Biometrics 000, 000–000 DOI: 000
000 0000

# Semi-Penalized Inference with False Discovery Rate Control in High-dimensional Linear Regression

**Jian Huang[1,5,*], Jin Liu[2], Shuangge Ma[3], Cun-Hui Zhang[4] and Yong Zhou[5]**

[1]Department of Statistics and Actuarial Science, University of Iowa, Iowa City, Iowa, U.S.A.

[2]Center of Quantitative Medicine, Duke-NUS Medical School,Singapore

[3]Department of Biostatistics, Yale University, New Haven, Connecticut, U.S.A.

[4]Department of Statistics and Biostatistics, Rutgers University, Piscataway, New Jersey, U.S.A.

[5]School of Statistics and Management, Shanghai University of Finance and Economics, Shanghai, China

*email: jian-huang@uiowa.edu

Summary: We propose a semi-penalized inference approach with direct false discovery rate control for variable selection and confidence interval construction in high-dimensional linear regression. With this approach, we first calculate semi-penalized estimators of the regression coefficients, which are shown to be asymptotically normal under a sparsity condition and other appropriate conditions. We then carry out selection by controlling the false discovery rate based on the distributions of these estimators. The approach provides an explicit assessment of the selection error and naturally leads to confidence intervals for the selected coefficients with a proper confidence statement. We conduct simulation studies to evaluate its finite sample performance and illustrate its application on a breast cancer gene expression data set. Our simulation studies and data example demonstrate that SPIDR is a useful method for high-dimensional statistical inference in practice.

Key words: Concave penalty; false discovery rate; selection error; sparsity; statistical inference; variable selection.

## 1. Introduction

Consider the linear regression model

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

where $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is a vector of response variables, $\boldsymbol{x}_j = (x_{1j}, \ldots, x_{nj})'$ is the $j$th vector of predictors, $\beta_j$ is the $j$th regression coefficient and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)'$ is a vector of error terms. Here $p$ is the number of predictors and $n$ is the sample size. Let $S = \{j : |\beta_j| > 0, 1 \leqslant j \leqslant p\}$ be the support of $\boldsymbol{\beta}$. We are interested in the high-dimensional case where $p \gg n$ and the model is sparse in the sense that the cardinality of $S$ is small relative to $n$. We propose a approach that formulates variable selection as a statistical inference problem based on semi-penalized inference with direct false discovery rate control. For brevity, we call the proposed method SPIDR.

There is now a substantial body of literature on penalized methods for variable selection. Several important penalty functions have been introduced. Examples include the least absolute shrinkage and selection operator (Lasso) or the $\ell_1$ penalty (Tibshirani, 1996), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li, 2000), and the minimum concave penalty (MCP) (Zhang, 2010). A common feature of these penalties is that they are capable of producing exact zero solutions, which automatically leads to variable selection. But they do not provide an error assessment of the selection results.

A different front in the area of high-dimensional statistics concerns the problem of large scale hypothesis testing. Since Benjamini and Hochberg (1995) introduced false discovery rate (FDR) for error assessment in multiple comparisons, it has become a widely used error measure in scientific investigations involving a large number of hypotheses, such as genomic studies with data from array-based technology (Storey and Tibshirani, 2003). Meinshausen and Bühlmann (2010) introduced stability selection that uses resampling to evaluate the probability of each variable being selected. It provided an upper bound for the expected number of falsely selected variables under an exchangeability condition on the predictors. Meinshausen, Meier and Bühlmann (2009) used sample splitting to obtain the $p$-values for the regression coefficients. However, these works are based on fully penalized estimators, but did not consider estimators with approximately normal distributions or directly calculate an error assessment of the selection results.

In this paper, we first propose a semi-penalized approach for estimating the regression coefficients. With this approach, we estimate the coefficients one at a time, and when we are estimating a given coefficient, we do not put penalty on it, but only penalize the other coefficients. The penalty is used to deal with the high dimensionality of the model, but not for variable selection. We formulate the problem of variable selection in the framework of large scale hypothesis testing based on semi-penalized estimators. This enables us to utilize the methods for multiple comparisons to assess the selection error. There are two aspects of SPIDR that are different from the existing penalized methods. First, SPIDR uses semi-penalized approach for estimating the regression coefficients. Second, the selection is done by controlling FDR calculated based on the distributions of the semi-penalized estimators. To study the theoretical properties of the SPIDR estimator, we introduce the concept of an ideal estimator, which is the estimator of a given coefficient (whether it is nonzero or zero) under the assumption that the support of the other coefficients is known. We use it as the gold standard in our theoretical analysis and show that the SPIDR estimator is ideal with high probability under a sparsity and other appropriate conditions. This implies that the SPIDR estimator is asymptotically normal. We also consider the factors that affect SPIDR selection, including the signal strength and the residual pairwise correlations among predictors.

Below, we first describe the SPIDR estimator. We then use a threshold rule for variable selection based on the SPIDR $z$-statistics and apply the approach for direct FDR control (Storey, 2002) to determine the selection rule. The details are given in Section 2, where we also point out that SPIDR naturally leads to confidence intervals for the selected coefficients with a proper confidence statement. In Section 3 we show that the SPIDR estimator equals an ideal estimator with high probability. We also discuss the grouping effect of SPIDR and consider factors contributing to it. In Section 4 we conduct simulation studies to evaluate the finite sample performance of SPIDR and demonstrate its application on a breast cancer gene expression data set. Section 5 includes some concluding remarks. Proofs of the theoretical results are given in the Appendix.

## 2. Method

### 2.1 *Semi-penalized estimation*

Let $\boldsymbol{\beta}_{-j} = (\beta_k, k \neq j, 1 \leqslant k \leqslant p)'$ be the vector of regression coefficients excluding $\beta_j$, and let $X_{-j} = (\boldsymbol{x}_k, k \neq j, 1 \leqslant k \leqslant p)$ be the design matrix excluding $\boldsymbol{x}_j$. Consider the semi-penalized criteria

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

where $\rho$ is a penalty function with a tuning parameter $\lambda \geqslant 0$. We focus on one coefficient, $\beta_j$, at a time and does not penalized it. The penalty is used for the purpose of regularizing the high-dimensional $\boldsymbol{\beta}_{-j}$. We use the MCP (Zhang, 2010),

$$\rho(t; \lambda) = \lambda \int_0^{|t|} \left(1 - \frac{x}{\gamma\lambda}\right)_+ dx, \tag{3}$$

where $\gamma$ is a given parameter that controls the concavity of $\rho$. Here $a_+ \equiv a1\{a > 0\}$ is the positive part of $a \in \mathbb{R}$. The Lasso and the hard threshold penalties can be considered two extremes of the MCP with $\gamma \to \infty$ and $\gamma \to 1$, respectively. We note that other penalized functions such as SCAD (Fan and Li, 2001) can also be used.

For a fixed $\lambda$, let

$$\hat{\boldsymbol{\beta}}_{(j)}(\lambda) = (\hat{\beta}_j(\lambda), \hat{\boldsymbol{\beta}}_{-j}(\lambda)) = \underset{\beta_j, \boldsymbol{\beta}_{-j}}{\operatorname{argmin}}\, L_j(\boldsymbol{\beta}; \lambda), 1 \leqslant j \leqslant p. \tag{4}$$

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

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

and

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

Let $\hat{S}_j = \{k : |\hat{\beta}_k(\lambda)| > 0, k \neq j\}$ be the set of nonzero elements in $\hat{\boldsymbol{\beta}}_{-j}$. We can write

$$\hat{\beta}_j(\lambda) = (\boldsymbol{x}_j'\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j'(\boldsymbol{y} - X_{\hat{S}_j}\hat{\boldsymbol{\beta}}_{\hat{S}_j}(\lambda)). \tag{7}$$

Here and in the sequel we use the notation $X_A = (\boldsymbol{x}_j : j \in A)$ and $\boldsymbol{\beta}_A = (\beta_j : j \in A)'$ for any $A \subset \{1,\ldots,p\}$. Take all the $\hat{\beta}_j(\lambda)$'s as a whole and denote it by $\hat{\boldsymbol{\beta}}(\lambda) = (\hat{\beta}_1(\lambda),\ldots,\hat{\beta}_p(\lambda))'$. For simplicity, we refer to $\hat{\beta}(\lambda)$ as a SPIDR estimator.

In comparison, the fully penalized criterion is

$$L(b;\lambda) = \frac{1}{2n}\|\boldsymbol{y} - \sum_{j=1}^{p}\boldsymbol{x}_j b_j\|^2 + \sum_{j=1}^{p}\rho(b_j;\lambda). \tag{8}$$

For a given $\lambda$, the solution to (8) is $\hat{b}(\lambda) = \operatorname{argmin}_b L(b;\lambda)$. Usually, a $\lambda = \hat{\lambda}$ is chosen based a data-driven procedure such as cross validation. Then $\hat{b}(\hat{\lambda})$ is the penalized estimator of $\boldsymbol{\beta}$. Since $\hat{b}(\hat{\lambda})$ can take exact zero value, the set $\hat{S}^* = \{j : |\hat{b}_j(\hat{\lambda})| > 0, 1 \leqslant j \leqslant p\}$ is taken as an estimator of $S$ based on the fully penalized criterion (8).

There is also an interesting connection between SPIDR and independence screening (Fan and Lv, 2008). Indeed, when $\lambda = \infty$ in (2), the SPIDR estimators $\hat{\beta}_1,\ldots,\hat{\beta}_p$ become the univariate least squares estimators of the regression coefficients.

[Figure 1 about here.]

We use an example to illustrate the basic properties of the solution paths $\hat{\boldsymbol{\beta}}_{(j)}(\lambda)$ and see how they differ from the fully penalized solution $\hat{b}(\lambda)$. Consider (1) with $(\beta_1,\ldots,\beta_6) = (3, 2, 1, -0.5, -1.0, -1.5)$, $\beta_j = 0, 7 \leqslant j \leqslant p$ and error distribution $N(0, 2.5^2)$. We set $n =$

$100, p = 1000$. The predictors are generated as follows:

$$x_{ij} = z_{ij} + au_{i1}, j = 1, \ldots, 4, \; x_{ij} = z_{ij} + au_{i2}, j = 5, \ldots, 8,$$

$$x_{ij} = z_{ij} + u_{i1}, j = 9, \ldots, 17, \; x_{ij} = z_{ij} + u_{i2}, j = 18, \ldots, 26, \; x_{ij} = z_{ij}, j = 27, \ldots, p,$$

where $\{z_{ij}, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p\}$ and $\{u_{ij} : 1 \leqslant i \leqslant n, j = 1, 2\}$ are independently generated random numbers from $N(0, 1)$. We 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 $a = 1, r = 0.5$.

The solution paths for $r = 0.25$ are shown in the top panel of Figure 1, where (a1) and (a2) show the Lasso and MCP paths, respectively; (a2)-(a5) show the SPIDR solution paths $\hat{\boldsymbol{\beta}}_{(1)}$, $\hat{\boldsymbol{\beta}}_{(2)}$ and $\hat{\boldsymbol{\beta}}_{(3)}$. The solid, dashed and dotted lines represent $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$, corresponding to $\beta_1 = 3$, $\beta_2 = 2$ and $\beta_3 = 1$, respectively. The bottom panel in Figure 1 shows the results for $r = 0.5$. The vertical lines are at the value of $\lambda$ chosen based on 5-fold cross validation. In (a1), $\log(\hat{\lambda}) = -0.47$, in (a2)-(a5), $\log(\hat{\lambda}) = -0.34$. In (b1), $\log(\hat{\lambda}) = -0.99$, in (b2)-(b5), $\log(\hat{\lambda}) = -0.27$.

This example illustrates two important features of the SPIDR estimator. First, the SPIDR estimator is more stable with respect to the change in the penalty parameter. This intuitively makes sense since $\hat{\beta}_j$ is not subject to penalization. Second, the SPIDR solution paths appear to be less severely impacted by the correlation among predictors comparing with Lasso and MCP. Indeed, it can be seen in Figure 1 (a1) and (b1) as correlation increases from 0.25 to 0.5, it becomes more difficult for Lasso and MCP to correctly select variables with smaller coefficients, but SPIDR is still able to identify such variables.

### 2.2 *Selection with direct false discovery rate control*

In this subsection, we first give a heuristic argument for the distributional property of $\hat{\boldsymbol{\beta}}$. We then use this property to define a selection rule based on directly controlling false discovery rate. We also discuss the confidence intervals of the selected coefficients that can be considered dual to the selection results.

For $A \subset \{1,\ldots,p\}$, denote the projection matrix onto the column space of $X_A$ by $P_A = X_A(X_A'X_A)^-X_A'$. Let $Q_{\hat{s}_j} = I - P_{\hat{s}_j}$ and let $\Sigma_{\hat{S}_j} = X_{\hat{S}_j}'X_{\hat{S}_j}/n$. Suppose the value of the penalty parameter $\lambda$ is chosen using cross validation. Let $\hat{\beta}_j = \hat{\beta}_j(\lambda)$. A useful alternative expression of (7) for $\hat{\beta}_j$ is

$$\hat{\beta}_j = (\boldsymbol{x}_j'Q_{\hat{S}_j}\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j'Q_{\hat{S}_j}\boldsymbol{y} + (\boldsymbol{x}_j'Q_{\hat{S}_j}\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j'X_{\hat{S}_j}\Sigma_{\hat{S}_j}^{-1}\dot{\rho}(\hat{\boldsymbol{\beta}}_{\hat{S}_j};\lambda), \tag{9}$$

where $\dot{\rho}(\hat{\boldsymbol{\beta}}_{\hat{S}_j};\lambda) \equiv (\dot{\rho}(\hat{\beta}_j;\lambda) : j \in \hat{S}_j)'$. We verify (9) in the Appendix. The second term on right-hand side represents the bias introduced by correlation between $\boldsymbol{x}_j$ and $X_{\hat{S}_j}$ and the penalization. If the correlation is small, then the bias is negligible. In general, if the nonzero coefficients are bigger than $\gamma\lambda$ and the estimator $\hat{\boldsymbol{\beta}}_{\hat{s}_j}$ is consistent so that $\hat{\beta}_j \geqslant \gamma\lambda$ for all $j \in \hat{S}_j$ with high probability, then since the derivative of MCP $\dot{\rho}(t;\lambda) = \lambda\{1 - |t|/(\gamma\lambda)\}_+\mathrm{sgn}(t)$, $\dot{\rho}(\hat{\boldsymbol{\beta}}_{\hat{S}_j};\lambda) = 0$ with high probability. In addition, if the estimator based on (4) is selection consistent in the sense that $\hat{S}_j$ equals $S_j \equiv \{k : \beta_k \neq 0, k \neq j\}$ with high probability, then

$$\hat{\beta}_j \approx (\boldsymbol{x}_j'Q_{S_j}\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j'Q_{S_j}\boldsymbol{y}, 1 \leqslant j \leqslant p. \tag{10}$$

In Section 3 we provide sufficient conditions under which the approximations in (10) hold simultaneously for all $1 \leqslant j \leqslant p$ with high probability. Under model (1), $\boldsymbol{y} = \boldsymbol{x}_j\beta_j + X_{S_j}\boldsymbol{\beta}_{S_j} + \varepsilon$, so we have $\hat{\beta}_j \approx \beta_j + (\boldsymbol{x}_j'Q_{S_j}\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j'Q_{S_j}\varepsilon$. It follows that $\hat{\beta}_j$ is consistent and asymptotically normal. Its variance can be consistently estimated by

$$\hat{\sigma}_j^2 = \hat{\sigma}^2(\boldsymbol{x}_j'Q_{\hat{s}_j}\boldsymbol{x}_j)^{-1}, \tag{11}$$

where $\hat{\sigma}^2$ is a consistent estimator of $\sigma^2$. We describe an approach for obtaining such an estimator in Section 4. The covariance between $\hat{\beta}_j$ and $\hat{\beta}_k$ can be consistently estimated by

$$\widehat{\mathrm{Cov}}(\hat{\beta}_j, \hat{\beta}_k) = \hat{\sigma}^2\frac{\boldsymbol{x}_j'Q_{\hat{s}_j}Q_{\hat{s}_k}\boldsymbol{x}_k}{(\boldsymbol{x}_j'Q_{\hat{s}_j}\boldsymbol{x}_j)(\boldsymbol{x}_k'Q_{\hat{s}_k}\boldsymbol{x}_k)}. \tag{12}$$

Thus $\hat{\beta} = (\hat{\beta}_1,\ldots,\hat{\beta}_p)'$ has an asymptotic multivariate normal distribution with mean

$(\beta_1, \ldots, \beta_p)'$ and covariance matrix specified by (11) and (12). This enables us to formulate the problem of variable selection into the framework of large scale hypothesis test.

We consider the $z$-statistics $z_j = \hat{\beta}_j/\hat{\sigma}_j, 1 \leqslant j \leqslant p$. We can think of variable selection as testing $p$ hypotheses $H_{0j} : \beta_j = 0, 1 \leqslant j \leqslant p$. For a given $t > 0$, we reject $H_{0j}$ if $|z_j| > t$, or equivalently, we select the $j$th variable if $|z_j| > t$. Therefore, the problem of variable selection becomes that of determining a threshold value according to a proper control of error. Let $R(t) = \sum_{j=1}^{p} 1\{|z_j| > t\}$ be the number of variables with $|z_j| > t$, and let $V(t) = \sum_{j=1}^{p} 1\{|z_j| > t, \beta_j = 0\}$ be the number of falsely selected variables. We can also write $V(t) = \sum_{j \in S^c} 1\{|z_j| > t\}$, where $S^c$ is the complement of $S$ in $\{1, \ldots, p\}$.

The false discovery proportion, or the proportion of the null variables among the selected ones for a given $t$ is

$$\mathrm{Fdp}(t) = \begin{cases} \frac{V(t)}{R(t)} & \text{if } R(t) > 0, \\ 0 & \text{if } R(t) = 0. \end{cases} \tag{13}$$

The FDR is defined to be $Q(t) = \mathrm{E}(\mathrm{Fdp}(t))$ (Benjamini and Hochberg, 1995). We seek a selection rule $R(t)$ by controlling $Q(t)$. However, since $Q(t)$ is an unknown population quantity, we need to estimate it in order to determine the threshold value. We can not directly use $\mathrm{Fdp}(t)$ as an estimator of $Q(t)$, since $V$ is unobservable. An approximation to $V(t)$ is by its expectation, $\mathrm{E}V(t) \approx 2|S^c|\Phi(-|t|)$, where $\Phi$ is the standard normal distribution function. In sparse models with $|S^c|/p \approx 1$, we further approximate $V(t)$ by $\hat{V}(t) = 2p\Phi(-|t|)$. This results in a first estimate of the FDR

$$\hat{Q}_0(t) = \begin{cases} \frac{\hat{V}(t)}{R(t)} & \text{if } R(t) > 0, \\ 0 & \text{if } R(t) = 0. \end{cases} \tag{14}$$

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

$$\hat{Q}(t) = \hat{Q}_0(t)\left[1 + 2A\frac{t\phi(t)}{\sqrt{2}\Phi(-t)}\right], \tag{15}$$

where $\hat{Q}_0(t)$ is given in (14), $\phi$ is the probability density function of $N(0,1)$. Here $A$ is a dispersion variable accounting for the correlation of the statistics $\hat{z}_j$, which can be estimated based on the their observed values. Methods for estimating $A$ are given in Efron (2007).

For $0 < q < 1$, let $\hat{t}_q$ be the value satisfying $\hat{Q}(\hat{t}_q) = q$, which is an estimator of $\tilde{t}_q$. The set of the indices of the selected variables is

$$\hat{S}_q = \{j : |z_j| \geqslant \hat{t}_q\}. \tag{16}$$

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

[Figure 2 about here.]

As an illustration of SPIDR selection, Figure 2 shows the $z$-statistics and $p$-values based on simulated data from the two models described in Examples 1 and 2 in Section 4. For comparison, we also include the selection results from the Lasso and MCP. In these two examples, there are 18 predictors with nonzero coefficients among a total of $p = 1000$ variables. Here the indices of the nonzero coefficients are randomly selected from 1 to $p$. The top panel in Figure 2 shows the results from a model with the largest pairwise correlation $r = 0.5$, where (a1) and (a2) show the Lasso and MCP selection results, the black dots represent predictors with nonzero coefficients; (a3) shows the SPIDR $z$-statistics, the two horizontal lines are drawn at the threshold values $\pm\hat{t}_q$ with $\hat{t}_q = 3.48$ and $q = 0.15$; and (a4) shows the negative $\log_{10}$ of the $p$ values based on the $z$ statistics, the horizontal line is drawn at $-\log_{10}(2\Phi(-\hat{t}_q)) = 3.30$. Plots (b1)-(b4) in the bottom panel show the results from Example 2 with $\hat{t}_q = 3.75$ in (b3), $-\log_{10}(2\Phi(-\hat{t}_q)) = 3.75$ in (b4) and the largest pairwise correlation $r = 0.8$.

By examining Figure 2, we see that SPIDR has better selection performance than Lasso and MCP for these two data sets. For $r = 0.5$, it has a smaller FDR and misses fewer non-null predictors. For $r = 0.8$, Lasso has zero FDR, but it misses 12 of the 18 non-null predictors. MCP has a higher FDR than SPIDR and misses 9 non-null predictors. It is interesting to note that the performance of SPIDR remains essentially unchanged as correlation increases from 0.5 to 0.8.

### 2.3 *Confidence intervals of selected coefficients*

The selection rule (16) directly leads to confidence intervals for the coefficients of the selected variables. The $1 - q$ level FDR-adjusted confidence intervals of the selected coefficients are

$$\hat{\beta}_j \pm \hat{t}_q \hat{\sigma}_j, j \in \hat{S}. \tag{17}$$

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

## 3. Statistical properties

### 3.1 *Ideal estimator*

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

$$(\tilde{\beta}_j, \tilde{\boldsymbol{\beta}}_{-j}) = \underset{\beta_j, \boldsymbol{\beta}_{-j}}{\operatorname{argmin}} \{\|\boldsymbol{y} - \boldsymbol{x}_j \beta_j - X_{-j} \boldsymbol{\beta}_{-j}\|^2 : \boldsymbol{\beta}_{S_j^c} = 0\}, 1 \leqslant j \leqslant p. \tag{18}$$

In particular, $\tilde{\beta}_j$ is an ideal estimator of $\beta_j$. We note that (18) is a counterpart of (2) without penalization assuming that the support of $\boldsymbol{\beta}_{-j}$ is known. It can be verified that an explicit expression of the ideal estimator $\tilde{\beta}_j$ is

$$\tilde{\beta}_j = \beta_j + (\boldsymbol{x}_j' Q_{S_j} \boldsymbol{x}_j)^{-1} \boldsymbol{x}_j' Q_{S_j} \varepsilon, \quad (j = 1, \ldots, p). \tag{19}$$

This expression of $\tilde{\beta}_j$ is parallel to (9).

By (19), $(\tilde{\beta}_1, \ldots, \tilde{\beta}_p)$ has a multivariate normal distribution with mean vector $\boldsymbol{\beta}$ and

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

We first state a result when the penalized criterion (8) is convex. This necessarily requires

$p < n$, but allows $p \to \infty$ as $n \to \infty$. Let $c_{\min} = \min\{c_j : 1 \leqslant j \leqslant p\}$, where $c_j$ is the smallest eigenvalue of $X'_{-j}Q_jX_{-j}/n$. Let $w^o = \max\{w^o_{jk} : k \in S_j, 1 \leqslant j \leqslant p\}$, 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 $\beta_* = \min\{|\beta^o_j| : \beta^o_j \neq 0, 1 \leqslant j \leqslant p\}$. 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) $\lambda \geqslant \sigma\sqrt{4\log p}\max_{j\leqslant p}\|\boldsymbol{x}_j\|/n$. Then,*

$$\mathrm{P}\{\cup_{j=1}^p(\hat{S}_j \neq S_j)\} \leqslant 3\epsilon \textit{ and } \mathrm{P}\{\cup_{j=1}^p(\hat{\beta}_j(\lambda) \neq \tilde{\beta}_j)\} \leqslant 3\epsilon.$$

This theorem shows that in the convex case, the SPIDR estimator equals the ideal estimator with high probability. As a consequence, it is asymptotically normal. The conditions are mild. The normality assumption on the errors is mainly used for bounding the tail probabilities of the error distribution. This assumption can be relaxed. Condition (a) guarantees that the SPIDR criteria in (2) are strictly convex to ensure unique solution. Condition (b) requires that the nonzero coefficients not be too small so that it is possible to separate them from zero in the presence of random noise. Condition (c) requires the penalty to be proportionally greater than the noise level to prevent false selection of null variables. For standardized predictors with $\|\boldsymbol{x}_j\|^2 = n$, this condition simplifies to $\lambda \geqslant \sigma\sqrt{(4/n)\log p}$. Conditions (b) and (c) are related, a bigger $\lambda$ requires a bigger $\beta^*$.

We now consider the high-dimensional cases where $p \gg n$ and the criteria (2) are non-convex. We require the sparse Riesz condition (SRC) (Zhang and Huang, 2008) on the the matrices $Q_jX$. Specifically, we assume there exist constants $0 < c_* \leqslant c^* < \infty$ and integer $d^* \geqslant |S|(K_* + 1)$ with $K_* = c^*/c_* - 1/2$ such that

$$0 < c_* \leqslant \|Q_jX_{A_j}u\|^2/n \leqslant c^* < \infty, \|u\|_2 = 1, \tag{20}$$

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

Theorem 2: *Suppose that $\varepsilon_1, \ldots, \varepsilon_n$ are independent and identically distributed as $N(0, \sigma^2)$. Also, suppose that (a) the SRC (20) holds with $\gamma \geqslant c_*^{-1}\sqrt{4 + c_*/c^*}$; (b) for a small $\epsilon > 0$, $\beta_* \geqslant \gamma 2\sqrt{c^*}\lambda + \sigma\sqrt{(2/n)w^o \log(p|S|/\epsilon)}$; (c) $\lambda \geqslant \sigma\sqrt{(4\log(p/\epsilon)}\max_{j\leqslant p}\|\boldsymbol{x}_j\|/n$. Then*

$$\mathrm{P}\{\cup_{j=1}^p(\hat{S}_j(\hat{\lambda}) \neq S_j)\} \leqslant 3\epsilon, \quad \textit{and} \;\; \mathrm{P}\{\cup_{j=1}^p(\hat{\beta}_j(\hat{\lambda}) \neq \tilde{\beta}_j)\} \leqslant 3\epsilon.$$

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

The SRC (20) ensures that the model is identifiable in a lower-dimensional space that contains the underlying model. When $p > n$, the smallest eigenvalue of $X_j' Q_j X_j/n$ is always zero. But the requirement $c_* > 0$ only concerns $d^* \times d^*$ diagonal submatrices of $X_j' Q_j X_j/n$. By examining the conditions (b) and (c), for standardized predictors with $\|\boldsymbol{x}_j\| = \sqrt{n}$, we can have $\log(p|S|/\epsilon) = o(n)$ or $p = \epsilon\exp(o(n))/|S|$. Thus for sparse models with $|S|$ small relative to $n$, Theorem 2 shows that the asymptotic idealness property of the SPIDR estimators continues to hold in high-dimensional settings under the SRC and other suitable conditions.

Theorems 1 and 2 are stated for fixed predictors. For random predictors, the conditions involving the predictors such as the SRC (20) need to hold with high probability.

### 3.2 *Impact of correlation*

The correlation structure of the predictors has a big impact on the selection results in penalized regression. In SPIDR, selection is based on the $z$-statistics $z_j, 1 \leqslant j \leqslant p$. Variables with similar $z$-statistic values will be selected or dropped together. So we consider the difference between $z_j$ and $z_k$ for $j \neq k$. Based on the asymptotic idealness property of SPIDR stated in Theorems 1 and 2, we can look at the ideal estimator from a large sample standpoint.

We first consider the notion of signal strength for measuring the importance of a predictor. Let $m_j = (\boldsymbol{x}_j' Q_{S_j}\boldsymbol{x}_j)^{1/2}$. The ideal estimator of $\beta_j$ can be written as $\tilde{\beta}_j = m_j^{-1}\boldsymbol{x}_j' Q_{S_j}\boldsymbol{y}$. The corresponding $z$-score is $\tilde{z}_j = m_j(\tilde{\beta}_j/\sigma)$. We define the signal strength of the $j$th predictor by $\psi_j = \mathrm{E}\tilde{z}_j = m_j(\beta_j/\sigma)$. The interpretation of $\psi_j$ is clear, it depends on the ratio of the $j$th coefficient over the error standard deviation and the length of $Q_{S_j}\boldsymbol{x}_j = \boldsymbol{x}_j - P_{S_j}\boldsymbol{x}_j$, the vector of residuals of $\boldsymbol{x}_j$ regressing on the variables in $S_j$. We refer to $\beta_j/\sigma$ as the base signal and $m_j$ as the signal multiplier.

In the extreme case where the signal multiplier $m_j$ is zero, that is, $\boldsymbol{x}_j$ is perfectly correlated with the variables in $S_j$, the signal strength of $\boldsymbol{x}_j$ is zero, no matter how large the base signal is. On the other hand, for a variable with a small to moderate base signal $\beta_j/\sigma$, its signal strength can still be relatively large if the signal multiplier is large.

We examine the effect of correlation by considering the mean squared difference $\mathrm{E}(\tilde{z}_j - \tilde{z}_k)^2$. It can be easily verified that

$$\mathrm{E}(\tilde{z}_j - \tilde{z}_k)^2 = (\psi_j - \psi_k)^2 + 2(1 - \mathrm{Cov}(\tilde{z}_j, \tilde{z}_k)), \tag{21}$$

where

$$\mathrm{Cov}(\tilde{z}_j, \tilde{z}_k) = \mathrm{Corr}(\tilde{\beta}_j, \tilde{\beta}_k) = \frac{\boldsymbol{x}_j' Q_{S_j} Q_{S_k} \boldsymbol{x}_k}{(\boldsymbol{x}_j' Q_{S_j} \boldsymbol{x}_j)^{1/2} (\boldsymbol{x}_k' Q_{S_k} \boldsymbol{x}_k)^{1/2}}.$$

So grouping is determined by the difference in signal strengthes and the predictor residual correlation between $Q_{S_j}\boldsymbol{x}_j$ and $Q_{S_k}\boldsymbol{x}_k$. It is not related to the usual pairwise correlations. Signal strength is a main factor in determining SPIDR selection. The pairwise correlations among predictors do not have an impact as big as in penalized selection. So two key quantities that affect SPIDR selection: the signal strength and pairwise predictor residual correlation.

The difficulty that Lasso has in the presence of high pairwise correlations had been pointed out by Zou and Hastie (2006). This is one of the main motivations for them to introduce the elastic net, which has a grouping effect by selecting or dropping strongly correlated predictors together. As discussed above, the grouping effect of SPIDR is different from that of elastic net. It depends on the signal strengths of the variables and residual correlations between predictors, but not the usual pairwise correlations.

## 4. Numerical studies

### 4.1 *Implementation*

To implement the proposed method, we need to determine the penalty parameter $\lambda$ and estimate the error variance $\sigma^2$. The former is needed for estimating the regression coefficients and the latter is required for computing the $z$-statistics based on the estimated regression coefficients.

We employ 5-fold cross validation for choosing $\lambda = \hat{\lambda}$ based on the fully penalized criterion (8) using the MCP (3) with $\gamma = 6$. This requires computing the solution path $\hat{b}(\lambda) = \operatorname{argmin}_b L(b; \lambda)$ for $\lambda$ in a properly specified interval. The R package *ncvreg* is used in the computation. This package implements a coordinate descent algorithm for penalized methods including the Lasso and MCP, and is available at `cran.r-project.org/web/packages/ncvreg` (Breheny and Huang, 2009). This $\hat{\lambda}$ is then used in calculating $\hat{\beta}_j = \hat{\beta}_j(\hat{\lambda}), 1 \leqslant j \leqslant p$ in (2). In this way, it is only necessary to calculate $\hat{\beta}_j$ at $\hat{\lambda}$. Conceptually, it is possible to choose a different $\lambda$ for each $\hat{\beta}_j$. However, this will substantially increase the computational cost, since it will involve calculating the whole solution path for each of the $p$ minimization problems in (2). Also, since $\hat{\beta}_j$ is not very sensitive to $\lambda$, choosing a $\hat{\lambda}$ based on (8) is reasonable.

For estimating $\sigma^2$, we use the following procedure. Let $\hat{b}(\hat{\lambda})$ be the MCP estimator with $\hat{\lambda}$ determined based on 5-fold cross validation. Let $\hat{S}$ be the set of the predictors with nonzero coefficients in $\hat{b}$. We randomly partition the dataset into two subsets $D_1$ and $D_2$ with equal sample sizes $n_1 = n_2 = n/2$. We use the first part to fit a model with variables in $\hat{S}$ and calculate the least squares estimate $\hat{b}^{(1)} = \operatorname{argmin}_{\boldsymbol{\beta}} \sum_{i \in D_1} (y_i - \sum_{j \in \hat{S}} x_{ij} b_j)^2$. Let

$$\hat{\sigma}^2 = \frac{1}{n_2 + |\hat{S}|} \sum_{i \in D_2} (y_i - \sum_{j \in \hat{S}} x_{ij} \hat{b}_j^{(1)})^2. \tag{22}$$

We show in the Appendix that this is a consistent estimator of $\sigma^2$. To smooth out the variations of the random partition, we repeat this process 10 times and take the average of the resulting $\hat{\sigma}^2$'s as the estimate of $\sigma^2$.

This procedure bears some resemblance to the cross-refitted method for variance estimation in Fan et al. (2012). But there are also important differences. Here we use the full dataset to select variables and then use a properly scaled prediction error for variance estimation. One reason for using the full dataset as opposed to using a subset is to achieve better selection results. Another reason is to take advantage of the fact that in choosing the penalty parameter $\hat{\lambda}$ based on (8) for calculating the SPIDR estimators based on (2), we have already computed the full penalized estimator. Thus the procedure described above does not incur any significant extra computational burden.

### 4.2 *Simulation studies*

We focus on the selection results of the SPIDR method in three models described below. Specifically, we look at the empirical FDR and FMR (false miss rate). For a given threshold value $t > 0$, let $U(t) = \sum_{j \in S} 1\{|z_j| > t\}$ be the number of selected variables in $S$. The false miss proportion is defined to be

$$\text{Fmp}(t) = \frac{|S| - U(t)}{|S|}.$$

Then the FMR at $t$ is $\text{E}(\text{Fmp}(t))$. As a comparison, we also look at the empirical FDR and FMR of the selection results based on the Lasso and MCP.

**Example 1.** We consider model (1) with $p = 1000$. The errors are independent and identically distributed as $N(0, \sigma^2)$ 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 $\{z_{ij}, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p\}$ and $\{u_{ij} : 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant 2\}$ 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 $\{19,\ldots,p\}$.

$$x_{ij} = z_{ij} + a_1 u_{i1}, j \in A_1,\ x_{ij} = z_{ij} + a_1 u_{i2}, j \in A_2,$$

$$x_{ij} = z_{ij} + a_2 u_{i1}, j \in A_3,\ x_{ij} = z_{ij} + a_2 u_{i2}, j \in A_4,$$

$$x_{ij} = z_{ij} + a_3 u_{i1} - a_3 u_{i2}, j \in A_5, x_{ij} = z_{ij}, j \notin \cup_{k=1}^{5} A_k,$$

where $a_1 = 1$, $a_2 = 0.5$ and $a_3 = 0.1$. In this model, there is correlation among predictors with nonzero coefficients as well as between such predictors and predictors with zero coefficients. For example, 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 $\Sigma$ is $\sigma_{jk} = 0.5^{|j-k|}$, $1 \leqslant j, k \leqslant p$.

[Figure 3 about here.]

Figure 3 shows the empirical FDR's and FMR's from 100 replications. For the SPIDR, the nominal FDR is set at $q = 0.15$. The top panel in Figure 3 shows the empirical false discovery rates for (a1) Example 1, (a2) Example 2 and (a3) Example 3, and the plots (b1)-(b3) in the bottom panel show the empirical false miss rates for these studies. Since it is difficult to assess the absolute performance of the SPIDR, we also include the selection results from the Lasso and MCP for comparison. The Lasso and MCP results are obtained at the penalty parameter value determined by 5-fold cross validation. In the plots, the results for Lasso, MCP and SPIDR are represented by the plus "+", cross "x" and open circle "○" signs, respectively.

Numerical summaries of Figure 3 are given in Table 1. As can be seen in the plots, there is a fair amount of variations in the false discovery rates. However, the average false discovery rate for SPIDR are close to the nominal level, as shown in Table 1. Overall, the SPIDR has smaller FDR and FMR than the Lasso and MCP in the three examples considered. In particular, in Example 2, where the correlation is high, the SPIDR has considerably smaller FDR and FMR than the Lasso and MCP.

[Table 1 about here.]

In Example 1, SPIDR has slightly higher PCS and slightly lower PFS than MCP. Both SPIDR and MCP perform better than Lasso in terms of PCS. In Example 2, SPIDR has considerably higher PCS and lower PFS than Lasso and MCP. In Example 3, SPIDR has higher PCS and lower PFS than Lasso and MCP, although for two predictors with smaller coefficients, all the methods have relatively low PCS.

In summary, the SPIDR has good performance in the examples considered here. It can achieve the nominal FDR control on average and tends to have smaller FMR than the Lasso and MCP. Especially, for the model in Example 2, which is a difficult case for the Lasso and MCP because of the high correlations among the predictors, the SPIDR still performs reasonably well.

### 4.3 *Breast cancer gene expression data*

We use the breast cancer data from The Cancer Genome Atlas (2012) project to illustrate the proposed method. In this dataset, tumour samples were assayed on several platforms. Here we focus on the gene expression data obtained using Agilent mRNA expression microarrays. In this dataset, 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.

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.

We start by looking at the Lasso and MCP solution paths together with 5-fold cross validation results, which are shown in Figure 4. The vertical lines are drawn at the values $\hat{\lambda}$ of the penalty parameter that achieve the smallest cross validation errors for Lasso and MCP, respectively. For the Lasso, $\log(\hat{\lambda}) = -2.96$, for the MCP, $\log(\hat{\lambda}) = -3.15$. The gray lines in Figure 4 (b) and (d) represent the standard deviations of the cross validation errors calculated based on 5-fold calculations. These plots show that for either Lasso or MCP, there is a unique point on the solution path that minimizes the cross validation error, which leads to a well-defined model.

[Figure 4 about here.]

The Lasso and MCP estimates at the cross-validated $\hat{\lambda}$ are shown in Figure 5 (a1) and (a2), the plus "+" and cross "x" signs represent genes selected by Lasso (24 genes) and MCP (48 genes). Figure 5 (a3) shows the SPIDR estimates, the circles "$\circ$" represent the selected genes (63 genes) with $q = 0.10$. The SPIDR z-statistics are shown in (a4), the cut-off values for selection corresponding to FDR level $q = 0.10$ are $\hat{t}_{0.10} = \pm 3.95$, which are indicated by two horizontal lines. Figure 5 (b1)-(b4) are parallel to (a1)-(a4), but now the overlaps between the methods are indicated. Figure 5 (b1) shows the overlap between the Lasso and SPIDR, the circles represent the genes that are also selected by SPIDR. Similarly, (b2) shows the overlap between the MCP and SPIDR. In (b3) and (b4), all the selected genes based on the three methods are indicated. As can be seen in (b4), genes with relatively large estimated coefficients based on Lasso or MCP are also selected by SPIDR, whereas those with small estimated coefficients tend to be deemed nonsignificant by SPIDR. There are large overlaps between the three methods. For example, 13 genes are selected by both Lasso and SPIDR, these same 13 genes are selected by all the three methods, and there are 24 genes selected by both MCP and SPIDR. One of the genes selected by all the three methods is CCDC56, it has the largest Lasso and MCP estimates and is also most significant based on SPIDR. This gene maps to human chromosome 17q21 and encodes the CCDC56 (coiled-coil domain containing 56) protein with 106 amino acid single-pass membranes. BRCA1 is located at 17q21-q24 and is in the neighborhood of CCDC56. Interestingly, another key tumour suppressor gene p53 also maps to chromosome 17.

[Figure 5 about here.]

On the other hand, there are genes not selected by the Lasso or MCP but selected by SPIDR. 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).

In this example we focus on illustrating the application of SPIDR. So we mainly highlight a few genes from the SPIDR analysis to confirm that it does reveal additional information from the data. A detailed description of the available biological functions of the selected genes is not included there, but can found from public database such as the website of National Center for Biotechnology Information (www.ncbi.nlm.nih.gov).

In Figure 6, plot (a) shows the histogram of the SPIDR $z$-statistics, the dashed curve represents the standard normal density function. The distribution of the SPIDR $z$-statistics has much heavier tail than the standard normal distribution and is slightly skewed to the right. This is due to the fact that some of the $z$-statistics are not from the null hypothesis. This can also be caused by correlation among $z$-statistics even if their marginal distributions are $N(0, 1)$. Such phenomenon has also been observed by Efron (2007) in the context of detecting differentially expressed genes using microarray data. This can also be clearly seen in the normal Q-Q plot (b). Plot (c) shows the negative $\log_{10}$ $p$-values for the SPIDR $z$-statistics. The cutoff for the negative $\log_{10}$ $p$-values for significance corresponding to FDR $q = 0.1$ is 4.10, which is represented by the horizontal line in the plot. For comparison, the $p$-values of the variables selected by Lasso and MCP are also indicated in the plot by plus "+" and cross "x" signs, respectively. Plot (d) shows the SPIDR confidence intervals for the selected coefficients.

[Figure 6 about here.]

Figure 6 provides a panel of useful summaries of the SPIDR analysis that can be used for statistical inference, including the distribution of z-statistics, the comparison with the normal distribution via Q-Q plot, the $p$-values and an indication of statistical significance according to a desired FDR control level, and the interval estimates of the selected effect sizes. These can be easily explained to the scientific investigators. It is best to use Figure 6 in combination with plots such as Figures 4 and 5 to give a clear view of the selection results along with tuning.

## 5. Discussion

SPIDR is built on two separate developments in high-dimensional statistics, penalized estimation and direct FDR control. It makes the connection between these two ideas and combines them for variable selection with an assessment of selection error. To study the theoretical property of the proposed SPIDR estimator, we introduced the concept of an ideal estimator and provided sufficient conditions under which the SPIDR estimator is ideal with high probability.

The proposed method can be extended in several directions. First, it can be applied to other regression models, including the generalized linear and Cox models. In these models, instead of using the quadratic loss in (2), we can use the negative log-likelihood or partial log-likelihood as the loss functions. Of course, detailed analysis of the theoretical properties of SPIDR in these models requires further work. Second, it is possible to consider the coefficients in groups and carry out the estimation one group at a time. In particular, SPIDR can be naturally extended to group selection problems with various types of group penalties, including the group Lasso and concave group penalties. However, in group selection, the definition of FDR needs to be modified accordingly. Third, the idea of SPIDR can be applied to semiparametric and nonparametric regression models such as the partially linear and generalized additive models.

The estimation of FDR with correlated statistics is a challenging problem. In addition to the difficulty caused by correlation, false discovery proportion is inherently variable in sparse models when the number of findings is relatively small. A small change in either the number of findings or the number of falsely selected variables can cause a big change in the proportion.

We used the method of Efron (2007), which is easy to implement and computationally efficient. Our simulation studies indicate that it can yield unbiased estimates, although the variability is relatively high. It would be interesting to develop methods tailored to the covariance structure given in (11) and (12).

We used the `R` package *ncvreg* to compute the SPIDR solutions. However, SPIDR appears especially suitable to be implemented in parallel, which should speed up the computation considerably. Thus it would be interesting to develop a more efficient implementation. In applications we recommend applying SPIDR in combination with penalized selection, as illustrated in the breast cancer data example in Section 4. In particular, it is helpful to present figures similar to Figures 4 to 6 to summarize the analysis results from both penalized selection and SPIDR. Our simulation studies and data example suggest that SPIDR is a useful addition to the existing methods for high-dimensional statistical inference in practice.

## References

Alhosin, M., Sharif, T., Mousli, M., Etienne-Selloum, Guy Fuhrmann, N., Schini-Kerth. V. B. and Christian Bronner (2011). Down-regulation of UHRF1, associated with re-expression of tumor suppressor genes, is a common feature of natural compounds exhibiting anti-cancer properties. *J. Experimental & Clinical Cancer Research*, 30:41.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289-300.

Benjamini, Y. and Yekutieli, D. (2005). False Discovery Rate Adjusted Multiple Confidence Intervals for Selected Parameters. *J. Amer. Statist. Assoc.*, **100**, 71-81.

Breheny, P. and Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression methods. *Ann. Appl. Statist.* **5**, 232-253.

Bühlmann P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, New York.

Efron, B. (2007). Correlation and large-scale simultaneous significance testing. *J. Amer. Statist. Assoc.*, **102**, 93-103.

Fan, J., Guo, S. and Hao, N. (2012). Variance Estimation Using Refitted Cross-validation in Ultrahigh Dimensional Regression. *J. R. Statist. Soc. B*, **74**, 37-65.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348-1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Statist. Soc. B*, **70**, 849-911.

Hayes, G. M., Carrigan, P. E., Beck, A. M. and Miller L. J. (2006). Targeting the RNA splicing machinery as a novel treatment strategy for pancreatic carcinoma. *Cancer Res.*, **66**, 3819-3827.

Hayes, G. M., Carrigan, P. E. and Miller, L. J. (2007). Serine-arginine protein kinase 1 overexpression is associated with tumorigenic imbalance in mitogen-activated protein kinase pathways in breast, colonic, and pancreatic carcinomas. *Cancer Res.*, **67**, 2072-2080.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *J. R. Statist. Soc. B*, **72**, 417-473.

Meinshausen, N., Meier, L. and Bühlmann, P. (2009). P-values for high-dimensional regression. *J. Amer. Statist. Assoc.*, **104**, 1671-1681.

Storey, J. D. (2002). A direct approach to false discovery rates. *J. R. Statist. Soc. B* **64**, 489-498.

Storey, J. D. and Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proc. Nat. Acad. Sci.*, **100**, 9440-9445.

The Cancer Genome Atlas Network (2012). Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61-70.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Statist. Soc. B* **58**, 267-288.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.* **38**, 894-942.

Zhang, C.-H. and Huang, J. (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *Ann. Statist.*, **36**, 1567-1594.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B.* **67**, 301-320.

*Submitted March* 2016

## Appendix

### *Technical details*

**Verification of (9).** The solution to (2) satisfies

$$X_{\hat{S}_j}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}_j\hat{\beta}_j - X_{\hat{S}_j}\hat{\boldsymbol{\beta}}_{\hat{S}_j}) = n\dot{\rho}(\hat{\boldsymbol{\beta}}_{\hat{S}_j};\lambda),$$

$$\boldsymbol{x}_j^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}_j\hat{\beta}_j - X_{\hat{S}_j}\hat{\boldsymbol{\beta}}_{\hat{S}_j}) = 0.$$

The first equation gives $\hat{\boldsymbol{\beta}}_{\hat{S}_j} = (X_{\hat{S}_j}^{\mathrm{T}}X_{\hat{S}_j})^{-1}X_{\hat{S}_j}^{\mathrm{T}}(\boldsymbol{y} - \boldsymbol{x}_j\hat{\beta}_i) - n(X_{\hat{S}_j}^{\mathrm{T}}X_{\hat{S}_j})^{-1}\dot{\rho}(\hat{\boldsymbol{\beta}}_{\hat{S}_j};\lambda)$. Thus $X_{\hat{S}_j}\hat{\boldsymbol{\beta}}_{\hat{S}_j} = P_{\hat{S}_j}(\boldsymbol{y}-\boldsymbol{x}_j\hat{\beta}_j)-X_{\hat{S}_j}\Sigma_{\hat{S}_j}^{-1}\dot{\rho}(\hat{\boldsymbol{\beta}}_{\hat{S}_j};\lambda)$. Substituting this expression into the second equation gives $\boldsymbol{x}_j^{\mathrm{T}}\{Q_{\hat{S}_j}(\boldsymbol{y}-\boldsymbol{x}_j\hat{\beta}_j)+X_{\hat{S}_j}\Sigma_{\hat{S}_j}^{-1}\dot{\rho}(\hat{\boldsymbol{\beta}}_{\hat{S}_j};\lambda)\} = 0$. It follows that $\hat{\beta}_j = (\boldsymbol{x}_j^{\mathrm{T}}Q_{\hat{S}_j}\boldsymbol{x}_j)^{-1}\boldsymbol{x}_j^{\mathrm{T}}(Q_{\hat{S}_j}\boldsymbol{y}+ X_{\hat{S}_j}\Sigma_{\hat{S}_j}^{-1}\dot{\rho}(\hat{\boldsymbol{\beta}}_{\hat{S}_j};\lambda)$. This verifies (9).

**Consistency of $\hat{\sigma}^2$ in (22).** Let $(\boldsymbol{y}^{(1)}, X_{\hat{S}}^{(1)})$ and $(\boldsymbol{y}^{(2)}, X_{\hat{S}}^{(2)})$ represent the data in the partitions $D_1$ and $D_2$ with predictors in $\hat{S}$, where $\hat{S}$ is the set of variables selected based on the full dataset. For simplicity, we set $n_1 = n_2 = n/2$. Then the least squares estimator based on $(\boldsymbol{y}^{(1)}, X_{\hat{S}}^{(1)})$ is $\hat{b}^{(1)} = (X_{\hat{S}}^{(1)\prime}X_{\hat{S}}^{(1)})^{-1}X_{\hat{S}}^{(1)\prime}\boldsymbol{y}^{(1)}$. Under the conditions of Theorem 2, $\hat{S} = S$ with probability tending to 1 (Zhang ,2010). Thus we can replace $\hat{S}$ by $S$ in showing the consistency here. Therefore, since $\boldsymbol{y}^{(1)} = X_S^{(1)}\boldsymbol{\beta}_S + \varepsilon^{(1)}$, we have

$$\hat{b}^{(1)} = \boldsymbol{\beta}_S + (X_S^{(1)\prime}X_S^{(1)})^{-1}X_S^{(1)\prime}\varepsilon^{(1)}. \tag{A.1}$$

It follows that

$$\mathrm{E}\|\boldsymbol{y}^{(2)} - X_S^{(2)}\hat{b}^{(1)}\|^2 = \mathrm{E}\|\varepsilon^{(2)} - X_S^{(2)}(\hat{b}^{(1)} - \boldsymbol{\beta}_S)\|^2 = n_2\sigma^2 + \mathrm{E}\|X_S^{(2)}(\hat{b}^{(1)} - \boldsymbol{\beta}_S)\|^2.$$

Here the cross product term vanishes because of the independence between $\varepsilon^{(2)}$ and $\{X^{(2)}, \hat{b}^{(1)}\}$. By (A.1), the independence between $X_S^{(1)}$ and $X_S^{(2)}$ and after some algebra,

$$|S|^{-1}\mathrm{E}\|X_S^{(2)}(\hat{b}^{(1)} - \boldsymbol{\beta}_S)\|^2 = \sigma^2\mathrm{trace}\{(X_S^{(1)\prime}X_S^{(1)}/(n_1|S|))^{-1}(X_S^{(2)\prime}X_S^{(2)}/(n_2|S|)\} \to \sigma^2.$$

Combining the above two equations we obtain

$$(n_2 + |S|)^{-1}\mathrm{E}\|\boldsymbol{y}^{(2)} - X_S^{(2)}\hat{b}^{(1)}\|^2 \to \sigma^2.$$

This proves the consistency of $\hat{\sigma}^2$.

**Proof of Theorem 1.** Let $B_j = \{\hat{\boldsymbol{\beta}}_{-j}(\lambda) \neq \tilde{\boldsymbol{\beta}}_{-j}$ or $\mathrm{sgn}(\hat{\boldsymbol{\beta}}_{-j}(\lambda)) \neq \mathrm{sgn}(\boldsymbol{\beta}_{-j})\}$. By the definition of $\tilde{\boldsymbol{\beta}}_{-j}$, we have

$$\tilde{\boldsymbol{\beta}}_{-j} = \underset{\boldsymbol{\beta}_{-j}}{\mathrm{argmin}}\{\frac{1}{2n}\|Q_j(\boldsymbol{y} - X_{-j}\boldsymbol{\beta}_{-j})\|^2, \boldsymbol{\beta}^o_{S_j^c} = 0\}. \tag{A.2}$$

Thus $\boldsymbol{x}_k' Q_j(\boldsymbol{y} - X_{-j}\tilde{\boldsymbol{\beta}}_{-j}) = 0$ for $k \in S_j$. Also, $\dot{\rho}(\hat{\beta}_{-j,k}; \lambda) = 0$ if $|\hat{\beta}_{-j,k}| \geqslant \gamma\lambda$, where $\hat{\beta}_{-j,k}$ is the $k$th element of $\hat{\boldsymbol{\beta}}_j$. Therefore, $\tilde{\boldsymbol{\beta}}_{-j}$ is a solution to (4) and $\mathrm{sgn}(\hat{\boldsymbol{\beta}}_{-j}) = \mathrm{sgn}(\boldsymbol{\beta}_{-j})$ in the intersection of

$$\Omega_{j1}(\lambda) = \left\{\max_{k \notin S_j} |\boldsymbol{x}_k' Q_j(\boldsymbol{y} - X_{-j}\tilde{\boldsymbol{\beta}}_{-j})|/n < \lambda_1\right\} \text{ and } \Omega_{j2}(\lambda) = \left\{\min_{k \in S_j} \mathrm{sgn}(\beta_k)\tilde{\beta}_{-j,k} > \gamma\lambda\right\}. \tag{A.3}$$

Thus $\mathrm{P}\{B_j\} \leqslant 1 - \mathrm{P}\{\Omega_{j1}(\lambda)\} + 1 - \mathrm{P}\{\Omega_{j2}(\lambda)\}$. Following the proof of Theorem 4 of Zhang (2010), we have $\mathrm{P}\{B_j\} \leqslant 3\epsilon/p$. Since $\{\hat{S}_j \neq S_j\} \subseteq B_j$,

$$\mathrm{P}\{\cup_{j=1}^p(\hat{S}_j \neq S_j)\} \leqslant \sum_{j=1}^p \mathrm{P}\{B_j\} \leqslant 3\epsilon.$$

Similarly,

$$\mathrm{P}\{\cup_{j=1}^p(\hat{\beta}_j(\lambda) \neq \hat{\beta}_j^o)\} \leqslant \sum_{j=2}^p \mathrm{P}\{B_j\} \leqslant 3\epsilon.$$

This completes the proof.

**Proof of Theorem 2.** For $m \geqslant 1$ and $B \subset \{1,\dots,p\} \setminus \{j\}$, let

$$\varsigma_j(\boldsymbol{v}; m, B) = \max\left\{ \frac{\|(P_A - P_B)\boldsymbol{v}\|}{(mn)^{1/2}} : B \subseteq A \subseteq \{1,\dots,p\} \setminus \{j\}, |A| = m + |B| \right\},$$

for $\boldsymbol{v} \in \mathbb{R}^n$, where $P_A$ is the orthogonal project matrix from $\mathbb{R}^n$ to the linear span of $\{Q_j \boldsymbol{x}_k : k \in A\}$. Let $\Omega_{3j}(\lambda) = \{\varsigma_j(\varepsilon; m^*, S_j) \leqslant \lambda\}$. Following the proof of Theorem 5 of Zhang (2011), we have

$$\mathrm{P}\{\hat{\boldsymbol{\beta}}_j \neq \tilde{\boldsymbol{\beta}}_j \text{ or } \mathrm{sgn}(\hat{\boldsymbol{\beta}}_j) \neq \mathrm{sgn}(\boldsymbol{\beta}_j)\} \leqslant \sum_{k=1}^{3} (1 - \mathrm{P}\{\Omega_{jk}(\lambda)\}),$$

where $\Omega_{jk}, k = 1, 2$ are defined in (A.3). This inequality and Theorem 5(ii) of Zhang (2011) imply $\mathrm{P}\{\hat{S}_j \neq S_j\} \leqslant 3\varepsilon/p$. Therefore, $\mathrm{P}\{\cup_{j=1}^{p}(\hat{S}_j(\lambda) \neq S_j)\} \leqslant 3\varepsilon$. Similarly, we have $\mathrm{P}\{\cup_{j=1}^{p}(\hat{\beta}_j(\lambda) \neq \tilde{\beta}_j)\} \leqslant 3\varepsilon$. This completes the proof.

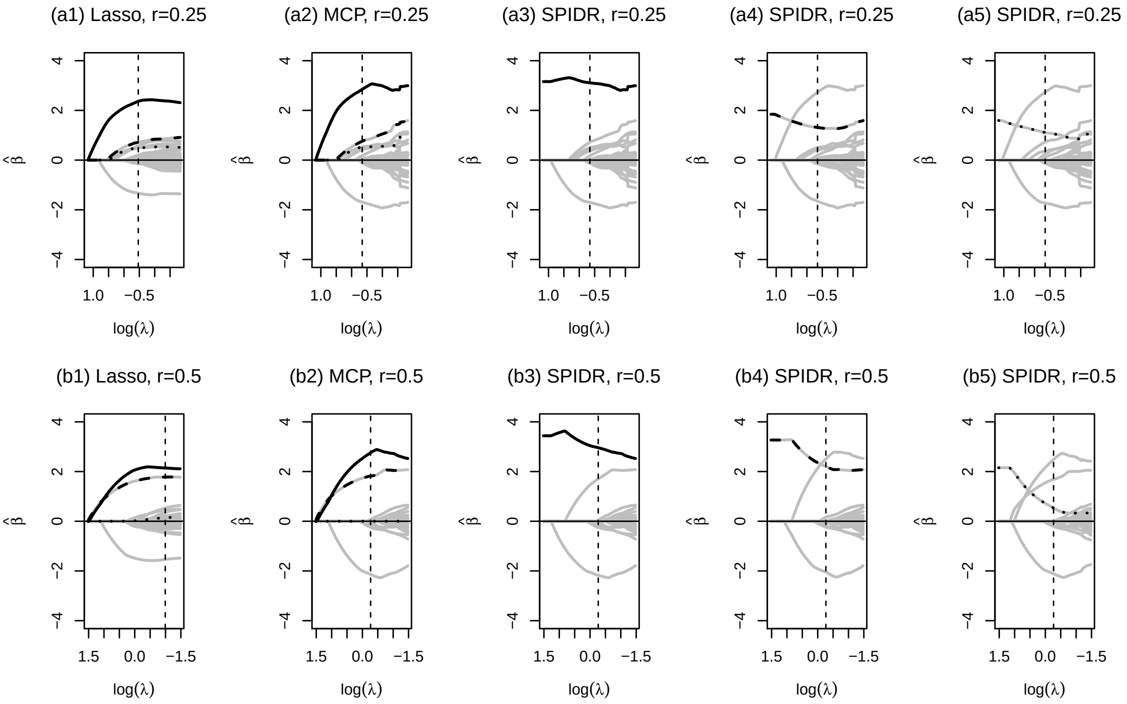

**Figure 1.** Lasso, MCP and SPIDR solution paths. The results for $r = 0.25$ are shown in the top panel (a1)-(a5), where (a1) and (a2) show the Lasso and MCP solution paths; (a3)-(a5) show the semi-MCP solution paths of $\hat{\boldsymbol{\beta}}_{(1)}$, $\hat{\boldsymbol{\beta}}_{(2)}$ and $\hat{\boldsymbol{\beta}}_{(3)}$. The solid, dashed and dotted lines represent the paths of $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\beta}_3$, corresponding to $\beta_1 = 3$, $\beta_2 = 2$ and $\beta_3 = 1$, respectively. The bottom panel (b1)-(b5) in Figure 1 shows the results for $r = 0.5$. The vertical lines are at the value of $\lambda$ chosen based on 5-fold cross validation.

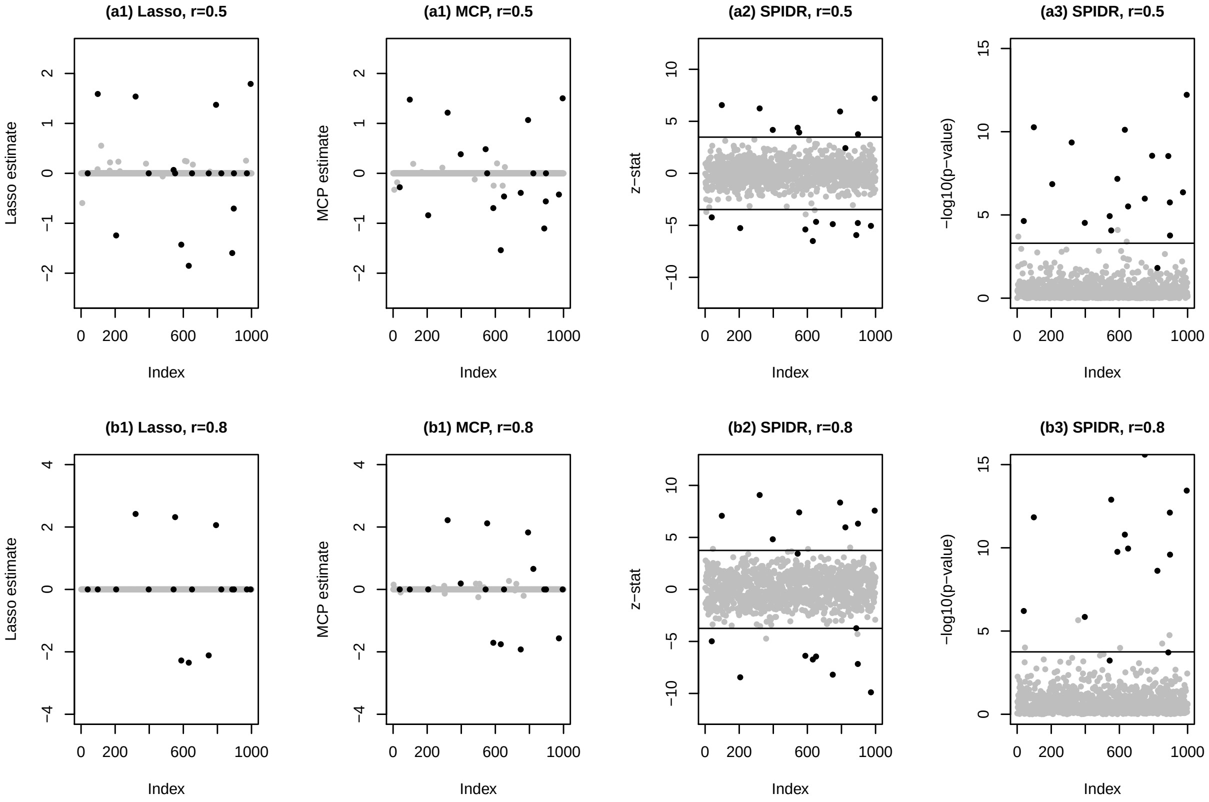

**Figure 2.** Selection results with $q = 0.15$ from the models in Examples 1 and 2. The top panel (a1)-(a4) shows the results from Example 1 with correlation $r = 0.5$. (a1) and (a2): the Lasso MCP selection results, the black dots represent predictors with nonzero coefficients; (a3): the $z$ statistics based on the SPIDR, the two horizontal lines are drawn at $\pm\hat{t}_q$. The bottom panel (b1)-(b4) shows the results from Example 2 with correlation $r = 0.8$.

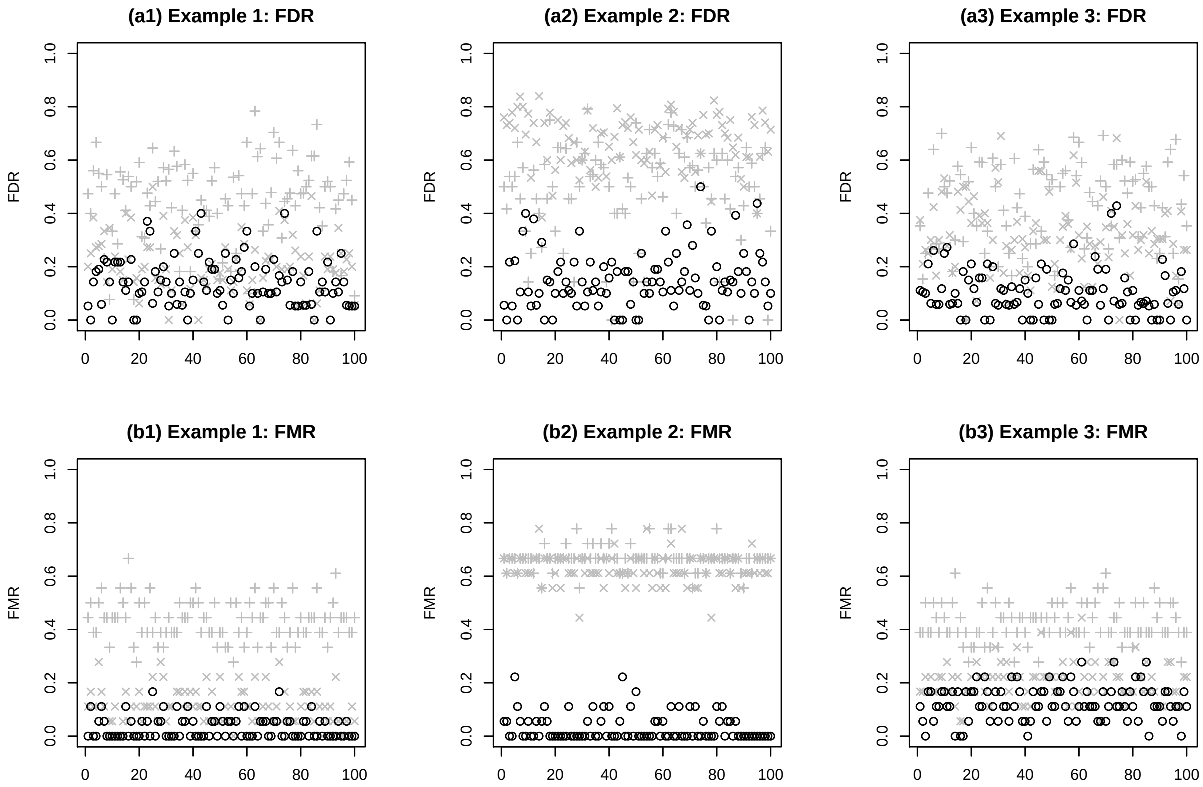

**Figure 3.** Top panel: False discovery rates from 100 replications for (a1) Example 1, (a2) Example 2 and (a3) Example 3. Bottom panel: False missing rates from 100 replications for (b1) Example 1, (b2) Example 2 and (b3) Example 3. The results for Lasso, MCP and SPIDR are represented by the plus "+", cross "x" and open circle "○" signs, respectively.

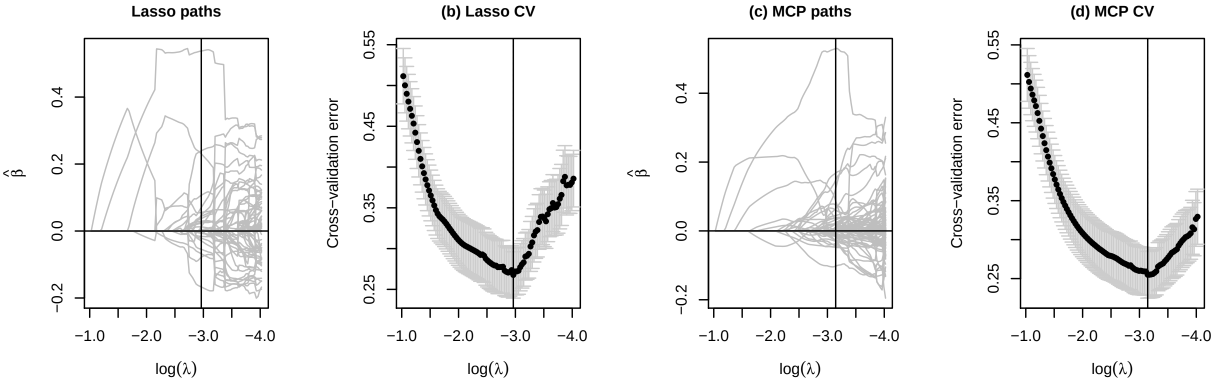

**Figure 4.** Breast cancer data. (a) Lasso solution paths; (b) Lasso cross validation results; (c) MCP solution paths; (d) MCP cross validation results.

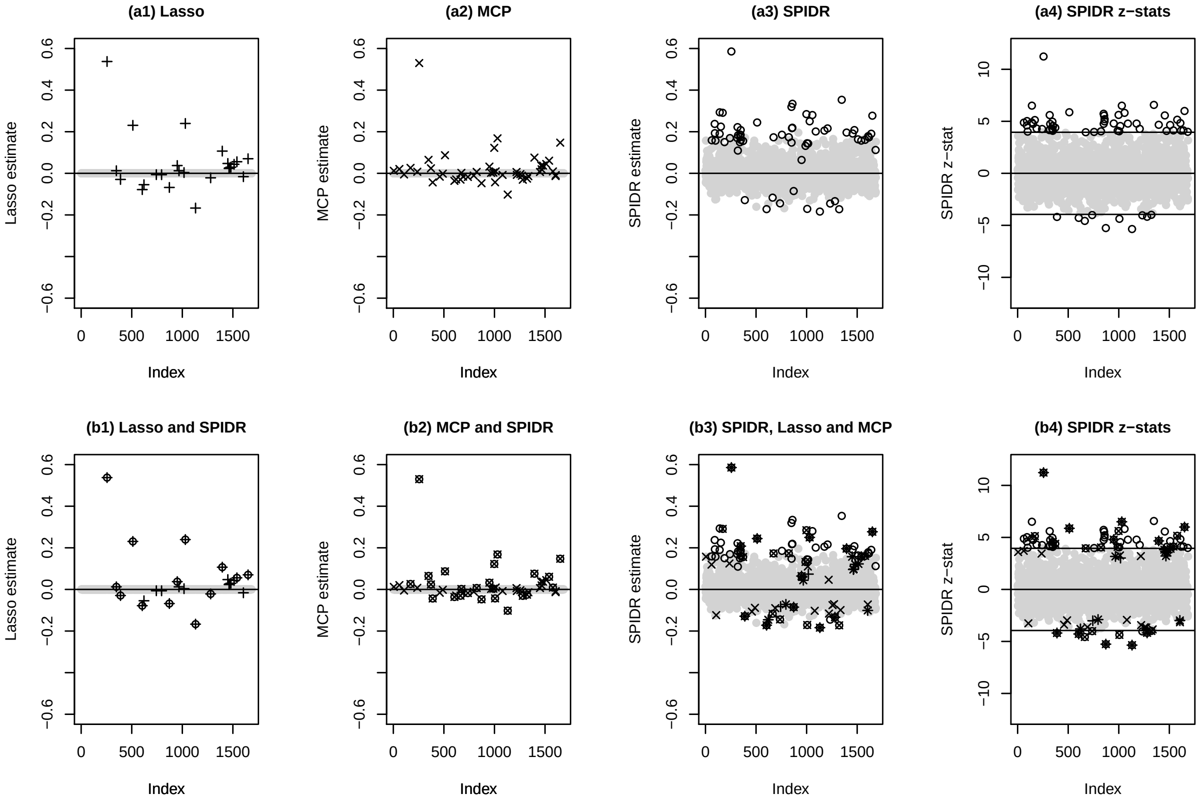

**Figure 5.** Breast cancer data. Lasso, MCP and SPIDR are represented by plus "+", cross "x", and circle "○", respectively. Top panel, (a1): Lasso estimates; (a2) MCP estimates; (a3) SPIDR estimates; (a4) SPIDR $z$-statistics. Bottom panel, (b1): Lasso and SPIDR overlap; (b2) MCP and SPIDR overlap; (b3) SPIDR estimates with Lasso and MCP selection results indicated; (b4) SPIDR $z$-statistics with Lasso and MCP selection results indicated.

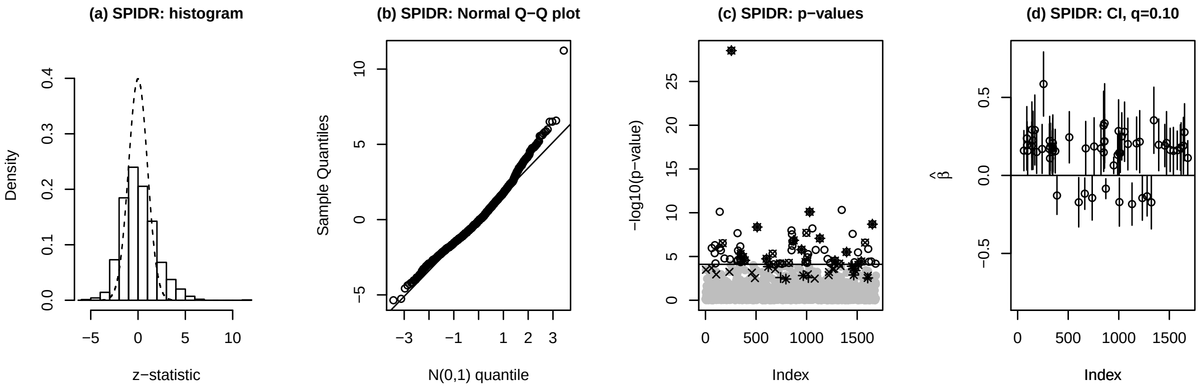

**Figure 6.** Breast cancer data. (a) Histogram of SPIDR $z$-statistics, the dashed curve represents the density function of $N(0,1)$; (b) Normal Q-Q plot for the SPIDR estimates; (c) SPIDR $p$-values; (d) SPIDR confidence intervals for the selected coefficients.

**Table 1**

*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) |