $\beta_i = 5\sqrt{2\log(m)} \approx 20.636$. Simulate 500 times. The non-zero variables are very sparse $(0\% - 1\%)$. This set up is the same as the Setup 1 for FDR control in Chapter 6.

Figure 5.2 shows the result for the high-dimensional simulation. All these four methods are able to control FWER at $\alpha = 0.05$. For the median and strong signal, all the four methods control the FWER very conservatively, while for the weak signal, our three methods get close to the threshold $\alpha = 0.05$ for 50 non-zero variables. However, all our three methods have much higher power than StrongStop, even when the signals are weak and non-separable from the noise.

In most of the practical problems, the number of erroneous selections should be taken into account and not only whether any error is made. It is not reasonable and practical to control whether there is a non-significant variable chosen when there are thousands or

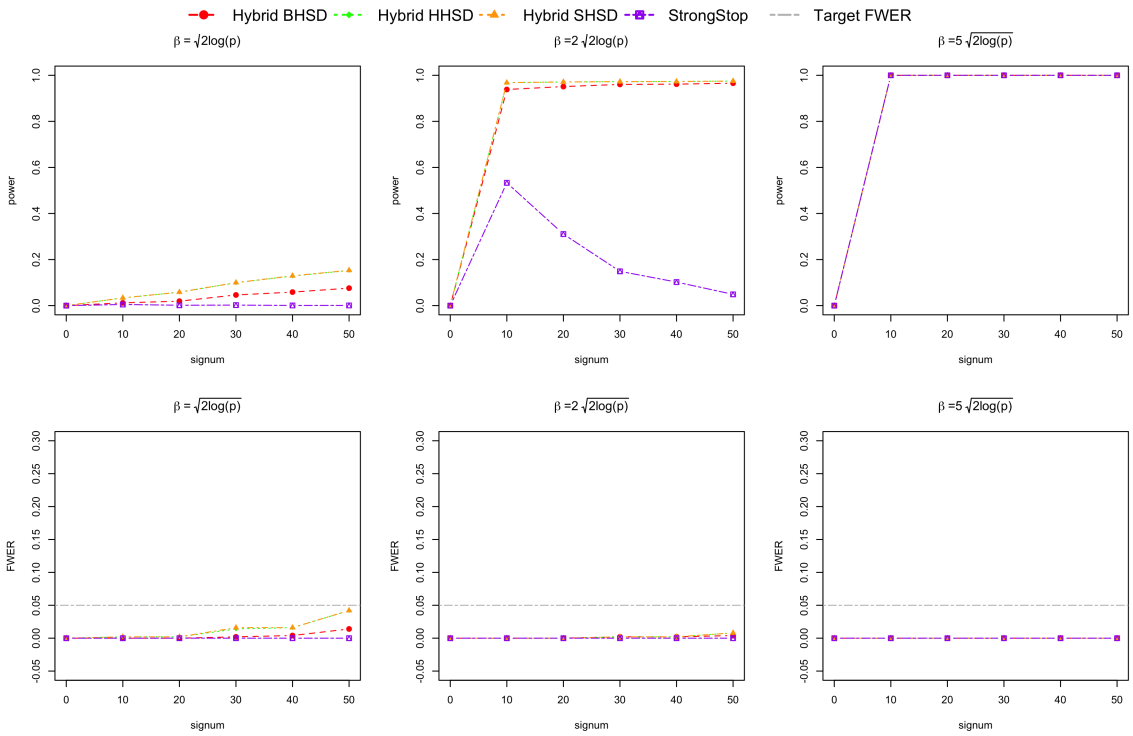


Figure 5.2: High-dimensional: Average power and observed FWER for the sequential multiple testing at α=0.05. Hybrid Bonferroni-Holm Step-Down, Hybrid Hochberg-Holm Step-Down, Hybrid Simes-Holm Step-Down and StrongStop methods are evaluated.

more variables in the model. When there are amount of hypotheses, controlling the rate of none error been made is very stringent, and will at the expense of power. Ever since the definition of false discovery rate was proposed, it becomes more desirable in large-scale multiple testing problems, especially in microarray gene expression studies. The concept of FDR not only provides the overall type I error rate control but also allows higher statistical power than FWER. Therefore, the error rate control for large-scale and high-dimensional data will be saved for FDR control.

Chapter 6

Hybrid sequential multiple testing to control FDR

False discovery rate (FDR) control is another multiple testing problem for type I error control, which is more often used now. FDR procedures are designed to control the expected proportion of incorrectly selected non-significant variables. FDR controlling procedures provide less stringent control compared to FWER procedures. It is equivalent to FWER when all the null hypotheses are true. FDR procedure is formally proposed by Benjamini (1995), followed by Benjamini and Yekutieli (2001) who considered the positive dependence property of the test statistics. Benjamini (1999) proposed the FDR Step-Down procedure comparing the ordered p-value with $\delta_k = 1 - \left[1 - \min(1, \frac{m}{m-k+1}q)\right]^{1/(m-k+1)}$.

6.1 FDR control procedure

In this section, the FDR control procedure is also based on our sequential multiple testing structure, which described in Section 3 and 4. The hypothesis and variables been tested at each step are inherited from Section 4. The composite p-values at each step have the increasing property and the order of the composite p-values is consistent with the order of the hypothesis. Similar as FWER control procedure in Section 5, in FDR control we implement the Benjamini-Liu (BL) Step-Down procedure in the sequential multiple testing structure, which gives the other main procedure:

Procedure 2 (Hybrid Bonferroni-BL Step-Down) *Define the critical values as:*

$$\delta^k = 1 - \left[1 - \min\left(1, \frac{m}{m-k+1}q\right)\right]^{\frac{1}{m}}.$$

We have $0 \leq \delta^1 \leq \delta^2 \leq \cdots \leq \delta^m \leq 1$.

Then the hybrid sequential multiple testing procedure 2 for FDR control work as follows:

Step 1: *test* $H_{01} : \beta_{\{1,\cdots,m\}\setminus A|_{y,X,k=1}} = 0$*;*

If $P^1 \leq \delta^1$*, reject* H_{01}*, continue to Step 2;*

Otherwise, fail to reject, stop; variable selected $\hat{D}_n = \{\emptyset\}$*;*

Step 2: *test* $H_{02} : \beta_{\{1,\cdots,m\}\setminus A|_{y,X,k=2}} = 0$*;*

If $P^2 \leq \delta^2$*, reject* H_{02}*, continue to Step 3;*

Otherwise, fail to reject, stop; $\hat{D}_n = \{1^{st}\}$*;*

...

Step k***:*** *test* $H_{0k} : \beta_{\{1,\cdots,m\}\setminus A|_{y,X,k}} = 0$*;*

If $P^k \leq \delta^k$*, reject* H_{0k}*, continue to Step* $k+1$*;*

Otherwise, fail to reject, stop; $\hat{D}_n = \{1^{st}, \cdots, (k-1)^{th}\}$*;*

...

Step m***:*** *test* $H_{0m} : \beta_{mth} = 0$*;*

If $P^m \leq \delta^m$*, reject* H_{0m}*, stop;* $\hat{D}_n = \{1, \cdots, m\}$*;*

Otherwise, fail to reject, stop; $\hat{D}_n = \{1^{st}, \cdots, (m-1)^{th}\}$*.*

Theorem 4 *The Hybrid Bonferroni-BL Step-Down (Hybrid BBLSD) procedure controls FDR at level* q*.*

Theorem 5 *If we use predetermined* s *hypotheses at each step for all the steps, and use Hybrid SD FDR procedure with the same critical values* $\delta^k = 1 - \left[1 - \min\left(1, \frac{m}{m-k+1}q\right)\right]^{\frac{1}{m}}$*, we also have* $FDR_s \leq q$*.*

Sometimes it may not be necessary to test all the hypotheses at each step, especially for ultra high-dimensional data. Firstly, it will reduce the computing resources; Secondly, the mean of the exponential family Exp(1), Exp(1/2), Exp(1/3), Exp(1/4) are $1, 1/2, 1/3, 1/4$, which decreases to 0, then it is more difficult to be significant; Moreover, the empirical distribution is deviate from the asymptotic distribution pointed by Fan and Ke (2014), similarly as G'Sell et al.(2015) suggested. Also because the lasso procedure chooses variables from most significant ones to the least significant ones, the latter variables are prone to be non-significant. If we choose s hypotheses at each step, say 20, the

Hybrid Bonferroni-BL Step-Down procedure also controls at q. Both proof of Theorem 4 and Theorem 5 are in Appendix.

6.2 Other methods

In this section, we introduce some recently proposed procedures for FDR control in selection inference and connect them with our procedures. Later on, comparisons will be conducted through simulations.

6.2.1 ForwardStop

ForwardStop (G'Sell et al. 2015) is proposed to control ordered FDR in variable selection. It is motivated by the problem of selecting stopping rules for model selection procedures such as the lasso, stepwise regression, and hierarchical clustering.

ForwardStop transforms the p-values and sets the rejection threshold at the largest k for which the first k transformed p-values have a small enough average. Let $\tilde{Y}_k = -log(1-p_k)$, where p_k is the p-value for single hypothesis under the null hypothesis and $\hat{k} = \max\{k \in \{1, \cdots, m\} : \frac{1}{k}\sum_{i=1}^{k} Y_i \leq q\}$.

ForwardStop scans the p-values in a forward manner: if $\frac{1}{k}\sum_{i=1}^{k} Y_i \leq \alpha$, then we know that we can reject the first $\hat{k}$ hypotheses regardless of the remaining p-values. Forward rejection rule is less robust to model misspecification comparing to StrongStop.

6.2.2 StrongStop

StrongStop (G'Sell et al. 2015) is proposed to control FWER and FDR. Under the global null, we could just as well obtain uniform order statistics q_i by summing from the back: let $\tilde{Y}_k = -log(p_k)$, $\tilde{Z}_k = \sum_{j=k}^{m} Y_j/j$, $\tilde{q}_k = e^{-\tilde{Z}_k}$. Reject hypotheses $1, \cdots, \hat{k}$ (select the first $\hat{k}$ variables), where $\hat{k} = \max\{k \in \{1, \cdots, m\} : \tilde{q}_k \leq \frac{qk}{m}\}$. Similarly as ForwardStop, it transforms the p-values to keep $\tilde{q}$'s in an increasing order.

If we apply the lasso covariance test, where $p_k = exp(-T_k)$, StrongStop can be simplified as $\tilde{Y}_k = T_k$, $\tilde{Z}_k = \sum_{j=k}^{m} T_j/j$ and $\tilde{q}_k = e^{-\tilde{Z}_k}$.

Unlike ForwardStop, StrongStop needs to look at the p-values corresponding to the last hypotheses before making any rejections. The weakness of StrongStop is that the decision to reject at kth step depends on the p-values after this variables. From intuition, it is not how we make decisions. Also if the p-values in later steps are larger than the uniform or deviate from the uniform distribution, it will suffer power loss and be unstable.

6.2.3 TailStop

TailStop (G'Sell et al. 2015) is an extension of StrongStop and is applied for covariance test of lasso. Under the asymptotic distribution of the covariance test, we have $jT_{k0+j} \sim \text{EXP}(1)$. Assume that we know k_0 and incorporate the StrongStop rule for the $m - k_0$ null test statistics, this would suggest a test based on

$$q_i^* = exp\left[-\sum_{j=i}^{m} \frac{\max\{1, j-k_0\}}{j} T_j\right].$$

The above equation depends on the knowledge of k_0. Now set $k_0 = 0$, giving

$$q_i^* = exp\left[-\sum_{j=i}^{m} T_j\right]. \tag{6.1}$$

TailStop would gain power comparing with StrongStop since it applies the asymptotical distribution of covariance test. However, the assumption will not be hold when the number of significant variables is large.

Now the TailStop works as follows: define q_i^* as (6.1). Reject hypotheses $1, \cdots, \hat{k}$ (select the first $\hat{k}$ variables), where $\hat{k} = \max\{k \in \{1, \cdots, m\} : q_k^* \leq \frac{qk}{m}\}$.

At the ith step, the pivotal statistics is $\sum_{j=i}^{m} T_j$. Compared with our hybrid sequential at the ith step, the p-value is measured by $p^i = \min\{p_j\}, j \in \{i, \cdots, m\}$, where $T_j \sim \text{EXP}(1/(j-i+1))$ and $p_i = 1 - pexp(T_j, \lambda = j-i+1) = exp(-(j-i+1)T_j)$. Thus, our pivotal statistics is $(j-i+1)T_j$. Interestingly, there is a connection between TailStop and Hybrid sequential pivotal statistics under the global null hypothesis, stated in the following Lemma.

Lemma 6 *Under the global null, the pivotal statistics for TailStop has the same distribution as the pivotal statistics for Hybrid sequential.*

Next is the summary distribution of the pivotal statistics for both two methods.

Assume $k_0 = 0$, at the 1^{st} step, and let $\text{TS}_1 = \sum_{j=1}^{m} T_j$ be short for the 1^{st} pivotal statistics for TailStop, then

$$f_{\text{TS}_1}(x) = me^{-x}(1 - e^{-x})^{m-1};$$

Let $\text{HS}_1 = \max_{j \in \{1, \cdots, m\}}\{jT_j\}$ be short for the 1^{st} pivotal statistics for Hybrid sequential

method, then

$$f_{\mathrm{HS}_1}(x) = me^{-x}(1-e^{-x})^{m-1}.$$

Assume $k_0 > 0$, at the (k_0+1)th step, let $\mathrm{TS}_{k_0+1} = \sum_{j=k_0+1}^{m} T_j$ be short for the (k_0+1)th pivotal statistics for TailStop, then

$$f_{\mathrm{TS}_{k_0+1}}(x) = m\binom{m-1}{k_0}e^{-(k_0+1)x}(1-e^{-x})^{m-(k_0+1)};$$

Let $\mathrm{HS}_{k_0+1} = \max_{j\in\{k_0+1,\cdots,m\}}\{(j-k_0)T_j\}$ be short for the (k_0+1)th pivotal statistics for Hybrid sequential method, then,

$$f_{\mathrm{HS}_{k_0+1}}(x) = (m-k_0)e^{-x}(1-e^{-x})^{m-(k_0+1)}.$$

The details of the calculation and Lemma 6 are showed in Appendix D.

From the density function of their statistics under the global null hypothesis, surprisingly they have the same distribution for different statistics while the realization of the two statistics will be different for each sampling of $T_j,\ j \in \{1,\ \cdots,\ m\}$. When the number of significant variables $k_0 > 0$, they have different density functions. However, when m is large, the difference of Hybrid sequential pivotal statistics are small among steps, while for TailStop, the density functions change dramatically among steps. The figure 6.1 next will show the density functions for both of the two pivotal statistics under the null hypothesis at $k_0 = 0,\ 1,\ 2,\ 5$ for $m = 100$.

In the first plot of Figure 6.1, the pivotal statistics for both method have the same distributions. Then the cutoff values will be the same to obtain the p-value, which means that those two methods are equivalent under the global null distribution.

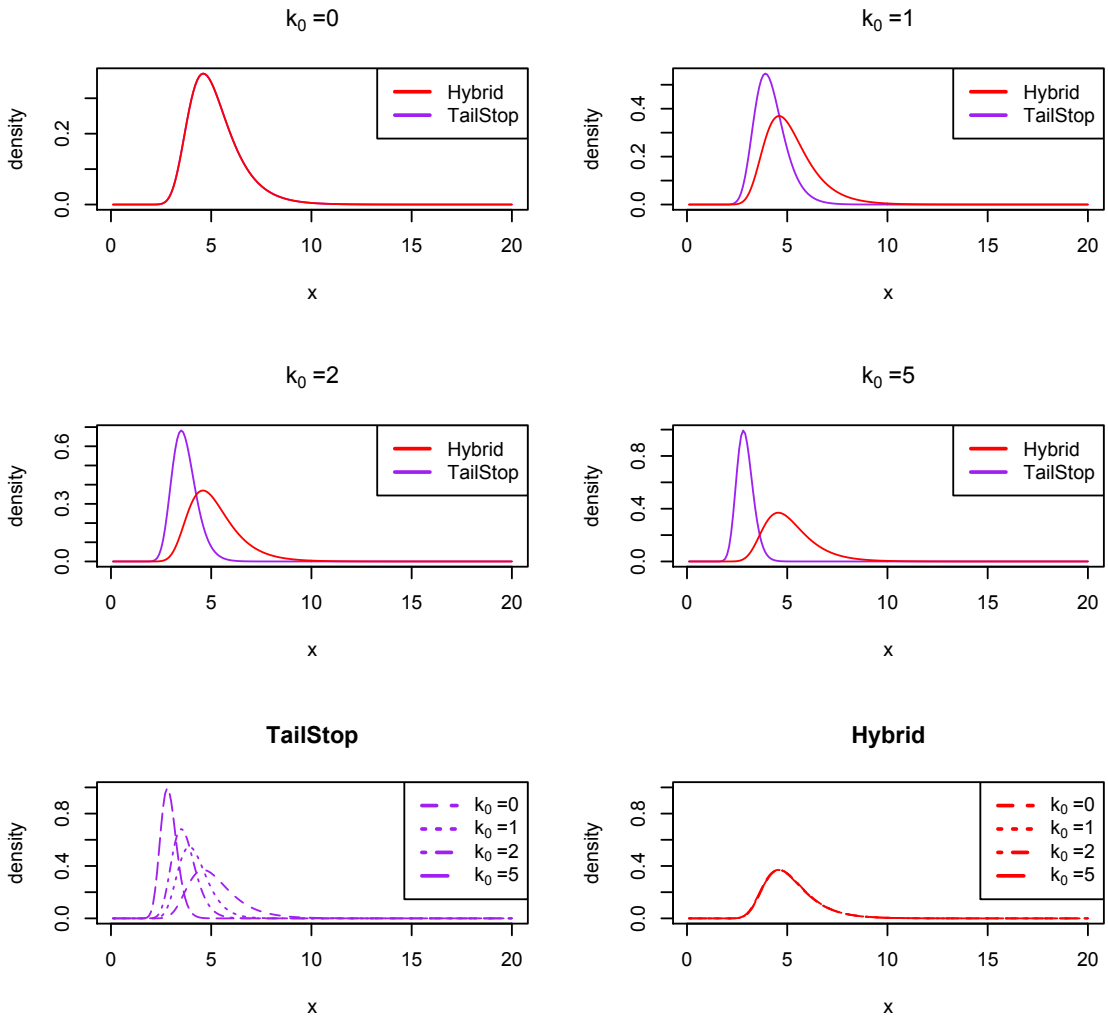


Figure 6.1: Density functions of pivotal statistics for TailStop and Hybrid sequential under null distribution at $k_0 = 0, 1, 2, 5$ for $m = 100$.

The next three plots in Figure 6.1 show when $k_0 > 0$, the density functions of the pivotal statistics for TailStop are sharper than Hybrid Sequential. For TailStop, $\text{TS}_2 = \sum_{j=2}^{m} T_j$, where $T_2 \sim \text{Exp}(\lambda = 2), T_3 \sim \text{Exp}(\lambda = 3)$. For Hybrid Sequential, $\text{HS}_2 = \max_{j \in \{2,\cdots,m\}}\{(j-1)T_j\}$, where $T_2 \sim \text{Exp}(\lambda = 1), T_3 \sim \text{Exp}(\lambda = 2)$. The null distributions are not the same for these two methods, and the density functions will have

smaller expectation for TailStop, where the assumption $k_0 = 0$ for TailStop deviate from the true distribution. It will be more complicated for the combination of pivotal statistics and cutoff p-values for both of these two procedures.

From last two subplots of Figure 6.1 where each method is compared with different k_0, the Hybrid sequential distributions are stable for different k_0. But the density functions of TailStop are getting sharp for larger k_0, which makes sense since the mean of summation of T_j is getting smaller. The assumption of $k_0 = 0$ in TailStop deviates more from the true distribution.

The TailStop procedure uses Benjamini-Hochberg (BH) cutoff value qk/m, here we denote it as TailStop-BH, while our Hybrid BBLSD applies adjusted Benjamini-Liu (BL) cutoff value δ^*. The cutoff value for BH is usually smaller than BL cutoff value, then the power potentially would be larger for TailStop-BH than our Hybrid BBLSD procedure. To make a fair comparison, we also use the cutoff value δ^* in TailStop, denoted as TailStop-BL, to compare the performance with our Hybrid BBLSD procedure in the next section.

6.2.4 SLOPE

The methodology of SLOPE (Bogdan et al. 2013, 2015) is inspired by the Benjamini-Hochberg (BHq) procedure for controlling the false discovery rate in multiple testing. For the FDR control under orthogonal design, the SLOPE solution reduces to

$$\min_{\beta \in R^m} \frac{1}{2}||\tilde{y} - \beta||_{l_2}^2 + \sum_{i=1}^{m} \lambda_i |\beta|_{(i)},$$

where $\tilde{y} = X'y \sim N(\beta, I_m)$. For testing the m hypotheses $H_i : \beta_i = 0$, the BHq step-up procedure proceeds as follows:

1) Sort the entries of $\tilde{y}$ in decreasing order of magnitude, $|\tilde{y}|_{(1)} \geq |\tilde{y}|_{(2)} \geq \cdots \geq |\tilde{y}|_{(m)}$ (this yields corresponding order hypotheses $H_{(1)}, \cdots, H_{(m)}$).

2) Find the largest index i such that

$$|\tilde{y}|_{(i)} > \lambda_{BH}(i) = \Phi^{-1}(1 - q_i), \quad q_i = q\frac{i}{2m},$$

where $\Phi^{-1}(\alpha)$ is the αth quantile of the standard normal distribution and q is a parameter in $[0, 1]$. Call this index i_{SU}.

3) Reject all $H_{(i)}$'s for which $i \leq i_{SU}$.

Under the orthogonal design, the SLOPE procedure controls FDR at level q.

Selecting a sequence $\{\lambda_i\}$ for use in SLOPE is an interesting research topic. In an orthogonal design, we can apply $\lambda_i = \lambda_{BH}(i)$. However, this choice is too liberal for general design, therefore, they proposed a method to adjust the regularizing sequence for the variance inflation. The estimate of nonzero components is roughly equal to

$$(X_S'X_S)^{-1}(X_S'y - \lambda_S),$$

where S is the support set $\{1, \cdots, k\}$ and $\lambda_S = (\lambda_{BH}(1), \cdots, \lambda_{BH}(k))'$ causing a bias approximately equal to

$$EX_S(\beta_S - \hat{\beta}_S) \approx X_S(X_S'X_S)^{-1}\lambda_S$$

For a gaussian design where the entries of X are i.i.d. $N(0, 1/n)$, the size of the variance inflation is

$$E(X_i'X_S(X_S'X_S)^{-1}\lambda_S)^2 = \frac{1}{n}\lambda_S'E(X_S'X_S)^{-1}\lambda_S = w(|S|) \cdot ||\lambda_S||_{l_2}^2, \quad w(k) = \frac{1}{n-k-1},$$

where the last equality uses the fact that the expected value of an inverse dimensional Wishart of dimension k with n degrees of freedom is equal to $I_k/(n-k-1)$.

They suggest the sequence of λ's described below denoted by λ_G since it is motivated by Gaussian designs. We start with $\lambda_G(1) = \lambda_{BH}(1)$. At the next stage, however, we need to account for the slight increase in variance so that we do not want to use $\lambda_{BH}(2)$ but rather

$$\lambda_G(2) = \lambda_{BH}(2)\sqrt{1 + w(1)\lambda_G(1)^2}.$$

Continuing, this gives

$$\lambda_G(i) = \lambda_{BH}(i)\sqrt{1 + w(i-1)\sum_{j<i}\lambda_G(j)^2}. \tag{6.2}$$

$\lambda_G(i)$ may no longer be decreasing. It would not make sense to use such a sequence to keep SLOPE convex, and let $k^* = k(n, m, q)$ be the location of the global minimum, they suggest to work with

$$\lambda_{G^*}(i) = \begin{cases} \lambda_G(i), & i \leq k^*, \\ \lambda_{k^*}, & i > k^*, \end{cases} \quad \text{with } \lambda_G(i) \text{ as in (6.2).}$$

The value k^* will play a role on the overall performance of the method.

When the number of non-nulls (the number of nonzero coefficients in the regression model) is not too large, SLOPE controls the FDR at a reasonable level. The above adjustment is informal and this holds provided that the columns of the design matrix are not strongly correlated. At the same time, the method has more power in detecting true regressors than the lasso and other competing methods.

6.3 Simulation results for FDR control

The simulations in this section are based on our proposed procedure and we compare its result with ForwardStop, StrongSop, TailStop-BH, TailStop-BL, and SLOPE for orthogonal and general design matrix.

First, in model (a - d) the orthogonal design matrices we considered for FDR simulation are the same as before for FWER. Let $n = 200,\ m = 20,\ \sigma = 1,\ q = 0.05$. Among the 20 variables, first 4 variables are significant and the following 16 are non-significant. We simulate 1000 times for different coefficients of true β with results showing in Table 6.1. Four different model coefficients are: a) $\beta = (8,\ 6,\ 4,\ 2,\ 0,\ \cdots,\ 0)$; b) $\beta = (10,\ 8,\ 6,\ 4,\ 0,\ \cdots,\ 0)$; c) $\beta = (6,\ 6,\ 6,\ 6,\ 0,\ \cdots,\ 0)$; d) $\beta = (4,\ 4,\ 4,\ 4,\ 0,\ \cdots,\ 0)$.

Table 6.1: Comparison among model a - d for FDR control.

	a		b		c		d	
	Power	FDR	Power	FDR	Power	FDR	Power	FDR
Hybrid BBLSD	0.6857	0.0141	0.9133	0.0112	0.9895	0.0100	0.7035	0.0160
ForwardStop	0.6382	0.0135	0.8337	0.0250	0.3157	0.0062	0.1765	0.0004
StrongStop	0.4983	0.0000	0.7077	0.0000	0.7133	0.0000	0.1573	0.0000
TailStop-BH	0.7632	0.0209	0.9702	0.0217	0.9990	0.0215	0.8705	0.0194
TailStop-BL	0.7342	0.0072	0.9525	0.0070	0.9980	0.0072	0.8007	0.0065
SLOPE	0.7920	0.0420	0.9788	0.0422	0.9992	0.0421	0.9067	0.0425

From Table 6.1 model a - d, we can clearly see that all the methods can control FDR at $q = 0.05$ and our Hybrid BBLSD procedure has the relatively higher power for all the four cases. ForwardStop and StrongStop have poor performance with equally weak signal (i.e. model c and d), with StrongStop better than ForwardStop when the signal is equally larger (i.e. model c). Both TailStop-BH and TailStop-BL have fairly large power, with BH cutoff value has higher power as expected. Among all the procedures, the SLOPE

has the highest power with FDR close to q for small scale cases and independent entries of the design matrix.

Figure 6.2 shows simulations in terms of different error rate, $q \in \{0.05,\ 0.10,\ 0.15,\ 0.20,\ 0.25\}$ for all the four true models. The non-zero coefficients are the same as model a - d. All the models have $m = 20$ variables with 4 significant variables and 16 non-significant ones. The simulation summarizes the performance of Hybrid BBLSD, ForwardStop, StrongStop, TailStop-BH and TailStop-BL, SLOPE procedures for sequential multiple testing over 1000 simulations.

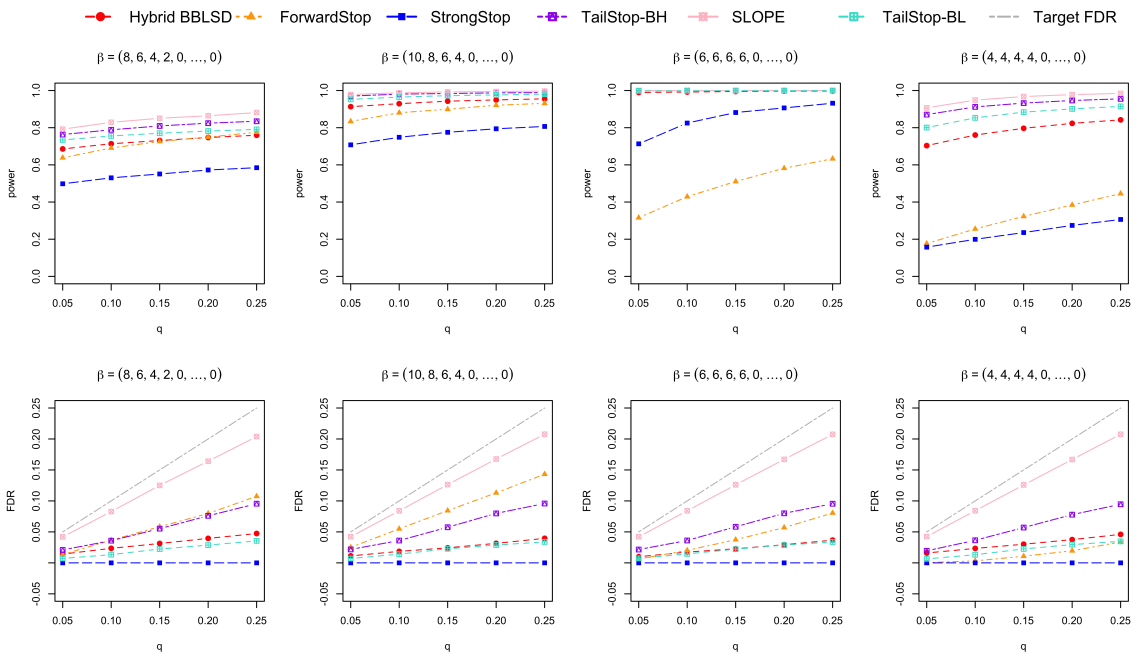


Figure 6.2: Average power and observed FDR for Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH, TailStop-BL and SLOPE procedures for $q \in \{0.05,\ 0.1,\ 0.15,\ 0.2,\ 0.25\}$.

All these methods are able to control FDR at their target level, with Hybrid Bonferroni-BL Step-Down having comparatively higher power. For equally weak signals,

ForwardStop and StrongStop have the lowest power, which is consistent with FWER control. SLOPE consistently has the highest power among all and with FDR close to the target q for $n = 200,\ m = 20$ orthogonal design.

Next, we implement our procedures for more general predictor matrix. The simulations show performance for different coefficients and different correlation structures. The covariates from model e and f are generated from $X \sim i.i.d.\ N(0, 1/n),\ \sigma = 1$. The covariates from model g and h are generated from multivariate normal distribution with correlation $cor(x_i, x_j) = 0.5^{|i-j|}$, $\sigma = 1$. All the models have $n = 200,\ m = 20$ and $q = 0.05$. Among the 20 variables, 4 variables are significant and 16 are non-significant. There are two kinds of different model coefficients. For model e and g, non-zero variables have coefficient $\beta = (10,\ 8,\ 6,\ 4,\ 0,\ \cdots,\ 0)$; For model f and h, non-zero variables have coefficitent $\beta = (6,\ 6,\ 6,\ 6,\ 0,\ \cdots,\ 0)$. The results of 1000 simulations are shown in Table 6.2.

Table 6.2: Comparison among model e - h for FDR control.

	e		f		g		h	
	Power	FDR	Power	FDR	Power	FDR	Power	FDR
Hybrid BBLSD	0.9113	0.0202	0.9812	0.0194	1.0000	0.0690	1.0000	0.0704
ForwardStop	0.7582	0.0185	0.3915	0.0090	1.0000	0.0413	0.9970	0.0479
StrongStop	0.6947	0.0000	0.7087	0.0000	1.0000	0.0000	1.0000	0.0000
TailStop-BH	0.9655	0.0238	0.9975	0.0260	1.0000	0.0316	1.0000	0.0395
TailStop-BL	0.9470	0.0087	0.9945	0.0116	1.0000	0.0118	1.0000	0.0156
SLOPE	0.9663	0.0367	0.9968	0.0401	1.0000	0.7614	1.0000	0.7610

From Table 6.2 model e - h, the FDR is higher when the covariates are more correlated and so does the power. For model e and f, our procedure controls FDR very well. The ForwardStop performs the worset for an independent design matrix (i.e. model f). The StrongStop is consistently better than ForwardStop for all the cases. Our methods

has fairly large power for all the cases and with slightly lower power than TailStop-BH and TailStop-BL. The SLOPE procedure loses control of FDR for the correlate design matrix (both model g and model h) and tends to choose more variables than supposed.

From these simulations, we clearly see SLOPE only works better for the orthogonal design and more independent entries from normal distributions, and the FDR will be hugely affected if there are moderate correlations among the design matrix. The possible reason in SLOPE is that the idea of the adjusted λ_i is derived marginally from the orthogonal analysis adjusting by the variance inflation factor, without considering the other related variables. Also, lasso-typed penalty for the variable selection is not an ideal choice to deal with the multicollinearity design matrix, so it is one of the reasons for lasso related variable selection procedures can lose control of FDR a little.

6.4 Simulation results for FDR control in large-scale and high-dimensional data

FDR control is more important and realistic for large scale and high-dimensional data than FWER. It allows the model to have false positives if there are too many variables in the model. A lot of research has changed gears towards an interest in FDR control instead of purely looking at FWER. Moreover, lasso and other high-dimensional variable selection is applicable to large scale data for either large n or large m, which the traditional statistical methods may not be able to handle. For high-dimensional data we often require sparsity. The simulations we run next are to show how sparsity affects the FDR control for different examples.

We have the following four setups for large scale and high-dimensional data and same four setups for correlated design matrix in our simulation study.

Setup 1: Let $n = 5000,\ m = 5000,\ n = m$. Consider general gaussian X, with $\mathbf{x_i} \sim$ *i.i.d.* $N(0, 1/n)$, $\sigma = 1$. The true significant number of variables $k_0 = 0,\ 10,\ 20,\ 30,\ 40,\ 50$. Non-zero variables for weak signal with coefficient $\beta_i = \sqrt{2\log(m)} \approx 4.127$; Non-zero variables for medium signal with coefficient $\beta_i = 2\sqrt{2\log(m)} \approx 8.255$; Non-zero variables for strong signal with coefficient $\beta_i = 5\sqrt{2\log(m)} \approx 20.636$. Simulate 500 times. Set $q = 0.01, 0.02$ and 0.05 for Setup 1.

Setup 2: Let $n = 5000,\ m = 5000,\ n = m$. Consider general gaussian X, with $\mathbf{x_i} \sim$ *i.i.d.*$N(0, 1/n)$, $\sigma = 1$. The true significant number of variables $k_0 = 0,\ 50,\ 100,\ 150,\ 200,\ 300$. Non-zero variables for weak signal with coefficient $\beta_i = \sqrt{2\log(m)} \approx 4.127$; Non-zero variables for medium signal with coefficient $\beta_i = 2\sqrt{2\log(m)} \approx 8.255$; Non-zero variables for strong signal with coefficient $\beta_i = 5\sqrt{2\log(m)} \approx 20.636$. Simulate 500 times of all hypotheses at each step. Set $q = 0.02$ and 0.05 for Setup 2 - 4.

Setup 3: Let $n = 5000,\ m = 10000,\ n = 0.5m$. Consider general gaussian X, with $\mathbf{x_i} \sim$ *i.i.d.* $N(0, 1/n)$, $\sigma = 1$. The true significant number of variables $k_0 = 0,\ 50,\ 100,\ 150,\ 200,\ 300$. Non-zero variables for weak signal with coefficient $\beta_i = \sqrt{2\log(m)} \approx 4.292$; Non-zero variables for medium signal with coefficient $\beta_i = 2\sqrt{2\log(m)} \approx 8.584$; Non-zero variables for strong signal with coefficient $\beta_i = 5\sqrt{2\log(m)} \approx 21.460$.

Setup 4: Let $n = 5000,\ m = 2500,\ n = 2m$. Consider general gaussian X, with $\mathbf{x_i} \sim$ *i.i.d.* $N(0, 1/n)$, $\sigma = 1$. The true significant number of variables $k_0 = 0,\ 50,\ 100,\ 150,\ 200,\ 300$. Non-zero variables for weak signal with coefficient $\beta_i = \sqrt{2\log(m)} \approx 3.956$;

Non-zero variables for medium signal with coefficient $\beta_i = 2\sqrt{2\log(m)} \approx 7.911$; Non-zero variables for strong signal with coefficient $\beta_i = 5\sqrt{2\log(m)} \approx 19.779$.

Setup corr1 - corr4: Same setup as Setup 1 - 4 correspondingly. The only difference is to consider entries of general gaussian $cor(x_i, x_{i+1}) = 0.5$, otherwise 0, and $\sigma = 1$.

Setup 1 - 2 (Setup corr1- corr2) are for large scale of $n = 5000$ and the non-zero variables are very sparse $(0\% - 1\%)$ and $(0\% - 6\%)$. And Setup 3 (Setup corr3) is for high-dimensional with $m = 10000 = 2n$ and the non-zero variables have sparsity $0\% - 3\%$. Setup 4 (Setup corr4) is for large scale with $n = 2m = 5000$ and the non-zero variables have sparsity $0\% - 12\%$. We choose non-zero variables with coefficient comparing with $\sqrt{2\log(m)}$, which is the smallest non-zero coefficient lasso can identify.

Figure 6.3 shows the result for Setup 1 when we set $q = 0.01$. All these six methods are able to control FDR at $q = 0.01$. For the weak signals, the power gets higher when there are more variables significant in the model, up to 24.0% for Hybrid BBLSD and 9.9% for TailStop-BH, 5.8% for TailStop-BL and 40.2% for SLOPE, while ForwardStop and StrongStop procedures are very poor. The TailStop-BL has smaller power than TailStop-BL as expected since it has smaller cutoff value. For the same Benjamini-Liu cutoff value comparison between Hybrid BBLSD and TailStop-BL, our Hybrid BBLSD has consistent higher power than TailStop-BL procedure. When the number of significant variables is 10, Hybrid BBLSD, TailStop-BH and TailStop-BL have similar power, with Hybrid BBLSD slightly lower. However, Hybrid BBLSD exceeds TailStop when the number of significant variables gets larger. Among all the procedures, SLOPE has the highest power. For the medium signals, the power gets close to 1 for Hybrid BBLSD, TailStop-BH, TailStop-BL and

SLOPE procedures. They are able to capture all the significant variables for the covariance drop. The performance of ForwardStop and StrongStop are still poor. For the strong signals, the power of Hybrid BBLSD, TailStop-BH, TailStop-BL and SLOPE are 1. The power of ForwardStop is higher than for weak and medium signal. For the StrongStop, the power decreases as the number of significant variables increases.

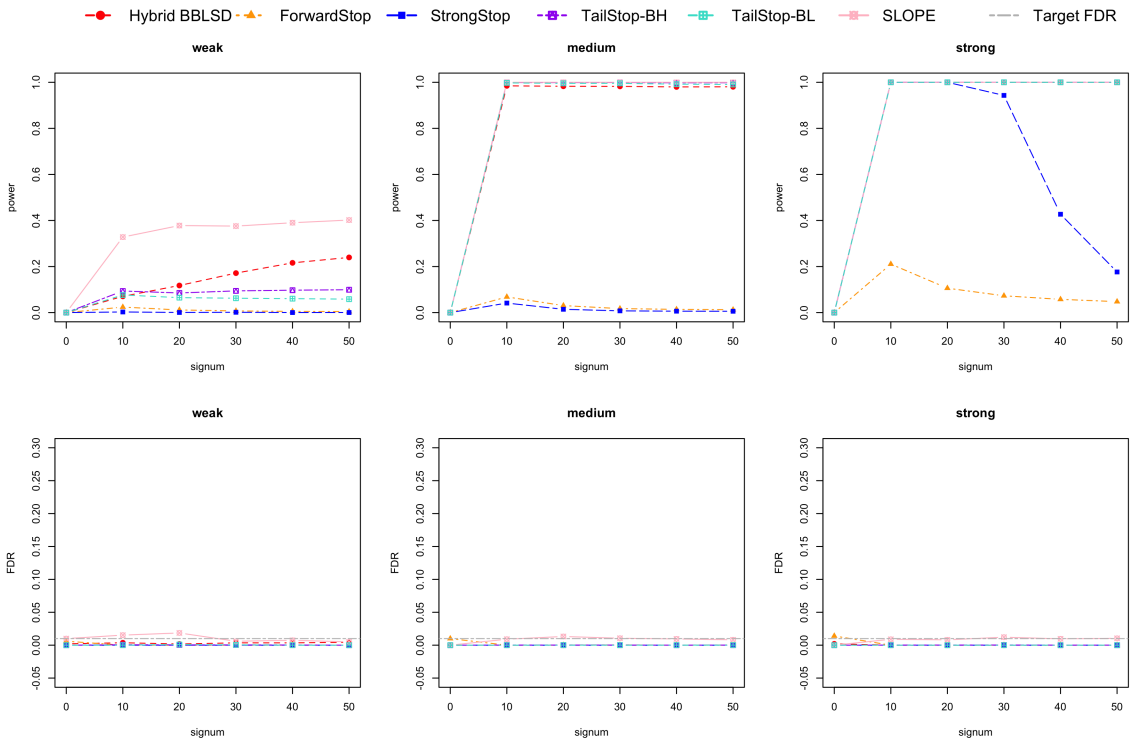
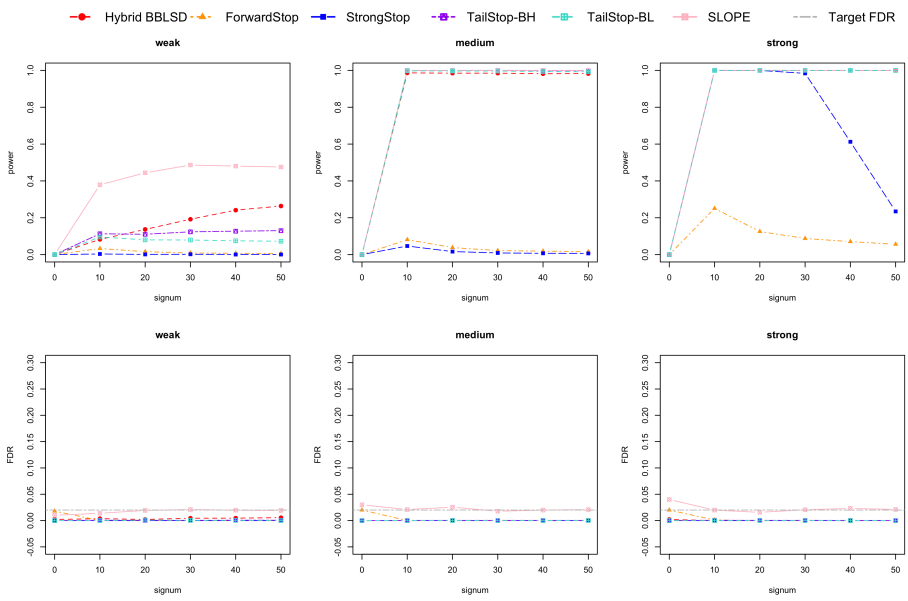
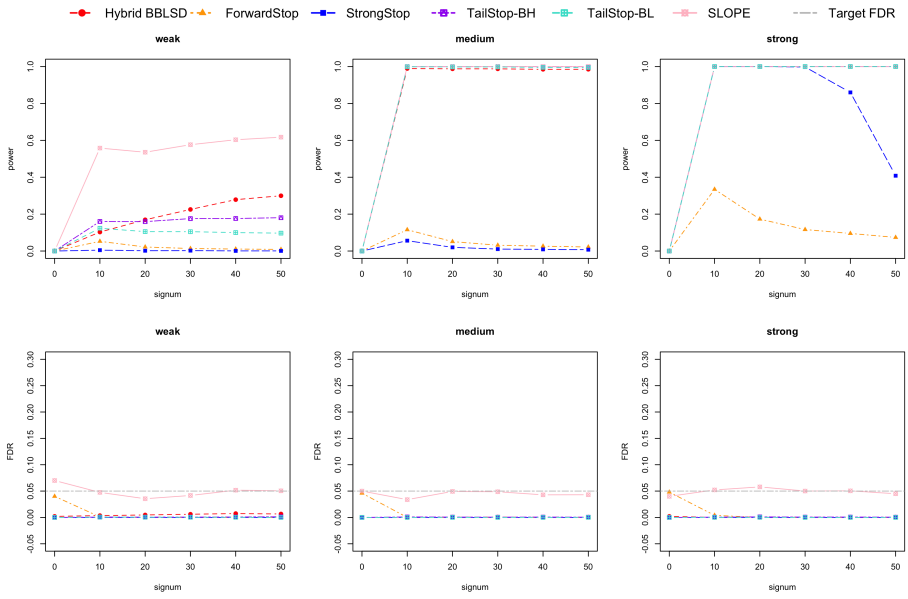


Figure 6.3: Setup 1: Average power and observed FDR for the sequential multiple testing at q=0.01. Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop, TailStop-BH, TailStop-BL and SLOPE procedures are evaluated.



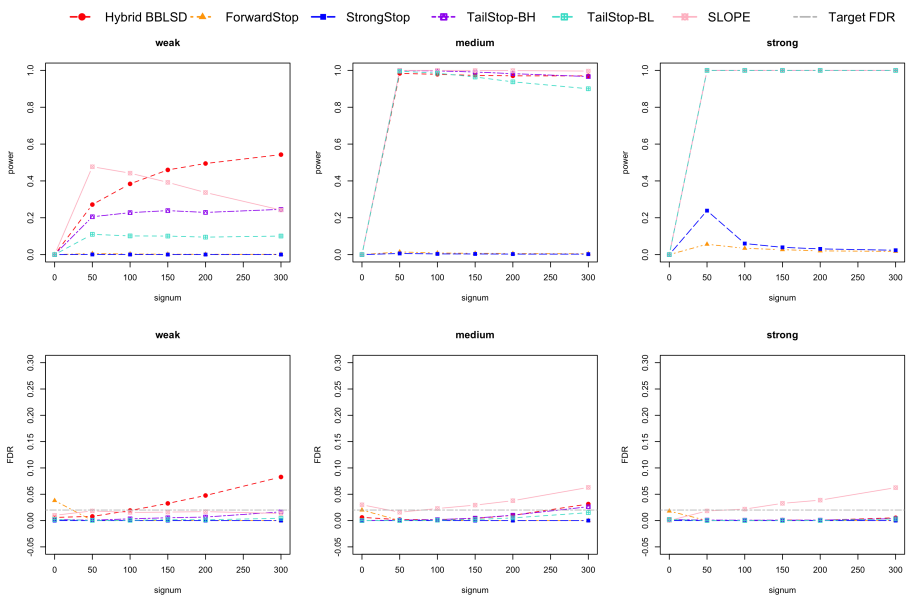
(a) Average power and observed FDR at q=0.02.



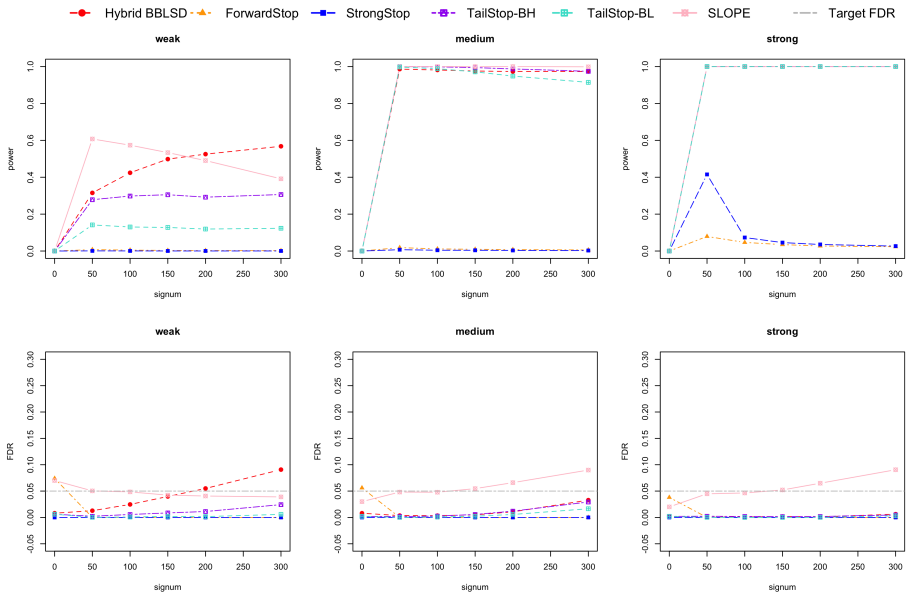
(b) Average power and observed FDR at q=0.05.

Figure 6.4: Setup 1: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH, TailStop-BL and SLOPE procedures are evaluated.

Figure 6.4 is the simulation results for Setup 1 at $q = 0.02$ and $q = 0.05$. For both of the plots, the calculated FDR are controlled at q, with the FDR of SLOPE around q and other covariance based procedures are much smaller than q. Therefore, the calculated power for SLOPE is higher than other procedures. The overall trend in Figure 6.4 is similar as Figure 6.3. In plot (a), for weak signal, the power gets higher when there are more variables significant in the model, up to 26.4% for Hybrid BBLSD and 13% for TailStop-BH, 7.3% for TailStop-BL and 47.6% for SLOPE, while ForwardStop and StrongStop procedure are very poor. In plot (b), for weak signal, the power gets higher when there are more variables significant in the model, up to 30.0% for Hybrid BBLSD and 18.1% for TailStop-BH, 9.7% for TailStop-BL and 61.7% for SLOPE, while ForwardStop and StrongStop procedure are very poor. When the number of significant variables is smaller, TailStop-BH is outperform than Hybrid BBLSD, when the number of significant variables gets larger than 10, our procedure has larger power than TailStop-BH. This finding is consistent with the assumption of $k_0 = 0$ for TailStop procedure. For both TailStop-BH and TailStop-BL, the assumption is deviate when the number of nun-zero variables is getting larger. Therefore, overall our Hybrid BBLSD sequential testing procedure is outperform than TailStop. All the four procedures except ForwardStop and StrongStop have power close to 1 for medium signals and have power 1 for strong signals.



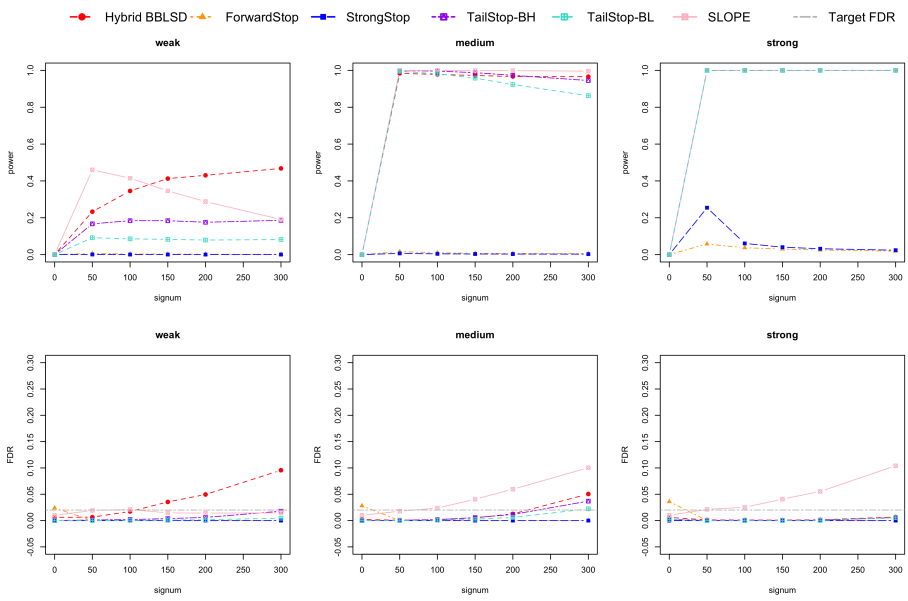
(a) Average power and observed FDR at q=0.02.



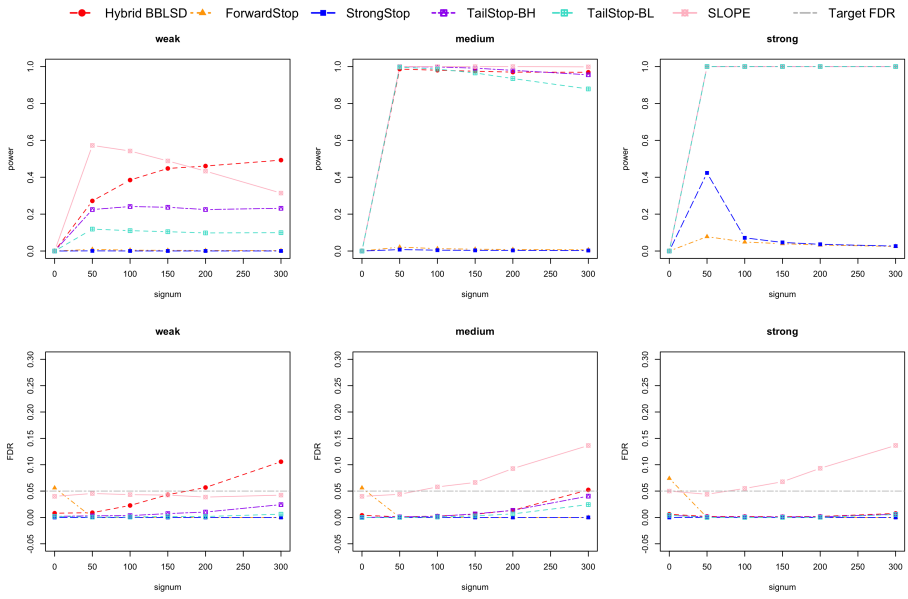
(b) Average power and observed FDR at q=0.05.

Figure 6.5: Setup 2: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH, TailStop-BL and SLOPE procedures are evaluated.

Figure 6.5 displays the results for Setup 2, which is an extension of Setup 1 for sparsity up to 6%. Overall, the power of Hybrid BBLSD increases as the number of significant variables increases. In plot (a), the power of Hybrid BBLSD increases up to 54.3% at 300 true significant variables for $q = 0.02$, 24.6% for TailStop-BH, 10.1% for TailStop-BL, and still ForwardStop and StrongStop have very small power. When there are 50 significant variables in the model, SLOPE has 47.7% of power, and decreases to 24.3% for 300 significant variables, but the average number of selected variables for SLOPE increases. In plot (b) it reaches 56.7% at 300 true significant variables for $q = 0.05$, while the power of SLOPE is decreasing, down to 39.2%. When the number of significant variables is 50, 31.5% power for Hybrid BBLSD and 60.7% for SLOPE, but the average number of selected variables for SLOPE increases. TailStop-BH has power up to 30.6% and TailStop-BL is 12.3%. Moreover, our Hybrid BBLSD has higher power than both TailStop-BH and TailStop-BL procedures. For The medium and strong signal, only ForwardStop and StrongStop procedures have very poor power while other procedures have power close to 1. On the other hand, the calculated FDR for almost all the procedures are controlled at q with Hybrid BBLSD is lightly larger when the number of significant variables is 300 for the weak signal. The value of the non-zero coefficient in the weak signal is the smallest value lasso can identify, therefore, there will be false discoveries occur early on the path (Su, Bogdan and Candes 2015). For the medium and strong signal cases, the Hybrid BBLSD procedure consistently controls FDR at q. However, the FDR of SLOPE exceeds q when the number of nun-zero variables exceed 200 and 300, which have more false positives for medium and strong signal. We should have more concerns for SLOPE since the signals have been strengthened.



(a) Average power and observed FDR at q=0.02.

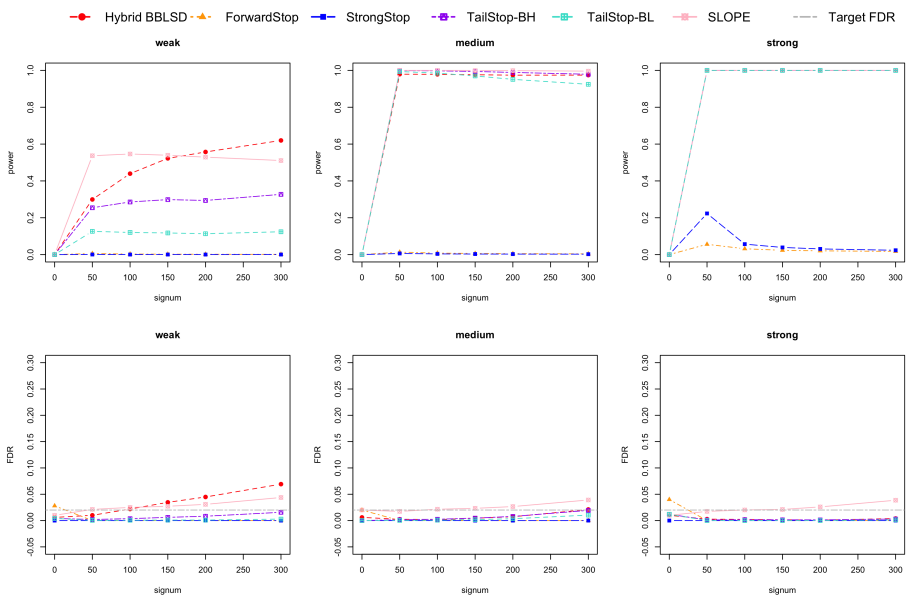


(b) Average power and observed FDR at q=0.05.

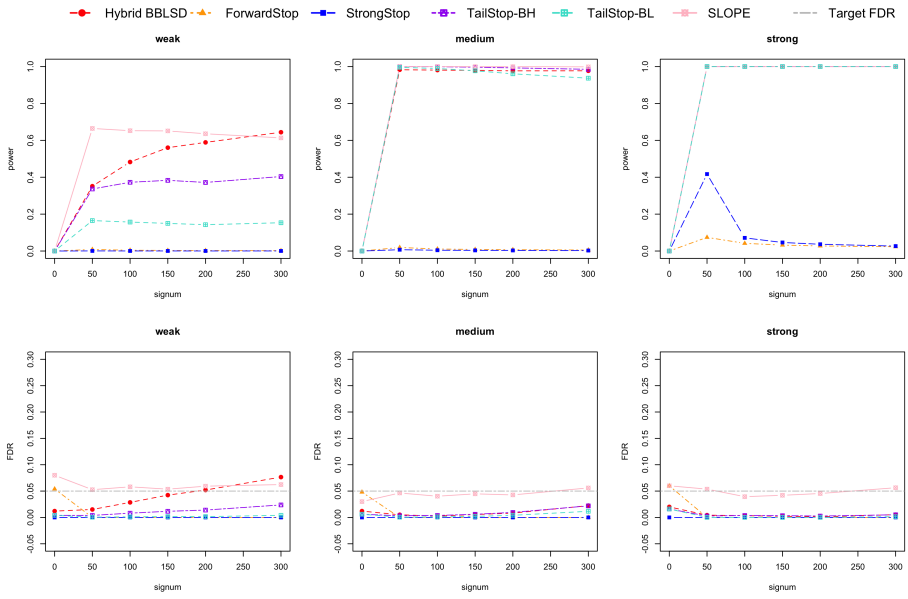
Figure 6.6: Setup 3: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH, TailStop-BL and SLOPE procedures are evaluated.

Figure 6.6 is for high dimensional data (i.e. $n = 0.5m$). Both Hybrid BBLSD and TailStop are able to use for $m > n$ case. FDR is slightly larger when there are more variables in the model since our procedure has the ability to choose more significant ones while still trying to control the error rate. Also this setup is high-dimensional, the assumptions will be slightly violated. The difference is that, for the weak signal, the performance of our procedure is consistent better than TailStop-BH and TailStop-BL procedures when the number of significant variables is larger than 50, which means the assumption of k_0 for TailStop is deviate more from the truth while our sequential procedure doesn't get affected. In plot (a), when the number of significant variables is 50, Hybrid BBLSD procedure can still capture around 23.3% of the true variables while SLOPE has power 40.0% for $q = 0.02$. When the number of significant variables reaches 300, Hybrid BBLSD has power of 46.8% while SLOPE is 19.1%, but the average selected number of variables increases. TailStop-BH reaches power up to 18.6% and 8.2% for TailStop-BL, while ForwardStop and StrongStop are very poor. In plot (b), when the number of significant variables is 50, Hybrid BBLSD procedure can still capture around 27.2% of the true variables while SLOPE has power 57.2% for $q = 0.05$. When the number of significant variables reaches 300, Hybrid BBLSD has power of 49.3% while SLOPE is 31.4%. TailStop-BH reaches power up to 23.2% and 10.0% for TailStop-BL, while ForwardStop and StrongStop are very poor.

On the other hand for FDR, Hybrid BBLSD exceeds q in the weak signal when the number of significant variable is 300, while it controls FDR at q when the signals are strengthened. However, the SLOPE has lost control at q when the number of nun-zero variables are higher than 200 in high-dimensional data.



(a) Average power and observed FDR at q=0.02.



(b) Average power and observed FDR at q=0.05.

Figure 6.7: Setup 4: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH, TailStop-BL and SLOPE procedures are evaluated.

Figure 6.7 is for large scale data as well with $n = 2m$ at $q = 0.02$ and $q = 0.05$. The results are very consistent with other large scale data set. In plot (a) for $q = 0.02$, the power reaches 62.0% for Hybrid BBLSD, 51.1% for SLOPE, 32.7% for TailStop-BH and 12.4% for TailStop-BL in weak signal when the number of significant variables is 300, while ForwardStop and StrongStop are still very low power. In plot (b), the power reaches 64.4% for Hybrid BBLSD, 61.3% for SLOPE, 40.4% for TailStop-BH and 15.3% for TailStop-BL. The trends of performance are very similar as the previous examples, and FDR are all controlled at q. The calculated FDR exceeds a little for Hybrid BBLSD in the weak signal and SLOPE for medium and strong signal when the number of significant variables are very large. Overall, for the independent entries of normal distributions, the performance of SLOPE is generally better than the covariance based procedures and Hybrid BBLSD is comparatively better for high-dimensional data, while ForwardStop and StrongStop is not recommended in these examples.

In reality, it is very common to have some of the variables correlated with each other. Next series of results are for the correlated design matrix in 500 simulations, and the correlation among any two adjacent variables is 0.5, and 0 otherwise.

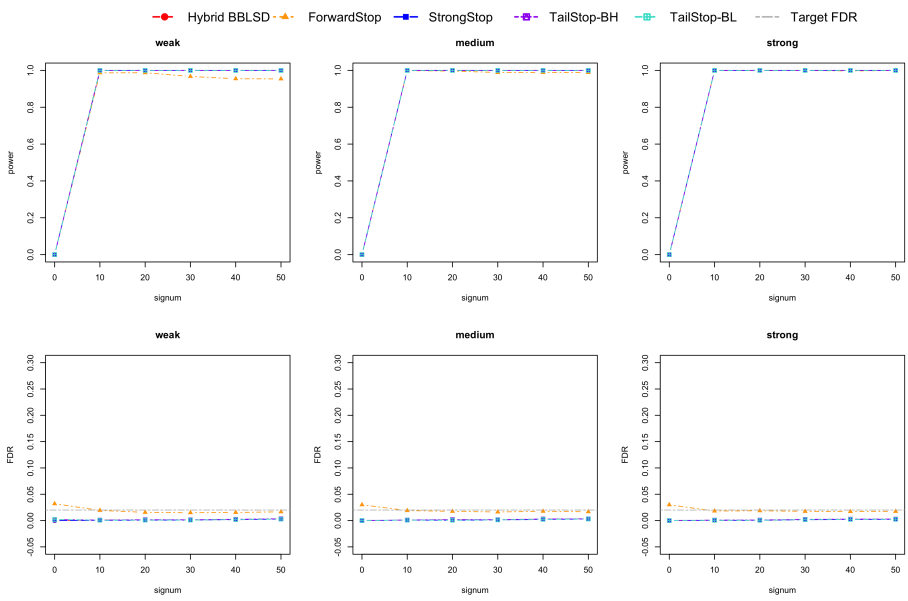
First of all, the computer resource and time needed for SLOPE are extremely expensive and the performance is bad when there are some large correlation (i.e. 0.5 here) in the design matrix. For all the setups, only Setup corr4 ($n = 5000,\ m = 2500$) reported the result within our walltime of 300 hours. From the simulation time cost, we can tell the solution of the SLOPE is highly affected by the total number of variables in the model. Secondly, Table 6.3 shows the simulation results for SLOPE in correlated design matrix

in Setup corr4. The number of significant variables ranges from 0 - 300 and lucky all the simulations are able to pick up all the non-zero variables, i.e. power is 1. However, the calculated FDR is close to 1 and the average selected variables are at least 2150 among all the 2500 variables, which means in this correlated design matrix, SLOPE tends to choose most of the variables in the model. When we decrease the correlation between adjacent variables to 0.2 or lower in the same correlation structure, we still get similar results. In summary, SLOPE only works well for independent entries of design matrix, besides, it will lose control of FDR and the algorithm is computational expensive for correlated case.

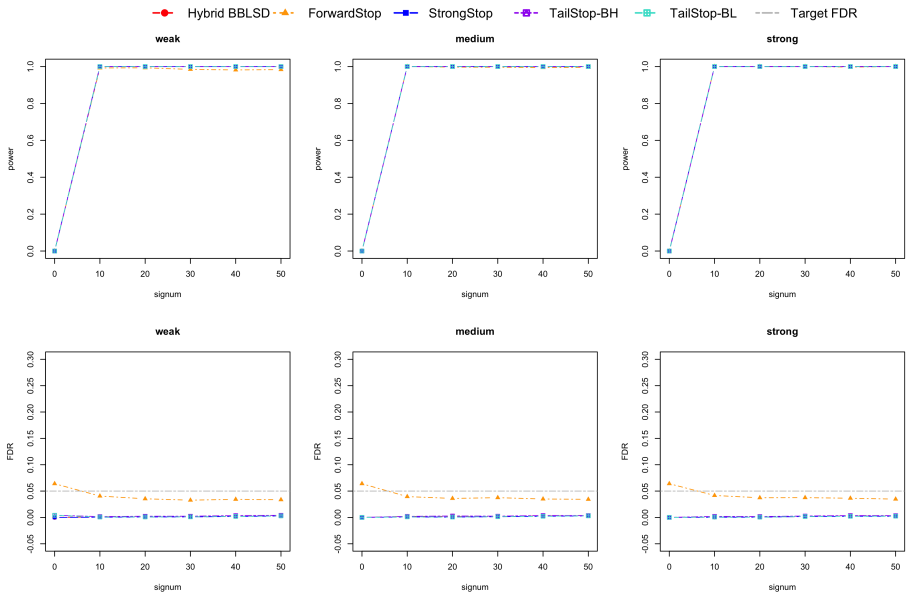
Table 6.3: The calculated FDR and average selected number of variables for SLOPE in Setup corr4.

		0	50	100	150	200	300
weak	FDR	1	0.9768	0.9538	0.9308	0.9080	0.8628
	ave selt	2154.25	2159.1	2162.91	2168.53	2173.61	2186.73
medium	FDR	1	0.9769	0.9537	0.9309	0.9080	0.8626
	ave selt	2154.09	2158.82	2162.91	2170.46	2174.9	2182.83
strong	FDR	1	0.9768	0.9537	0.9306	0.9080	0.8626
	ave selt	2150.63	2157.98	2161.9	2169.46	2174.1	2184.23

Therefore, we exclude the SLOPE in the simulation results for Setup corr1 - corr4, displaying in Figure 6.8 - Figure 6.11. In each of the setup, we evaluated the performance of our Hybrid BBLSD, ForwardStop, StrongStop, TailStop-BH and TailStop-BL methods for $q = 0.02$ and 0.05.

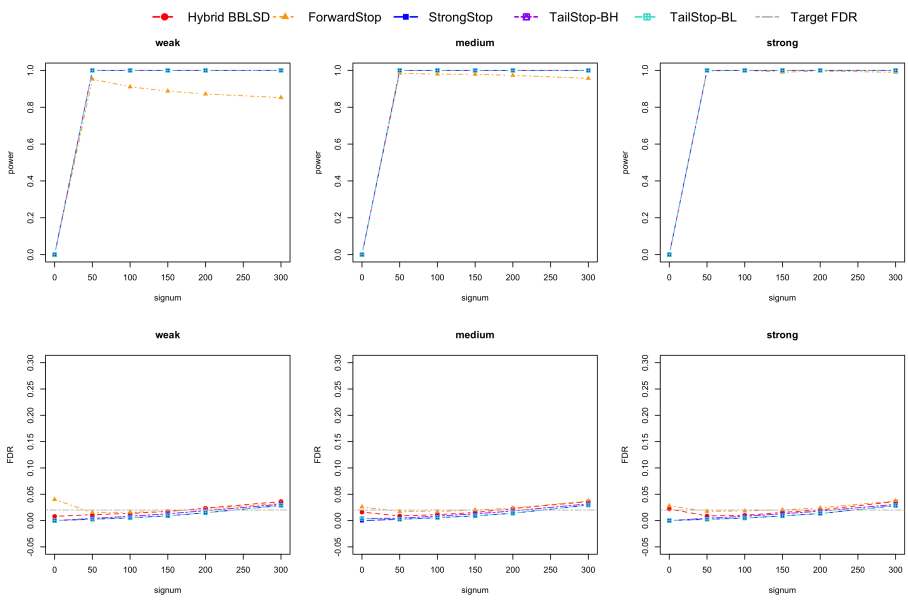


(a) Average power and observed FDR at q=0.02.

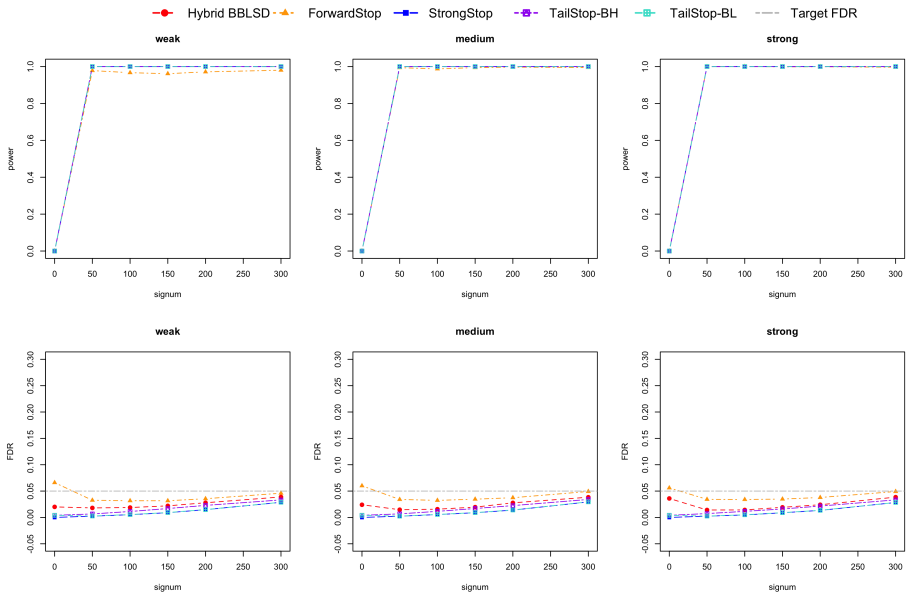


(b) Average power and observed FDR at q=0.05.

Figure 6.8: Setup corr1: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH and TailStop-BL procedures are evaluated.

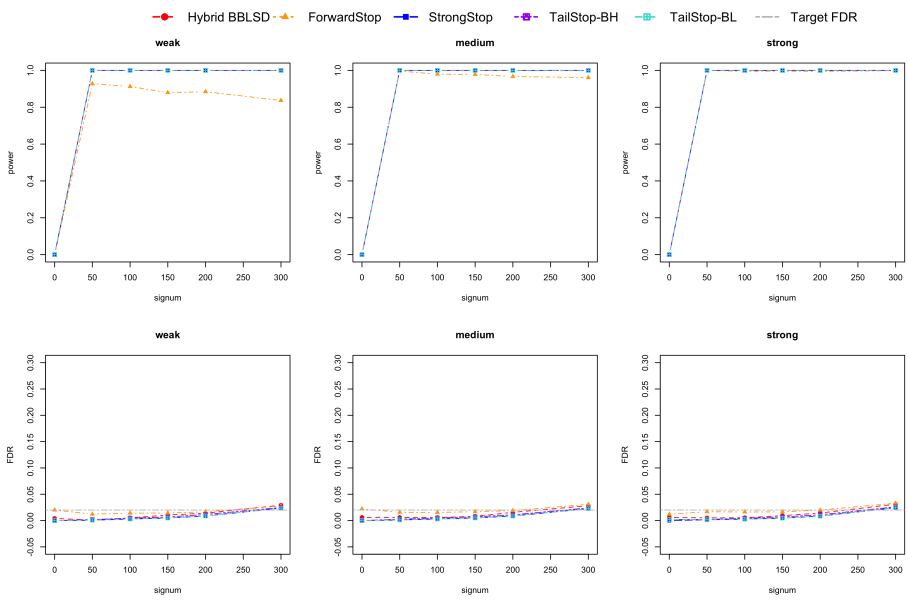


(a) Average power and observed FDR at q=0.02.

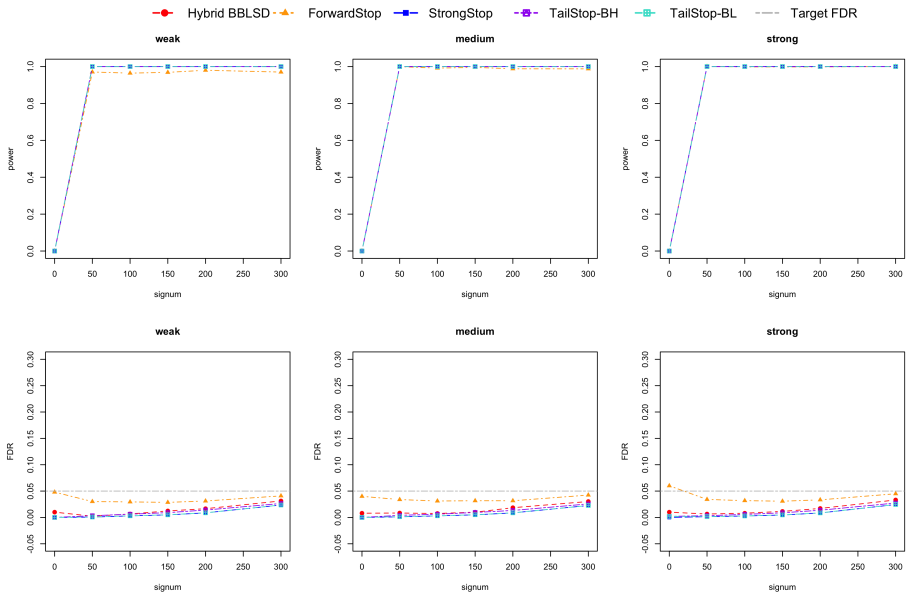


(b) Average power and observed FDR at q=0.05.

Figure 6.9: Setup corr2: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH and TailStop-BL procedures are evaluated.

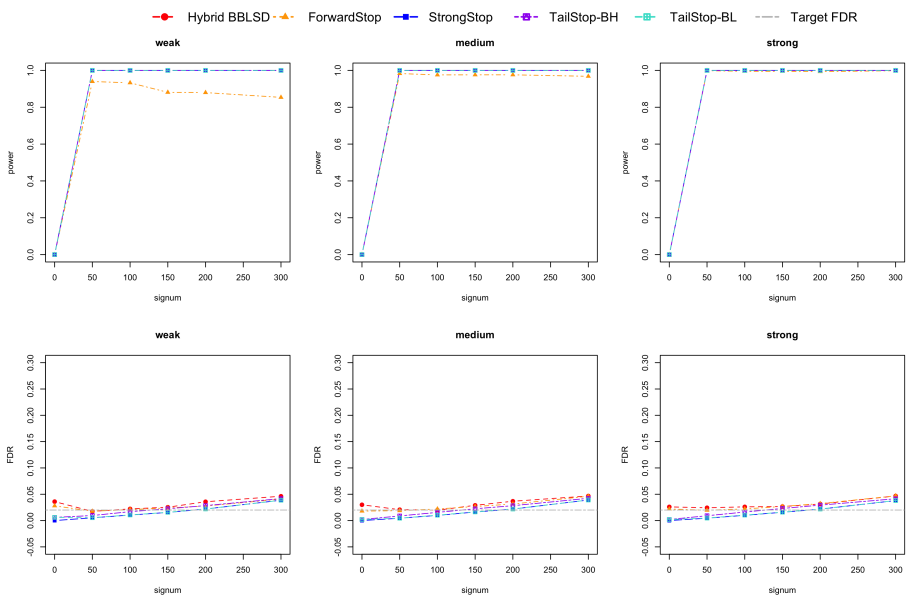


(a) Average power and observed FDR at q=0.02.

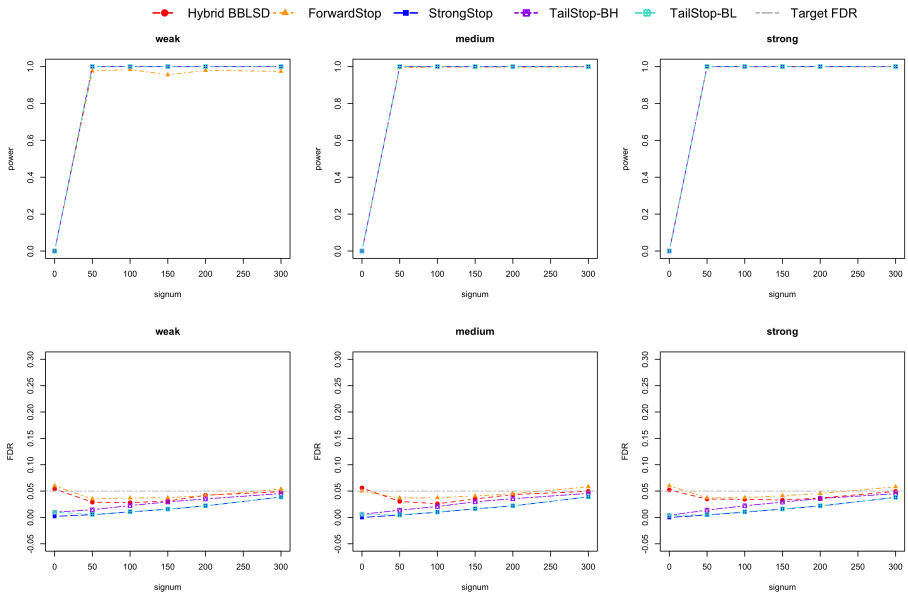


(b) Average power and observed FDR at q=0.05.

Figure 6.10: Setup corr3: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH and TailStop-BL procedures are evaluated.



(a) Average power and observed FDR at q=0.02.



(b) Average power and observed FDR at q=0.05.

Figure 6.11: Setup corr4: Hybrid Bonferroni-BL Step-Down, ForwardStop, StrongStop, TailStop-BH and TailStop-BL procedures are evaluated.

Figure 6.8 shows the performance for Setup corr1, in which $n = m = 5000$ and the sparsity is $0\% - 1\%$. All the lasso and covariance test based procedures perform well even for the weak signal. All the methods control FDR within their targeted level with ForwardStop having the FDR around q. The truth is the covariance drop for all the methods can clearly separate well among the null and non-null variable groups, thus, all the methods can correctly choose the significant variables without choosing more false positives. The power of ForwardStop is slightly lower than others. The performance for all the procedures are similar for $q = 0.02$ as for $q = 0.05$.

The simulation results for Setup corr2 are illustrated in Figure 6.9, in which $n = m = 5000$ and the sparsity is $0\% - 6\%$. The power of all the procedures except ForwardStop for all signals are 1. The power of ForwardStop is lower, and it is getting higher power when the signal is stronger. In terms of calculated error, FDR increases as the number of significant variable increases. The FDR is slightly over the targeted 0.02 when the number of non-zero variables is 300. In plot (b), the FDR is controlled at $q = 0.05$ for all the procedures.

Figure 6.10 displays the performance for Setup corr3 high dimensional data, in which $n = 5000$, $m = 10000$ and sparsity is $0\% - 3\%$. All the covariance test based procedures perform very well for high-dimensional data and the FDR is controlled at q.

Figure 6.11 shows the performance for Setup corr4 for $n = 5000$, $m = 2500$. The trend of power and FDR are similar as the previous results. The calculated FDR for our Hybrid BBLSD is slightly closer to the targeted q than StrongStop, TailStop-BH and TailStop-BL procedures.

Compared to the independent entries of $N(0, 1/n)$ in Setup 1 - 4, the overall covariance test based methods have higher power for the correlated design matrix. The covariance test statistic has significant drop between null and non-null variables, therefore, the procedures are able to pick out the significant variables more easily. For smaller correlation coefficient (0.2 or 0.1) among the adjacent variables, the calculated FDR is lower and all controlled at q.

Overall, for large scale and high dimensional data, TailStop also performed well, and our proposed procedure has the best performance and better than TailStop. ForwardStop and StrongStop are not very recommended for the large scaled data. And for SLOPE, it performs very well in more independent variables while needs to be adjusted to use for correlated design matrix.

Chapter 7

Real data example

7.1 Diabetes study

The diabetes data is from diabetes study (Efron 2004). Ten baseline variables, age, sex, body mass index, average blood pressure, and six blood serum measurements (tc, ldl, hdl, tch, ltg, glu) were obtained for each of $n = 442$ diabetes patients, as well as the response of interest, a quantitative measure of disease progression one year after baseline. Data is also available in R database.

Applying the lasso variable selection, we can get the sequence of lasso moves of the diabetes data:

Table 7.1: Lasso path for Diabetes study.

	bmi	ltg	map	hdl	sex	glu	tc	tch	ldl	age	hdl	hdl
Var	3	9	4	7	2	10	5	8	6	1	-7	7
Step	1	2	3	4	5	6	7	8	9	10	11	12

And the lasso path according the moves is showing in Figure 7.1. The y-axis is the standardized coefficients of the variables, and the x-axis is the L1 norm of the coefficient vector, as a fraction of the maximal L1 norm. Also we can check the coefficients for all the ten variables when the tuning parameters decreases until a certain point and there are changes in the active set, see Table 7.2. At the beginning, none of the variables enter the model. When λ decrease, more variables enter the model. At $k = 1$ and $\lambda = 949.43$, none of the variable in the model, after λ decrease smaller than 949.43, the first variable, 'bmi', starts to join in the active set, thus, we can test the variable 'bmi' if it is signinificant. At $k = 2$, the second variable 'ltg' joins the model when λ is smaller than 889.31. $A =$'bmi' is the active set before λ_2.

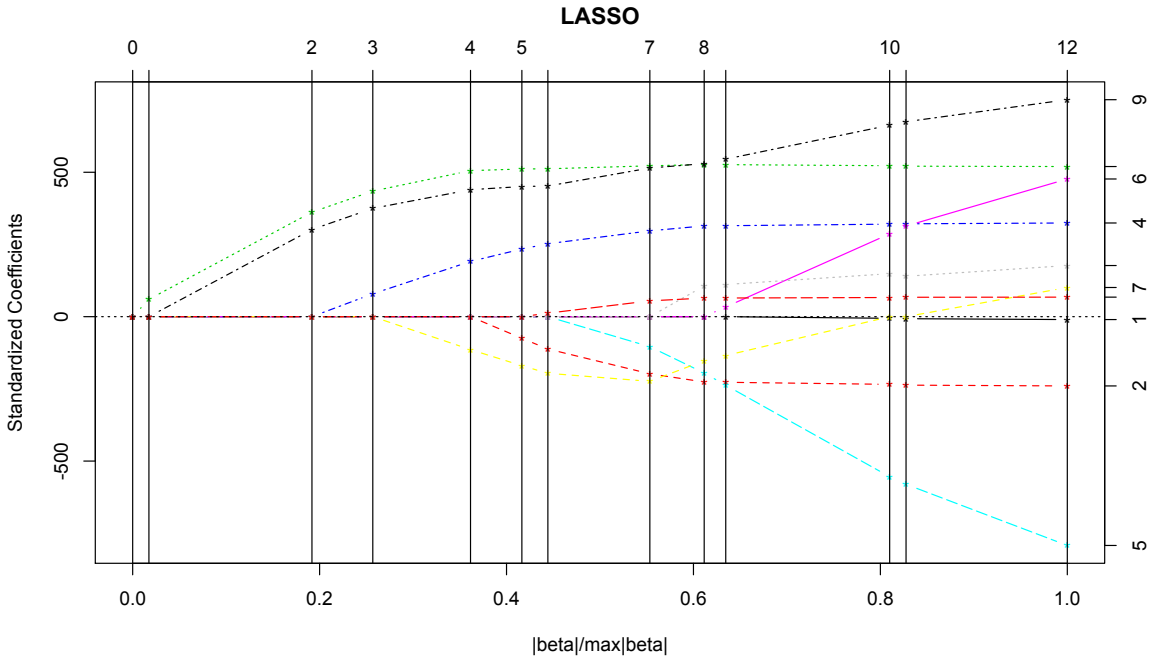


Figure 7.1: Lasso path for Diabetes study.

Table 7.2: Coefficients at each steps for Diabetes study.

	lambda	age	sex	bmi	map	tc	ldl	hdl	tch	ltg	glu
1	949.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2	889.31	0.00	0.00	60.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	452.9	0.00	0.00	361.89	0.00	0.00	0.00	0.00	0.00	301.77	0.00
4	316.07	0.00	0.00	434.76	79.24	0.00	0.00	0.00	0.00	374.92	0.00
5	130.13	0.00	0.00	505.66	191.27	0.00	0.00	-114.10	0.00	439.66	0.00
6	88.78	0.00	-74.92	511.35	234.15	0.00	0.00	-169.71	0.00	450.67	0.00
7	68.96	0.00	-111.98	512.04	252.53	0.00	0.00	-196.05	0.00	452.39	12.08
8	19.98	0.00	-197.76	522.26	297.16	-103.95	0.00	-223.93	0.00	514.75	54.77
9	5.47	0.00	-226.13	526.89	314.39	-195.11	0.00	-152.48	106.34	529.92	64.49
10	5.09	0.00	-227.18	526.39	314.95	-237.34	33.63	-134.60	111.38	545.48	64.61
11	2.18	-5.72	-234.40	522.65	320.34	-554.27	286.74	0.00	148.90	663.03	66.33
12	1.31	-7.01	-237.10	521.08	321.55	-580.44	313.86	0.00	139.86	674.94	67.18
13	0	-10.01	-239.82	519.84	324.39	-792.18	476.75	101.04	177.06	751.28	67.63

Using the covariance test to test each entering variable in the active set, we check the covariance drop in the lasso path displaying in Table 7.3.

Table 7.3: Covariance drop in Diabetes study.

	bmi	ltg	map	hdl	sex	glu	tc	tch	ldl	age
Drop in covariance	19.51	50.80	5.57	6.02	4.88	0.16	3.29	0.61	0.14	0.02

Here we only apply the FDR control procedure. Our Hybrid Bonferroni-BL Step-Down procedure gives first 6 variables, which are 'bmi', 'ltg', 'map', 'hdl', 'sex' and 'glu' in lasso path. Both ForwardStop and SLOPE choose the first 5 variables, while StrongStop choses the first 3 variables in the lasso path. And TailStop has one more variable 'tc' in its model. In reality, the glucose level is normally higher for diabetes patients. Doctors usually use some tests to find out if the patients have diabetes, such as fasting plasma glucose test or oral glucose tolerance test. Therefore, it is reasonable to consider 'gul' as an important factor for diabetes.

7.2 Framingham heart study:

Cardiovascular disease (CVD) is the leading cause of death and serious illness in the United States. The objective of the Framingham Heart study was to identify the common factors or characteristics that contribute to CVD. Over the years, the Framingham study led to identification of major CVD risk factors-high blood pressure, high blood cholesterol, smoking, obesity, diabetes, and physical inactivity. Our data include 2384 subjects from original generation and 3777 subjects from third generation, with total 6161 subjects. Consider HDL cholesterol levels as the response variable and SNPs located in each chromo-

some as the genetic factors. The histogram of HDL cholesterol level displays in Figure 7.2, which is skewed right. The three possible allele combinations are coded as $X = (0, 1, 2)$. Other covariates include age, sex and body mass.

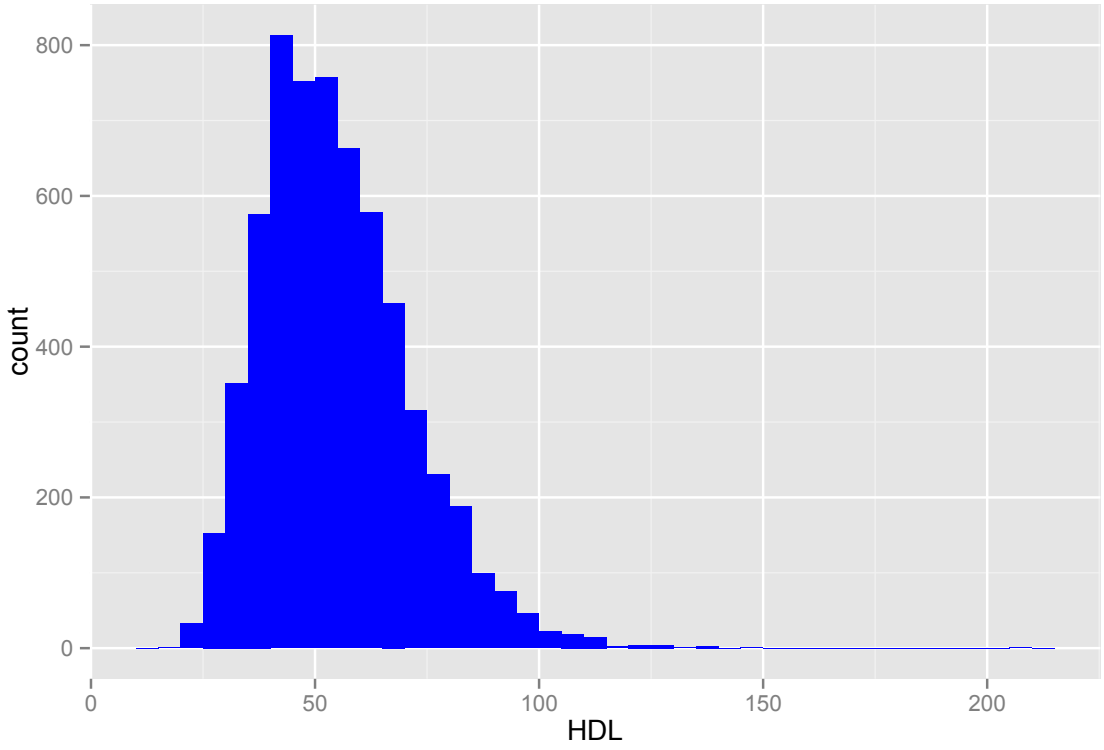


Figure 7.2: Histogram of HDL for Framingham study.

We use the cross validation to estimate the σ^2 and the missing value is imputed by its mean. We use our proposed Hybrid Bonferroni-BL Step-Down procedure and compare it with other procedures to control FDR. The results of the significant variables identified by those procedures are displayed in Table 7.4, and the chromosomes in which don't have significant SNPs are not included here.

Table 7.4: Framingham heart study results.

	No. SNPs	Hybrid BBLSD	ForwardStop	StrongStop	TailStop-BH	TailStop-BL
Chr1	40055	ss66520830	ss66520830 ss66517982	none	ss66520830 ss66517982	ss66520830 ss66517982
Chr2	41266	none	ss66451524	none	none	none
Chr3	33690	none	ss66303954	none	none	none
Chr4	32219	none	ss66485909	none	none	none
Chr8	27632	none	none	none	ss66508871 ss66143264	ss66508871
Chr10	28391	none	ss66135879	none	ss66135879	ss66135879
Chr15	14275	ss66125757 ss66378055 ss66271313	none	none	none	none
Chr16	15266	ss66166798	ss66166798	ss66166798	ss66166798	ss66166798
Chr17	11232	ss66079776	ss66079776	none	ss66079776	ss66079776
Chr18	14832	ss66488334	ss66488334	none	ss66488334	ss66488334
Chr21	7085	none	ss66114419 ss66113490	none	ss66114419	ss66114419

Among the SNPs, there are correlation exist pair-wisely. For example, in chromosome 21, the maximum correlation is close to 1. Here are the summaries for absolute value of the pair-wise correlation: 42% of the correlations are between $0 - 0.01$, 20% are between $0.01 - 0.02$, 25% are between $0.02 - 0.05$, 0.37% are between $0.1 - 0.2$, 0.25% are between $0.2 - 0.5$, and 0.1% are larger than 0.5. SLOPE fails to selected effective SNPs within 200 hours walltime while our procedure takes less than half an hour in chromosome 21. Therefore, we exclude the results of SLOPE procedure in Table 7.4.

In chromosome 15, the SNPs we find link to gene LIPC. Metaanalysis reports indicate the association between LIPC and HDL cholesterol levels, and conclude that variation at the LIPC locus influences HDL metabolism and coronary artery disease (Feitosa et al. 2009). In chromosome 16, we find gene CEPT. As HDL can alleviate atherosclerosis and other cardiovascular diseases, and certain disease states such as the metabolic syndrome feature low HDL, pharmacological inhibition of CETP is being studied as a method of improving HDL levels. Individuals possessing a genetic deficiency for CETP have higher HDL cholesterol and lower LDL cholesterol and may have a reduced risk for developing cardiovascular disease (Ruggeri 2005). In chromosome 17, we detect Myh10 which can cause major defects in heart development (Ma and Adelstein 2014). Other SNPs are also very challenging for researchers to investigate.

Chapter 8

Conclusions

In this dissertation, based on the lasso variable selection, we introduced the hybrid sequential multiple testing procedures for sequential problem. Hybrid Bonferroni-Holm Step-Down procedure was proposed to control FWER and Hybrid Bonferroni-BL Step-Down procedure was introduced for FDR control. All the work was build on the sequential multiple hypotheses testing and we showed that these procedures can control FWER and FDR at desired level respectively. Furthermore, our procedures are robust for most models, especially when the signals are equally weak for the orthogonal design matrix, and for large scale and high-dimensional data. The simulations and comparison with other methods suggests our hypothesis testing structure and our hybrid methods are more preferred to implement. Also, some results of our procedure applied on the diabetes study and framingham heart study have been validated by literatures.

Sequential multiple testing and other selective inference are very interesting topics and researchers are still continuing to work on this area for selection inference. Our further

work will mainly focus on high dimension implementation for FDR control using sequential multiple testing procedures. We mentioned that this sequential multiple testing structure works for covariance test of SCAD and MCP. Later performance will be compared between different variables selection methods for FDR control. It is also meaningful to expand our procedure to general explanatory matrix and in general linear models.

Bibliography

[1] A Ahrens, A. and Bhattacharjee, A. (2015). Two-step lasso estimation of the spatial weights matrix. *Econometrics*. 3(1), 128–155.

[2] Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*. 19 (4), 716–723.

[3] Barnhart, H. X. and Williamson, J. M. (1998). Goodness-of-fit tests for GEE modeling with binary responses. *Biometrics,* 54, 720–729.

[4] Benjamini, Y. and Gavrilov, Y. (2009). A simple forward selection procedure based on false discovery rate control. *Ann. Appl. Stat.* 3, 179–198.

[5] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society* 57, 289–300.

[6] Benjamini, Y. and Liu, W. (1999). A step-down multiple hypotheses testing procedure that controls the false discovery rate under independence. *Journal of Statistical Planning and Inference* 82, 163–170.

[7] Benjamini, Y and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics* 29, 1165–1188.

[8] Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*. 41(2), 802–837.

[9] Bogdan, M., Van den Berg, E., Sabatti, C., Su, W. and Candes E. (2015). SLOPE - adaptive variable selection via convex optimization. *The Annals of Applied Statistics*. 9(3), 1103–1140.

[10] Bogdan, M., Van den Berg, E., Su, W. and Candes E. (2013). Statistical estimation and testing via the sorted l_1 norm. *arXiv:1310.1969*.

[11] Buhlmann, P. (2012). Statistical significance in high-dimensional linear models. *arXiv:1202.1377*.

[12] Buja, A. and Brown, L. (2014). Discussion: 'A Significant Test for the Lasso'. *The Annals of Statistics*. 42 (2), 509–517.

[13] Chen, S., Donoho, D., and Saunders, M. (2001). Atomic decomposition by basis pursuit, *SIAM Review.* 43, 129–159.

[14] Cheng, M., Honda, T. and Zhang, J. (2014). Forward variable selection for sparse ultra-high dimensional varying coefficient models. *arXiv:1410.6556*.

[15] Dezeure, R., Buhlmann, P., Meier, L. and Meinshausen, N. (2015). High-dimensional inference: Confidence intervals, p-values and r-software hdi. *Statistical science*. 30(4), 533–558.

[16] Dmitrienko, A., Tamhane, A.C. (2007). Gatekeeping procedures with clinical trial applications. *Pharmaceutical Statistics.* 6, 171–180.

[17] Donoho, D. (2004). For most large underdetermined systems of equations, the minimal 1-Norm solution is the sparsest solution *technical report, Stanford University, Dept. of Statistics.*

[18] Donoho, D., and Elad, M. (2002). Optimally sparse representation in general (nonorthogonal) dictionaries via 1-Norm minimizations. *Proceedings of the National Academy of Science USA* 1005, 2197–2202.

[19] Donoho, D., and Huo, X. (2002). Uncertainty principles and ideal atomic decompositions. *IEEE Transactions on Information Theory* 47, 2845–2863.

[20] Donoho, D., Johnstone, I., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: asymptopia? (with discussion) *Journal of the Royal Statistical Society, Ser. B* 57, 301–337.

[21] Dunnett, C.W., Tamhane, A.C. (1991). Step-down multiple tests forcomparing treatments with a control in unbalanced one-way layouts. *Statistics in Medicine.* 10, 939–947.

[22] EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* 81, 461–470.

[23] Efron, B. (2004). Large-Scale Simultaneous Hypothesis Testing. *Journal of the American Statistical Association*. 99, 96–104.

[24] Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Journal of the American Statistical Association* 87, 451–457.

[25] Fan, J. and Ke, Z. (2014). Discussion: a significance test for the lasso. *The Annals of Statistics* 42, 483–492.

[26] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348–1360.

[27] Fan, J., Samworth, R. and Wu, Y. (2009). Ultrahigh dimensional variable selection: beyond the linear model. *Journal of Machine Learning Research.* 10, 2013–2038.

[28] Feitosa, M. F., Myers, R. H., Pankow, J. S., Province, M. A., and Boreckia, I. B. (2009) LIPC variants in the promoter and intron 1 modify HDL-C levels in a sex-specific fashion. *Atherosclerosis.* 204 (1), 171–177.

[29] Fithian, W., Sun, D. and Taylor, J. (2014). Optimal inference after model selection. *arXiv:1410.2597.*

[30] Frank, I. E. and Friedman, J.H. (1993). A Statistical View of Some Chemometrics Regression Tools, *Technometrics.* 35(2), 109–135.

[31] G'Sell, M. G., Hastie, T. and Tibshirani, R. (2013). False variable selection rates in regression. *arXiv:1302.2303.*

[32] G'Sell, M.G., Taylor, J. and Tibshirani, R. (2013). Adaptive testing for the graphical lasso. *arXiv:1307.4765.*

[33] G'Sell, M. G., Wager, S., Chouldechova, A., and Tibshirani, R. (2015). Sequential selection procedures and false discovery rate control. *arXiv:1309.5352.*

[34] Guo, W., Sarkar, S. K. and Peddada, S. D. (2010). Controlling False Discoveries in Multidimensional Directional Decisions, with Applications to Gene Expression Data on Ordered Categories. *Biometrics.* 66, 485–492.

[35] Hastie, T. and Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning, 2nd Ed. *Springer.*

[36] Hastie, T., Tibshirani, R. and Wainwright, M. (2015). Statistical Learning with Sparsity: Lasso and Its Generalizations. *Chapman and Hall, London.*

[37] Hochberg, Y. (1998). A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75, 800–802.

[38] Hoerl, A.E. and Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems, *Technometrics.*12, 55–67.

[39] Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6, 65–70.

[40] Hommel, G. (1989). A comparison of two modified Bonferroni procedures. *Biometrika.* 76, 624–625.

[41] Huang, Y. and Hsu, J. (2007) Hochberg's step-up method: cutting corners off Holm's step-down method. *Biometrika* 94: 965–975.

[42] Knight, K. and Fu, W. J. (2000). Asymptotics for Lasso-type estimators. *The Annals of Statistics* 28:1356–1378.

[43] Lee, J., Sun, D., Sun, Y., and Taylor, J. (2013). Exact Post-Selection Inference With Application To The Lasso. *arXiv:1311.6238*.

[44] Lee, J. and Taylor, J. (2014). Exact post model selection inference for marginal screening. *arXiv:1402.5596*.

[45] Leng, C., Lin, Y. and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statist. Sinica*. 16, 1273–1284.

[46] Lin, D., Foster, D. P. and Ungar, L. H. (2011). VIF regression: A fast regression algorithm for large data. *Journal of the American Statistical Association*. 106, 232–247.

[47] Loftus, J. R. and Taylor, J. E. (2014). A significance test for forward stepwise model selection. *arXiv:1405.3920*.

[48] Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics* 42, 413–468.

[49] Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014). Correction to rejoinder to "A significance test for the Lasso". *The Annals of Statistics*. 42 (5), 2138–2139.

[50] Javanmard, A. and Montanari, A. (2013). Confidence intervals and hypothesis testing for high-dimensional regression. *arXiv:1306.3171*.

[51] Javanmard, A. and Montanari, A. (2014). Hypothesis testing in high-dimensional regression under the Gaussian random design model: Asymptotic theory. *arXiv:1301.4240*.

[52] Javanmard, A. and Montanari, A. (2015). On Online Control of False Discovery Rate. *arXiv:1502.06197*.

[53] Ma, X. and Adelstein, R.S. (2014). A point mutation in Myh10 causes major defects in heart development and body wall closure. *Circulation: Cardiovascular Genetics*. 7(3), 257-265.

[54] Mallows, C. L. (1973). Some Comments on CP. *Technometrics*. 15(4), 661–675.

[55] Mandozzi, J. and Buhlmann, P. (2015). A sequential rejection testing method for high-dimensional regression with correlated variables. *arXiv:1502.03300*.

[56] Meinshausen, N. and Buhlmann, P. (2006). Consistent neighbourhood selection for high-dimensional graphs with the Lasso. *The Annals of Statistics* 34(3), 1436–1462.

[57] Meinshausen, N. and Buhlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B* 74, 417–473.

[58] Meinshausen, N., Meier, L. and Buhlmann, P. (2009). p-Values for high-dimensional regression. *Journal of the American Statistical Association*. 104, 1671–1681.

[59] Ning, Y. and Liu, H. (2014). A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *arXiv:1412.8765*.

[60] Reid, S., Taylor, J. and Tibshirani, R. (2014). Post-selection point and interval estimation of signal sizes in Gaussian samples. *arXiv:1405.3340*.

[61] Ruggeri, R. B. (2005). Cholesteryl ester transfer protein: pharmacological inhibition for the modulation of plasma cholesterol levels and promising target for the prevention of atherosclerosis. *Current Topics in Medicinal Chemistry*. 5(3), 257-264

[62] Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*. 6(2) 461–464.

[63] Shah, R. and Samworth, R. (2013). Variable selection with error rate control: another look at stability selection. *Journal of the Royal Statistical Society: Series B Stat. Methodol.* 75, 55–80.

[64] Simes, R. J. (1986). An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73, 751–754.

[65] Stone, M. (1974). Cross-Validatory Choice and Assessment of Statistical Predictions, *Journal of the Royal StatisticalSociety. Series B (Methodological)*. 36(2), 111–147.

[66] Stone, J. D. (1974). A direct approach to false discovery rates. *J. R. Statist. Soc. B*. 64, 479–498.

[67] Su, W., Bogdan, M. and Candes, E. (2015). False Discoveries Occur Early on the Lasso Path. *arXiv:1511.01957*.

[68] Su, W. and Candes, E. (2015). SLOPE is adaptive to unknown sparsity and asymptotically minimax. *arXiv:1503.08393*.

[69] Taylor, J., Loftus, J. and Tibshirani, R. (2013). Tests in adaptive regression via the Kac-Rice formula. *arXiv:1308.3020*.

[70] Taylor, J., Lockhart, R., Tibshirani, R. J. and Tibshirani, R. (2014). Post-selection adaptive inference for Least Angle Regression and the Lasso. *arXiv:1401.3889*.

[71] Taylor, J. and Tibshirani, R. (2015). Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*. 112(25), 7629–7634.

[72] Tibshirani, R. J., Taylor, J., Lockhart, R., Tibshirani, R. (2014). Exact post-selection inference for forward stepwise and least angle regression. *http://arXiv:1401.3889*.

[73] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*. 58, 267–288.

[74] Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.* 7, 1456–1490.

[75] Van de Geer, S., Buhlmann, P., Ritov, Y. and Dezeure, R. (2014). On Asymptotically Optimal Confidence Regions And Tests For High-Dimensional Models. *The annals of statistics*. 42(3), 1166–1202.

[76] Voorman, A., Shojaie, A., and Witten, D. (2014). Inference in high dimensions with the penalized score test. *arXiv:1401.2678.*

[77] Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *The Annals of Statistics*. 37, 2178–2201.

[78] Wiens, B. (2003). A fixed-sequence Bonferroni procedure for testing multipleendpoints. *Pharmaceutical Statistics.* 2, 211–215.

[79] Wu, Y., Boos, D. D. and Stefanski, L. A. (2007). Controlling variable selection by the addition of pseudovariables. *Journal of the American Statistical Association*. 102, 235–243.

[80] Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* 38, 894–942.

[81] Zhang, C. and Zhang, S. S. (2014). Confidence intervals for low-dimensional parameters with high-dimensional data. *Journal of the Royal Statistical Society: Series B* 76, 217–242.

[82] Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research.* 7, 2541–2563.

[83] Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society. Series B (Methodological)*. 67(2), 301–320.

[84] Zou, H., Hastie, T. and Tibshirani, R. (2007). On the "degrees of freedom" of the lasso. *The Annals of Statistics*. 35, 2173–2192.

[85] Zou, H. and Li, R. (2008). One-step Sparse Estimates in Nonconcave Penalized Likelihood Models. *The Annals of Statistics*. 36(5), 1509–1533.

Appendix A

Derive the exact distribution in Chapter 3

Derive the exact distribution of covariance test statistics T_1, T_2 for orthogonal matrix and their joint distribution. We have $T_k = V_k(V_k - V_{k+1})$.

1. distribution of $T_1 = V_1(V_1 - V_2)$

$V_1 \geq V_2 \geq \cdots \geq V_m$ are the order statistic for $V \sim F(x)$, *i.i.d.* χ_1 variates (V_1 is the mth, V_2 is the $(m-1)$th in the order statistics). Then the cdf and pdf of V are

$$F(x) = (2\Phi(x) - 1)I(x > 0), \qquad f(x) = 2\phi(x)I(x > 0).$$

By the distribution of the order statistics, the joint distribution of V_1, V_2 is:

$$g(V_1, V_2) = \frac{m!}{((m-1)-1)!(m-(m-1)-1)!(m-m)!} F^{m-2}(V_2)(F(V_1) - F(V_2))^0 F(V_1)^0 f(V_2) f(V_1)$$

$$= m(m-1)F^{m-2}(V_2) f(V_2) f(V_1)$$

$$= m(m-1)(2\Phi(V_2) - 1)^{m-2} 2\phi(V_2) 2\phi(V_1) I(0 \le V_2 \le V_1)$$

$$= 4m(m-1)(2\Phi(V_2) - 1)^{m-2} \phi(V_2)\phi(V_1) I(0 \le V_2 \le V_1).$$

To find the distribution of $T_1 = V_1(V_1 - V_2)$, set $T = (V_1 - V_2)$, then

$$\begin{cases} V_1 = T_1/T \\ V_2 = T_1/T - T \end{cases} \Rightarrow J = \begin{vmatrix} 1/T & -T_1/T^2 \\ 1/T & -T_1/T^2 - 1 \end{vmatrix} = 1/T.$$

So the joint distribution of $T_1,\ T$ is

$$f(T_1, T) = \frac{1}{T} 4m(m-1)(2\Phi(T_1/T - T) - 1)^{m-2} \phi(T_1/T - T)\phi(T_1/T) I(0 \le T_1) I(0 \le T).$$

In order to find the distribution of $T_1 = V_1(V_1 - V_2)$, just integral out T, then we get

$$f(T_1) = \int_0^\infty \frac{1}{T} 4m(m-1)(2\Phi(T_1/T - T) - 1)^{m-2} \phi(T_1/T - T)\phi(T_1/T) dT.$$

2. Distribution of $T_2 = V_2(V_2 - V_3)$ (similar as T_1)

By the distribution of order statistics, the joint distribution of V_2, V_3 is:

$$\begin{aligned} g(V_2,V_3) &= \frac{m!}{((m-2)-1)!((m-1)-(m-2)-1)!(m-(m-1))!}\times \\ &\quad F^{m-3}(V_3)(F(V_2)-F(V_3))^0F(V_2)^1f(V_2)f(V_3) \\ &= m(m-1)F^{m-3}(V_3)F(V_2)f(V_2)f(V_3) \\ &= m(m-1)(2\Phi(V_2)-1)^{m-3}(2\Phi(V_2)-1)2\phi(V_2)2\phi(V_3)I(0\le V_3\le V_2) \\ &= 4m(m-1)(2\Phi(V_2)-1)^{m-3}(2\Phi(V_2)-1)\phi(V_2)\phi(V_3)I(0\le V_3\le V_2). \end{aligned}$$

To find the distribution of $T_2 = V_2(V_2 - V_3)$, set $T = (V_2 - V_3)$, then

$$\begin{cases} V_2 = T_2/T \\ V_3 = T_2/T - T \end{cases} \Rightarrow J = \begin{vmatrix} 1/T & -T_2/T^2 \\ 1/T & -T_2/T^2 - 1 \end{vmatrix} = 1/T.$$

So the joint distribution of T_2, T is

$$f(T_2,T) = \frac{1}{T}4m(m-1)(2\Phi(\frac{T_2}{T}-T)-1)^{m-3}(2\Phi(\frac{T_2}{T})-1)\phi(\frac{T_2}{T}-T)\phi(\frac{T_2}{T})I(0\le T_2)I(0\le T).$$

In order to find the distribution of $T_2 = V_2(V_2 - V_3)$, just integral out T, then we get

$$f(T_2) = \int_0^\infty \frac{1}{T}4m(m-1)(2\Phi(\frac{T_2}{T})-1)^{m-3}(2\Phi(\frac{T_2}{T})-1)\phi(\frac{T_2}{T}-T)\phi(\frac{T_2}{T})dT.$$

3. Joint distribution of $T_1 = V_1(V_1 - V_2), T_2 = V_2(V_2 - V_3)$, set $T = V_1$, then

$$\begin{cases} V_1 = T \\ V_2 = T - T_1/T \\ V_3 = T - T_1/T - \frac{T_2}{T - T_1/T} \end{cases} \Rightarrow J = \begin{vmatrix} 1 & 0 & 0 \\ 1 + 1/T^2 & -\frac{1}{T} & 0 \\ * & * & \frac{T}{T^2 - T_1} \end{vmatrix} = \frac{1}{T^2 - T_1}.$$

By the distribution of ordering statistics, the joint distribution of V_1, V_2, V_3 is:

$$g(V_1,V_2,V_3) = \frac{m!}{((m-2)-1)!((m-1)-(m-2)-1)!(m-(m-1)-1)!(m-m)!}\times$$
$$F^{m-3}(V_3)(F(V_2)-F(V_3))^0(F(V_1)-F(V_2))^0F(V_1)^0f(V_1)f(V_2)f(V_3)$$
$$= m(m-1)(m-2)F^{m-3}(V_3)f(V_1)f(V_2)f(V_3)$$
$$= m(m-1)(m-2)(2\Phi(V_3)-1)^{m-3}2\phi(V_1)2\phi(V_2)2\phi(V_3)I(0\le V_3\le V_2\le V_1)$$
$$= 8m(m-1)(m-2)(2\Phi(V_3)-1)^{m-3}\phi(V_1)\phi(V_2)\phi(V_3)I(0\le V_3\le V_2\le V_1).$$

So the joint distribution of T, T_1, T_2 is

$$f(T,T_1,T_2) = \frac{1}{T^2-T_1}8m(m-1)(m-2)(2\Phi(T-\frac{T_1}{T}-\frac{TT_2}{T^2-T_1})-1)^{m-3}\times$$
$$\phi(T)\phi(T-T_1/T)\phi(T-\frac{T_1}{T}-\frac{TT_2}{T^2-T_1})I(0\le T_2)I(0\le T_1)I(0\le T).$$

In order to find the distribution of T_1, T_2, just integral out T, then we have

$$f(T_1,T_2) = \int_0^\infty \frac{1}{T^2-T_1}8m(m-1)(m-2)(2\Phi(T-\frac{T_1}{T}-\frac{TT_2}{T^2-T_1})-1)^{m-3}\times$$
$$\phi(T)\phi(T-\frac{T_1}{T})\phi(T-\frac{T_1}{T}-\frac{TT_2}{T^2-T_1})dT.$$

Appendix B

Proof of Theorems in Chapter 5

Proof of Lemma 2. Suppose the first k_0 variables are significant and the order in the true path from most significant to less significant is $x_1,\ x_2,\ \cdots,\ x_{k_0}$.

Next we are going to prove for any $k < k_0,\ P^k < P^{k+1}$, where $P^k = \min\{P_1^k,\ P_2^k, \cdots, P_{m-k+1}^k\}$ and P_i^k is the p-value of $H_{0k,i} : \beta_{k+i-1} = 0$, which in the form of test statistics in null distribution is $T_{k+i-1} \sim \text{Exp}(1/i)$.

Write all the hypotheses in terms of the distribution of T_k at kth step.

At kth step: $H_{0k} : T_k \sim \text{Exp}(1),\ T_{k+1} \sim \text{Exp}(1/2),\ T_{k+2} \sim \text{Exp}(1/3),\ \cdots,\ T_m \sim \text{Exp}(1/(m-k+1))$;

Then the according p-values are, $P_1^k = P(T > T_{k|\text{Exp}(1)}),\ P_2^k = P(T > T_{k+1|\text{Exp}(1/2)}),\ P_3^k = P(T > T_{k+2|\text{Exp}(1/3)}),\ \cdots,\ P_{m-k+1}^k = P(T > T_{m|\text{Exp}(1/(m-k+1))})$.

Similarly, at $(k+1)$th step: $H_{0(k+1)} : T_{k+1} \sim \text{Exp}(1),\ T_{k+2} \sim \text{Exp}(1/2),\ \cdots,\ T_m \sim \text{Exp}(1/(m-k))$;

Then, $P_1^{k+1} = P(T > T_{k+1|\text{Exp}(1)})$, $P_2^{k+1} = P(T > T_{k+2|\text{Exp}(1/2)})$, $\cdots$, $P_{m-k}^{k+1} = P(T > T_{m|\text{Exp}(1/(m-k))})$.

From the p-value of the exponential definition, we know

$$P_2^k = P(T > T_{k+1|\text{Exp}(1/2)}) < P(T > T_{k+1|\text{Exp}(1)}) = P_1^{k+1};$$

$$P_3^k = P(T > T_{k+2|\text{Exp}(1/3)}) < P(T > T_{k+2|\text{Exp}(1/2)}) = P_2^{k+1};$$

$$\cdots$$

$$P_{m-k+1}^k = P(T > T_{m|\text{Exp}(1/(m-k+1))}) < P(T > T_{m|\text{Exp}(1/(m-k))}) = P_{m-k}^{k+1}.$$

Using the same test statistics, the p-value is smaller for null distribution which has smaller mean. Therefore, $\min\{P_1^k, P_2^k, \cdots, P_{m-k}^k, P_{m-k+1}^k\} \leq \min\{P_2^k, P_3^k, \cdots, P_{m-k+1}^k\} < \min\{P_1^{k+1}, P_2^{k+1}, \cdots, P_{m-k}^{k+1}\}$, i.e. $P^k < P^{k+1}$. ■

Proof of Theorem 3. Let $I_0 = D^c$ be the set of indices corresponding to the true null hypotheses, having m_0 members. D is denoted as the true significant variables in the model. Then $m_0 = m - |D|$. Let k be the first selected variable which follows the true hypothesis. So $k-1$ is the last correctly variables and $k - 1 + m_0 \leq m$. From here, we get $\frac{1}{m-k+1} \leq \frac{1}{m_0}$.

Since variable k is been rejected at step k, then

$$\begin{aligned} Pr(P^k \leq \frac{\alpha/m}{m-k+1}) &= 1 - \left(1 - \frac{\alpha/m}{m-k+1}\right)^{m-k+1} \\ &\leq 1 - \left(1 - \frac{\alpha/m}{m_0+1}\right)^{m_0+1} \\ &\leq \alpha/m. \end{aligned}$$

The first inequality holds since $1-\left(1-\frac{\alpha}{m-k+1}\right)^{m-k+1}$ is a decreasing function of $m-k+1$, it will get the maximum value at $m_0+1(m-k+1>m_0)$. The last inequality holds when $m_0=0$. The inequality above means if we wrongly reject a true hypothesis, there has to be a true hypothesis with p-value at most $\frac{\alpha}{m}$ at each step.

Then define $V \geq 1 = \left\{P_i \leq \frac{\alpha}{m} \text{ for some } i \in I_0\right\}$. Whatever the set of true hypotheses I_0 is, we have $Pr(V \geq 1) \leq \frac{m_0}{m}\alpha$ by the Bonferroni inequalities. Therefore, FWER $\leq \frac{m_0}{m}\alpha \leq \alpha$. ■

Appendix C

Proof of Theorems and Lemmas in Chapter 6

Proof of Theorem 4. In our proposed Hybrid BBLSD procedure, define

$$\delta^k = 1 - \left[1 - \min\left(1, \frac{m}{m-k+1}q\right)\right]^{\frac{1}{m}}$$

as the critical values at step k. Let m_0 be the number of true null hypotheses. Assume D is the true significant variables and $\hat{D}_n$ is the choosen variables by lasso path. Suppose V is the number of falsely rejected variables and R is the number of rejected, then $V = |\hat{D}_n \cap D^c|$ and $R = |\hat{D}_n|$; Also m_1 is denoted as the number of false null hypotheses, then $m_1 = |D| = m - m_0$.

Let proportion of false discovery be $Q = \frac{V}{R}$, and false discovery rate is

$$Q_e = FDR = E(Q) = E\left(\frac{V}{R \vee 1}\right).$$

If $m_0 = 0$, then $V = 0$, then $Q = 0, Q_e = 0 \leq q$.

If $m_0 = m$, then $V = R$, then

$$\begin{aligned} Q_e = E(I_{V>0}) &= \Pr(V > 0) = \Pr(P^1 \leq \delta^1) \\ &= 1 - \left[\Pr(P_1^1 > \delta^1)\right]^m = 1 - (1 - \delta^1)^m = q. \end{aligned}$$

Therefore, we assume in the rest of the proof that $1 \leq m_0 \leq m - 1$. Denote $p^{1'}, \cdots, p^{m_1'}$ be the p-value corresponding to m_1 false hypotheses and $p^{1\star}, \cdots, p^{m_0\star}$ be the p-value corresponding to m_0 true hypotheses. Define the expectation of Q conditioning on $p^{1'}, \cdots, p^{m_1'}$.

$$Q_e(p^{1'}, \cdots, p^{m_1'}) = E(Q|p^{1'}, \cdots, p^{m_1'})$$

Next show $Q_e(p^{1'}, \cdots, p^{m_1'}) \leq q$.

Define $0 \leq S \leq m$ be the largest j satisfying $p^{(1)'} \leq \delta^1, \cdots, p^{(j)'} \leq \delta^j$; $S = 0$ if $p^{(1)'} > \delta^1$.

Now,

$$\begin{aligned} Q_e(p^{1'}, \cdots, p^{m_1'}) &= E\left(\frac{V}{R} I_{V>0} | p^{1'}, \cdots, p^{m_1'}\right) \\ &\leq E\left(\frac{V}{S+V} I_{V>0} | p^{1'}, \cdots, p^{m_1'}\right) \\ &\leq \frac{m_0}{S+m_0} E\left(I_{V>0} | p^{1'}, \cdots, p^{m_1'}\right) \\ &= \frac{m_0}{S+m_0} \Pr(\min(p^{1\star}, \cdots, p^{m_0\star}) \leq \delta^{S+1}) \\ &\leq \frac{m_0}{S+m_0} \Pr(P^1 \leq \delta^{S+1}) \\ &= \frac{m_0}{S+m_0} \Pr(\min\{p_i^1\} \leq \delta^{S+1}) \end{aligned}$$

$$= \frac{m_0}{S+m_0}[1-(1-\delta^{S+1})^m]$$
$$= \frac{m_0}{S+m_0}\min\{1, \frac{m}{m-S}q\})$$
$$\leq \frac{m_0 m}{(S+m_0)(m-S)}q$$
$$\leq q.$$

Therefore, FDR$\leq q$. ■

Proof of Theorem 5. At each step, we use the first s hypotheses and use the composite $P^{k_s} = \min\{P_1^k, \cdots, P_s^k\}$, and we have $P^{k_s} \leq P^k$ using all the hypotheses. Then when compare with each $\Pr(P^{k_s} \leq \delta^k)$, we also have $Pr(P^{k_s} \leq \delta^k) = Pr(P^k \leq \delta^k)$. If $m_0 = m$, then $V = R$, $Q_e = E(I_{V>0}) = \Pr(V > 0) = \Pr(P^{1_s} \leq \delta^1) \leq \Pr(P^1 \leq \delta^1) = q$. For $1 \leq m_0 \leq m-1$, the expectation of Q conditioning on $p^{1'}, \cdots, p^{m'_1}$.

$$Q_e(p^{1'_s}, \cdots, p^{m'_{1_s}}) = E\left(\frac{V}{R} I_{V>0} | p^{1'_s}, \cdots, p^{m'_{1_s}}\right)$$
$$\leq \frac{m_0}{S+m_0}\Pr(\min\{p_i^{1_s}\} \leq \delta^{S+1})$$
$$\leq \frac{m_0}{S+m_0}\Pr(\min\{p_i^1\} \leq \delta^{S+1})$$
$$= \frac{m_0}{S+m_0}\min\{1, \frac{m}{m-S}q\})$$
$$\leq \frac{m_0 m}{(S+m_0)(m-S)}q$$
$$\leq q.$$

Therefore, FDR$_s \leq q$. ■

Proof of Lemma 6 Derive the exact distribution of pivotal statistics for TailStop and Hybrid Sequential procedures.

Before we proof Lemma 6, we need to use another Lemma for the distribution of sum of independent exponential distribution.

Lemma 7 *Let* $(X_i)_{i=1,\cdots,m}, m \geq 2$*, be independent exponential random variables with pairwise distinct respective parameters* λ_i*. Then the density of their sum is*

$$f_{X_1+X_2+\cdots+X_m}(x) = \left[\prod_{i=1}^{m} \lambda_i\right] \sum_{j=1}^{m} \frac{e^{-\lambda_j x}}{\prod_{k=1,k\neq j}^{m}(\lambda_k - \lambda_j)}, \qquad x > 0$$

For TailStop, assume $k_0 = 0$. At 1^{st} step, $\mathrm{TS}_1 = \sum_{j=1}^{m} T_j$ is the pivotal statistic, where $T_j \sim \mathrm{EXP}(\lambda = j)$. Thus, by Lemma 7, the density function of TS_1 is

$$\begin{aligned}
f_{\mathrm{TS}_1}(x) &= \left[\prod_{i=1}^{m} \lambda_i\right] \sum_{j=1}^{m} \frac{e^{-\lambda_j x}}{\prod_{k=1,k\neq j}^{m}(\lambda_k - \lambda_j)}, \qquad x > 0 \\
&= \left[\prod_{i=1}^{m} i\right] \sum_{j=1}^{m} \frac{e^{-jx}}{\prod_{k=1,k\neq j}^{m}(k - j)} \\
&= m! \sum_{j=1}^{m} \frac{e^{-jx}}{\prod_{k=1}^{j-1}(k-j)\prod_{k=j+1}^{m}(k-j)} \\
&= m! \sum_{j=1}^{m} \frac{e^{-jx}}{(-1)^{j-1}(j-1)!(m-j)!} \\
&= \sum_{j=1}^{m} \frac{m! e^{-jx}}{(-1)^{j-1}(j-1)!(m-j)!} \\
&= m \sum_{j=1}^{m} (-1)^{j-1} \binom{m-1}{j-1} e^{-jx} \\
&= m e^{-x} \sum_{j=1}^{m} (-1)^{j-1} \binom{m-1}{j-1} e^{-(j-1)x}
\end{aligned}$$

$$= me^{-x}\sum_{j=0}^{m-1}(-1)^j\binom{m-1}{j}e^{-jx}$$

$$= me^{-x}(1-e^{-x})^{m-1}.$$

For Hybrid Sequential, at, the 1^{st} step, $\mathrm{HS}_1 = \max_{j\in\{1,\cdots,m\}}(jT_j)$ is the pivotal statistic, where $T_j \sim \mathrm{EXP}(\lambda = j)$ and $jT_j \sim \mathrm{EXP}(1)$, which is the equivalent to the mth order statistics. By the density function of order statistics, then the density function of HS_1 is

$$f_{\mathrm{HS}_1}(x) = mF(x)^{m-1}f(x) = me^{-x}(1-e^{-x})^{m-1}.$$

which has the same distribution as TS_1. But the realization of them are not the same.

Generalized to $k_0 > 0$ case, and suppose the selection has move forward to step k_0+1. $\mathrm{TS}_{k_0+1} = \sum_{j=k_0+1}^{m} T_j$, where $T_{k_0+1} \sim \mathrm{Exp}(\lambda = k_0+1), T_{k_0+2} \sim \mathrm{Exp}(\lambda = k_0+2), \cdots$. Then the density function of TS_{k_0+1} is

$$\begin{aligned}
f_{\mathrm{TS}_{k_0+1}}(x) &= \left[\prod_{i=k_0+1}^{m}\lambda_j\right]\sum_{j=k_0+1}^{m}\frac{e^{-\lambda_j x}}{\prod_{k=k_0+1,k\neq j}^{m}(\lambda_k-\lambda_j)}, \qquad x>0\\
&= \frac{m!}{k_0!}\sum_{j=k_0+1}^{m}\frac{e^{-jx}}{(-1)^{j-(k_0+1)}(j-(k_0+1))!(m-j)!}\\
&= \frac{m!}{k_0!(m-(k_0+1))!}\sum_{j=1}^{m}\frac{(m-(k_0+1))!e^{-jx}}{(-1)^{j-(k_0+1)}(j-(k_0+1))!(m-j)!}\\
&= m\binom{m-1}{k_0}e^{-(k_0+1)x}\sum_{j=k_0+1}^{m}(-1)^{j-(k_0+1)}\binom{m-(k_0+1)}{j-(k_0+1)}e^{-(j-(k_0+1))x}\\
&= m\binom{m-1}{k_0}e^{-(k_0+1)x}(1-e^{-x})^{m-(k_0+1)}.
\end{aligned}$$

For Hybrid Sequential, $\mathrm{HS}_{k_0+1} = \max_{j\in\{k_0+1,\cdots,m\}}((j-k_0)T_j)$, where $T_{k_0+1} \sim \mathrm{EXP}(\lambda = 1)$ and $T_{k_0+2} \sim \mathrm{EXP}(\lambda = 2)$.

The above HS_{k_0+1} is the equivalent to $\mathrm{HS}_{k_0+1} = \max_{j\in\{1,\cdots,m-(k_0+1)\}}(jT_j)$ and $T_j \sim \mathrm{Exp}(1)$, then HS_{k_0+1} is the $m-(k_0+1)$th order statistics. By the density function of order statistics, then the density function of HS_{k_0+1} is

$$f_{\mathrm{HS}_{k_0+1}}(x) = (m-k_0)F(x)^{m-k_0-1}f(x) = me^{-x}(1-e^{-x})^{m-(k_0+1)}.$$

At step k_0+1, pivotal statistics for TailStop and Hybrid sequential don't share the same density functions.