SCIENCE CHINA
Mathematics

• ARTICLES •

April 2019 Vol. 62 No. 4: 771–782
https://doi.org/10.1007/s11425-016-9186-x

Variable selection procedures from multiple testing

Baoxue Zhang[1,*], Guanghui Cheng[2], Chunming Zhang[3] & Shurong Zheng[2]

[1]*School of Statistics, Capital University of Economics and Business, Beijing* 100070*, China;*
[2]*School of Mathematics and Statistics and Key Laboratory of Applied Statistics of Ministry of Education, Northeast Normal University, Changchun* 130024*, China;*
[3]*Department of Statistics, University of Wisconsin-Madison, Madison, WI* 53706*, USA*

*Email: zhangbaoxue@cueb.edu.cn, chenggh*845*@nenu.edu.cn, cmzhang@stat.wisc.edu, zhengsr@nenu.edu.cn*

Received December 7, 2016; accepted October 18, 2017; published online June 8, 2018

Abstract Variable selection has played an important role in statistical learning and scientific discoveries during the past ten years, and multiple testing is a fundamental problem in statistical inference and also has wide applications in many scientific fields. Significant advances have been achieved in both areas. This study attempts to find a connection between the adaptive LASSO (least absolute shrinkage and selection operator) and multiple testing procedures in linear regression models. We also propose procedures based on multiple testing methods to select variables and control the selection error rate, i.e., the false discovery rate. Simulation studies demonstrate that the proposed methods show good performance relative to controlling the selection error rate under a wide range of settings.

Keywords variable selection, multiple testing, adaptive LASSO, false discovery rate, linear regression

MSC(2010) 47N30, 62F03

Citation: Zhang B X, Cheng G H, Zhang C M, et al. Variable selection procedures from multiple testing. Sci China Math, 2019, 62: 771–782, https://doi.org/10.1007/s11425-016-9186-x

1 Introduction

The classical linear regression model is written as follows:

$$Y = \beta_0 + X_1\beta_1 + \cdots + X_p\beta_p + \varepsilon, \tag{1.1}$$

where Y is the response variable, $(X_1, \dots, X_p)$ are the potential explanatory variables, and ε is noise with mean zero and variance σ^2. To increase prediction accuracy and facilitate model interpretation, many methods have been developed to exclude insignificant predictors. Prior to 1990, traditional methods were used for model selection, such as the Akaike information criterion, the Bayesian information criterion, and stepwise selection techniques. In 1996, Tibshirani [16] proposed the LASSO, a simultaneous estimation and variable selection method that solves the l_1-penalized regression problem of finding $\{\beta_j\}$ to minimize

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_j x_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|,$$

[*] Corresponding author

where $(y_1, \ldots, y_n)$ and $(x_{1j}, \ldots, x_{nj})$, $j = 1, \ldots, p$ are the observations of Y and X_j, respectively. Note that extensions to the LASSO are described in the literature [19,20,23,24]. To address the inconsistency of the LASSO, the adaptive LASSO, which achieves an oracle property, has been developed [23]. Many other popular methods, such as smoothly clipped absolute deviation [10], least angle regression [8], boosting [3], the Dantzig selector [5], and nearly isotonic regression [17] have been proposed.

Benjamini and Hochberg [2] proposed a framework to control the expected proportion of false rejections in a multiple hypothesis testing problem, and it has been claimed that this framework gains more applicable power for calling for false discovery rate (FDR) control rather than the family-wise error rate (FWER). Compared to the FWER, the FDR is a less conservative quantity to control, particularly with a large number of tests. Recently, multiple hypothesis testing for FDR control has received significant attention. Storey [14] first proposed a point estimate for the FDR in realistic applications and improved Benjamini and Hochberg's procedure by estimating the number of true null hypotheses, which appears to be more effective, flexible, and powerful. Zhang [21] considered mean and median filters in their FDR_L procedure to alleviate the "lack of identification" phenomenon of the FDR procedure that occurs with large-scale imaging data. Other FDR studies can be found in the literature [7, 9, 11, 15, 22]. Variable selection methods that use a multiple testing procedure can also be found in the literature [4, 12]. These studies achieved consistent variable selection via a pre-specified level α that tends to zero as sample sizes tend to infinity in high-dimensional linear regression.

Although many methods for variable selection in linear models can select predictors due to an oracle property, to the best of our knowledge, most methods cannot control the selection error rate under finite sample sizes using a selection tuning parameter. However, there are some exceptions. For example, Wasserman and Roeder [18] controlled the FWER for regression coefficients using a t-statistic based on multi-stage screening, and Meinshausen et al. [13] split data into two parts and applied adjusted p-values to control the FWER and FDR when the sample size n tends to infinity. Barber and Candes [1] constructed a knockoff-filter method to control the FDR under finite sample settings. However, the knockoff-filter appears to be inefficient for the non-sparse case (i.e., when the proportion of non-null effect β_j is large). Moreover, when the sample size satisfies $p < n < 2p$, the knockoff-filter method forms a $(2p - n)$-dimensional vector $\boldsymbol{y}'$ with independent and identically distributed (i.i.d.) components from $N(0, \hat{\sigma}^2)$, where $\hat{\sigma}^2$ estimates noise level σ. Then, the knockoff-filter method augments the response vector $\boldsymbol{y} = (y_1, \ldots, y_n)$ with $\boldsymbol{y}'$ and the design matrix $\boldsymbol{X}$ with $2p - n$ rows of zeros, which leads to a linear model with $2p$ observations. Therefore, the estimate $\hat{\sigma}$ has a significant effect on knockoff performance. Differing from previous studies, this study attempts to find a connection between the adaptive LASSO and multiple testing procedures in linear regression models. In addition, we also attempt to select variables and control the selection error rate via multiple testing methods.

The remainder of this paper is organized as follows. Section 2 describes our motivation, establishes the connection between the adaptive LASSO and multiple hypothesis testing procedures in linear models, and proposes testing procedures to control the FDR for variable selection. Section 3 extends the testing procedures to control the FDR for variable selection in generalized linear models. Simulation results are given in Section 4, and Section 5 provides a practical data analysis. Conclusions and suggestions for future work are given in Section 6.

2 Multiple testing procedures for variable selection in linear models

2.1 Motivation

Assume that data $\{(y_i, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n\}$ for n individuals are from the following linear model:

$$y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i, \tag{2.1}$$

without loss of generality, and the intercept term is set to zero. For simplicity, we have

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.2}$$

where $\boldsymbol{Y} = (y_1, \dots, y_n)^{\mathrm{T}}$, $y_1, \dots, y_n$ are independent response samples, and $\boldsymbol{X}_{+j} = (x_{1j}, \dots, x_{nj})^{\mathrm{T}}$ $(j = 1, \dots, p)$, $\boldsymbol{X} = (\boldsymbol{X}_{+1}, \dots, \boldsymbol{X}_{+p})$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^{\mathrm{T}}$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^{\mathrm{T}}$, and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. random samples with mean 0 and variance σ^2. In this paper, we consider the case of $n > p$ with fixed p. Then, $\tilde{\boldsymbol{\beta}} = (\tilde{\beta}_1, \dots, \tilde{\beta}_p)^{\mathrm{T}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}$ is the least square estimator of the parameter vector $\boldsymbol{\beta}$ and $\hat{\sigma}^2 = (\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\tilde{\boldsymbol{\beta}})/(n-p)$ is an unbiased estimator of σ^2. Let $\mathcal{A}$ be the set of indices satisfying $\mathcal{A} = \{j \mid \beta_j \neq 0, j = 1, \dots, p\}$, which indicates the true active set.

Then, we show a connection between the adaptive LASSO and multiple testing procedures. By transformation (2.2), we have

$$\sqrt{n}\tilde{\boldsymbol{\beta}} = \sqrt{n}\boldsymbol{\beta} + \sqrt{n}\tilde{\boldsymbol{\varepsilon}}, \tag{2.3}$$

where $\tilde{\boldsymbol{\varepsilon}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon}$. Note that $\tilde{\boldsymbol{\varepsilon}}$ is distributed with mean $\underline{0}$ and covariance matrix $\sigma^2 n^{-1}\boldsymbol{C}_n^{-1}$ conditioned on $\boldsymbol{X}$, where $\boldsymbol{C}_n = n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$. Furthermore, assume that $\boldsymbol{C}_n = (C_{ij}^n)_{i,j=1}^p = n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \xrightarrow{p} \boldsymbol{C}$, $\boldsymbol{C}$ is a positive definite matrix, and the inverse $\boldsymbol{C}_n^{-1} = (C_n^{ij})_{i,j=1}^p \xrightarrow{p} \boldsymbol{C}^{-1}$. Then, the proposed estimate is given as follows:

$$\hat{\boldsymbol{\beta}}^t = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\left\{\boldsymbol{\beta} : f(\boldsymbol{\beta}) = n\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 + \lambda_n \sum_{j=1}^p w_j|\beta_j|\right\}, \tag{2.4}$$

where $\|\cdot\|$ denotes the L_2 norm, $\{w_j, j = 1, \dots, p\}$ are weights, and λ_n is a non-negative tuning parameter. By Kuhn-Tucker conditions, we have

$$\frac{\partial f(\boldsymbol{\beta})}{\partial \beta_j} = 2n(\beta_j - \tilde{\beta}_j) + \lambda_n \cdot w_j \cdot \operatorname{sgn}(\beta_j) = 0$$

for $j = 1, \dots, p$, where $\operatorname{sgn}(\cdot)$ is the sign function. Then, we have $\beta_j > 0 \Leftrightarrow n\tilde{\beta}_j/w_j > \lambda_n/2$ and $\beta_j < 0 \Leftrightarrow n\tilde{\beta}_j/w_j < -\lambda_n/2$. It follows that

$$n|\tilde{\beta}_j|/w_j > \lambda_n/2 \Leftrightarrow \beta_j \neq 0.$$

For simplicity, weights w_j are taken as $w_j = C_n^{jj}\hat{\sigma}^2/|\tilde{\beta}_j|$. Then, the adaptive LASSO selects the variables from (2.4) as follows: $\hat{\mathcal{A}} = \{j \mid n|\tilde{\beta}_j|^2/(C_n^{jj}\hat{\sigma}^2) > \lambda_n/2\}$.

In fact, for the multiple testing problem

$$H_{0j} : \beta_j = 0 \quad \text{vs.} \quad H_{1j} : \beta_j \neq 0, \quad j = 1, \dots, p,$$

the selected predictors also satisfy $\{j \mid t_{(j)}^2 > c_j\}$, where $t_j^2 = n|\tilde{\beta}_j|^2/(C_n^{jj}\hat{\sigma}^2)$, and $t_1^2, \dots, t_j^2$ are ordered as $t_{(1)}^2, \dots, t_{(j)}^2$, and c_j is determined by a previously reported procedure [2] to control FDR because

$$\sqrt{n}\tilde{\beta}_j/\sqrt{C_n^{jj}\hat{\sigma}^2} \xrightarrow{d} N(0, 1)$$

under H_{0j}. Then, it can be seen that the adaptive LASSO and the multiple testing procedure are very similar. The difference is related to the tuning parameter λ_n and threshold c_j, where the tuning parameter λ_n is determined by a cross-validation method and c_j is determined by a multiple testing procedure where FDR is controlled at a pre-specified level α. It is well known that the threshold $c_j = \Phi^2(1 - \frac{j\alpha}{2p})$ for the Benjamini-Hochberg (BH) procedure when the statistic $\sqrt{n}\tilde{\beta}_j/\sqrt{C_n^{jj}\hat{\sigma}^2}$ is approximated to Gaussian distribution; therefore, if we determine that λ_n is equal to c_j after ordering statistic $n|\tilde{\beta}_j|^2/(C_n^{jj}\hat{\sigma}^2)$, there is equivalence between (2.4) and the BH procedure.

Therefore, we propose methods to select predictors based on multiple testing procedures.

2.2 Multiple testing procedures for variable selection in linear models

First, we review multiple testing procedures to control the FDR, i.e., the BH [2], Storey [14], and Z-mean [21] methods. Here, let $p_1, \dots, p_p$ be p-values from the following multiple testing problem:

$$H_{0j} : \beta_j = 0 \quad \text{vs.} \quad H_{1j} : \beta_j \neq 0. \tag{2.5}$$

Without loss of generality, we assume that i_j satisfies $p_{i_j} = p_{(j)}$, where $(i_1, i_2, \dots, i_p)$ is a permutation of $(1, \dots, p)$ and $p_{(k)}$ is the k-th largest p-value. Then, the BH procedure at level α finds the greatest k such that $p_{(k)} \leqslant \frac{k}{p}\alpha$ and rejects all H_{0i_j} for $j = 1, \dots, k$.

The BH procedure represents a new perspective for the multiple-hypothesis testing error measure problem, which is a sequential p-value method to control the FDR at $(m_0/p)\alpha$ (m_0 denotes the number of true null hypotheses). However, a weakness of the BH procedure is that the error rate is controlled for all values of m_0 simultaneously without using any information about m_0 in the data. Storey [14] proposed a point estimate of the FDR and provided finite and large sample results for consideration relative to realistic applications. In addition, Storey presented an estimate of m_0. The procedure at level α can be described as follows:

- Give a pre-specified $0 < \omega < 1$, e.g., $\omega = 0.1$;
- Define the threshold as follows:

$$t_\alpha(\widehat{\mathrm{FDR}}_\omega) = \sup\{0 \leqslant t \leqslant 1 : \widehat{\mathrm{FDR}}_\omega(t) \leqslant \alpha\},$$

where

$$\widehat{\mathrm{FDR}}_\omega(t) = \frac{\hat{m}_0 t}{\{R(t) \vee 1\}}, \quad W(\omega) = p - R(\omega), \quad R(t) = \#\{p_i : p_i \leqslant t\}, \quad \text{and} \quad \hat{m}_0 = W(\omega)/\{(1-\omega)\};$$

- Reject H_{0j} if $p_j \leqslant t_\alpha(\widehat{\mathrm{FDR}}_\omega)$.

As can be seen, the above Storey procedure is essentially a threshold-based approach for multiple testing problems, where the data-driven threshold $t_\alpha(\widehat{\mathrm{FDR}}_\omega)$ plays an important role. In addition, a null hypothesis is rejected if the corresponding p-value is less than or equal to the threshold $t_\alpha(\widehat{\mathrm{FDR}}_\omega)$. It can also be seen that $t_\alpha(\widehat{\mathrm{FDR}}_\omega)$ depends on both the estimates $\widehat{\mathrm{FDR}}_\omega(t)$ and the control level α, and $t_\alpha(\widehat{\mathrm{FDR}}_\omega)$ is a nondecreasing function of α. This indicates that, when α is reduced to less than $\inf_{0<t\leqslant 1}\widehat{\mathrm{FDR}}_\omega$, the threshold $t_\alpha(\widehat{\mathrm{FDR}}_\omega)$ will drop to zero in the Storey procedure. Accordingly, the Storey FDR procedure can only reject hypotheses with p-values that are exactly equal to zero. This phenomenon is referred to as "lack of identification". Zhang [21] proposed using a mean filter in the FDR_L procedure to alleviate the "lack of identification" phenomenon that occurs in the FDR procedure. The Z-mean procedure at level α is described as follows:

- Give size k, where k is a positive integer;
- Let N_i be a set of indices of the neighborhood points of p_i with size k;
- By the mean filter, obtain $p_i^* = \mathrm{mean}(\{p_j : j \in N_i\})$;
- Define the threshold as follows:

$$t_\alpha(\widehat{\mathrm{FDR}}_L) = \sup\{0 \leqslant t \leqslant 1 : \widehat{\mathrm{FDR}}_L(t) \leqslant \alpha\},$$

where $\widehat{\mathrm{FDR}}_L(t) = \frac{W^*(\omega)\widehat{G}^*(t)}{\{R^*(t)\vee 1\}\{1-\widehat{G}*(\omega)\}}$, $W^*(\omega) = p - R^*(\omega)$, $R^*(t) = \#\{p_i^* : p_i^* \leqslant t\}$, and $\widehat{G}^*(t)$ is the empirical distribution function of $p_1^*, \dots, p_p^*$, which is constructed as follows:

$$\widehat{G}^*(t) = \begin{cases} \dfrac{\sum_{i=1}^p I\{p_i^* \geqslant (1-t)\}}{2\sum_{i=1}^p (p_i^* > 0.5) + \sum_{i=1}^p I(p_i^* = 0.5)}, & \text{if } 0 \leqslant t \leqslant 0.5, \\ 1 - \dfrac{\sum_{i=1}^p I\{p_i^* \geqslant t\}}{2\sum_{i=1}^p (p_i^* > 0.5) + \sum_{i=1}^p I(p_i^* = 0.5)}, & \text{if } 0.5 < t \leqslant 1; \end{cases}$$

- Reject H_{0j} satisfying $p_j^* \leqslant t_\alpha(\widehat{\mathrm{FDR}}_L)$.

Compared with the Storey procedure, the original p-values in the Storey method are replaced by "local aggregated" p^*-values.

In the following, we proposed three new selection procedures.

As $\sqrt{n}\tilde{\beta}_j/\sqrt{C_n^{jj}\hat{\sigma}^2} \xrightarrow{d} N(0,1)$ under H_{0j}, p-values can be defined as follows:

$$p_j = \mathrm{P}(T > |t_j|), \quad j = 1, \dots, p,$$

where $T \xrightarrow{d} N(0,1)$ and t_i is the observation of $\sqrt{n}\tilde{\beta}_j/\sqrt{C_n^{jj}\hat{\sigma}^2}$. Then, the permutation of p-values is

$$p_{(1)} \leqslant \cdots \leqslant p_{(p)}.$$

Without loss of generality, assume that i_j satisfies $p_{i_j} = p_{(j)}$, where $(i_1, i_2, \ldots, i_p)$ is a permutation of $(1, \ldots, p)$. Then, the BH selection procedure at level α can be described as follows:

- By the BH procedure to control the FDR, find the k-th largest p-value such that $p_{(k)} \leqslant \frac{k}{p}\alpha$;
- Let estimate $\hat{\mathcal{A}}$ of $\mathcal{A}$ be $\{i_j, j = 1, \ldots, k\}$.

The Storey selection procedure at level α can be described as follows:

- By the Storey procedure to control the FDR, obtain the threshold $t_\alpha(\widehat{\mathrm{FDR}}_\omega)$;
- Let estimate $\hat{\mathcal{A}}$ of $\mathcal{A}$ be $\{j : p_j \leqslant t_\alpha(\widehat{\mathrm{FDR}}_\omega)\}$.

The Z-mean selection procedure at level α can be described as follows:

- By the Z-mean procedure to control the FDR, obtain the threshold $t_\alpha(\widehat{\mathrm{FDR}}_L)$;
- Let estimate $\hat{\mathcal{A}}$ of $\mathcal{A}$ be $\{i : p_i^* \leqslant t_\alpha(\widehat{\mathrm{FDR}}_L)\}$.

Theorem 2.1. *If $n^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} \xrightarrow{p} \boldsymbol{C}$, where $\boldsymbol{C}$ is a diagonal matrix, then the BH selection procedure at level α satisfies that the error rate selecting insignificant predictors as significant predictors at level α asymptotically approaches $(m_0/p)\alpha \leqslant \alpha$, where m_0 is the number of zero β_j $(j = 1, \ldots, p)$.*

Proof. Note that $\sqrt{n}(\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}) \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{C}_n^{-1})$. Then, let $\tilde{T}_j = \sqrt{n}\tilde{\beta}_j/\sqrt{C_n^{jj}}\hat{\sigma}$ and $T_j = \sqrt{n}\tilde{\beta}_j/\sqrt{C_n^{jj}}\sigma$. Here, $\tilde{T}_j - T_j = \sqrt{n}\tilde{\beta}_j/\sqrt{C_n^{jj}}(\frac{1}{\hat{\sigma}} - \frac{1}{\sigma})$. As $\hat{\sigma}$ is a consistent estimator of σ, then $\tilde{T}_j - T_j \xrightarrow{p} 0$ as $n \to \infty$. Moreover, we have $\mathrm{Cov}(T_i, T_j) = C_n^{ij}/\sqrt{C_n^{ii}C_n^{jj}}$. When $\boldsymbol{C}$ is a diagonal matrix, for any $i \neq j$, $\mathrm{Cov}(T_i, T_j) = 0$ as $n \to \infty$, we know that $T_1, \ldots, T_p$ are asymptotically mutually independent due to normality. Then, $(\tilde{T}_1, \ldots, \tilde{T}_p)$ has the same asymptotic distribution as $(T_1, \ldots, T_p)$ and the vector's components are asymptotically independent. Define $p_j = \mathrm{P}(\tilde{T}_j > |\tilde{t}_j|)$. Thus, the p-values $p_1, \ldots, p_p$ are asymptotically mutually independent. By the proof of [2], we find that the error rate selecting insignificant predictors as significant predictors among the total rejections asymptotically approaches $(m_0/p)\alpha \leqslant \alpha$, where m_0 is the number of zero β_j $(j = 1, \ldots, p)$. □

Remark 2.1. Theorem 2.1 shows that we can control the FDR with a pre-specified α under a diagonal covariance structure from the BH multiple testing procedure, which is a limitation in real applications. Therefore, Theorem 2.2 indicates that, under an arbitrary dependence structure, the BH procedure demonstrates a consistent variable selection property; however, the asymptotic variance of the non-zero β_j is greater than that reported in a previous study [23].

2.3 Theoretical properties of new selection of the smoothing parameter based on FDR

Theorem 2.2. *As $\lambda_n/\sqrt{n} \to 0, \lambda_n \to +\infty$. Assume that $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \xrightarrow{p} \boldsymbol{C}$ with a positive matrix $\boldsymbol{C}$. Then, under model* (2.1)*, we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}^t_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}) \xrightarrow{d} N(0, \sigma^2\boldsymbol{C}_{11\cdot 2}^{-1}),$$

where $\mathcal{A}$ is the index set satisfying $\mathcal{A} = \{j \mid \beta_j \neq 0, j = 1, \ldots, p\}$, $\boldsymbol{C}_{11\cdot 2} = \boldsymbol{C}_{11} - \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21}$, and

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}_{11} & \boldsymbol{C}_{12} \\ \boldsymbol{C}_{21} & \boldsymbol{C}_{22} \end{pmatrix}.$$

Proof. To prove asymptotic normality, let $\boldsymbol{\beta} + \frac{\boldsymbol{u}}{\sqrt{n}}$ substitute $\boldsymbol{\beta}$ in function $f(\boldsymbol{\beta})$, which is defined in (2.4); thus, we have

$$f_n(\boldsymbol{u}) = n\left\|\tilde{\boldsymbol{\beta}} - \left(\boldsymbol{\beta} + \frac{\boldsymbol{u}}{\sqrt{n}}\right)\right\|^2 + \lambda_n \sum_{j=1}^p w_j \left|\beta_j + \frac{u_j}{\sqrt{n}}\right|,$$

$\hat{\boldsymbol{u}}_n = \mathrm{argmin} f_n(\boldsymbol{u})$. Then, $\hat{\boldsymbol{\beta}}^t = \boldsymbol{\beta} + \frac{\hat{\boldsymbol{u}}_n}{\sqrt{n}}$ and $\hat{\boldsymbol{u}}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}^t - \boldsymbol{\beta})$. Note that $f_n(\boldsymbol{u}) - f_n(\underline{0}) = V_n(\boldsymbol{u})$, where

$$V_n(\boldsymbol{u}) = \boldsymbol{u}^{\mathrm{T}}\boldsymbol{u} - 2\sqrt{n}\tilde{\boldsymbol{\epsilon}}^{\mathrm{T}}\boldsymbol{u} + \lambda_n \sum_{j=1}^p w_j \left(\left|\beta_j + \frac{u_j}{\sqrt{n}}\right| - |\beta_j|\right).$$

If $\beta_j \neq 0$, then $\sqrt{n}(|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|) \to u_j \mathrm{sgn}(\beta_j)$ and $w_j \xrightarrow{p} \frac{C_{jj}\sigma^2}{\beta_j}$. By Slutsky's theorem, we have $\frac{\lambda_n}{\sqrt{n}} w_j \sqrt{n}(|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|) \to 0$. If $\beta_j = 0$, then $\sqrt{n}(|\beta_j + \frac{u_j}{\sqrt{n}}| - |\beta_j|) \to |u_j|$ and $\lambda_n \frac{C_{jj}\sigma^2}{\sqrt{n}|\tilde{\beta}_j|} \to \infty$ as $\sqrt{n}|\tilde{\beta}_j| = O_p(1)$. Thus, by Slutsky's theorem, $V_n(\boldsymbol{u}) \xrightarrow{d} V(\boldsymbol{u})$ for every $\boldsymbol{u}$ where

$$V(\boldsymbol{u}) = \begin{cases} \boldsymbol{u}_{\mathcal{A}}^{\mathrm{T}} \boldsymbol{u}_{\mathcal{A}} - 2\boldsymbol{u}_{\mathcal{A}}^{\mathrm{T}} \sqrt{n} \tilde{\boldsymbol{\epsilon}}_{\mathcal{A}}^{\mathrm{T}}, & \text{if} \quad u_j = 0, \quad \forall\, j \notin \mathcal{A}, \\ \infty, & \text{otherwise.} \end{cases}$$

The unique minimum of V is $(\sqrt{n}\tilde{\boldsymbol{\epsilon}}_{\mathcal{A}}, \boldsymbol{0})$, and we know that $\sqrt{n}\tilde{\boldsymbol{\epsilon}} \xrightarrow{d} N(\boldsymbol{0}, \sigma^2 \boldsymbol{C}^{-1})$ where

$$\boldsymbol{C}^{-1} = \begin{pmatrix} \boldsymbol{C}_{11}^{-1} + \boldsymbol{C}_{11}^{-1}\boldsymbol{C}_{12}\boldsymbol{C}_{22\cdot 1}^{-1}\boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1} & -\boldsymbol{C}_{11}^{-1}\boldsymbol{C}_{12}\boldsymbol{C}_{22\cdot 1}^{-1} \\ -\boldsymbol{C}_{22\cdot 1}^{-1}\boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1} & \boldsymbol{C}_{22\cdot 1}^{-1} \end{pmatrix},$$

$\boldsymbol{C}_{22\cdot 1} = \boldsymbol{C}_{22} - \boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}\boldsymbol{C}_{12}$, and $\boldsymbol{C}_{11\cdot 2}^{-1} = (\boldsymbol{C}_{11} - \boldsymbol{C}_{12}\boldsymbol{C}_{22}^{-1}\boldsymbol{C}_{21})^{-1} = \boldsymbol{C}_{11}^{-1} + \boldsymbol{C}_{11}^{-1}\boldsymbol{C}_{12}\boldsymbol{C}_{22\cdot 1}^{-1}\boldsymbol{C}_{21}\boldsymbol{C}_{11}^{-1}$. Thus, we obtain $\hat{\boldsymbol{u}}_{n\mathcal{A}} \xrightarrow{d} N(0, \sigma^2 \boldsymbol{C}_{11\cdot 2}^{-1})$ and

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{t} - \boldsymbol{\beta}_{\mathcal{A}}) \xrightarrow{d} N(0, \sigma^2 \boldsymbol{C}_{11\cdot 2}^{-1}).$$

We can also demonstrate that $\mathrm{P}(j \in \hat{\mathcal{A}}) \to 0$ $(\forall\, j \in \mathcal{A}^c)$ by the following algebraic calculations. If $j \in \hat{\mathcal{A}}$, we have $2n(\tilde{\beta}_j - \beta_j) = \lambda_n w_j$ and $\lambda_n w_j / \sqrt{n} \to \infty$ because $\sqrt{n}(\tilde{\beta}_j - \beta_j)$ converges to normal distribution. Thus, we have $\mathrm{P}(j \in \hat{\mathcal{A}}) \leqslant \mathrm{P}(2\sqrt{n}(\tilde{\beta}_j - \beta_j) = \lambda_n w_j/\sqrt{n}) = 0$ and $\mathrm{P}(\hat{\mathcal{A}} = \mathcal{A}) = 1$. □

Lemma 2.1 (See [23, $\hat{\boldsymbol{\beta}}_{\mathcal{A}}$ in Equation (4)])**.** *As $\lambda_n/\sqrt{n} \to 0, \lambda_n n^{-(\gamma-1)/2} \to +\infty$, where $w_j = 1/|\hat{\beta}_j(ols)|^{\gamma}$ with the ordinary least square estimate $\hat{\beta}_j(\mathrm{ols})$. Here, assume that $\frac{1}{n}\boldsymbol{X}'\boldsymbol{X} \xrightarrow{p} \boldsymbol{C}$ with a positive matrix $\boldsymbol{C}$. Then, we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}}) \xrightarrow{d} N(0, \sigma^2 \boldsymbol{C}_{11}^{-1}).$$

Remark 2.2. Our estimator $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{t} - \boldsymbol{\beta}_{\mathcal{A}})$ has greater asymptotic variance than the adaptive estimator $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}} - \boldsymbol{\beta}_{\mathcal{A}})$ because $\boldsymbol{C}_{11\cdot 2}^{-1} > \boldsymbol{C}_{11}^{-1}$. Furthermore, when $\boldsymbol{C}$ is diagonal, the proposed procedure is equivalent to the adaptive LASSO [23].

3 Multiple testing procedures for variable selection in generalized linear models

A generalized linear model assumes that the data of n individuals have the following density function:

$$f(y_i \,|\, \boldsymbol{x}_i, \boldsymbol{\beta}) = \exp\left(\frac{y_i \cdot \theta_i - b(\theta_i)}{\phi/\tau_i} + c(y_i, \phi)\right), \tag{3.1}$$

where $b'(\theta_i) = \mu_i = \mathrm{E}(Y_i)$ and $\eta_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} = g(\mu_i)$ for $i = 1, \ldots, n$. Let the maximum likelihood estimate of $\boldsymbol{\beta}$ be $\tilde{\boldsymbol{\beta}}(\mathrm{glm}) = (\tilde{\beta}_1(\mathrm{glm}), \ldots, \tilde{\beta}_p(\mathrm{glm}))^{\mathrm{T}}$, i.e.,

$$\tilde{\boldsymbol{\beta}}(\mathrm{glm}) = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Q}\boldsymbol{z},$$

where $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}$, $\boldsymbol{Q}$ is the diagonal matrix with entries $q_i = \tau_i/[b''(\theta_i)(d\eta_i/d\mu_i)^2]$ and $\boldsymbol{z} = (z_1, \ldots, z_n)^{\mathrm{T}}$ with $z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\frac{d\eta_i}{d\mu_i}$, $\hat{\eta}_i = \boldsymbol{x}_i^{\mathrm{T}}\tilde{\boldsymbol{\beta}}(\mathrm{glm})$, and $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$. Then, we have

$$\sqrt{n}\boldsymbol{D}^{-1/2}(\tilde{\boldsymbol{\beta}}(\mathrm{glm}) - \boldsymbol{\beta}) \xrightarrow{d} N_p(\boldsymbol{0}_p, I_p),$$

where $\boldsymbol{D} = (D_{ij})_{i,j=1}^{p} = (n^{-1}\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{X})^{-1}\phi$. Let the multiple testing problems be

$$H_{0j} : \beta_j = 0 \quad \text{vs.} \quad H_{1j} : \beta_j \neq 0, \quad j = 1, \ldots, p.$$

Then, we have

$$\frac{\sqrt{n}\tilde{\beta}_j(\mathrm{glm})}{\sqrt{D_{jj}}} \overset{H_{0j}}{\to} N(0,1).$$

Let $t_j = \sqrt{n}\tilde{\beta}_j(\mathrm{glm})/\sqrt{D_{jj}}$ from the observations $(y_j, \boldsymbol{x}_j), j = 1, \ldots, n$. Then, the p-values are defined as

$$p_j = \mathrm{P}(T \geqslant |t_j|), \quad j = 1, \ldots, p,$$

where $T \xrightarrow{d} N(0,1)$. Without loss of generality, assume that i_j satisfies $p_{i_j} = p_{(j)}$ where $(i_1, i_2, \ldots, i_p)$ is a permutation of $(1, \ldots, p)$. Then, the BH selection procedure at level α can be described as follows:

- By the BH procedure to control the FDR, find the greatest k such that $p_{(k)} \leqslant \frac{k}{p}\alpha$;
- Let estimate $\hat{\mathcal{A}}$ of $\mathcal{A}$ be $\{i_j, j = 1, \ldots, k\}$.

The Storey selection procedure at level α is described as follows:

- By the Storey procedure to control the FDR, obtain the threshold $t_\alpha(\widehat{\mathrm{FDR}}_\omega)$;
- Let estimate $\hat{\mathcal{A}}$ of $\mathcal{A}$ be $\{j : p_j \leqslant t_\alpha(\widehat{\mathrm{FDR}}_\omega)\}$.

The Z-mean selection procedure at level α is described as follows:

- By the Z-mean procedure to control the FDR, obtain the threshold $t_\alpha(\widehat{\mathrm{FDR}}_L)$;
- Let estimate $\hat{\mathcal{A}}$ of $\mathcal{A}$ be $\{i : p_i^* \leqslant t_\alpha(\widehat{\mathrm{FDR}}_L)\}$.

Note that $\tilde{\beta}_{(k)}(\mathrm{glm})$ is the estimator that corresponds to $p_{(k)}$. Then, $\frac{|\tilde{\beta}_{(k)}(\mathrm{glm})|}{\sqrt{\mathrm{var}(\tilde{\beta}_{(k)}(\mathrm{glm}))\cdot\frac{n-1}{n-p-1}}}$ can be considered as a threshold for the BH procedure. When

$$\frac{|\tilde{\beta}_j(\mathrm{glm})|}{\sqrt{\mathrm{var}(\tilde{\beta}_j(\mathrm{glm}))}} > \frac{|\tilde{\beta}_{(k)}(\mathrm{glm})|}{\sqrt{\mathrm{var}(\tilde{\beta}_{(k)}(\mathrm{glm}))}} \quad \text{or} \quad \frac{|\tilde{\beta}_j(\mathrm{glm})|^2}{\mathrm{var}(\tilde{\beta}_j(\mathrm{glm}))} > \frac{|\tilde{\beta}_{(k)}(\mathrm{glm})|^2}{\mathrm{var}(\tilde{\beta}_{(k)}(\mathrm{glm}))}, \tag{3.2}$$

then β_j is not zero significantly. The selected tuning parameter based on the FDR is given as

$$\lambda_n = \frac{2|\tilde{\beta}_{(k)}|^2}{\mathrm{var}(\tilde{\beta}_{(k)})}. \tag{3.3}$$

Theorem 3.1. *If $\boldsymbol{D}$ tends to a diagonal matrix, then the BH selection procedure at level α satisfies that the error rate used to select insignificant predictors as significant predictors at level α asymptotically approaches $(m_0/p)\alpha \leqslant \alpha$, where m_0 is the number of zero β_j.*

Theorem 3.2. *As $\lambda_n/\sqrt{n} \to 0, \lambda_n \to +\infty$, assume that $(n^{-1}\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{X})^{-1}\phi \xrightarrow{p} \boldsymbol{I}^{-1}$ with a positive definite $\boldsymbol{I}$ where $\boldsymbol{Q}$ is the diagonal matrix with entries $q_i = \tau_i/[b''(\theta_i)(d\eta_i/d\mu_i)^2]$, and $\tilde{\boldsymbol{\beta}}$ in (2.4) is replaced by $\tilde{\boldsymbol{\beta}}(\mathrm{glm})$. Then, we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}(\mathrm{glm})_{\mathcal{A}}^t - \boldsymbol{\beta}_{\mathcal{A}}) \xrightarrow{d} N(0, \boldsymbol{I}_{11\cdot 2}^{-1}),$$

where $\mathcal{A}$ is the index set satisfying $\mathcal{A} = \{j \,|\, \beta_j \neq 0, j = 1, \ldots, p\}$, $\boldsymbol{I}_{11\cdot 2}^{-1} = (\boldsymbol{I}_{11} - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\boldsymbol{I}_{21})^{-1}$, $\boldsymbol{I}^{-1} = (\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{X})^{-1}\phi = \left(\begin{smallmatrix}\boldsymbol{I}_{11} & \boldsymbol{I}_{12} \\ \boldsymbol{I}_{21} & \boldsymbol{I}_{22}\end{smallmatrix}\right)^{-1}$, $\tilde{\boldsymbol{\beta}}(\mathrm{glm}) = (\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Q}\boldsymbol{z}$ and $\boldsymbol{z}$ is the response vector with entries $z_i = \hat{\eta}_i + (y_i - \hat{\mu}_i)\frac{d\eta_i}{d\mu_i}$ with $\hat{\eta}_i = \boldsymbol{x}_i'\tilde{\boldsymbol{\beta}}(\mathrm{glm})$ and $\hat{\mu}_i = g^{-1}(\hat{\eta}_i)$.

Proof. We have

$$\boldsymbol{I}^{-1} = \begin{pmatrix} \boldsymbol{I}_{11}^{-1} + \boldsymbol{I}_{11}^{-1}\boldsymbol{I}_{12}\boldsymbol{I}_{22\cdot 1}^{-1}\boldsymbol{I}_{21}\boldsymbol{I}_{11}^{-1} & -\boldsymbol{I}_{11}^{-1}\boldsymbol{I}_{12}\boldsymbol{I}_{22\cdot 1}^{-1} \\ -\boldsymbol{I}_{22\cdot 1}^{-1}\boldsymbol{I}_{21}\boldsymbol{I}_{11}^{-1} & \boldsymbol{I}_{22\cdot 1}^{-1} \end{pmatrix},$$

where $\boldsymbol{I}_{22\cdot 1} = \boldsymbol{I}_{22} - \boldsymbol{I}_{21}\boldsymbol{I}_{11}\boldsymbol{I}_{12}$ and $\boldsymbol{I}_{11}^{-1} + \boldsymbol{I}_{11}^{-1}\boldsymbol{I}_{12}\boldsymbol{I}_{22\cdot 1}^{-1}\boldsymbol{I}_{21}\boldsymbol{I}_{11}^{-1} = \boldsymbol{I}_{11\cdot 2}^{-1}$. Using a procedure similar to that of Theorem 2.2, we obtain

$$\sqrt{n}(\hat{\boldsymbol{\beta}}(\mathrm{glm})_{\mathcal{A}}^t - \boldsymbol{\beta}_{\mathcal{A}}) \xrightarrow{d} N(0, \boldsymbol{I}_{11\cdot 2}^{-1}).$$

This completes the proof. □

Table 1 Empirical FDR and powers of BH, Storey, Z-mean, and knockoff SDP for the test $H_{0j} : \beta_j = 0,\ j = 1, \ldots, p$ when m_1 is 50 or 100

Method	(m_1, α)	ρ	eFDR	Power	Method	(m_1, α)	eFDR	Power
BH	$(50, 0.05)$	0	0.047	0.315	BH	$(50, 0.1)$	0.092	0.441
		0.3	0.037	0.196			0.086	0.301
		0.5	0.037	0.070			0.090	0.129
		0.7	0.041	0.014			0.073	0.025
	$(100, 0.05)$	0	0.046	0.421		$(100, 0.1)$	0.085	0.545
		0.3	0.040	0.285			0.086	0.402
		0.5	0.038	0.117			0.094	0.212
		0.7	0.030	0.017			0.087	0.039
Storey	$(50, 0.05)$	0	0.047	0.316	Storey	$(50, 0.1)$	0.096	0.447
		0.3	0.036	0.195			0.091	0.313
		0.5	0.033	0.069			0.087	0.135
		0.7	0.029	0.009			0.074	0.027
	$(100, 0.05)$	0	0.049	0.435		$(100, 0.1)$	0.093	0.562
		0.3	0.044	0.296			0.094	0.419
		0.5	0.039	0.120			0.100	0.221
		0.7	0.021	0.015			0.087	0.042
Z-mean	$(50, 0.05)$	0	0.068	0.814	Z-mean	$(50, 0.1)$	0.118	0.871
		0.3	0.105	0.721			0.171	0.796
		0.5	0.155	0.472			0.321	0.588
		0.7	0.292	0.164			0.327	0.190
	$(100, 0.05)$	0	0.057	0.865		$(100, 0.1)$	0.103	0.916
		0.3	0.082	0.787			0.145	0.842
		0.5	0.110	0.560			0.164	0.656
		0.7	0.195	0.182			0.248	0.227
SDP	$(50, 0.05)$	0	0.045	0.432	SDP	$(50, 0.1)$	0.089	0.597
		0.3	0.036	0.323			0.083	0.369
		0.5	0.025	0.150			0.054	0.261
		0.7	0.008	0.132			0.032	0.166
	$(100, 0.05)$	0	0.037	0.436		$(100, 0.1)$	0.068	0.584
		0.3	0.023	0.301			0.057	0.445
		0.5	0.008	0.172			0.047	0.269
		0.7	0.001	0.086			0.012	0.175

Lemma 3.1 (See [23, $\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\text{glm})$ in Equation (11)])**.** *Here,*

$$\lambda_n/\sqrt{n} \to 0, \quad \lambda_n n^{-(\gamma-1)/2} \to +\infty,$$

where $w_j = 1/|\hat{\beta}_j(\text{MLE})|^{\gamma}$ *with the maximum likelihood estimate* $\hat{\beta}_j(\text{MLE})$*. Thus, under some mild regularity conditions, we have*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\text{glm}) - \boldsymbol{\beta}_{\mathcal{A}}) \xrightarrow{d} N(0, \boldsymbol{I}_{11}^{-1}).$$

This lemma is the same as [23, Theorem 4]; thus, the mild regularity conditions and the details of the proof can be found in the literature [23, Theorem 4].

Remark 3.1. If the link function g is canonical, the Fisher information matrix $\boldsymbol{I}$ is the same as that in the literature [23]. Thus, our estimator $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\text{glm})^t - \boldsymbol{\beta}_{\mathcal{A}})$ has greater asymptotic variance than the adaptive estimator $\sqrt{n}(\hat{\boldsymbol{\beta}}_{\mathcal{A}}(\text{glm}) - \boldsymbol{\beta}_{\mathcal{A}})$ because

$$(\boldsymbol{I}_{11} - \boldsymbol{I}_{12}\boldsymbol{I}_{22}^{-1}\boldsymbol{I}_{21})^{-1} > \boldsymbol{I}_{11}^{-1}.$$

4 Simulation studies

Here, we discuss simulation studies for variable selection using some FDR control methods. First, we consider the following linear regression model:

$$Y = \beta_1 X_1 + \cdots + \beta_{m_1} X_{m_1} + \beta_{m_1+1} X_{m_1+1} + \cdots + \beta_p X_p + \epsilon,$$

where

$$(X_1, \ldots, X_p) \sim N(\mathbf{0}, \mathbf{\Sigma}_\rho), \quad \epsilon \sim N(0,1) \quad \text{and} \quad \mathbf{\Sigma}_\rho = (\rho^{|i-j|})_{i,j=1}^p.$$

For simplicity, we take the following values:

$$\beta_1 = \cdots = \beta_{m_1} = \beta$$

and $\beta_{m_1+1} = \cdots = \beta_p = 0$. The samples are

$$\{(y_i, x_{i1}, \ldots, x_{ip}), i = 1, \ldots, n\},$$

Table 2 Empirical FDR and powers of BH, Storey, Z-mean, and knockoff SDP for the test $H_{0j} : \beta_j = 0$, $j = 1, \ldots, p$ when m_1 is 300 or 500

Method	(m_1, α)	ρ	eFDR	Power	Method	(m_1, α)	eFDR	Power
BH	$(300, 0.05)$	0	0.034	0.604	BH	$(300, 0.1)$	0.068	0.712
		0.3	0.035	0.477			0.069	0.611
		0.5	0.038	0.253			0.071	0.384
		0.7	0.034	0.043			0.063	0.107
	$(500, 0.05)$	0	0.025	0.685		$(500, 0.1)$	0.050	0.793
		0.3	0.025	0.569			0.050	0.699
		0.5	0.024	0.343			0.050	0.497
		0.7	0.025	0.076			0.051	0.179
Storey	$(300, 0.05)$	0	0.047	0.657	Storey	$(300, 0.1)$	0.094	0.772
		0.3	0.046	0.533			0.090	0.666
		0.5	0.045	0.294			0.087	0.437
		0.7	0.039	0.054			0.077	0.128
	$(500, 0.05)$	0	0.045	0.779		$(500, 0.1)$	0.088	0.869
		0.3	0.042	0.669			0.085	0.794
		0.5	0.037	0.434			0.077	0.598
		0.7	0.031	0.108			0.067	0.233
Z-mean	$(300, 0.05)$	0	0.053	0.940	Z-mean	$(300, 0.1)$	0.097	0.965
		0.3	0.064	0.886			0.104	0.926
		0.5	0.067	0.703			0.111	0.791
		0.7	0.089	0.287			0.126	0.400
	$(500, 0.05)$	0	0.047	0.968		$(500, 0.1)$	0.088	0.984
		0.3	0.050	0.928			0.089	0.960
		0.5	0.051	0.788			0.084	0.893
		0.7	0.057	0.394			0.083	0.508
SDP	$(300, 0.05)$	0	0.024	0.308	SDP	$(300, 0.1)$	0.063	0.572
		0.3	0.009	0.241			0.045	0.394
		0.5	0.006	0.114			0.016	0.311
		0.7	4.5E–4	0.061			0.008	0.126
	$(500, 0.05)$	0	0.019	0.265		$(500, 0.1)$	0.056	0.583
		0.3	0.009	0.195			0.035	0.398
		0.5	0.008	0.089			0.014	0.216
		0.7	3.2E–4	0.029			0.006	0.061

where n is the sample size. The parameter configuration is as follows: $\beta = 3.5$, $n = 3{,}000$, $p = 1{,}000$, m_1 equals 50, 100, 300, 500, respectively, and ρ equals 0, 0.3, 0.5, 0.7, respectively. Then, the set of indices of significant covariates for the response is $\mathcal{A} = \{1, 2, \ldots, m_1\}$. Thus, the multiple testing problem can be constructed as follows:

$$H_{0j} : \beta_j = 0, \quad j = 1, \ldots, p.$$

Tables 1 and 2 show the simulation results for the power and empirical FDR (eFDR), where the power is the proportion of false hypotheses correctly rejected among the total number of false hypotheses. Four methods are compared: structured semidefinite program (SDP), BH, Storey, and Z-mean, where SDP is the knockoff method proposed by Barber and Candes [1] with SDP construction, and BH, Storey, and Z-mean are defined as above. Here, the parameters ω and k in the Storey and Z-mean methods are set to 0.1 and 3, respectively. The number of simulations is 200, and the nominal level is taken as $\alpha = 5\%$ and $\alpha = 10\%$.

From Tables 1 and 2, we can observe the following:

• The Z-mean method has greater powers than the SDP, BH, and Storey methods;

• When the number of nonzero coefficients m_1 is large, the Z-mean method behaves very well;

• The Z-mean method behaves better in the case of small $|\rho|$ than that of large $|\rho|$, i.e., when the correlation between covariates is small, the Z-mean method controls the FDR and demonstrates good power performance;

• All methods control the FDR well under weak dependence of covariates. When the correlation $|\rho|$ becomes large, the SDP, BH, and Storey methods do control the FDR; however, they are very conservative;

• In the sparse case, the SDP method demonstrates greater powers than the BH and Storey methods. In contrast, the BH and Storey methods show greater powers than the SDP method.

5 Practical applications

As a practical application, we apply five methods: SDP, equi-correlated knockoffs (EQU), BH, Storey, and Z-mean to select motifs, where EQU is the knockoff method [1] with equivalent construction. The motif data are taken from http://stat.math.ethz.ch/Research-Reports/Other-Manuscripts/buhlmann/motif-spellman.RData. The data consist of the expression ratios (treated vs. control) of $n = 4{,}443$ genes and the motif-matching scores of $p = 2{,}155$ candidate motifs from Saccharomyces cerevisiae. The motif-matching scores are organized into a matrix of dimension $n \times p$, with each element x_{gm} in the matrix being the matching score of motif m to the promoter of gene g. The $\log_2$(expression ratio) of different genes are put into a vector $\boldsymbol{Y}$, where Y_g is the $\log_2$(expression ratio) of gene g.

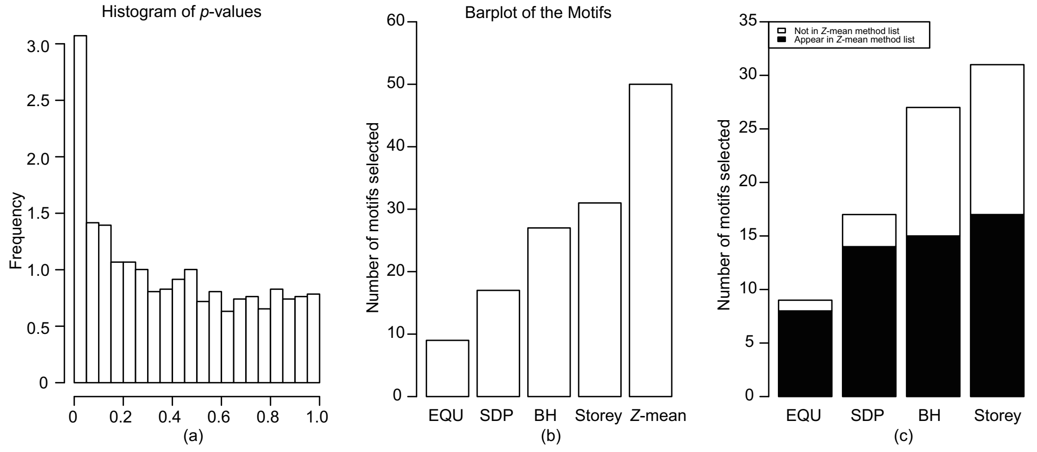

Figure 1 (a) Histogram of p-values of 918 motifs; (b) number of the selected motifs by SDP, EQU, BH, Storey, and Z-mean; (c) overlap of motifs between Z-mean and the other four methods

Conlon et al. [6] first used linear regression modeling for each candidate motif by regressing the gene expression ratios against their promoter region's matching scores for each candidate motif. Differing from [6], we consider all promoter region matching scores as covariates in the full linear regression model.

Before selecting these motifs, we first subtract some motifs that are highly correlated with each other but poorly correlated with the response expression $\boldsymbol{Y}$. Here, we eliminate variables whose correlation is greater than 0.7, leaving 918 motifs. The SDP, EQU, BH, Storey, and Z-mean methods are applied to the motif data with a test size of $\alpha = 5\%$. The results are given in Figure 1. As can be seen, the frequency of p-values close to zero appears very large, which indicates that a non-sparse model may be considered. It can be shown that the BH, Storey, and Z-mean methods select more motifs than the knockoff methods. For example, the Z-mean method selects 50 motifs, and the SDP and EQU methods select 17 and nine motifs, respectively. The BH and Storey methods select 31 and 27 motifs, respectively. EQU's motifs are included in SDP's motifs, and the SDP and Z-mean methods have the same 16 motifs. More than one-half of the motifs selected by the BH and Storey methods are included in the motifs selected by the Z-mean method.

6 Conclusion

This paper has established a connection between multiple testing procedures and the adaptive LASSO, and has proposed BH, Storey, and Z-mean methods based on multiple testing via controlled FDR for variable selection. Simulation results demonstrate that all of these methods can control the FDR, and the Z-mean method shows greater powers than the BH, Storey, and SDP methods under weak dependence. However, the drawback of the Z-mean method is that empirical FDR results are greater than the nominal level α compared to the other three methods when the covariates are dependent. In addition, the SDP results are much more conservative when the proportion of non-null effect β_j is large. Moreover, we require that dimension p is less than the sample size n, and p is assumed to be fixed. Therefore, one focus of future work will be to extend our idea to the case of $p > n$. It is known that controlling the FDR under dependence of the test statistic is challenging; thus, we would also like to control the FDR under an arbitrary dependence structure from a multiple testing procedure.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 11671268, 11522105, and 11690012). The authors thank the reviewers for their constructive comments, which helped us improve this manuscript substantially.

References

1 Barber R F, Candes E. Controlling the false discovery rate via knockoffs. Ann Statist, 2015, 43: 2055–2085

2 Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc Ser B Stat Methodol, 1995, 57: 289–300

3 Buhlmann P, van de Geer S. Statistics for High-Dimensional Data: Methods, Theory and Applications. New York: Springer, 2011

4 Bunea F, Wegkamp M H, Auguste A. Consistent variable selection in high dimensional regression via multiple testing. J Statist Plann Inference, 2006, 136: 4349–4364

5 Candes E J, Tao T. The Dantzig selector: Statistical estimation when p is much larger than n. Ann Statist, 2007, 35: 2313–2351

6 Conlon E M, Liu X S, Lieb J D, et al. Integrating regulatory motif discovery and genome-wide expression analysis. Proc Natl Acad Sci USA, 2003, 100: 3339–3344

7 Efron B. Correlation and large-scale simultaneous sigfinicance testing. J Amer Statist Assoc, 2007, 102: 93–103

8 Efron B, Hastie T, Johnstone I, et al. Least angle regression. Ann Statist, 2004, 32: 407–489

9 Fan J Q, Han X, Gu W J. Estimating false discovery proportion under arbitrary covariance testing. J Amer Statist Assoc, 2012, 107: 1019–1035

10 Fan J Q, Li R Z. Variable selection via noncave penalized likelihood and its oracle properties. J Amer Statist Assoc, 2001, 96: 1348–1360

11 Ferreira J, Zwinderman A. On the Benjamini-Hochberg method. Ann Statist, 2006, 34: 1827–1849

12 Furmańczyk K. On some stepdown procedures with application to consistent variable selection in linear regression. Statistics, 2015, 49: 614–628

13 Meinshausen N, Meier L, Bühlmann P. P-values for high-dimensional regression. J Amer Statist Assoc, 2009, 104: 1671–1681

14 Storey J D. A direct approach to false discovery rates. J R Stat Soc Ser B Stat Methodol, 2002, 64: 479–498

15 Storey J D, Taylor J E, Siegmund D. Strong control, conservative point estimation and simultaneous conservative consistency of false discovery rates: A unified approach. J R Stat Soc Ser B Stat Methodol, 2004, 66: 187–205

16 Tibshirani R. Regression shrinkage and selection via the LASSO. J R Stat Soc Ser B Stat Methodol, 1996, 58: 267–288

17 Tibshirani R, Hoefling H, Tibshirani R. Nearly isotonic regression. Technometrics, 2011, 53: 54–61

18 Wasserman L, Roeder K. High dimensional variable selection. Ann Statist, 2009, 37: 2178–2201

19 Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol, 2006, 68: 49–67

20 Yuan M, Lin Y. Model selection and estimation in the Gaussian graphical model. Biometrika, 2007, 94: 19–35

21 Zhang C M. Assessing mean and median filters in multiple testing for large-scale imaging data. TEST, 2014, 23: 51–71

22 Zhang C M, Fan J Q, Yu T. Multiple testing via FDRL for large-scale imaging data. Ann Statist, 2011, 39: 613–642

23 Zou H. The adaptive LASSO and its oracle properties. J Amer Statist Assoc, 2006, 476: 1418–1429

24 Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol, 2005, 67: 301–320