Semi-Penalized Inference with Direct FDR Control

Jian Huang
University of Iowa

April 4, 2016

The problem

Consider the linear regression model

$$\boldsymbol{y} = \sum_{j=1}^{p} \boldsymbol{x}_j \beta_j + \boldsymbol{\varepsilon}, \tag{1}$$

where $\boldsymbol{y} \in \mathbb{R}^n$, $\boldsymbol{x}_j \in \mathbb{R}^n$, $\boldsymbol{\varepsilon} \in \mathbb{R}^n$, and β_j is the jth regression coefficient, Here $p \gg n$. Let $S = \{j : |\beta_j| > 0, 1 \le j \le p\}$ be the support of $\boldsymbol{\beta}$. Suppose the model is sparse in the sense that $|S| \ll n$.

Our goals are to

- Estimate S and provide an error assessment in terms of false discovery rate (FDR).
- Construct confidence intervals for the selected coefficients.
- We propose a method, Semi-Penalized Inference with Direct fdR control (SPIDR).

Background: penalized methods

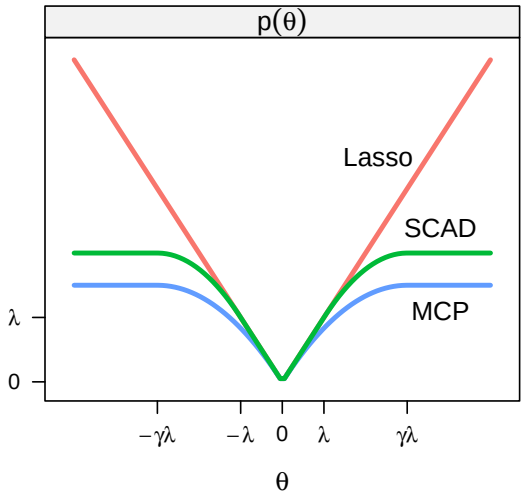
There has been much work on penalized methods for variable selection and estimation in high-dimensional settings, e.g.,

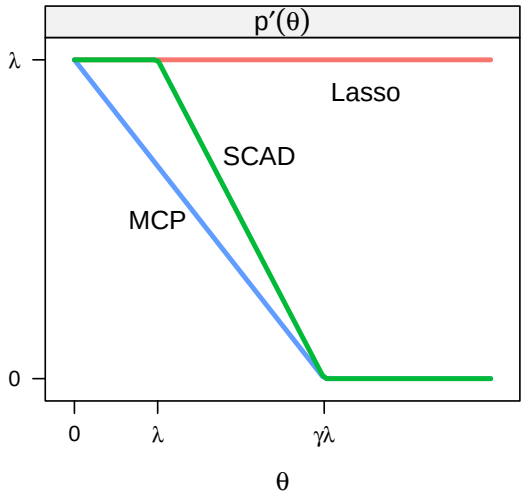
- ℓ_1-penalized method or LASSO: Tibshirani (1996), Chen, Saunders and Donoho (1998).
- Bridge: Frank and Friedman (1993)
- Smoothly clipped absolute deviation (SCAD) penalty: Fan (1997); Fan and Li (2001)
- Minimax concave penalty (MCP): Zhang (2010)
- Others .....

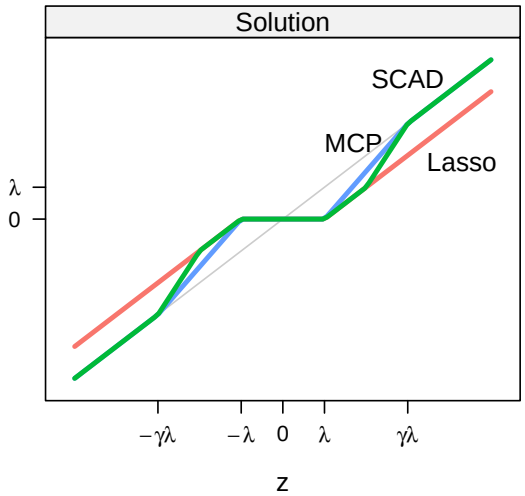
These methods are based on a (fully) penalized criterion

$$L(\boldsymbol{\beta}; \lambda) = \frac{1}{2n} \|\boldsymbol{y} - X\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} \rho(\beta_j; \lambda).$$

LASSO, SCAD and MCP







Penalized selection

The (fully) penalized criterion

$$L(\boldsymbol{b};\lambda) = \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{b}\|^2 + \sum_{j=1}^{p} \rho(b_j;\lambda).$$

Let $\widehat{\boldsymbol{b}}(\lambda) = \arg\min_{\boldsymbol{b}} L(\boldsymbol{b};\lambda)$. Usually, a $\lambda = \hat{\lambda}$ is chosen based a data-driven procedure such as cross validation. Then $\widehat{\boldsymbol{b}}(\hat{\lambda})$ is the penalized estimator of $\boldsymbol{\beta}$.

The set

$$\hat{S}^* = \{j : |\widehat{b}_j(\hat{\lambda})| > 0, 1 \le j \le p\}$$

is taken as an estimator of S.

A different approach is to select a $\hat{\lambda}$ such that the FDR of $\hat{S}^*$ is controlled at a given level.

LASSO, SCAD and MCP

- LASSO (Tibshirani, 1996; Chen, Donoho and Saunders, 1998):

$$\rho(t;\lambda) = \lambda|t|.$$

- SCAD (Fan and Li 2001):

$$\rho(t;\lambda,\gamma) = \begin{cases} \lambda|t|, & |t| \le \lambda \\ \frac{\gamma\lambda|t| - 0.5(t^2+\lambda^2)}{\gamma-1}, & \lambda < |t| \le \gamma\lambda \\ \frac{\lambda^2(\gamma^2-1)}{2(\gamma-1)}, & |t| > \gamma\lambda, \end{cases} \quad , \gamma > 2.$$

- MCP (Zhang 2010):

$$\rho(t;\lambda,\gamma) = \begin{cases} \lambda|t| - \frac{|t|^2}{2\gamma}, & |t| \le \lambda\gamma \\ \lambda^2\gamma/2, & |t| > \lambda\gamma \end{cases} \quad , \gamma > 1.$$

Semi-penalized estimation

- Let $\boldsymbol{\beta}_{-j} = (\beta_k, k \neq j, 1 \leq k \leq p)'$ and $X_{-j} = (\boldsymbol{x}_k, k \neq j, 1 \leq k \leq p)$.
- Consider the semi-penalized criterions

$$L_j(\boldsymbol{\beta}; \lambda) = \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{x}_j \beta_j - X_{-j} \boldsymbol{\beta}_{-j}\|^2 + \sum_{k \neq j} \rho(\beta_k; \lambda), 1 \leq j \leq p, \quad (2)$$

  where ρ is a penalty function with a tuning parameter $\lambda \geq 0$.
- For a fixed λ, let

$$\hat{\boldsymbol{\beta}}_{(j)}(\lambda) = (\widehat{\beta}_j(\lambda), \hat{\boldsymbol{\beta}}_{-j}(\lambda)) = \underset{\beta_j, \boldsymbol{\beta}_{-j}}{\arg\min}\, L_j(\boldsymbol{\beta}; \lambda), 1 \leq j \leq p. \quad (3)$$

- Let $\hat{\boldsymbol{\beta}}(\lambda) = (\widehat{\beta}_1(\lambda), \ldots, \widehat{\beta}_p(\lambda))'$.

Semi-penalized solution paths

For a fixed λ, let $\hat{\boldsymbol{\beta}}_{(j)}(\lambda) = (\widehat{\beta}_j(\lambda), \hat{\boldsymbol{\beta}}_{-j}(\lambda))$ be the value that minimizes the jth penalized criterion in (2), that is,

$$\hat{\boldsymbol{\beta}}_{(j)}(\lambda) = (\widehat{\beta}_j(\lambda), \hat{\boldsymbol{\beta}}_{-j}(\lambda)) = \underset{\beta_j, \boldsymbol{\beta}_{-j}}{\arg\min}\, L_j(\boldsymbol{\beta}; \lambda), 1 \le j \le p. \tag{4}$$

Let $Q_j = I - \boldsymbol{x}_j(\boldsymbol{x}_j'\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j'$. It can be easily verified that

$$\hat{\boldsymbol{\beta}}_{-j}(\lambda) = \underset{\boldsymbol{\beta}_{-j}}{\arg\min} \frac{1}{2n}\|Q_j(\boldsymbol{y} - X_{-j}\boldsymbol{\beta}_{-j}\|^2 + \sum_{k \ne j} \rho(\beta_k; \lambda), \tag{5}$$

and

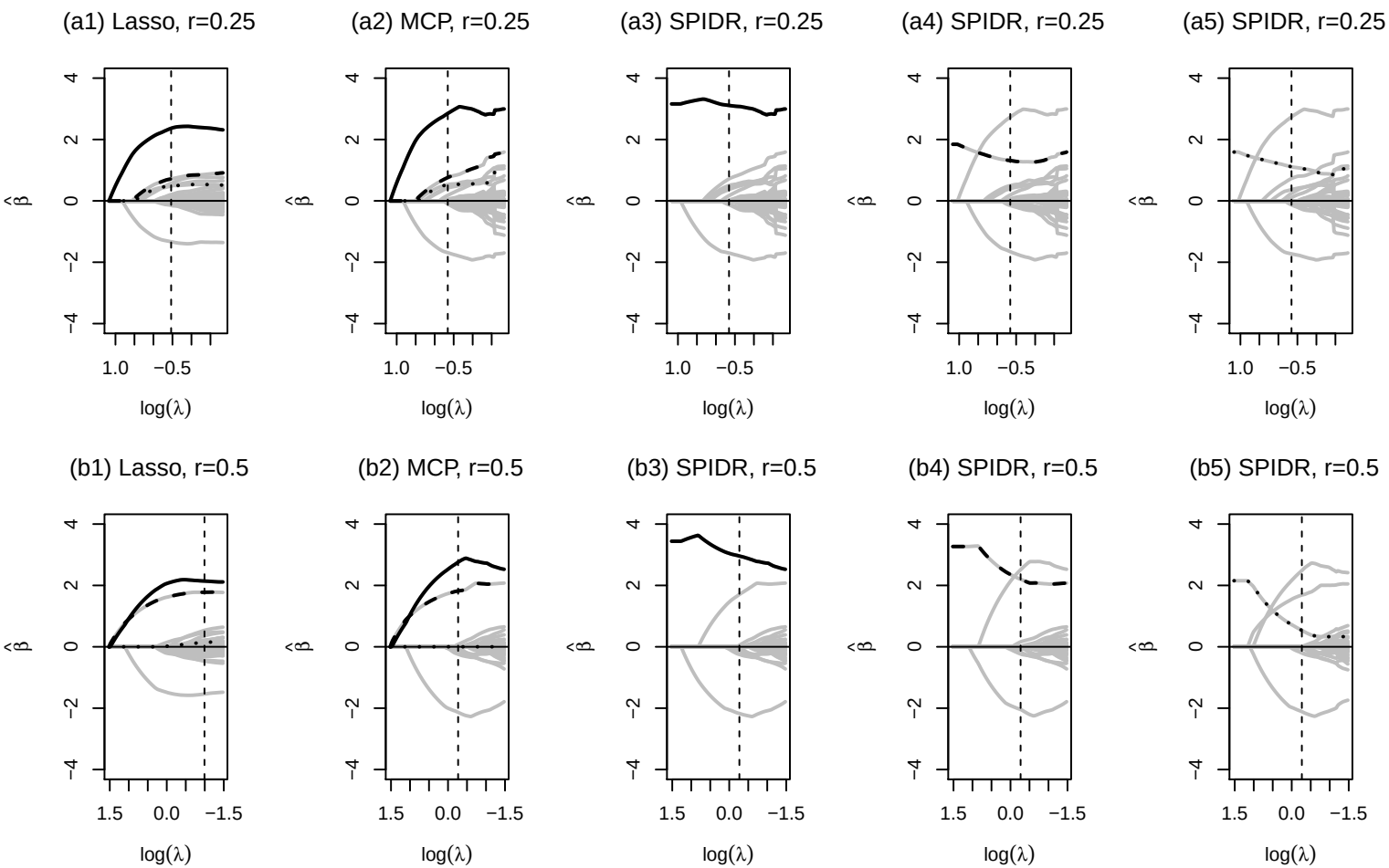
$$\widehat{\beta}_j(\lambda) = \underset{\beta_j}{\arg\min}\, \|\boldsymbol{y} - X_{-j}\hat{\boldsymbol{\beta}}_{-j} - \boldsymbol{x}_j\beta_j\|^2 = (\boldsymbol{x}_j'\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j'(\boldsymbol{y} - X_{-j}\hat{\boldsymbol{\beta}}_{-j}(\lambda)). \tag{6}$$

Semi-penalized solution paths: Example

- Consider the linear model (1) with $(\beta_1, \ldots, \beta_6) = (3, 2, 1, -0.5, -1.0, -1.5)$, $\beta_j = 0, 7 \le j \le p$ and $\varepsilon_i \sim N(0, 2.5^2)$.
- Set $n = 100, p = 1000$.
- Let $\{z_{ij}, 1 \le i \le n, 1 \le j \le p\}$ and $\{u_{ij} : 1 \le i \le n, j = 1, 2\}$ be independently generated random numbers from $N(0, 1)$. The predictors are

$$
\begin{aligned}
x_{ij} &= z_{ij} + a u_{i1}, j = 1, \ldots, 4, \\
x_{ij} &= z_{ij} + a u_{i2}, j = 5, \ldots, 8, \\
x_{ij} &= z_{ij} + u_{i1}, j = 9, \ldots, 17, \\
x_{ij} &= z_{ij} + u_{i2}, j = 18, \ldots, 26, \\
x_{ij} &= z_{ij}, j = 27, \ldots, p.
\end{aligned}
$$

- Consider two values of a, $a = \sqrt{1/3}$ and $a = 1$. The strength of the correlation between the predictors are determined by a. The maximum correlation is $r = a^2/(1 + a^2)$. So for $a = \sqrt{1/3}$, $r = 0.25$ and for $a = 1, r = 0.5$.


(a1) Lasso, r=0.25
(a2) MCP, r=0.25
(a3) SPIDR, r=0.25
(a4) SPIDR, r=0.25
(a5) SPIDR, r=0.25
(b1) Lasso, r=0.5
(b2) MCP, r=0.5
(b3) SPIDR, r=0.5
(b4) SPIDR, r=0.5
(b5) SPIDR, r=0.5
β̂
log(λ)
1.0
−0.5
1.5
0.0
−1.5
4
2
0
−2
−4

Semi-penalized selection with direct FDR control

For $A \subset \{1, \ldots, p\}$, let $P_A = X_A(X_A'X_A)^- X_A'$ and $Q_{\hat{S}_j} = I - P_{\hat{S}_j}$. Let $\Sigma_{\hat{S}_j} = X_{\hat{S}_j}' X_{\hat{S}_j}/n$. Let $\widehat{\beta}_j = \widehat{\beta}_j(\lambda)$. It can be shown that

$$\widehat{\beta}_j = (\boldsymbol{x}_j' Q_{\hat{S}_j} \boldsymbol{x}_j)^{-1} \boldsymbol{x}_j' [Q_{\hat{S}_j} \boldsymbol{y} - X_{\hat{S}_j} \Sigma_{\hat{S}_j}^{-1} \dot{\rho}(\hat{\boldsymbol{\beta}}_{\hat{S}_j}; \lambda)], \tag{7}$$

where $\dot{\rho}(\hat{\boldsymbol{\beta}}_{\hat{S}_j}; \lambda) \equiv (\dot{\rho}(\widehat{\beta}_j; \lambda) : j \in \hat{S}_j)'$. By the oracle property of $\hat{\boldsymbol{\beta}}_{(j)}$ under suitable conditions,

$$\widehat{\beta}_j \approx \beta_j + (\boldsymbol{x}_j' Q_{S_j} \boldsymbol{x}_j)^{-1} \boldsymbol{x}_j' Q_{S_j} \boldsymbol{\varepsilon}.$$

It follows that $\widehat{\beta}_j$ is consistent and asymptotically normal. Its variance can be consistently estimated by

$$\widehat{\sigma}_j^2 = \widehat{\sigma}^2 (\boldsymbol{x}_j' Q_{\hat{S}_j} \boldsymbol{x}_j)^{-1}, \tag{8}$$

where $\widehat{\sigma}^2$ is a consistent estimator of σ^2. The covariance between $\widehat{\beta}_j$ and $\widehat{\beta}_k$ can be consistently estimated by

$$\widehat{\mathrm{Cov}}(\widehat{\beta}_j, \widehat{\beta}_k) = \widehat{\sigma}^2 \frac{\boldsymbol{x}_j' Q_{\hat{S}_j} Q_{\hat{S}_k} \boldsymbol{x}_k}{(\boldsymbol{x}_j' Q_{\hat{S}_j} \boldsymbol{x}_j)(\boldsymbol{x}_k' Q_{\hat{S}_k} \boldsymbol{x}_k)}. \tag{9}$$

Therefore, $\widehat{\beta} = (\widehat{\beta}_1, \ldots, \widehat{\beta}_p)'$ has an asymptotic multivariate normal distribution with mean $(\beta_1, \ldots, \beta_p)'$ and covariance matrix specified by (8) and (9).

Semi-penalized selection with direct FDR control

- Consider the z-statistics

$$z_j = \widehat{\beta}_j / \widehat{\sigma}_j, 1 \leq j \leq p.$$

- We can think of variable selection as testing p hypotheses $H_{0j} : \beta_j = 0, 1 \leq j \leq p$. For a given $t > 0$, we reject H_{0j} if $|z_j| > t$, or equivalently, we select the jth variable if $|z_j| > t$.
- The problem of variable selection becomes that of determining a threshold value according to a proper control of error.

Semi-penalized selection with direct FDR control

- Let $R(t) = \sum_{j=1}^{p} 1\{|z_j| > t\}$ be the number of variables with $|z_j| > t$,
- Let $V(t) = \sum_{j=1}^{p} 1\{|z_j| > t, \beta_j = 0\}$ be the number of falsely selected variables.
- The false discovery proportion for a given t is

$$\mathrm{Fdp}(t) = \begin{cases} \frac{V(t)}{R(t)} & \text{if } R(t) > 0, \\ 0 & \text{if } R(t) = 0. \end{cases} \tag{10}$$

- The FDR is defined to be $Q(t) = \mathrm{E}(\mathrm{Fdp}(t))$ (Benjamini and Hochberg (1995)).

Semi-penalized selection with direct FDR control

- A simple estimator of the FDR is

$$\hat{Q}_0(t) = \frac{\mathrm{E}V(t)}{R(t)}.$$

  For independent test statistics, $\hat{Q}_0$ is a good estimator of Q.

- However, for correlated statistics, Efron (2007) demonstrated that $\hat{Q}_0$ can give grossly misleading estimate of FDR and proposed an improved estimator. For two-sided tests, this estimator is

$$\hat{Q}(t) = \hat{Q}_0(t)\left[1 + 2A\frac{t\phi(t)}{\sqrt{2}\Phi(-t)}\right], \tag{11}$$

  where ϕ is the probability density function of $N(0,1)$. Here A is a dispersion variable accounting for the correlation of the statistics $\hat{z}_j$. Methods for estimating A are given in Efron (2007).

Semi-penalized selection with direct FDR control

- For $0 < q < 1$, let $\hat{t}_q$ be the value satisfying

$$\hat{Q}(\hat{t}_q) = q.$$

- The set of the selected variables is

$$\hat{S}_q = \{j : |z_j| \geq \hat{t}_q\}. \tag{12}$$

- By construction, the FDR of $\hat{S}_q$ is approximately controlled at the level q.

Direct FDR control in the context of multiple comparisons was proposed by Storey (2002).

Confidence intervals for selected coefficients

- The selection rule (12) directly leads to confidence intervals for the coefficients of the selected variables.
- The $1-q$ level FDR-adjusted confidence intervals for the selected coefficients are

$$\widehat{\beta}_j \pm \hat{t}_q \widehat{\sigma}_j, j \in \hat{S}. \tag{13}$$

- The interpretation is that the expected proportion of these intervals that do not cover their respective parameters is q (Benjamini and Yekutieli (2005)).

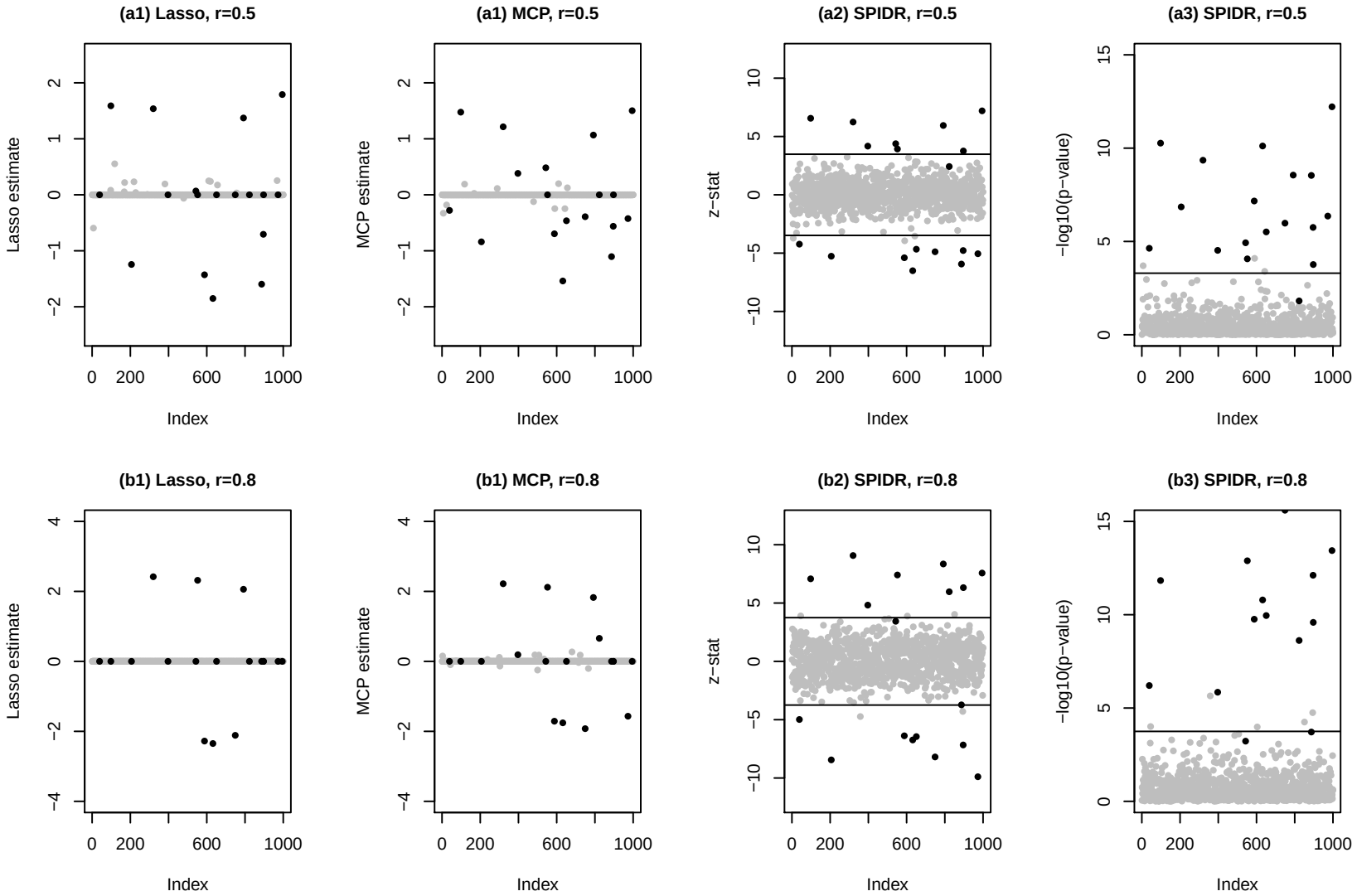


Figure : Selection results with $q = 0.15$ from the models in Examples 1 and 2.

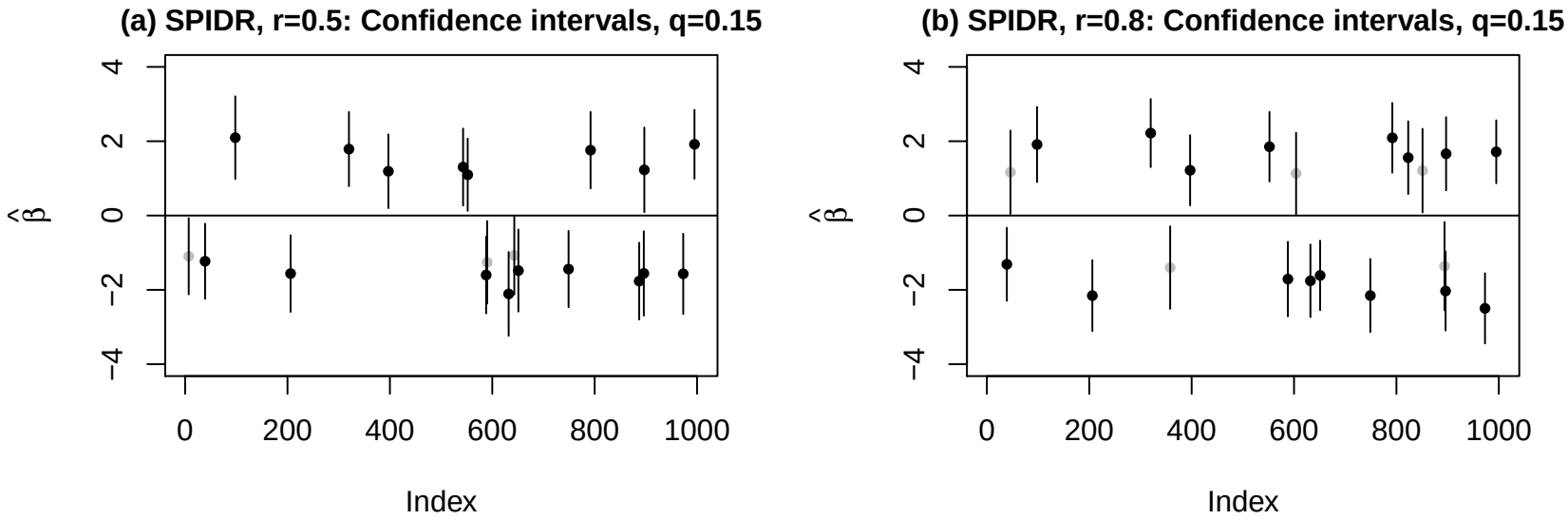


Figure : The confidence intervals of the selected coefficients with the FDR level $q = 0.15$. The gray dots indicate false

Simulation studies: Implementation

- Computation: Use *ncvreg* (Breheny and Huang 2009).
- Penalty parameter selection: 5-fold cross validation based on the fully penalized criterion. Then the selected λ is used in the semi-penalized criterions.

Variance Estimation

1. Let $\widehat{\boldsymbol{b}}(\hat{\lambda})$ be the MCP estimator with $\hat{\lambda}$ determined based on 5-fold cross validation.
2. Let $\hat{S}^*$ be the set of the predictors with nonzero coefficients in $\widehat{\boldsymbol{b}}$. Randomly partition the dataset into two subsets D_1 and D_2 with equal sample sizes $n_1 = n_2 = n/2$.
3. Use the first part to fit a model with variables in $\hat{S}^*$ and calculate the least squares estimate

$$\widehat{\boldsymbol{b}}^{(1)} = \arg\min_{\boldsymbol{b}} \sum_{i \in D_1} (y_i - \sum_{j \in \hat{S}^*} x_{ij} b_j)^2.$$

A consistent estimator of σ^2 is

$$\widehat{\sigma}^2 = \frac{1}{n_2 + |\hat{S}|} \sum_{i \in D_2} (y_i - \sum_{j \in \hat{S}^*} x_{ij} \widehat{b}_j^{(1)})^2.$$

Variance estimation is an important problem in high-dimensional regression. We refer to Fan, Guo and Hao (2012) and Sun and Zhang (2012) for the discussions on this problem and other approaches.

Example 1. We consider model (1) with $p = 1000$.

- The errors are i.i.d. $N(0, \sigma^2)$ with $\sigma = 3$.
- The first $q = 18$ coefficients are nonzero with values

$$(\beta_1, \ldots, \beta_{18}) = (1, 1, 1, .8, .8, .8, .6, .6, .6, -.6, -.6, -.6, -.8, -.8, -.8, -1, -1, -1).$$

  The sample size $n = q^2/2 = 162$. The remaining coefficients are zero.

- The predictors are generated as follows. Let $\{z_{ij}, 1 \le i \le n, 1 \le j \le p\}$ and $\{u_{ij} : 1 \le i \le n, 1 \le j \le 2\}$ be independently generated random numbers from $N(0, 1)$. Let $A_1 = \{1, \ldots, 9\}$ and $A_2 = \{10, \ldots, 18\}$ be the sets of predictors with nonzero coefficients. Let A_3, A_4 and A_5 be different sets of 50 indices randomly chosen from $\{19, \ldots, p\}$. Set

$$x_{ij} = z_{ij} + a_1 u_{i1}, j \in A_1, \; x_{ij} = z_{ij} + a_1 u_{i2}, j \in A_2,$$
$$x_{ij} = z_{ij} + a_2 u_{i1}, j \in A_3, \; x_{ij} = z_{ij} + a_2 u_{i2}, j \in A_4,$$
$$x_{ij} = z_{ij} + a_3 u_{i1} - a_3 u_{i2}, j \in A_5, x_{ij} = z_{ij}, j \notin \cup_{k=1}^5 A_k,$$

  where $a_1 = 1$, $a_2 = 0.5$ and $a_3 = 0.1$.

- The correlation of the predictors in A_1 is $r_{11} = a_1^2/(1 + a_1^2) = 0.5$ and the correlation between the predictors in A_1 and A_3 is $r_{13} = a_1 a_2/(\sqrt{1 + a_1^2}\sqrt{1 + a_2^2}) = 0.32$.

Example 2.

- The generating model is the same as that in Example 1, except $a_1 = 2$.
- Now there is stronger correlation among the predictors. For example, the correlation between the predictors in A_1 is $r_{11} = 0.8$ and the correlation between the predictors in A_1 and A_3 is $r_{13} = 0.40$.

Example 3.

- The generating model is the same as that in Example 1, except now the predictors are generated from a multivariate normal distribution $N(0, \Sigma)$, where the (j, k)th element of the covariance matrix Σ is $\sigma_{jk} = 0.5^{|j-k|}$, $1 \leq j, k \leq p$.

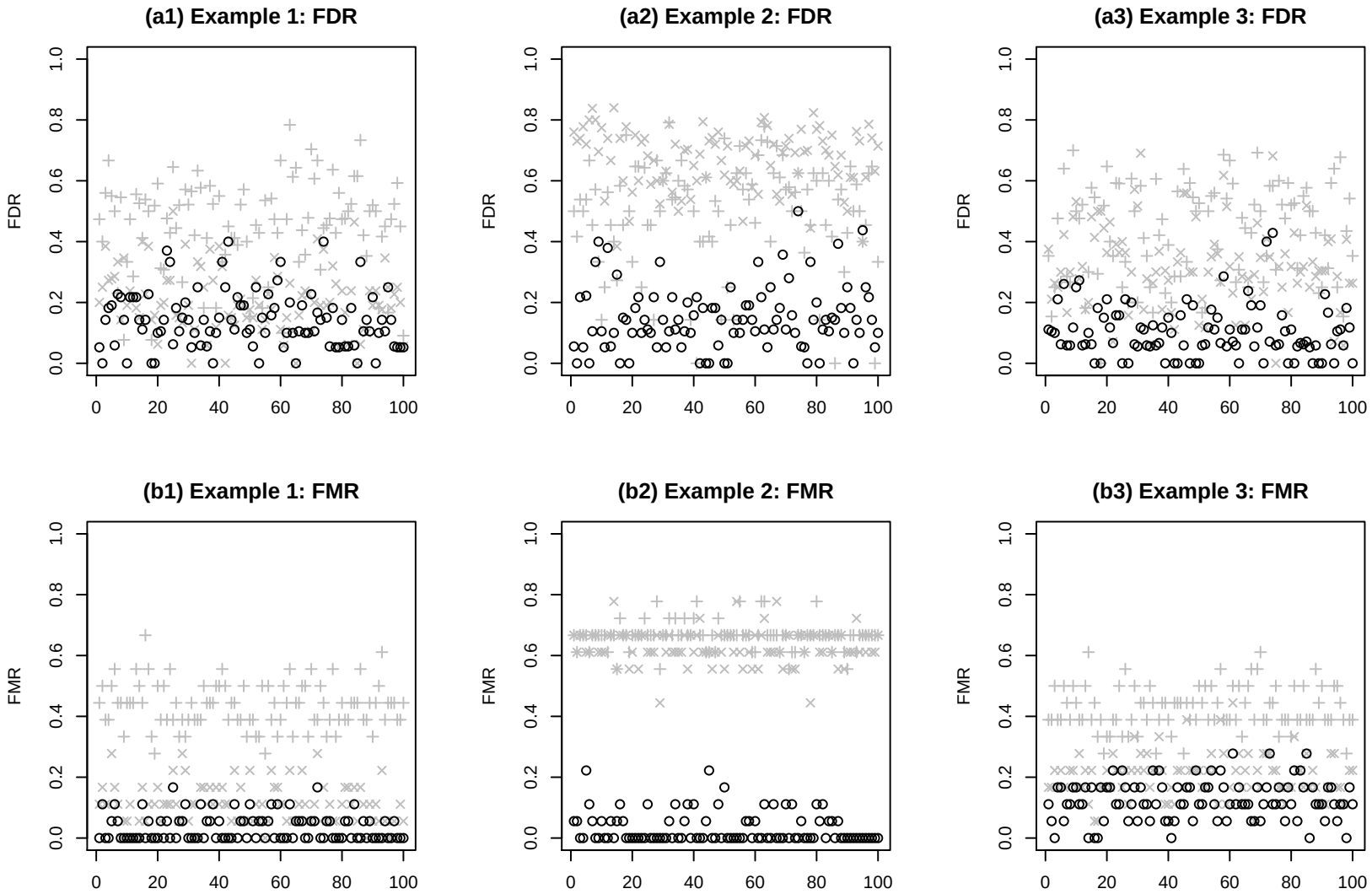


Figure : The results for Lasso, MCP and SPIDR are represented by the plus "+", cross "x" and open circle "○" signs, respectively.

Table : Simulation study. NVS, number of variables selected; FDR, false discovery rate; FMR, false miss rate, averaged over 100 replications with standard deviations in parentheses, for Examples 1 to 3.

Method	NVS	FDR	FMR
		Example 1	
SPIDR	20.52 (2.78)	0.14 (0.09)	0.03 (0.04)
MCP	20.66 (2.83)	0.21 (0.10)	0.11 (0.07)
Lasso	19.97 (5.92)	0.45 (0.15)	0.43 (0.07)
		Example 2	
SPIDR	20.90 (3.54)	0.15 (0.11)	0.03 (0.05)
MCP	21.53 (5.76)	0.67 (0.10)	0.63 (0.06)
Lasso	13.22 (4.08)	0.50 (0.17)	0.66 (0.04)
		Example 3	
SPIDR	17.75 (2.44)	0.10 (0.08)	0.12 (0.06)
MCP	22.05 (5.42)	0.32 (0.14)	0.20 (0.07)
Lasso	19.63 (6.19)	0.43 (0.16)	0.42 (0.07)

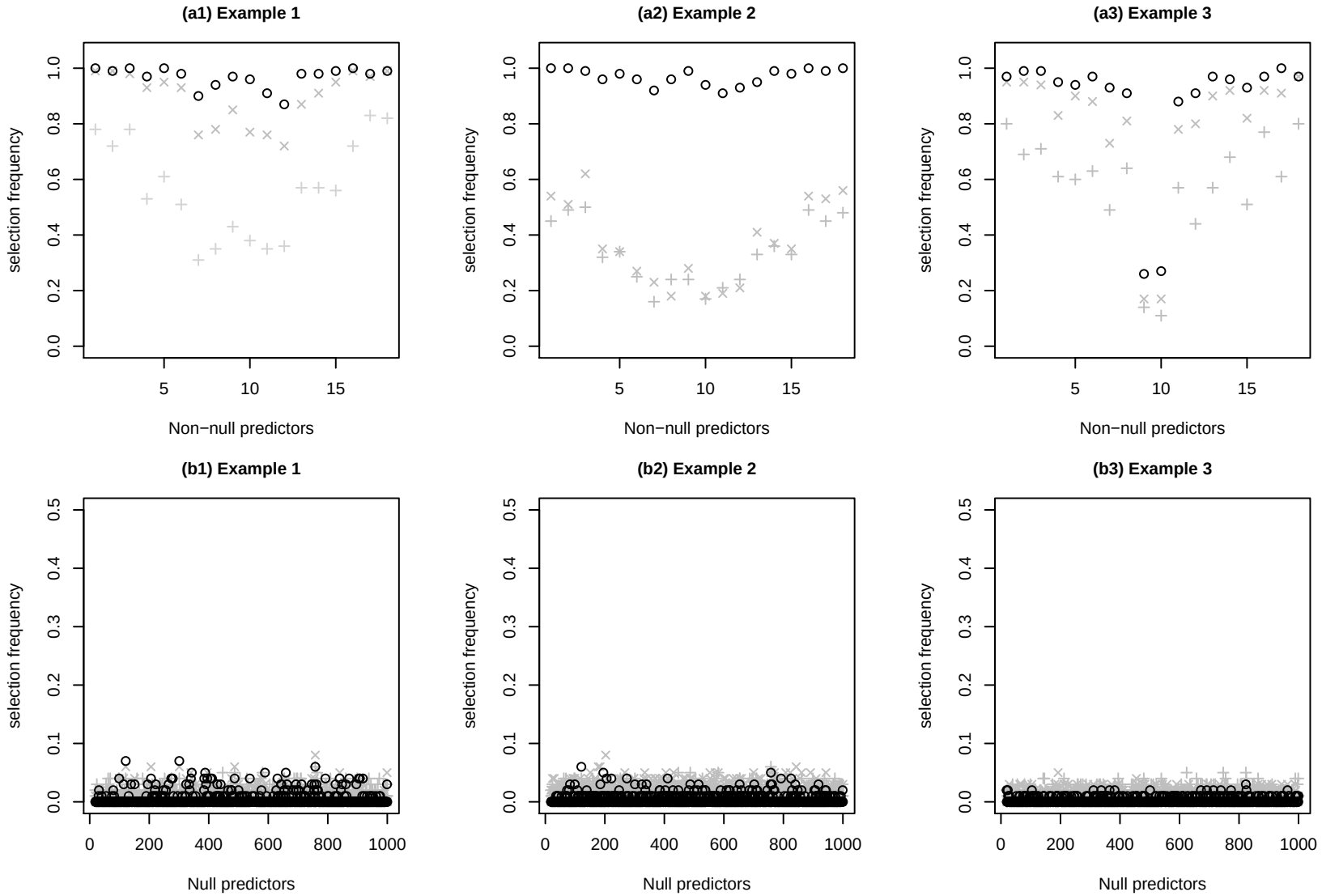


Figure : Percentages of variables being selected based on 100 replications for Examples 1-3. The results for Lasso, MCP and SPIDR are represented by the plus "+", cross "x" and open circle "○" signs, respectively.

Theoretical properties: *ideal* estimator

- Let $S_j = \{k : \beta_k \neq 0, k \neq j, 1 \leq k \leq p\}$ and let S_j^c be the complement of S_j in $\{1, \ldots, p\}$. We define the *ideal* estimator by

$$(\widetilde{\beta}_j, \widetilde{\boldsymbol{\beta}}_{-j}) = \underset{\beta_j, \boldsymbol{\beta}_{-j}}{\arg\min}\{\|\boldsymbol{y} - \boldsymbol{x}_j\beta_j - X_{-j}\boldsymbol{\beta}_{-j}\|^2 : \boldsymbol{\beta}_{S_j^c} = 0\}, 1 \leq j \leq p. \quad (14)$$

  In particular, $\widetilde{\beta}_j$ is an ideal estimator of β_j.

- An expression of $\widetilde{\beta}_j$ parallel to (7) is

$$\widetilde{\beta}_j = \beta_j + (\boldsymbol{x}_j' Q_{S_j} \boldsymbol{x}_j)^{-1} \boldsymbol{x}_j' Q_{S_j} \boldsymbol{\varepsilon}, 1 \leq j \leq p.$$

- $(\widetilde{\beta}_1, \ldots, \widetilde{\beta}_p)$ has a multivariate normal distribution with mean $\boldsymbol{\beta}$ and

$$\text{Var}(\widetilde{\beta}_j) = \sigma^2 (\boldsymbol{x}_j' Q_{S_j} \boldsymbol{x}_j)^{-1} \text{ and } \text{Cov}(\widetilde{\beta}_j, \widetilde{\beta}_k) = \sigma^2 \frac{\boldsymbol{x}_j' Q_{S_j} Q_{S_k} \boldsymbol{x}_k}{(\boldsymbol{x}_j' Q_{S_j} \boldsymbol{x}_j)(\boldsymbol{x}_k' Q_{S_k} \boldsymbol{x}_k)}.$$

Theoretical properties: convex case ($p < n$)

- Let $c_{\min} = \min\{c_j : 1 \le j \le p\}$, where c_j is the smallest eigenvalue of $X'_{-j}Q_jX_{-j}/n$. Let $w^o = \max\{w^o_{jk} : k \in S_j, 1 \le j \le p\}$, where $(w^o_{jk}, k \in S_j)$ are the diagonal elements of $(X'_{S_j}Q_jX_{S_j}/n)^{-1}$.
- Denote the smallest nonzero coefficient by $\beta_* = \min\{|\beta^o_j| : \beta^o_j \neq 0, 1 \le j \le p\}$.
- Denote the cardinality of S by $|S|$.

Theorem 1. Suppose that $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed as $N(0, \sigma^2)$. Also, suppose that (a) $\gamma > 1/c_{\min}$; (b) for a small $\epsilon > 0$, $\beta_* > \gamma\lambda + \sigma\sqrt{(2/n)w^o \log(p|S|/\epsilon)}$; and (c) $\lambda \ge \sigma\sqrt{4\log p}\max_{j\le p}\|\boldsymbol{x}_j\|/n$. Then,

$$\mathrm{P}\{\cup_{j=1}^p(\hat{S}_j \neq S_j)\} \le 3\epsilon \text{ and } \mathrm{P}\{\cup_{j=1}^p(\widehat{\beta}_j(\lambda) \neq \widetilde{\beta}_j)\} \le 3\epsilon.$$

Theoretical properties: nonconvex case ($p \gg n$)

We require the sparse Riesz condition (SRC, Zhang and Huang (2008)) on the the matrices $Q_j X$.

Sparse Riesz condition: there exist constants $0 < c_* \le c^* < \infty$ and integer $d^* \ge |S|(K_* + 1)$ with $K_* = c^*/c_* - 1/2$ such that

$$0 < c_* \le \|Q_j X_{A_j} \boldsymbol{u}\|^2/n \le c^* < \infty, \|\boldsymbol{u}\|_2 = 1, \tag{15}$$

for every $A_j \subset \{1, \ldots, p\} \setminus \{j\}$ with $|A_j \cup S_j| \le d^*$, for all $1 \le j \le p$.

Theorem 2. Suppose that $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed as $N(0, \sigma^2)$. Also, suppose that

1. the SRC (15) holds with $\gamma \ge c_*^{-1}\sqrt{4 + c_*/c^*}$;
2. for a small $\epsilon > 0$, $\beta_* \ge 2\gamma\sqrt{c^*}\lambda + \sigma\sqrt{(2/n)w^o \log(p|S|/\epsilon)}$;
3. $\lambda \ge \sigma\sqrt{(4\log(p/\epsilon)}\max_{j \le p}\|\boldsymbol{x}_j\|/n$.

Then

$$\mathrm{P}\{\cup_{j=1}^p(\hat{S}_j(\hat{\lambda}) \ne S_j)\} \le 3\epsilon, \quad \text{and} \quad \mathrm{P}\{\cup_{j=1}^p(\widehat{\beta}_j(\hat{\lambda}) \ne \widetilde{\beta}_j)\} \le 3\epsilon.$$

Therefore, $\mathrm{P}\{\cup_{j=1}^p(\hat{S}_j(\hat{\lambda}) \ne S_j)\} \to 0$ and $\mathrm{P}\{\cup_{j=1}^p(\widehat{\beta}_j(\hat{\lambda}) \ne \widetilde{\beta}_j)\} \to 0$ as $\epsilon \to 0$.

Breast Cancer Data

- We use the breast cancer gene expression data from The Cancer Genome Atlas (2012) project to illustrate the proposed method. In this dataset, tumour samples were assayed on several platforms.
- The expression measurements of 17814 genes, including BRCA1, from 536 patients are available at `http://cancergenome.nih.gov/`.
- BRCA1 is the first gene identified that increases the risk of early onset breast cancer. Because BRCA1 is likely to interact with many other genes, including tumor suppressors and regulators of the cell division cycle, it is of interest to find genes with expression levels related to that of BRCA1. These genes may be functionally related to BRCA1 and are useful candidates for further studies.

Breast Cancer Data

We only include genes with sufficient expression levels and variations across the subjects in the analysis. So we first do an initial screen according to the following requirements:

- (a) the coefficient of variation is greater than 1;
- (b) the standard deviation is greater than 0.6;
- (c) the marginal correlation coefficient with BRCA1 is greater than 0.1.

A total of 1685 genes passed these screening steps. These are the genes included in the model.

Breast cancer data: Lasso and MCP solution paths

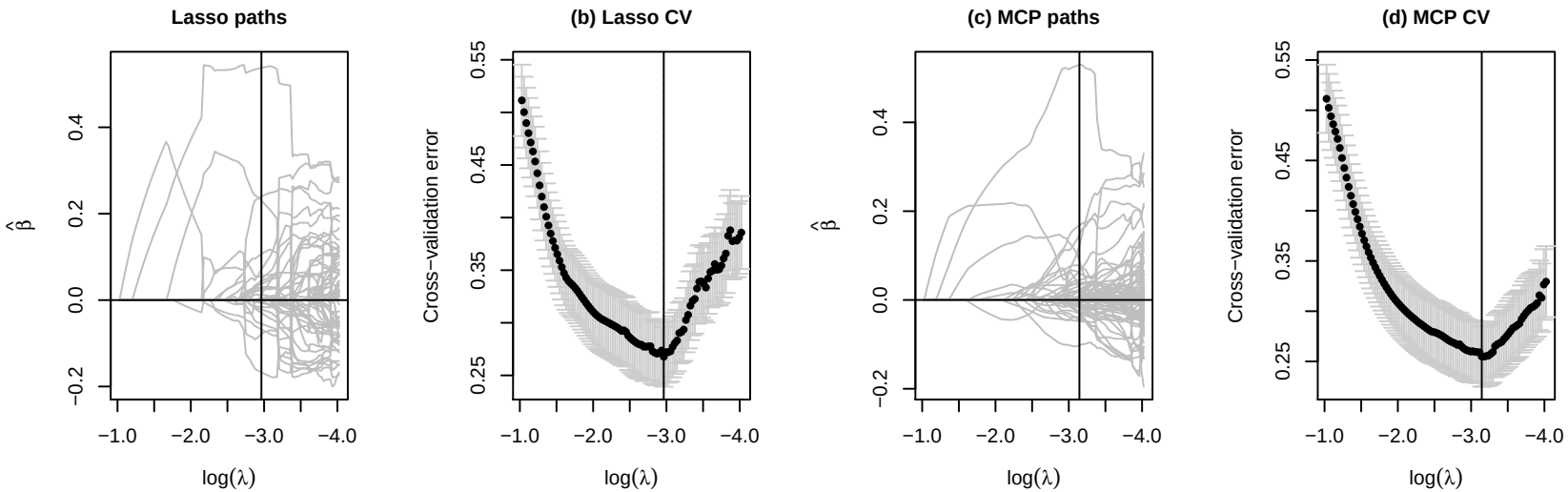


Figure : Breast cancer data. (a) Lasso solution paths; (b) Lasso cross validation results; (c) MCP solution paths; (d) MCP cross validation results.

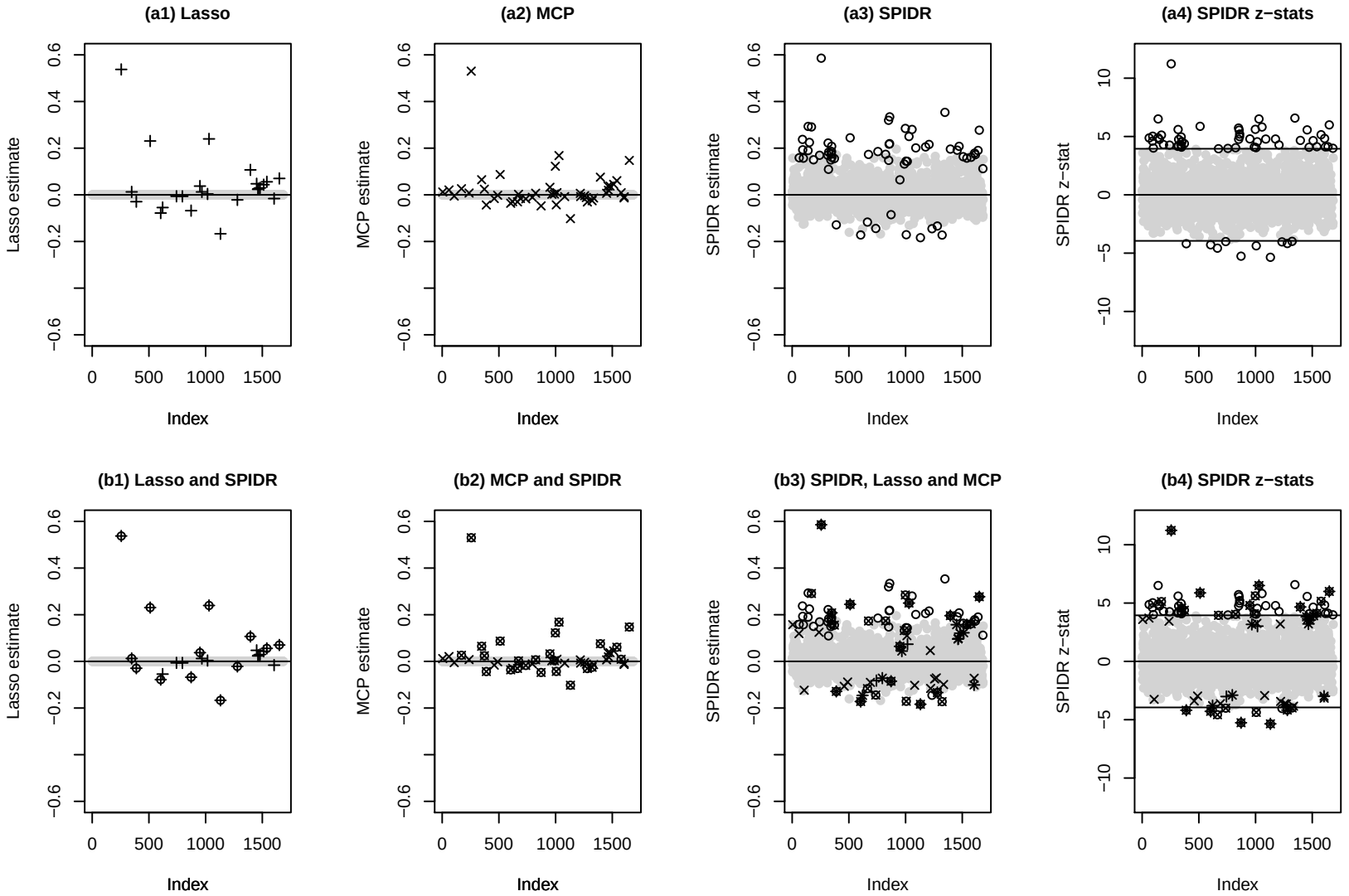


Figure : Breast cancer data. Lasso, MCP and SPIDR are represented by plus "+", cross "x", and circle "○", respectively.

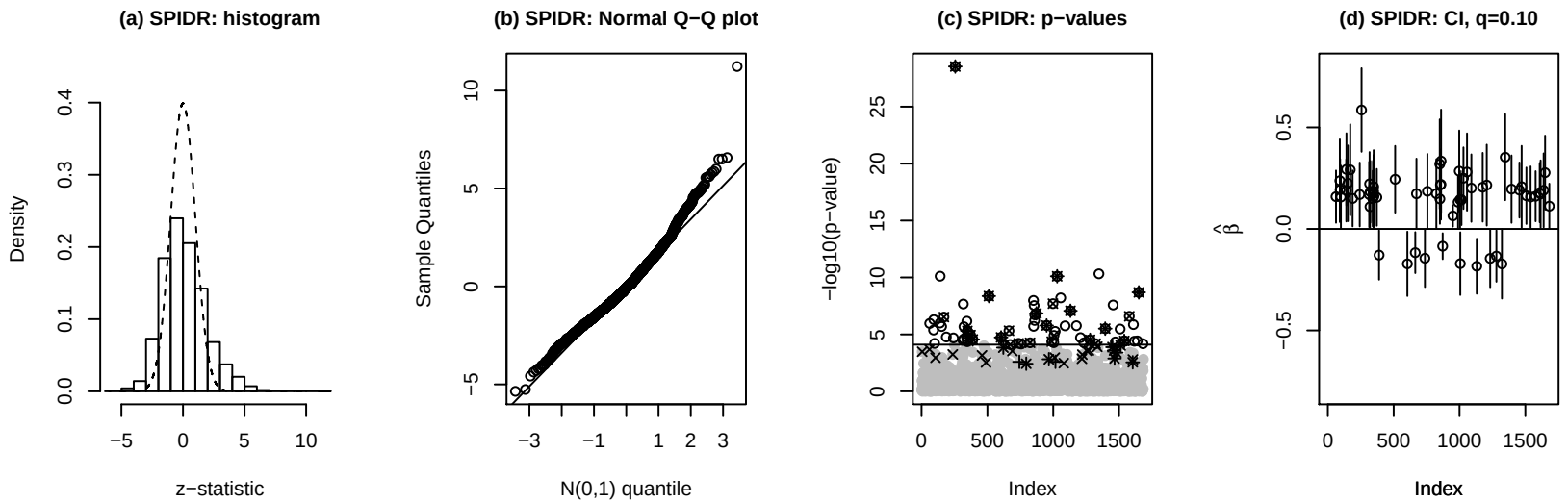


Figure : Breast cancer data. (a) Histogram of SPIDR z-statistics, the dashed curve represents the density function of $N(0,1)$; (b) Normal Q-Q plot for the SPIDR estimates; (c) SPIDR p-values; (d) SPIDR confidence intervals for the selected coefficients.

Some findings

There are genes not selected by the Lasso or MCP but selected by SPIDR, for example,

- An interesting one is gene UHRF1, which plays a major role in the G1/S transition and functions in the p53-dependent DNA damage checkpoint. Multiple transcript variants encoding different isoforms have been found for this gene (www.ncbi.nlm.nih.gov). UHRF1 is a putative oncogenic factor over-expressed in several cancers, including the bladder and lung cancers. It has been reported that UHRF1 is responsible for the repression of BRCA1 gene in sporadic breast cancer through DNA methylation (Alhosin et al. (2011)).
- Another interesting finding based on SPIDR is a gene called SRPK1. This gene is upregulated in breast cancer and its expression level is proportional to the tumor grade. Targeted SRPK1 treatment appears to be a promising way to enhance the effectiveness of chemotherapeutics drugs (Hayes et al. (2006, 2007)).
- Other interesting findings include several genes (CDC6, CDC20, CDC25C and CDCA2) that play key roles in the regulation of cell division and interact with several proteins at multiple points in the cell cycle (www.ncbi.nlm.nih.gov).

Concluding Remarks

- SPIDR is built on two important developments in high-dimensional statistics, penalized estimation and direct FDR control. It makes the connection between these two ideas and combines them in the context of variable selection.
- The idea of SPIDR can be applied to other models or more general loss functions.
- The estimation of FDR with correlated statistics is a challenging problem. We used the method of Efron (2007), which is easy to implement and computationally efficient. Other methods can be used in estimating the FDR in the presence of correlation, for example, the method of Fan et al. (2013). It would also be particularly interesting to develop methods tailored to the covariance structure given in (8) and (9).
- In the implementation, we used the R package *ncvreg* to compute the SPIDR solutions. It is useful to develop efficient algorithms specifically designed for SPIDR.
- Finally, in applications we recommend applying SPIDR in combination with penalized selection, as illustrated in the breast cancer data example.

Homework problem.

Consider the linear regression model with two predictors:

$$\boldsymbol{y} = \boldsymbol{x}_1\beta_1 + \boldsymbol{x}_2\beta_2 + \boldsymbol{\varepsilon},$$

where $\boldsymbol{y} \in \mathbb{R}^n, \boldsymbol{x}_1 \in \mathbb{R}^n, \boldsymbol{x}_2 \in \mathbb{R}^n$ and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$. The least squares estimator of (β_1, β_2) is

$$(\widehat{\beta}_1^{\mathsf{LS}}, \widehat{\beta}_2^{\mathsf{LS}}) = \arg\min_{\beta_1,\beta_2} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{x}_1\beta_1 - \boldsymbol{x}_2\beta_2\|^2.$$

The fully penalized estimator is

$$(\widetilde{\beta}_1(\lambda), \widetilde{\beta}_2(\lambda)) = \arg\min_{\beta_1,\beta_2} \left\{ \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{x}_1\beta_1 - \boldsymbol{x}_2\beta_2\|^2 + \sum_{j=1}^{2} \rho_{\mathsf{MCP}}(\beta_j; \lambda, \gamma) \right\}.$$

(1) The SPIDR solution $\widehat{\beta}_1(\lambda)$ of β_1 is obtained from

$$(\widehat{\beta}_1(\lambda), \widehat{\beta}_2(\lambda)) = \arg\min_{\beta_1,\beta_2} \left\{ \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{x}_1\beta_1 - \boldsymbol{x}_2\beta_2\|^2 + \rho_{\mathsf{MCP}}(\beta_2; \lambda, \gamma) \right\}.$$

Let $a = \boldsymbol{x}_2' Q_{\boldsymbol{x}_1} \boldsymbol{x}_2/n$. Show that

$$\widehat{\beta}_2(\lambda) = \begin{cases} \mathsf{sgn}(\widehat{\beta}_2^{LS}) \frac{(|\widehat{\beta}_2^{\mathsf{LS}}| - (\lambda/a))_+}{1 - \frac{1}{a\gamma}}, & \text{if } |\widehat{\beta}_2^{\mathsf{LS}}| \le \gamma\lambda, \\ \widehat{\beta}_2^{\mathsf{LS}}, \text{ if } |\widehat{\beta}_2^{\mathsf{LS}}| > \gamma\lambda. \end{cases}$$

Then

$$\widehat{\beta}_1(\lambda) = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'(\boldsymbol{y} - \boldsymbol{x}_2\widehat{\beta}_2(\lambda)).$$

(2) Write an `R` program to compute the solution paths of $\widehat{\beta}_1(\lambda)$ and $\widetilde{\beta}_1(\lambda)$.

(3) Design a simulation study to compare the solution paths of $\widehat{\beta}_1(\lambda)$ and $\widetilde{\beta}_1(\lambda)$.