Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

# Forward Regression for Ultra-High Dimensional Variable Screening

Hansheng Wang

**To cite this article:** Hansheng Wang (2009) Forward Regression for Ultra-High Dimensional Variable Screening, Journal of the American Statistical Association, 104:488, 1512-1524, DOI: 10.1198/jasa.2008.tm08516

**To link to this article:** http://dx.doi.org/10.1198/jasa.2008.tm08516

# Forward Regression for Ultra-High Dimensional Variable Screening

Hansheng WANG

Motivated by the seminal theory of Sure Independence Screening (Fan and Lv 2008, SIS), we investigate here another popular and classical variable screening method, namely, forward regression (FR). Our theoretical analysis reveals that FR can identify all relevant predictors consistently, even if the predictor dimension is substantially larger than the sample size. In particular, if the dimension of the true model is finite, FR can discover all relevant predictors within a finite number of steps. To practically select the "best" candidate from the models generated by FR, the recently proposed BIC criterion of Chen and Chen (2008) can be used. The resulting model can then serve as an excellent starting point, from where many existing variable selection methods (e.g., SCAD and Adaptive LASSO) can be applied directly. FR's outstanding finite sample performances are confirmed by extensive numerical studies.

KEY WORDS: Adaptive LASSO; BIC; Forward regression; LASSO; SCAD; Screening consistency; Ultra-High dimensional predictor.



## 1. INTRODUCTION

Modern scientific research very often encounters datasets with tens of thousands of variables. Nevertheless, only very few of them are believed to be truly relevant to the response. Thus, the problem of variable selection with an ultra-high dimensional predictor becomes a problem of fundamental importance (Fan and Li 2006). A lot of research efforts have been devoted to this subject during the past several years (Fan and Peng 2004; Zou and Hastie 2005; Candes and Tao 2007; Fan and Lv 2008; Huang, Horowitz, and Ma 2008; Zou and Zhang 2009).

Obviously, the traditional method of best subset selection is computationally infeasible for high-dimensional data. As a result, various shrinkage methods have gain a lot of popularity during the past decade. Those methods include but are not limited to: the nonnegative garrotte (Breiman 1995; Yuan and Lin 2007), the LASSO (Tibshirani 1996), bridge regression (Fu 1998; Huang, Horowitz, and Ma 2008), the SCAD (Fan and Li 2001; Fan and Peng 2004), the elastic net (Zou and Hastie 2005; Zou and Zhang 2009), the Adaptive LASSO (Zou 2006; Zhang and Lu 2007), the penalized one-step estimator (Zou and Li 2008), and others. All those methods are very useful and can be formulated as penalized optimization problems, which could be selection consistent, if the sample size is much larger than the predictor dimension (Fan and Peng 2004; Huang, Horowitz, and Ma 2008; Zou and Zhang 2009).

However, if the predictor dimension is much larger than the sample size, the story changes drastically. Consider for example those useful methods with nonconvex objective functions (e.g., bridge regression, the SCAD, etc.). With the predictor dimension much larger than the sample size, computationally how to optimize those nonconvex objective functions remains a nontrivial task (Hunter and Li 2005; Chen and Chen 2008). Efficient algorithms (Efron et al. 2004, LARS) do exist for LASSO-type methods (e.g., the LASSO and the elastic net), where the objective functions are strictly convex. However, those methods are not selection consistent under a general design condition (Leng, Lin, and Wahba 2006; Zhao and Yu 2006; Zou 2006; Jia and Yu 2008). Although the Adaptive LASSO is selection consistent under a general design condition (Zou 2006; Huang, Ma, and Zhang 2007; Zhang and Lu 2007; Zou and Li 2008; Zou and Zhang 2009), it requires the sample size to be much larger than the predictor dimension. In fact, the Adaptive LASSO cannot be directly implemented if the predictor dimension is larger than the sample size. This is because the adaptive weights utilized by the Adaptive LASSO rely on a consistent initial estimator for the unknown regression coefficients, which is usually not possible in the high dimensional setup.

Then, how to conduct model selection in an ultra-high dimensional setup? One reasonable solution is variable screening. By doing so, the original ultra-high dimensional problem might be greatly simplified into a low-dimensional one. Thereafter, many standard variable selection methods (e.g., the SCAD, the Adaptive LASSO, etc.) can be implemented easily. To this end, Fan and Lv (2008) developed the important theory of Sure Independence Screening (SIS), which is potentially useful for ultra-high dimensional data analysis.

The outstanding performance of SIS stimulates us to investigate another very popular yet classical variable screening method, that is, forward regression (FR). As one type of important greedy algorithms, FR's theoretical properties have been considered in the past literature; see, for example, Donoho and Stodden (2006) and Barron and Cohen (2008). Nevertheless, to our best knowledge, none of those pioneer research efforts has rigorously established FR's screening consistency, under an ultra-high dimensional setup; see (3.1) in Section 3.2 for a rigorous definition about screening consistency. In other words, we attempt to provide here a unique additional contribution to the existing literature by establishing FR's screening consistency property, under an ultra-high dimensional setup. More specifically, we show that (both theoretically and numerically) FR can identify all relevant predictors consistently, even if the predictor dimension is considerably larger than the sample size. In particular, we find that, if the dimension of the true model is

Hansheng Wang is an Associate Professor of Statistics, from Guanghua School of Management, Peking University, Beijing, 100871, P. R. China (E-mail: *hansheng@gsm.pku.edu.cn*). The author is very grateful to the Editor, the AE, and two anonymous referees for many insightful comments and helpful advices, which lead to a substantially improved manuscript. The author is particularly grateful to the AE for many detailed editorial suggestions, which led to a better presentation of the paper. This research is supported in part by a NSFC grant 10771006 and also a grant from Microsoft Research Asia.


Journal of the American Statistical Association
December 2009, Vol. 104, No. 488, Theory and Methods
DOI: 10.1198/jasa.2008.tm08516

finite, FR might discover all relevant predictors within a finite number of steps. To practically select the "best" candidate from the models generated by FR, we find that the recently proposed BIC criterion of Chen and Chen (2008) can be used. The resulting model can then serve as an excellent starting point, from where many existing variable selection methods can be applied directly. FR's outstanding finite sample performances are confirmed by extensive numerical studies.

The rest of the article is organized as the follows. Next section introduces the model notations and the FR method. The theoretical properties are discussed in Section 3. Numerical studies are reported in Section 4. Lastly, the article is concluded with a short discussion in Section 5. All technical proofs are left to the Appendix.

## 2. FORWARD REGRESSION METHOD

### 2.1 The Model and Notations

Let $(\mathbf{X}_i, Y_i)$ be the observation collected from the $i$th subject $(1 \le i \le n)$, where $Y_i \in \mathbb{R}^1$ is the response and $\mathbf{X}_i = (X_{i1}, \ldots, X_{id})^\top \in \mathbb{R}^d$ is the ultra-high dimensional predictor with $d \gg n$ and $\operatorname{cov}(X) = \mathbf{\Sigma}$. Without loss of generality, it is assumed that $\operatorname{var}(Y_i) = 1$, $E(X_{ij}) = 0$, and $\operatorname{var}(X_{ij}) = 1$. To model the regression relationship, we assume further

$$Y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \varepsilon_i, \tag{2.1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)^\top \in \mathbb{R}^d$ is the regression coefficient and $\varepsilon_i$ is the random noise, with mean 0 and finite variance $\sigma^2$. $X_{ij}$ is referred to as a relevant (irrelevant) predictor, if $\beta_j \neq 0$ $(\beta_j = 0)$. Next, we use a generic notation $\mathcal{M} = \{j_1, \ldots, j_{d^*}\}$ to denote an arbitrary model with $X_{ij_1}, \ldots, X_{ij_{d^*}}$ as relevant predictors. We then define the full model as $\mathcal{F} = \{1, \ldots, d\}$, and the true model as $\mathcal{T} = \{j : \beta_j \neq 0\}$. For convenience, we use $|\mathcal{M}|$ to denote the number of predictors contained in model $\mathcal{M}$ (i.e., the model size). One knows immediately $|\mathcal{F}| = d$ and we assume $|\mathcal{T}| = d_0$. For an arbitrary candidate model $\mathcal{M}$, we use notation $\mathbf{X}_{i(\mathcal{M})} = \{X_{ij} : j \in \mathcal{M}\}$ to denote the subvector of $\mathbf{X}_i$ corresponding to $\mathcal{M}$. Similarly, if we define $\mathfrak{X} = (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n \times d}$ as the design matrix, $\mathfrak{X}_{(\mathcal{M})}$ is then the subdesign matrix corresponding to $\mathcal{M}$. Lastly, let $\mathfrak{Y} = (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ be the response vector.

### 2.2 The FR Algorithm

Under the assumption that $\mathcal{T}$ exists, our main objective is to discover all relevant predictors consistently. To this end, we consider the following FR algorithm.

*Step 1* (Initialization). Set $\mathcal{S}^{(0)} = \emptyset$.

*Step 2* (Forward regression).

(2.1) *Evaluation*. In the $k$th step $(k \ge 1)$, we are given $\mathcal{S}^{(k-1)}$. Then, for every $j \in \mathcal{F} \setminus \mathcal{S}^{(k-1)}$, we construct a candidate model $\mathcal{M}_j^{(k-1)} = \mathcal{S}^{(k-1)} \cup \{j\}$. We then compute $\mathrm{RSS}_j^{(k-1)} = \mathfrak{Y}^\top \{\mathbf{I}_n - \tilde{\mathbf{H}}_j^{(k-1)}\}\mathfrak{Y}$, where

$$\tilde{\mathbf{H}}_j^{(k-1)} = \mathfrak{X}_{(\mathcal{M}_j^{(k-1)})} \big\{\mathfrak{X}_{(\mathcal{M}_j^{(k-1)})}^\top \mathfrak{X}_{(\mathcal{M}_j^{(k-1)})}\big\}^{-1} \times \mathfrak{X}_{(\mathcal{M}_j^{(k-1)})}^\top$$

is a projection matrix and $\mathbf{I}_n \in \mathbb{R}^{n \times n}$ is the identity matrix.

(2.2) *Screening*. We then find

$$a_k = \arg \min_{j \in \mathcal{F} \setminus \mathcal{S}^{(k-1)}} \mathrm{RSS}_j^{(k-1)}$$

and update $\mathcal{S}^{(k)} = \mathcal{S}^{(k-1)} \cup \{a_k\}$ accordingly.

*Step 3* (Solution path). Iterating Step 2 for $n$ times, which leads a total of $n$ nested candidate models. We then collect those models by a solution path $\mathbb{S} = \{\mathcal{S}^{(k)} : 1 \le k \le n\}$ with $\mathcal{S}^{(k)} = \{a_1, \ldots, a_k\}$.

## 3. THEORETICAL PROPERTIES

### 3.1 The Technical Conditions

To gain theoretical insights about the FR algorithm, the following standard technical conditions are needed.

(C1) *Normality assumption*. Assume that both $X$ and $\varepsilon$ follow normal distributions.

(C2) *Covariance matrix*. $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ represent, respectively, the smallest and largest eigenvalues of an arbitrary positive definite matrix $A$. We assume that there exist two positive constant $0 < \tau_{\min} < \tau_{\max} < \infty$, such that $2\tau_{\min} < \lambda_{\min}(\mathbf{\Sigma}) \le \lambda_{\max}(\mathbf{\Sigma}) < 2^{-1}\tau_{\max}$.

(C3) *Regression coefficients*. We assume that $\|\boldsymbol{\beta}\| \le C_{\boldsymbol{\beta}}$ for some constant $C_{\boldsymbol{\beta}} > 0$ and $\beta_{\min} \ge \nu_{\boldsymbol{\beta}} n^{-\xi_{\min}}$ for some $\xi_{\min} > 0$, where $\|\cdot\|$ denotes the standard $L_2$ norm and $\beta_{\min} = \min_{j \in \mathcal{T}} |\beta_j|$.

(C4) *Divergence speed of $d$ and $d_0$*. There exists constants $\xi$, $\xi_0$, and $\nu$, such that $\log d \le \nu n^{\xi}$, $d_0 \le \nu n^{\xi_0}$, and $\xi + 6\xi_0 + 12\xi_{\min} < 1$.

Note that the normality assumption (C1) has been popularly assumed in the past literature to facilitate theory development (Fan and Lv 2008; Zhang and Huang 2008). Condition (C2) together with condition (C1) assures the Sparse Riesz Condition (SRC), as defined by Zhang and Huang (2008); see Lemma 1 in Appendix A for a more detailed discussion. One can verify that model (2.1)'s the signal-to-noise ratio $\sigma^{-2}\operatorname{var}(\mathbf{X}_i^\top \boldsymbol{\beta})$ might diverge to infinity if $\|\boldsymbol{\beta}\| \to \infty$, which seems unlikely to happen in real practice. Thus, condition (C3) requires that $\|\boldsymbol{\beta}\| \le C_{\boldsymbol{\beta}}$ for some constant $C_{\boldsymbol{\beta}}$. Furthermore, conditions (C3) and (C4) together also put a constraint on the minimal size of the nonzero regression coefficient. Intuitively, if some of the nonzero coefficients converge to zero too fast, they cannot be identified consistently; see Fan and Peng (2004). Lastly, by condition (C4), we allow the predictor dimension $d$ to diverge to infinity at an exponentially fast speed, which implies that the predictor dimension can be substantially larger than the sample size $n$ (Fan and Lv 2008).

### 3.2 The Screening Consistency

Note that, it is unrealistic to require $\mathcal{T} \in \mathbb{S}$, because this is not guaranteed even in the fixed dimension situation. However, it is indeed possible to have $\mathcal{T} \subset \mathcal{S}^{(k)}$ for some $1 \le k \le n$ (Fan and Lv 2008). Otherwise, there exists at least one relevant predictor completely missed by the solution path $\mathbb{S}$. To facilitate an easy discussion, we define the solution path $\mathbb{S}$ to be screening consistent, if

$$P\big(\mathcal{T} \subset \mathcal{S}^{(k)} \in \mathbb{S} \text{ for some } 1 \le k \le n\big) \to 1. \tag{3.1}$$

To investigate (3.1), we define $K = 2\tau_{\max}\nu C_{\boldsymbol{\beta}}^2\tau_{\min}^{-2}\nu_{\boldsymbol{\beta}}^{-4}$, which is a constant independent of $n$. Let $[t]$ be the smallest integer no less than $t$. Then, FR's screening consistency can be formally established by the following theorem.

*Theorem 1.* Under model (2.1) and conditions (C1)–(C4), we have as $n \to \infty$

$$P\big(\mathcal{T} \subset \mathcal{S}^{([K\nu n^{2\xi_0+4\xi_{\min}}])}\big) \to 1.$$

By Theorem 1 we know that, with probability tending to one, the FR algorithm can detect all relevant predictors within $O(n^{2\xi_0+4\xi_{\min}})$ steps, which is a number much smaller than the sample size $n$ under the condition (C4). In particular, if the dimension of the true model is finite with $\xi_0 = \xi_{\min} = 0$, only a finite number of steps are needed to discover the entire relevant variable set.

### 3.3 A BIC Criterion

Let $\mathcal{M}$ be an arbitrary candidate model with $|\mathcal{M}| \le n$. Define $\hat{\sigma}^2_{(\mathcal{M})} = n^{-1}\,\mathrm{RSS}(\mathcal{M})$, where $\mathrm{RSS}(\mathcal{M}) = \mathfrak{Y}^\top\{\mathbf{I}_n - \mathbf{H}_{(\mathcal{M})}\}\mathfrak{Y}$ and $\mathbf{H}_{(\mathcal{M})} = \mathfrak{X}_{(\mathcal{M})}\{\mathfrak{X}_{(\mathcal{M})}^\top\mathfrak{X}_{(\mathcal{M})}\}^{-1}\mathfrak{X}_{(\mathcal{M})}^\top$. As one can see, the solution path $\mathbb{S}$ contains a total of $n$ models. For a real application, one has to decide which model should be used for further variable selection. To this end, we consider the following BIC criterion (Chen and Chen 2008),

$$\mathrm{BIC}(\mathcal{M}) = \log\hat{\sigma}^2_{(\mathcal{M})} + n^{-1}|\mathcal{M}|(\log n + 2\log d), \qquad (3.2)$$

which was recently proposed by Chen and Chen (2008) with sound Bayesian motivation. Let $\hat{m} = \arg\min_{1\le m\le n}\mathrm{BIC}(\mathcal{S}^{(m)})$ and $\widehat{\mathcal{S}} = \mathcal{S}^{(\hat{m})}$. As theoretically proved by Chen and Chen (2008), the BIC criterion (3.2) is selection consistent if $d = O(n^\alpha)$ for some $\alpha > 0$; see theorem 1 in Chen and Chen (2008). However, whether it is still selection consistent in a ultra-high dimensional setup with $\log d = O(n^{\xi_0})$ is unknown. Furthermore, as we mentioned before, it is even possible that $\mathcal{T} \notin \mathbb{S}$. Thus, it is unrealistic to require $\widehat{\mathcal{S}}$ to be selection consistent, that is, $P(\widehat{\mathcal{S}} = \mathcal{T}) \to 1$. However, the following theorem says that $\widehat{\mathcal{S}}$ is indeed screening consistent.

*Theorem 2.* Under model (2.1) and conditions (C1)–(C4), then as $n \to \infty$

$$P(\mathcal{T} \subset \widehat{\mathcal{S}}) \to 1.$$

## 4. NUMERICAL STUDIES

### 4.1 Preliminaries

Extensive simulation studies have been conducted to evaluate the finite sample performance of FR under various stopping rules. For comparison purpose, the performances of SIS, Iterative SIS (Fan and Lv 2008, ISIS), and LARS (Efron et al. 2004) were also examined. Thus, we have a total of four variable screening methods (i.e., SIS, ISIS, LARS, and FR). For a given screening method and its screened model, we use NONE to represent the model without using further variable selection. In contrast, we use the notations ALASSO and SCAD to represent, respectively, the models further selected by Adaptive LASSO and SCAD with the BIC criterion (3.2). The size of the SIS model is fixed to be $[n/\log n]$; see Fan and Lv (2008). For the ISIS method, a total of $[\log n - 1]$ ISIS steps are conducted and $[n/\log n]$ variables are screened in each iterative step. The number of relevant variable kept in each step is determined by NONE, ALASSO, or SCAD, in together with the BIC criterion (3.2). Lastly, the size of both FR-NONE and LARS-NONE models are also determined by (3.2).

For each parameter setup, a total of 200 simulation replications are conducted. The Theoretical $R^2 \doteq \mathrm{var}(\mathbf{X}_i^\top\boldsymbol{\beta})/\mathrm{var}(Y_i)$ is given by 30%, 60%, or 90%. Let $\hat{\boldsymbol{\beta}}_{(k)} = (\hat{\beta}_{1(k)}, \ldots, \hat{\beta}_{d(k)})^\top \in \mathbb{R}^d$ be the estimator realized in the $k$th simulation replication, by one particular method (e.g., FR-NONE). Then, the model selected by $\hat{\boldsymbol{\beta}}_{(k)}$ is given by $\widehat{\mathcal{S}}_{(k)} = \{j : |\hat{\beta}_{j(k)}| > 0\}$ and the corresponding Average Model Size $= 200^{-1}\sum_k|\widehat{\mathcal{S}}_{(k)}|$. Recall $\mathcal{T}$ represents the true model, we evaluate the Coverage Probability $= 200^{-1}\sum_k I(\widehat{\mathcal{S}}_{(k)} \supset \mathcal{T})$, which measures how likely all relevant variables will be discovered by one particular screening method. To characterizes the method's capability in producing sparse solutions, we define

$$\text{Percentage of Correct Zeros (\%)} = \frac{100\%}{d - d_0}\left(\frac{1}{200}\sum_{k=1}^{200}\sum_{j=1}^{d} I\big(\hat{\beta}_{j(k)} = 0\big) \times I(\beta_j = 0)\right).$$

To characterizes the method's underfitting effect, we further define

$$\text{Percentage of Incorrect Zeros (\%)} = \frac{100\%}{d_0}\left(\frac{1}{200}\sum_{k=1}^{200}\sum_{j=1}^{d} I\big(\hat{\beta}_{j(k)} = 0\big) \times I(\beta_j \ne 0)\right).$$

If all sparse solutions are correctly identified for all irrelevant predictors and no sparse solution is mistakenly produced for all relevant variables, the true model is perfectly identified, that is, $\widehat{\mathcal{S}}_{(k)} = \mathcal{T}$. To gauge such a performance, we define the Percentage of Correctly Fitted (%) $= 200^{-1}\sum_k I(\widehat{\mathcal{S}}_{(k)} = \mathcal{T})$. Lastly, to evaluate the usefulness of $\hat{\boldsymbol{\beta}}_{(k)}$ for out-of-sample forecasting, we define

$$\text{Out-of-Sample } R^2 = 100\% \times \left\{1 - \frac{\sum(Y_i^* - \mathbf{X}_i^{*\top}\hat{\boldsymbol{\beta}}_{(k)})^2}{\sum(Y_i^* - \bar{Y}^*)^2}\right\},$$

where $\bar{Y}^* = n^{-1}\sum Y_i^*$ and $(\mathbf{X}_i^*, Y_i^*)$ are the testing data, which are independent of the training data $(\mathbf{X}_i, Y_i)$ but has the same sample size.

### 4.2 The Simulation Models

For a reliable numerical comparison, we considered the following six simulation models with the detailed results summarized in Tables 1–6.

*Example 1* (Independent predictors). This is an example borrowed from Fan and Lv (2008) with $(n, d, d_0) = (200, 10{,}000, 8)$. $\mathbf{X}_i$ is generated independently according to a standard multivariate normal distribution. Thus, different predictors are mutually independent. The $j$th $(1 \le j \le d_0)$ nonzero coefficient of $\boldsymbol{\beta}$ is given by $\beta_j = (-1)^{U_j}(4\log n/\sqrt{n} + |Z_j|)$, where $U_j$ is a binary random variable with $P(U_j = 1) = 0.4$ and $Z_j$ is a standard normal random variable.

Table 1. Example 1 with $(n, d, d_0) = (200, 10{,}000, 8)$

| Screening method | Selection method | Coverage probability (%) | Percentage of correct zeros (%) | Percentage of incorrect zeros (%) | Percentage of correctly fitted (%) | Average model size | Out of sample $R^2$ (%) Mean | Out of sample $R^2$ (%) SE |
|---|---|---|---|---|---|---|---|---|
| | | | Theoretical $R^2 = 30\%$ | | | | | |
| SIS | NONE | 40.6 | 99.7 | 59.4 | 0.0 | 38.0 | −16.4 | 0.69 |
| | ALASSO | 3.1 | 100.0 | 96.9 | 0.0 | 0.3 | 1.8 | 0.32 |
| | SCAD | 3.4 | 100.0 | 96.6 | 0.0 | 0.3 | 2.0 | 0.35 |
| ISIS | NONE | 44.1 | 98.1 | 55.9 | 0.0 | 190.0 | −40.1 | 1.27 |
| | ALASSO | 3.1 | 100.0 | 96.9 | 0.0 | 0.3 | 1.8 | 0.32 |
| | SCAD | 3.4 | 100.0 | 96.6 | 0.0 | 0.3 | 2.0 | 0.35 |
| LARS | NONE | 0.1 | 100.0 | 99.9 | 0.0 | 0.0 | 0.2 | 0.07 |
| | ALASSO | 0.1 | 100.0 | 99.9 | 0.0 | 0.0 | 0.2 | 0.07 |
| | SCAD | 0.1 | 100.0 | 99.9 | 0.0 | 0.0 | 0.2 | 0.07 |
| FR | NONE | 3.4 | 100.0 | 96.6 | 0.0 | 0.3 | 1.9 | 0.35 |
| | ALASSO | 3.4 | 100.0 | 96.6 | 0.0 | 0.3 | 1.9 | 0.35 |
| | SCAD | 3.4 | 100.0 | 96.6 | 0.0 | 0.3 | 1.9 | 0.35 |
| | | | Theoretical $R^2 = 60\%$ | | | | | |
| SIS | NONE | 71.0 | 99.7 | 29.0 | 0.0 | 38.0 | 21.8 | 0.72 |
| | ALASSO | 42.6 | 100.0 | 57.4 | 1.0 | 3.4 | 33.9 | 0.96 |
| | SCAD | 37.4 | 100.0 | 62.6 | 0.0 | 3.0 | 30.7 | 0.90 |
| ISIS | NONE | 81.3 | 98.2 | 18.7 | 0.0 | 190.0 | 4.9 | 1.00 |
| | ALASSO | 43.3 | 100.0 | 56.7 | 1.5 | 3.5 | 34.3 | 0.98 |
| | SCAD | 38.3 | 100.0 | 61.7 | 0.5 | 3.1 | 31.1 | 0.93 |
| LARS | NONE | 3.6 | 100.0 | 96.4 | 0.0 | 0.3 | 4.6 | 0.76 |
| | ALASSO | 3.5 | 100.0 | 96.5 | 0.0 | 0.3 | 4.6 | 0.75 |
| | SCAD | 3.5 | 100.0 | 96.5 | 0.0 | 0.3 | 4.6 | 0.75 |
| FR | NONE | 47.9 | 100.0 | 52.1 | 2.5 | 3.9 | 36.9 | 0.90 |
| | ALASSO | 47.9 | 100.0 | 52.1 | 2.5 | 3.9 | 36.9 | 0.90 |
| | SCAD | 47.9 | 100.0 | 52.1 | 2.5 | 3.9 | 36.9 | 0.90 |
| | | | Theoretical $R^2 = 90\%$ | | | | | |
| SIS | NONE | 85.1 | 99.7 | 14.9 | 0.0 | 38.0 | 73.3 | 0.74 |
| | ALASSO | 84.9 | 100.0 | 15.1 | 24.5 | 6.8 | 82.4 | 0.44 |
| | SCAD | 84.5 | 100.0 | 15.5 | 23.0 | 6.9 | 82.1 | 0.46 |
| ISIS | NONE | 99.8 | 98.2 | 0.2 | 0.0 | 190.0 | 69.9 | 0.66 |
| | ALASSO | 99.6 | 100.0 | 0.4 | 95.5 | 8.0 | 89.5 | 0.12 |
| | SCAD | 99.4 | 100.0 | 0.6 | 84.5 | 8.1 | 89.2 | 0.14 |
| LARS | NONE | 95.5 | 100.0 | 4.5 | 37.0 | 8.5 | 86.5 | 0.87 |
| | ALASSO | 95.5 | 100.0 | 4.5 | 85.5 | 7.6 | 86.9 | 0.88 |
| | SCAD | 95.4 | 100.0 | 4.6 | 76.0 | 7.8 | 86.8 | 0.88 |
| FR | NONE | 99.9 | 100.0 | 0.1 | 97.5 | 8.0 | 89.6 | 0.10 |
| | ALASSO | 99.9 | 100.0 | 0.1 | 97.5 | 8.0 | 89.6 | 0.10 |
| | SCAD | 99.9 | 100.0 | 0.1 | 97.5 | 8.0 | 89.6 | 0.10 |

*Example 2* (Autoregressive correlation). We consider in this example an autoregressive type correlation structure. Such type of correlation structure might be useful if a natural order exists among the predictors. As a consequence, the predictors with large distances in order are expected to be mutually independent approximately. To this end, we revise an example from Tibshirani (1996) with $(n, d, d_0) = (200, 40{,}000, 3)$. Specifically, $\mathbf{X}_i$ is generated from a multivariate normal distribution with mean 0 and $\text{cov}(X_{ij_1}, X_{ij_2}) = 0.5^{|j_1 - j_2|}$. The 1st, 4th, and 7th components of $\boldsymbol{\beta}$ are given by 3, 1.5, and 2, respectively. Other components of $\boldsymbol{\beta}$ are fixed to be 0.

*Example 3* (Compound symmetry). Compound symmetry is another very useful correlation structure. By compound symmetry, all predictors are equally correlated with each other. To compare FR and SIS under compound symmetric correlation structure, we borrow the following example from Fan and Lv (2008) with $(n, d, d_0) = (75, 5000, 3)$. Specifically, $\mathbf{X}_i$ is generated such that $\text{var}(X_{ij}) = 1$ and $\text{var}(X_{ij_1}, X_{ij_2}) = 0.5$ for any $j_1 \neq j_2$. The nonzero coefficients of $\boldsymbol{\beta}$ are fixed to be 5.

*Example 4* (A challenging case). To further test the finite sample performance of FR, we create here a rather challenging example. Specifically, we fix $(n, d, d_0) = (300, 10{,}000, 5)$ and $\beta_j = 2j$ for every $1 \leq j \leq d_0$. We simulate independently

Table 2. Example 2 with $(n, d, d_0) = (200, 40{,}000, 3)$

| Screening method | Selection method | Coverage probability (%) | Percentage of correct zeros (%) | Percentage of incorrect zeros (%) | Percentage of correctly fitted (%) | Average model size | Out of sample $R^2$ (%) Mean | Out of sample $R^2$ (%) SE |
|---|---|---|---|---|---|---|---|---|
| | | | Theoretical $R^2 = 30\%$ | | | | | |
| SIS | NONE | 87.8 | 99.9 | 12.2 | 0.0 | 38.0 | −7.1 | 0.70 |
| | ALASSO | 32.7 | 100.0 | 67.3 | 0.0 | 1.0 | 14.9 | 0.66 |
| | SCAD | 38.2 | 100.0 | 61.8 | 0.5 | 1.1 | 17.1 | 0.58 |
| ISIS | NONE | 88.2 | 99.5 | 11.8 | 0.0 | 190.0 | −23.2 | 0.97 |
| | ALASSO | 32.7 | 100.0 | 67.3 | 0.0 | 1.0 | 14.9 | 0.66 |
| | SCAD | 38.2 | 100.0 | 61.8 | 0.5 | 1.1 | 17.1 | 0.58 |
| LARS | NONE | 7.8 | 100.0 | 92.2 | 0.0 | 0.2 | 3.7 | 0.51 |
| | ALASSO | 7.8 | 100.0 | 92.2 | 0.0 | 0.2 | 3.7 | 0.51 |
| | SCAD | 7.8 | 100.0 | 92.2 | 0.0 | 0.2 | 3.7 | 0.51 |
| FR | NONE | 40.2 | 100.0 | 59.8 | 0.0 | 1.2 | 17.7 | 0.61 |
| | ALASSO | 40.2 | 100.0 | 59.8 | 0.0 | 1.2 | 17.7 | 0.61 |
| | SCAD | 40.2 | 100.0 | 59.8 | 0.0 | 1.2 | 17.7 | 0.61 |
| | | | Theoretical $R^2 = 60\%$ | | | | | |
| SIS | NONE | 99.8 | 99.9 | 0.2 | 0.0 | 38.0 | 31.9 | 0.49 |
| | ALASSO | 93.2 | 100.0 | 6.8 | 77.0 | 2.8 | 57.1 | 0.39 |
| | SCAD | 92.7 | 100.0 | 7.3 | 78.0 | 2.8 | 57.0 | 0.42 |
| ISIS | NONE | 99.8 | 99.5 | 0.2 | 0.0 | 190.0 | 19.5 | 0.69 |
| | ALASSO | 93.2 | 100.0 | 6.8 | 77.0 | 2.8 | 57.1 | 0.39 |
| | SCAD | 92.8 | 100.0 | 7.2 | 78.5 | 2.8 | 57.1 | 0.42 |
| LARS | NONE | 92.5 | 100.0 | 7.5 | 76.5 | 2.9 | 56.3 | 0.57 |
| | ALASSO | 89.7 | 100.0 | 10.3 | 75.0 | 2.7 | 55.9 | 0.58 |
| | SCAD | 90.0 | 100.0 | 10.0 | 76.0 | 2.7 | 56.0 | 0.58 |
| FR | NONE | 94.2 | 100.0 | 5.8 | 82.5 | 2.8 | 57.5 | 0.39 |
| | ALASSO | 94.2 | 100.0 | 5.8 | 82.5 | 2.8 | 57.5 | 0.39 |
| | SCAD | 94.2 | 100.0 | 5.8 | 82.5 | 2.8 | 57.5 | 0.39 |
| | | | Theoretical $R^2 = 90\%$ | | | | | |
| SIS | NONE | 100.0 | 99.9 | 0.0 | 0.0 | 38.0 | 84.1 | 0.15 |
| | ALASSO | 100.0 | 100.0 | 0.0 | 100.0 | 3.0 | 89.6 | 0.09 |
| | SCAD | 100.0 | 100.0 | 0.0 | 100.0 | 3.0 | 89.6 | 0.09 |
| ISIS | NONE | 100.0 | 99.5 | 0.0 | 0.0 | 190.0 | 79.9 | 0.18 |
| | ALASSO | 100.0 | 100.0 | 0.0 | 100.0 | 3.0 | 89.6 | 0.09 |
| | SCAD | 100.0 | 100.0 | 0.0 | 100.0 | 3.0 | 89.6 | 0.09 |
| LARS | NONE | 100.0 | 100.0 | 0.0 | 89.5 | 3.1 | 89.6 | 0.09 |
| | ALASSO | 100.0 | 100.0 | 0.0 | 100.0 | 3.0 | 89.6 | 0.09 |
| | SCAD | 100.0 | 100.0 | 0.0 | 100.0 | 3.0 | 89.6 | 0.09 |
| FR | NONE | 100.0 | 100.0 | 0.0 | 100.0 | 3.0 | 89.6 | 0.09 |
| | ALASSO | 100.0 | 100.0 | 0.0 | 100.0 | 3.0 | 89.6 | 0.09 |
| | SCAD | 100.0 | 100.0 | 0.0 | 100.0 | 3.0 | 89.6 | 0.09 |



$Z_i = (Z_{ij}) \in \mathbb{R}^d$ and $W_i = (W_{ij}) \in \mathbb{R}^d$ from a standard multivariate normal distribution. Next, we generate $\mathbf{X}_i$ according to $X_{ij} = (Z_{ij} + W_{ij})/\sqrt{2}$ for every $1 \le j \le d_0$ and $X_{ij} = (Z_{ij} + \sum_{j'=1}^{d_0} Z_{ij'})/2$ for every $d_0 < j \le d$. A simple Monte Carlo computation reveals that the correlation coefficient of $X_{i1}$ and $Y_i$ (recall $X_{i1}$ is the first relevant predictor) is much smaller than that of $X_{ij}$ and $Y_i$ for every $j > d_0$ (recall $X_{ij}$ is an irrelevant predictor for every $j > d_0$). Consequently, it would be extremely difficult to discover (for example) $X_{i1}$ as a relevant predictor.

*Example 5* (Normality assumption). Note that our theory replies on the assumption that $\varepsilon_i$ is normally distributed; see condition (C1) in Section 3.1. Thus, it is of interest to test FR's finite performance against nonnormally distributed $\varepsilon_i$. Thus, Example 1 is replicated but with both $X_{ij}$ and $\varepsilon_i$ generated independently from a standardized exponential distribution, that is, $\exp(1) - 1$.

*Example 6* (Diverging $d_0$). Note that the true model size $d_0$ is fixed for the previous five examples. To further demonstrate the competitive performance of FR, we consider the situation, where $d_0$ is also diverging (Zou and Zhang 2009). Specifically, we replicate Example 1 with the Theoretical $R^2 = 75\%$ and $(d, d_0) = (10{,}000, [\sqrt{n}])$, where the sample size $n$ is fixed to be 200, 400, and 800.

Table 3. Example 3 with $(n, d, d_0) = (75, 5000, 3)$

| Screening method | Selection method | Coverage probability (%) | Percentage of correct zeros (%) | Percentage of incorrect zeros (%) | Percentage of correctly fitted (%) | Average model size | Out of sample $R^2$ (%) Mean | Out of sample $R^2$ (%) SE |
|---|---|---|---|---|---|---|---|---|
| | | | Theoretical $R^2 = 30\%$ | | | | | |
| SIS | NONE | 12.7 | 99.6 | 87.3 | 0.0 | 18.0 | 5.7 | 1.28 |
| | ALASSO | 2.5 | 100.0 | 97.5 | 0.0 | 0.9 | 4.7 | 0.87 |
| | SCAD | 1.5 | 100.0 | 98.5 | 0.0 | 0.7 | 3.3 | 0.75 |
| ISIS | NONE | 15.8 | 98.6 | 84.2 | 0.0 | 72.0 | −147.1 | 13.42 |
| | ALASSO | 2.7 | 100.0 | 97.3 | 0.0 | 1.0 | 4.8 | 0.88 |
| | SCAD | 1.5 | 100.0 | 98.5 | 0.0 | 0.8 | 4.7 | 0.80 |
| LARS | NONE | 0.0 | 100.0 | 100.0 | 0.0 | 0.0 | −0.1 | 0.16 |
| | ALASSO | 0.0 | 100.0 | 100.0 | 0.0 | 0.0 | −0.1 | 0.16 |
| | SCAD | 0.0 | 100.0 | 100.0 | 0.0 | 0.0 | −0.1 | 0.16 |
| FR | NONE | 2.2 | 100.0 | 97.8 | 0.0 | 1.0 | 3.0 | 0.95 |
| | ALASSO | 2.2 | 100.0 | 97.8 | 0.0 | 1.0 | 3.0 | 0.95 |
| | SCAD | 2.2 | 100.0 | 97.8 | 0.0 | 1.0 | 3.0 | 0.95 |
| | | | Theoretical $R^2 = 60\%$ | | | | | |
| SIS | NONE | 41.3 | 99.7 | 58.7 | 0.0 | 18.0 | 38.1 | 0.86 |
| | ALASSO | 16.0 | 100.0 | 84.0 | 0.0 | 1.5 | 28.2 | 1.08 |
| | SCAD | 9.0 | 100.0 | 91.0 | 0.0 | 1.4 | 23.7 | 1.08 |
| ISIS | NONE | 51.3 | 98.6 | 48.7 | 0.0 | 72.0 | −120.0 | 50.03 |
| | ALASSO | 16.0 | 100.0 | 84.0 | 0.0 | 1.6 | 28.4 | 1.05 |
| | SCAD | 9.3 | 100.0 | 90.7 | 0.0 | 1.6 | 27.5 | 0.90 |
| LARS | NONE | 0.8 | 100.0 | 99.2 | 0.0 | 0.0 | 0.6 | 0.39 |
| | ALASSO | 0.8 | 100.0 | 99.2 | 0.0 | 0.0 | 0.6 | 0.39 |
| | SCAD | 0.8 | 100.0 | 99.2 | 0.0 | 0.0 | 0.6 | 0.39 |
| FR | NONE | 13.5 | 100.0 | 86.5 | 0.5 | 1.4 | 25.9 | 1.13 |
| | ALASSO | 13.5 | 100.0 | 86.5 | 0.5 | 1.4 | 25.9 | 1.13 |
| | SCAD | 13.5 | 100.0 | 86.5 | 0.5 | 1.4 | 25.9 | 1.13 |
| | | | Theoretical $R^2 = 90\%$ | | | | | |
| SIS | NONE | 82.7 | 99.7 | 17.3 | 0.0 | 18.0 | 80.3 | 0.51 |
| | ALASSO | 80.8 | 100.0 | 19.2 | 53.0 | 2.7 | 82.6 | 0.71 |
| | SCAD | 73.3 | 100.0 | 26.7 | 46.5 | 2.9 | 78.1 | 1.18 |
| ISIS | NONE | 98.2 | 98.6 | 1.8 | 0.0 | 72.0 | 53.0 | 2.41 |
| | ALASSO | 89.5 | 100.0 | 10.5 | 73.0 | 3.0 | 85.3 | 0.68 |
| | SCAD | 77.3 | 100.0 | 22.7 | 52.0 | 3.1 | 79.9 | 1.10 |
| LARS | NONE | 27.2 | 100.0 | 72.8 | 9.5 | 1.0 | 25.3 | 2.78 |
| | ALASSO | 27.2 | 100.0 | 72.8 | 24.5 | 0.8 | 25.5 | 2.80 |
| | SCAD | 27.2 | 100.0 | 72.8 | 24.5 | 0.8 | 25.5 | 2.80 |
| FR | NONE | 94.3 | 100.0 | 5.7 | 76.0 | 3.2 | 86.9 | 0.65 |
| | ALASSO | 94.2 | 100.0 | 5.8 | 89.5 | 3.0 | 87.1 | 0.67 |
| | SCAD | 94.2 | 100.0 | 5.8 | 89.0 | 3.0 | 87.1 | 0.65 |



### 4.3 Simulation Results

Based on the simulation results as summarized in Tables 1–6, we are able to draw the following conclusions.

Firstly, in the low signal-to-noise ratio situation (i.e., Theoretical $R^2 = 30\%$), no method performs well in terms of the prediction accuracy. The resulting out-of-sample $R^2$ is always substantially smaller than its optimal value, that is, 30%. In many situations (e.g., Examples 1, 3, and 5), this number is even less than 5%. However, comparatively speaking, the three FR methods' out-of-sample $R^2$ are the best. It is noteworthy that although SIS-NONE and/or ISIS-NONE always demonstrate the best coverage probabilities, those outstanding performances are achieved at the costs of very large average model size. As a consequence, their out-of-sample forecasting accuracies are poor, with even negative out-of-sample $R^2$ in many situations.

Secondly, for all the simulation models we considered, the coverage probability of ISIS-NONE is always larger than that of SIS-NONE, which corroborates the numerical findings of Fan and Lv (2008). Some times, the difference might as huge as $99.8 - 15.5 = 84.3\%$; see, for example, Example 4 with Theoretical $R^2 = 90\%$. However, compared with SIS-NONE, the better coverage probability of ISIS-NONE is obtained by sacrificing a much larger average model size. As

Table 4. Example 4 with $(n, d, d_0) = (300, 10{,}000, 5)$

| Screening method | Selection method | Coverage probability (%) | Percentage of correct zeros (%) | Percentage of incorrect zeros (%) | Percentage of correctly fitted (%) | Average model size | Out of sample $R^2$ (%) Mean | Out of sample $R^2$ (%) SE |
|---|---|---|---|---|---|---|---|---|
| | | | Theoretical $R^2 = 30\%$ | | | | | |
| SIS | NONE | 10.6 | 99.5 | 89.4 | 0.0 | 53.0 | 11.5 | 0.44 |
| | ALASSO | 10.6 | 100.0 | 89.4 | 0.0 | 1.4 | 12.9 | 0.45 |
| | SCAD | 6.1 | 100.0 | 93.9 | 0.0 | 1.1 | 9.0 | 0.55 |
| ISIS | NONE | 78.5 | 97.4 | 21.5 | 0.0 | 265.0 | −73.8 | 2.31 |
| | ALASSO | 14.9 | 100.0 | 85.1 | 0.0 | 1.6 | 14.4 | 0.49 |
| | SCAD | 8.4 | 100.0 | 91.6 | 0.0 | 1.4 | 10.2 | 0.57 |
| LARS | NONE | 2.0 | 100.0 | 98.0 | 0.0 | 0.2 | 1.4 | 0.35 |
| | ALASSO | 1.9 | 100.0 | 98.1 | 0.0 | 0.1 | 1.3 | 0.32 |
| | SCAD | 2.0 | 100.0 | 98.0 | 0.0 | 0.2 | 1.4 | 0.35 |
| FR | NONE | 10.9 | 100.0 | 89.1 | 0.0 | 1.3 | 13.1 | 0.49 |
| | ALASSO | 10.9 | 100.0 | 89.1 | 0.0 | 1.3 | 13.1 | 0.49 |
| | SCAD | 10.9 | 100.0 | 89.1 | 0.0 | 1.3 | 13.1 | 0.49 |
| | | | Theoretical $R^2 = 60\%$ | | | | | |
| SIS | NONE | 13.4 | 99.5 | 86.6 | 0.0 | 53.0 | 29.2 | 0.62 |
| | ALASSO | 13.4 | 100.0 | 86.6 | 0.0 | 1.7 | 30.7 | 0.72 |
| | SCAD | 13.3 | 100.0 | 86.7 | 0.0 | 2.9 | 29.9 | 0.90 |
| ISIS | NONE | 93.0 | 97.4 | 7.0 | 0.0 | 265.0 | −5.9 | 1.24 |
| | ALASSO | 52.1 | 100.0 | 47.9 | 0.0 | 4.3 | 49.6 | 0.46 |
| | SCAD | 48.2 | 100.0 | 51.8 | 0.0 | 5.5 | 48.0 | 0.50 |
| LARS | NONE | 13.8 | 100.0 | 86.2 | 0.0 | 1.6 | 14.8 | 1.58 |
| | ALASSO | 13.7 | 100.0 | 86.3 | 0.0 | 0.8 | 14.6 | 1.55 |
| | SCAD | 13.8 | 100.0 | 86.2 | 0.0 | 1.0 | 14.7 | 1.57 |
| FR | NONE | 54.4 | 100.0 | 45.6 | 0.0 | 3.5 | 51.7 | 0.54 |
| | ALASSO | 54.4 | 100.0 | 45.6 | 0.0 | 3.3 | 51.5 | 0.54 |
| | SCAD | 54.4 | 100.0 | 45.6 | 0.0 | 3.5 | 51.7 | 0.54 |
| | | | Theoretical $R^2 = 90\%$ | | | | | |
| SIS | NONE | 15.5 | 99.5 | 84.5 | 0.0 | 53.0 | 48.9 | 0.82 |
| | ALASSO | 15.5 | 100.0 | 84.5 | 0.0 | 1.9 | 49.3 | 0.94 |
| | SCAD | 15.4 | 100.0 | 84.6 | 0.0 | 4.1 | 51.1 | 0.91 |
| ISIS | NONE | 99.8 | 97.4 | 0.2 | 0.0 | 265.0 | 65.5 | 0.38 |
| | ALASSO | 79.8 | 100.0 | 20.2 | 0.0 | 6.5 | 87.9 | 0.13 |
| | SCAD | 64.9 | 99.9 | 35.1 | 0.0 | 8.5 | 82.2 | 0.57 |
| LARS | NONE | 41.2 | 99.9 | 58.8 | 0.0 | 8.3 | 53.5 | 2.69 |
| | ALASSO | 41.2 | 100.0 | 58.8 | 3.0 | 2.4 | 53.4 | 2.69 |
| | SCAD | 41.2 | 100.0 | 58.8 | 3.0 | 2.6 | 53.4 | 2.69 |
| FR | NONE | 94.2 | 100.0 | 5.8 | 48.5 | 5.2 | 89.4 | 0.09 |
| | ALASSO | 94.2 | 100.0 | 5.8 | 70.5 | 4.8 | 89.4 | 0.09 |
| | SCAD | 94.2 | 100.0 | 5.8 | 70.5 | 4.8 | 89.4 | 0.09 |



a consequence, ISIS-NONE seldom yields better prediction accuracy than SIS-NONE, in terms of the out-of-sample $R^2$. In fact, in most situations, the out-of-sample $R^2$ ISIS-NONE is considerably smaller than that of SIS-NONE. As a consequence, no clear improvement was demonstrated by ISIS-SCAD, as compared with SIS-SCAD. Similar phenomena also happens between the ISIS-ALASSO and SIS-ALASSO estimators.

Lastly, as the signal-to-noise ratio (i.e., the Theoretical $R^2$) increases (i.e., Theoretical $R^2 \geq 60\%$), every method's out-of-sample forecasting accuracy improves. Overall speaking, we find that the FR methods' (i.e., FR-NONE, FR-ALASSO, FR-SCAD) finite sample performances are always very competitive. In most situations, their out-of-sample $R^2$s are the best. Furthermore, we find that the FR methods' percentages of correct zeros are always 100% while the percentage of incorrect zeros are close to 0. Those simulation results numerically confirmed the screening consistency of the FR methods and the BIC criterion (3.2). As a directly consequence, FR's resulting average model Sizes are very close to the size of true model (i.e., $d_0$), and the percentages of correctly fitted are close to 100%. In most situations, the FR methods' percentages of correctly fitted are the best, and the difference observed between the FR methods and their competitors are very often substan-

Table 5. Example 5 with $(n, d, d_0) = (200, 10{,}000, 8)$

| Screening method | Selection method | Coverage probability (%) | Percentage of correct zeros (%) | Percentage of incorrect zeros (%) | Percentage of correctly fitted (%) | Average model size | Out of sample $R^2$ (%) Mean | Out of sample $R^2$ (%) SE |
|---|---|---|---|---|---|---|---|---|
| | | | Theoretical $R^2 = 30\%$ | | | | | |
| SIS | NONE | 40.1 | 99.7 | 59.9 | 0.0 | 38.0 | −15.2 | 0.88 |
| | ALASSO | 4.9 | 100.0 | 95.1 | 0.0 | 0.4 | 1.8 | 0.36 |
| | SCAD | 4.9 | 100.0 | 95.1 | 0.0 | 0.5 | 1.3 | 0.38 |
| ISIS | NONE | 46.6 | 98.1 | 53.4 | 0.0 | 190.0 | −37.3 | 1.42 |
| | ALASSO | 4.9 | 100.0 | 95.1 | 0.0 | 0.4 | 1.8 | 0.36 |
| | SCAD | 4.9 | 100.0 | 95.1 | 0.0 | 0.5 | 1.3 | 0.38 |
| LARS | NONE | 0.2 | 100.0 | 99.8 | 0.0 | 0.0 | 0.2 | 0.10 |
| | ALASSO | 0.2 | 100.0 | 99.8 | 0.0 | 0.0 | 0.2 | 0.10 |
| | SCAD | 0.2 | 100.0 | 99.8 | 0.0 | 0.0 | 0.2 | 0.10 |
| FR | NONE | 6.1 | 100.0 | 93.9 | 0.0 | 0.6 | 1.4 | 0.47 |
| | ALASSO | 6.1 | 100.0 | 93.9 | 0.0 | 0.6 | 1.4 | 0.47 |
| | SCAD | 6.1 | 100.0 | 93.9 | 0.0 | 0.6 | 1.4 | 0.47 |
| | | | Theoretical $R^2 = 60\%$ | | | | | |
| SIS | NONE | 69.4 | 99.7 | 30.6 | 0.0 | 38.0 | 22.5 | 0.88 |
| | ALASSO | 45.3 | 100.0 | 54.8 | 0.5 | 3.7 | 35.3 | 1.03 |
| | SCAD | 33.8 | 100.0 | 66.2 | 0.0 | 2.9 | 27.9 | 0.97 |
| ISIS | NONE | 81.8 | 98.2 | 18.2 | 0.0 | 190.0 | 5.8 | 1.20 |
| | ALASSO | 46.8 | 100.0 | 53.2 | 2.0 | 3.8 | 35.9 | 1.06 |
| | SCAD | 34.6 | 100.0 | 65.4 | 0.0 | 2.9 | 28.1 | 0.99 |
| LARS | NONE | 4.4 | 100.0 | 95.6 | 0.0 | 0.3 | 4.7 | 0.73 |
| | ALASSO | 4.3 | 100.0 | 95.8 | 0.0 | 0.3 | 4.6 | 0.71 |
| | SCAD | 4.3 | 100.0 | 95.8 | 0.0 | 0.3 | 4.6 | 0.71 |
| FR | NONE | 49.0 | 100.0 | 51.0 | 2.0 | 4.1 | 36.5 | 1.11 |
| | ALASSO | 49.0 | 100.0 | 51.0 | 2.5 | 4.1 | 36.5 | 1.11 |
| | SCAD | 49.0 | 100.0 | 51.0 | 2.5 | 4.1 | 36.5 | 1.11 |
| | | | Theoretical $R^2 = 90\%$ | | | | | |
| SIS | NONE | 85.3 | 99.7 | 14.7 | 0.0 | 38.0 | 73.5 | 0.76 |
| | ALASSO | 85.0 | 100.0 | 15.0 | 25.0 | 6.8 | 82.0 | 0.50 |
| | SCAD | 83.8 | 100.0 | 16.3 | 16.5 | 7.1 | 81.3 | 0.53 |
| ISIS | NONE | 99.7 | 98.2 | 0.3 | 0.0 | 190.0 | 70.3 | 0.71 |
| | ALASSO | 99.6 | 100.0 | 0.4 | 86.0 | 8.1 | 89.3 | 0.17 |
| | SCAD | 98.1 | 100.0 | 1.9 | 55.0 | 8.3 | 88.6 | 0.23 |
| LARS | NONE | 90.3 | 100.0 | 9.7 | 29.0 | 8.0 | 83.0 | 1.28 |
| | ALASSO | 90.3 | 100.0 | 9.8 | 76.0 | 7.2 | 83.4 | 1.29 |
| | SCAD | 90.3 | 100.0 | 9.8 | 60.0 | 7.4 | 83.3 | 1.29 |
| FR | NONE | 100.0 | 100.0 | 0.0 | 76.0 | 8.3 | 89.3 | 0.16 |
| | ALASSO | 100.0 | 100.0 | 0.0 | 76.0 | 8.3 | 89.3 | 0.16 |
| | SCAD | 100.0 | 100.0 | 0.0 | 76.0 | 8.3 | 89.3 | 0.16 |



tial; see, for example, Tables 1, 3, 4, and 5 with the Theoretical $R^2 = 90\%$.

As a cautionary note, we should not claim FR as the only good method for variable screening. However, our extensive simulation studies do confirm that FR is a very promising method, as compared with its competitors. As a consequence, it should be useful for real applications.

### 4.4 A Supermarket Dataset

To further demonstrate the usefulness of the FR methods in real practice, we present here a supermarket dataset with $(n, d) = (464, 6398)$. Each record corresponds to a daily observation collected from a major supermarket located in northern China. The response of interest is the number of customers in one particular day. Each predictor corresponds to one particular product's sale volume on that day. The supermarket manager is interested in knowing which product's sale volume is mostly correlated with the number of customers, after controlling for the effect of other products. Thus, the regression model (2.1) is useful. Due to confidentiality reasons, both the response and predictors have been standardized to be zero mean and unit variance.

To compare different methods' finite sample performances, we randomly split the data into a training dataset (with 400 ob-

Table 6. Example 6 with $(R^2, d, d_0) = (75\%, 10{,}000, [\sqrt{n}])$

| Screening method | Selection method | Coverage probability (%) | Percentage of correct zeros (%) | Percentage of incorrect zeros (%) | Percentage of correctly fitted (%) | Average model size | Out of sample $R^2$ (%) Mean | SE |
|---|---|---|---|---|---|---|---|---|
| | | | Sample size $n = 200$ | | | | | |
| SIS | NONE | 40.0 | 99.7 | 60.0 | 0.0 | 38.0 | 14.0 | 0.82 |
| | ALASSO | 6.5 | 100.0 | 93.5 | 0.0 | 1.3 | 8.8 | 0.74 |
| | SCAD | 5.8 | 100.0 | 94.2 | 0.0 | 1.2 | 7.7 | 0.69 |
| ISIS | NONE | 50.1 | 98.2 | 49.9 | 0.0 | 190.0 | −0.5 | 0.99 |
| | ALASSO | 6.5 | 100.0 | 93.5 | 0.0 | 1.3 | 8.8 | 0.74 |
| | SCAD | 5.8 | 100.0 | 94.2 | 0.0 | 1.2 | 7.7 | 0.69 |
| LARS | NONE | 0.2 | 100.0 | 99.8 | 0.0 | 0.0 | 0.3 | 0.15 |
| | ALASSO | 0.2 | 100.0 | 99.8 | 0.0 | 0.0 | 0.3 | 0.15 |
| | SCAD | 0.2 | 100.0 | 99.8 | 0.0 | 0.0 | 0.3 | 0.15 |
| FR | NONE | 7.1 | 100.0 | 92.9 | 0.0 | 1.4 | 8.8 | 0.74 |
| | ALASSO | 7.1 | 100.0 | 92.9 | 0.0 | 1.4 | 8.8 | 0.74 |
| | SCAD | 7.1 | 100.0 | 92.9 | 0.0 | 1.4 | 8.8 | 0.74 |
| | | | Sample size $n = 400$ | | | | | |
| SIS | NONE | 54.8 | 99.5 | 45.2 | 0.0 | 67.0 | 34.0 | 0.56 |
| | ALASSO | 22.1 | 100.0 | 77.9 | 0.0 | 6.4 | 28.7 | 0.82 |
| | SCAD | 17.6 | 100.0 | 82.4 | 0.0 | 5.1 | 23.7 | 0.77 |
| ISIS | NONE | 75.8 | 96.9 | 24.2 | 0.0 | 335.0 | 25.1 | 0.60 |
| | ALASSO | 22.3 | 100.0 | 77.7 | 0.0 | 6.5 | 28.8 | 0.84 |
| | SCAD | 17.6 | 100.0 | 82.4 | 0.0 | 5.2 | 23.8 | 0.78 |
| LARS | NONE | 0.6 | 100.0 | 99.4 | 0.0 | 0.2 | 1.3 | 0.32 |
| | ALASSO | 0.6 | 100.0 | 99.4 | 0.0 | 0.2 | 1.3 | 0.32 |
| | SCAD | 0.6 | 100.0 | 99.4 | 0.0 | 0.2 | 1.3 | 0.32 |
| FR | NONE | 24.2 | 100.0 | 75.8 | 0.0 | 7.1 | 30.2 | 0.84 |
| | ALASSO | 24.1 | 100.0 | 75.9 | 0.0 | 7.0 | 30.1 | 0.83 |
| | SCAD | 24.2 | 100.0 | 75.8 | 0.0 | 7.1 | 30.2 | 0.84 |
| | | | Sample size $n = 800$ | | | | | |
| SIS | NONE | 70.5 | 99.1 | 29.5 | 0.0 | 120.0 | 51.7 | 0.30 |
| | ALASSO | 52.9 | 100.0 | 47.1 | 0.0 | 21.2 | 57.0 | 0.38 |
| | SCAD | 49.3 | 100.0 | 50.8 | 0.0 | 20.1 | 54.1 | 0.46 |
| ISIS | NONE | 92.5 | 93.1 | 7.5 | 0.0 | 720.0 | 40.3 | 0.32 |
| | ALASSO | 54.9 | 100.0 | 45.1 | 0.0 | 22.0 | 58.1 | 0.39 |
| | SCAD | 51.0 | 100.0 | 49.0 | 0.0 | 20.9 | 55.0 | 0.49 |
| LARS | NONE | 3.9 | 100.0 | 96.1 | 0.0 | 1.6 | 7.3 | 0.94 |
| | ALASSO | 3.8 | 100.0 | 96.2 | 0.0 | 1.5 | 7.3 | 0.93 |
| | SCAD | 3.8 | 100.0 | 96.2 | 0.0 | 1.5 | 7.3 | 0.93 |
| FR | NONE | 59.0 | 100.0 | 41.0 | 0.0 | 23.6 | 60.1 | 0.39 |
| | ALASSO | 58.9 | 100.0 | 41.1 | 0.0 | 23.6 | 60.0 | 0.40 |
| | SCAD | 59.0 | 100.0 | 41.0 | 0.0 | 23.6 | 60.1 | 0.39 |



servations) and a testing dataset (with 64 observations). Next, denote by $\hat{\boldsymbol{\beta}}$ the estimator produced by one particular method (e.g., the FR-ALASSO estimator) on the training dataset. Similar to the simulation study, we evaluate its prediction accuracy by the out-of-sample $R^2$. Because the response has been standardized to have unit variance, the out-of-sample $R^2$ can be defined as $R^2 = 100(1 - 64^{-1}\|\mathcal{Y}^* - \mathcal{X}^*\hat{\boldsymbol{\beta}}\|^2)$, where $\mathcal{Y}^*$ and $\mathcal{X}^*$ are the response vector and the design matrix of the testing data, respectively. For a reliable evaluation, this experiment is randomly replicated for a total of 200 times.

The detailed results are summarized in Table 7. We find that, in terms of the prediction accuracy, the SIS-NONE demonstrated the best performance, with the mean out-of-sample $R^2 = 90\%$. Thus, if the forecasting accuracy is the only concern, SIS-NONE is the best choice for this particular dataset. However, if interpretation is also of importance, then the model selected by SIS-NONE is less attractive. This is because simultaneously interpret a total of $67 = [400/\log 400]$ products (as selected by SIS-NONE) could be quite a challenging task for the supermarket manager. As a result, those models further selected by ALASSO and SCAD with much smaller sizes might be more relevant for real practice.

For those models further selected by ALASSO and SCAD, the two FR models (i.e., FR-ALASSO and FR-SCAD) achieved

Table 7. The supermarket dataset. The $p$-value is generated by a two-sided pairwise $t$-test. It compares every screening-estimation method's (e.g., FR-SCAD) mean out-of-sample $R^2$ against that of the SIS-NONE

| Screening method | Estimation method | Out-of-sample $R^2$ (%) | | | The model size | |
|---|---|---|---|---|---|---|
| | | Mean | SE | $p$-value | Mean | SE |
| SIS | NONE | 90 | (1.9) | – | 67 | (0.0) |
| | ALASSO | 85 | (2.8) | 0.06 | 10 | (1.4) |
| | SCAD | 84 | (2.9) | 0.03 | 10 | (1.4) |
| ISIS | NONE | 80 | (4.5) | 0.03 | 335 | (0.0) |
| | ALASSO | 86 | (2.7) | 0.15 | 11 | (1.6) |
| | SCAD | 87 | (2.4) | 0.29 | 11 | (1.4) |
| LARS | NONE | 86 | (2.6) | 0.08 | 12 | (2.4) |
| | ALASSO | 84 | (3.1) | 0.03 | 8 | (1.1) |
| | SCAD | 85 | (3.0) | 0.07 | 9 | (2.0) |
| FR | NONE | 88 | (2.3) | 0.43 | 14 | (1.8) |
| | ALASSO | 88 | (2.4) | 0.33 | 12 | (1.6) |
| | SCAD | 88 | (2.4) | 0.35 | 12 | (1.6) |

the best prediction accuracy with the mean out-of-sample $R^2 = 88\%$, which is a number slightly worse than 90% of the SIS-NONE. However, the average model size is considerably smaller and can be as small as 12. Thus, by sacrificing 2% out-of-sample $R^2$, the model's interpretability can be much improved. In fact, the $p$-values generated from pairwise $t$-tests (comparing the two FR methods against SIS-NONE) can be as large as 33% (see the 5th column in Table 7), which indicates that such a 2% difference is not even statistically significant. Lastly, we note that the performances of ISIS-ALASSO, ISIS-SCAD, and LARS-NONE are slight worse than those of the FR methods, but also comparable, with the mean out-of-sample $R^2 \geq 86\%$. Once again, we do not claim FR as the only good method for variable screening. However, we do find it very useful (as a high dimensional predictor screening method) for this particular dataset.

## 5. CONCLUDING REMARKS

We investigate here a very popular yet classical variable screening method, that is, FR. We show both theoretically and numerically that FR can discover all relevant predictors consistently, even if the predictor dimension is substantially larger than the sample size. Our preliminary experience with FR is quite encouraging. Nevertheless, as we mentioned before, we should not claim FR as the only good method for variable screening. In fact, as we observed in the supermarket dataset, the best prediction accuracy was achieved by SIS-NONE, with of course a much larger model size. Next, how to investigate similar algorithms under a general loss function, for example, generalized linear model (McCullagh and Nelder 1989), and/or a semiparametric context, for example, single-index model (Kong and Xia 2007), are interesting topics for future study.

## APPENDIX A: A LEMMA AND ITS PROOF

To prove Theorem 1, the following lemma is needed. For convenience, we define $\hat{\mathbf{\Sigma}} = n^{-1}\mathfrak{X}^\top\mathfrak{X}$. Lastly, recall that $\hat{\mathbf{\Sigma}}_{(\mathcal{M})}$ and $\mathbf{\Sigma}_{(\mathcal{M})}$ are the submatrices of $\hat{\mathbf{\Sigma}}$ and $\mathbf{\Sigma}$ (corresponding to $\mathcal{M}$), respectively.

*Lemma 1.* Assume conditions (C1), (C2), (C4), and $m = O(n^{2\xi_0+4\xi_{\min}})$. Then, with probability tending to one, we have

$$\tau_{\min} \leq \min_{|\mathcal{M}|\leq m} \lambda_{\min}\{\hat{\mathbf{\Sigma}}_{(\mathcal{M})}\} \leq \max_{|\mathcal{M}|\leq m} \lambda_{\max}\{\hat{\mathbf{\Sigma}}_{(\mathcal{M})}\} \leq \tau_{\max}. \quad \text{(A.1)}$$

*Proof.* Let $\mathbf{r} = (r_1, \ldots, r_d)^\top \in \mathbb{R}^d$ be an arbitrary $d$-dimensional vector and $\mathbf{r}_{(\mathcal{M})}$ be the subvector corresponding to $\mathcal{M}$. By condition (C2), we know immediately

$$\begin{aligned} 2\tau_{\min} &\leq \min_{\mathcal{M}\subset\mathcal{F}} \inf_{\|\mathbf{r}_{(\mathcal{M})}\|=1} \mathbf{r}_{(\mathcal{M})}^\top \mathbf{\Sigma}_{(\mathcal{M})}\mathbf{r}_{(\mathcal{M})} \\ &\leq \max_{\mathcal{M}\subset\mathcal{F}} \sup_{\|\mathbf{r}_{(\mathcal{M})}\|=1} \mathbf{r}_{(\mathcal{M})}^\top \mathbf{\Sigma}_{(\mathcal{M})}\mathbf{r}_{(\mathcal{M})} \leq 2^{-1}\tau_{\max}. \end{aligned}$$

Thus, the desired conclusion (A.1) is implied by

$$P\Big(\max_{|\mathcal{M}|\leq m} \sup_{\|\mathbf{r}_{(\mathcal{M})}\|=1} |\mathbf{r}_{(\mathcal{M})}^\top\{\hat{\mathbf{\Sigma}}_{(\mathcal{M})} - \mathbf{\Sigma}_{(\mathcal{M})}\}\mathbf{r}_{(\mathcal{M})}| > \epsilon\Big) \to 0, \quad \text{(A.2)}$$

where $\epsilon > 0$ is an arbitrary positive number. The left-hand side of (A.2) is bounded by

$$\leq \sum_{|\mathcal{M}|\leq m} P\Big(\sup_{\|\mathbf{r}_{(\mathcal{M})}\|=1} |\mathbf{r}_{(\mathcal{M})}^\top\{\hat{\mathbf{\Sigma}}_{(\mathcal{M})} - \mathbf{\Sigma}_{(\mathcal{M})}\}\mathbf{r}_{(\mathcal{M})}| > \epsilon\Big). \quad \text{(A.3)}$$

Note that, for any $\mathcal{M}$ with $|\mathcal{M}| \leq m$, we have

$$\begin{aligned} &|\mathbf{r}_{(\mathcal{M})}^\top\{\hat{\mathbf{\Sigma}}_{(\mathcal{M})} - \mathbf{\Sigma}_{(\mathcal{M})}\}\mathbf{r}_{(\mathcal{M})}| \\ &\leq \sum_{j_1,j_2\in\mathcal{M}} |r_{j_1}| \times |r_{j_2}| \times |\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}| \\ &\leq \max_{1\leq j_1,j_2\leq d} |\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}| \sum_{j_1,j_2\in\mathcal{M}} |r_{j_1}| \times |r_{j_2}| \\ &= \max_{1\leq j_1,j_2\leq d} |\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}| \Big(\sum_{j\in\mathcal{M}} |r_j|\Big)^2 \\ &\leq |\mathcal{M}| \max_{1\leq j_1,j_2\leq d} |\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}| \\ &\leq m \max_{1\leq j_1,j_2\leq d} |\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}|. \end{aligned}$$

Consequently, the right-hand side of (A.3) can be bounded further by

$$\begin{aligned} &\leq \sum_{|\mathcal{M}|\leq m} P\Big(\max_{1\leq j_1,j_2\leq d} |\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}| > \frac{\epsilon}{m}\Big) \\ &\leq \sum_{|\mathcal{M}|\leq m} \sum_{1\leq j_1,j_2\leq d} P\Big(|\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}| > \frac{\epsilon}{m}\Big). \end{aligned} \quad \text{(A.4)}$$

Then, by conditions (C1), (C2), and lemma A3 in Bickel and Levina (2008), there exists constants $C_1 > 0$ and $C_2 > 0$, such that that $P(|\hat{\sigma}_{j_1j_2} - \sigma_{j_1j_2}| > \epsilon) \leq C_1\exp(-C_2n\epsilon^2)$. Apply this inequality back to the right-hand side of (A.4). Furthermore, note that the number of models satisfying $|\mathcal{M}| \leq m$ is no more than $d^{(m+1)}$. Then, the right-hand side of (A.4) can be further bounded by

$$\begin{aligned} &\leq d^{(m+1)} \cdot d^2 \cdot C_1\exp(-C_2n\epsilon^2m^{-2}) \\ &= C_1\exp\{(m+3)\log d - C_2n\epsilon^2m^{-2}\}. \end{aligned} \quad \text{(A.5)}$$

Note that, $m + 3 \leq \nu_m n^{2\xi_0+4\xi_{\min}}$ for some constant $\nu_m > 0$ by the lemma assumption, while $\log d \leq \nu n^\xi$ by (C4). Thus, the right side of (A.5) can be further bounded by

$$\begin{aligned} &\leq C_1\exp[\nu_m\nu n^{\xi+2\xi_0+4\xi_{\min}} - C_2\epsilon^2\nu_m^{-2}n^{1-4\xi_0-8\xi_{\min}}] \\ &= C_1\exp\big[\nu_m\nu n^{\xi+2\xi_0+4\xi_{\min}}(1 - C_2\epsilon^2\nu_m^{-3}\nu^{-1}n^{1-\xi-6\xi_0-12\xi_{\min}})\big]. \end{aligned}$$

By condition (C4), we have $\xi + 6\xi_0 + 12\xi_{\min} < 1$. Thus, the right-hand side of the above equation converges to 0 as $n \to \infty$. This proves (A.2) and completes the proof.

## APPENDIX B: PROOF OF THEOREM 1

The theorem conclusion can be proved in a total of $d_0$ (i.e., no more than $\nu n^{\xi_0}$) steps. In every step, we prove that, within $Kn^{\xi_0+4\xi_{\min}}$ steps, at least one new relevant predictor will be identified by the FR algorithm, conditional on those already included. Because $d_0 \le \nu n^{\xi_0}$ by condition (C2), we thus know all relevant predictors will be identified by the FR algorithm within $K\nu n^{2\xi_0+4\xi_{\min}}$ steps. Because the proof of each step is similar, we will provide details for the first step only.

For the first FR step, we have $k = 0$, $\mathcal{S}^{(0)} = \emptyset$, and $|\mathcal{S}^{(0)}| \le K\nu n^{2\xi_0+4\xi_{\min}}$. Thus, Lemma 1 can be applied. To prove the desired conclusion, we assume that no relevant predictor has been identified in the first $k$ steps, given the existence of $\mathcal{S}^{(0)}$. We then evaluate how likely at least one relevant predictor will be discovered in the next, that is, $(k+1)$'s, step. To this end, we consider what happens if the predictor selected by the $(k+1)$'s step is still an irrelevant one. We then have the following relationship:

$$\begin{aligned}\Omega(k) &\doteq \mathrm{RSS}\big(\mathcal{S}^{(k)}\big) - \mathrm{RSS}\big(\mathcal{S}^{(k+1)}\big)\\ &= \big\|\mathbf{H}^{(k)}_{a_{k+1}}\{\mathbf{I}_n - \mathbf{H}_{(\mathcal{S}^{(k)})}\}\mathfrak{Y}\big\|^2, \end{aligned} \tag{B.1}$$

where $\mathbf{H}^{(k)}_j = \mathfrak{X}^{(k)}_j \mathfrak{X}^{(k)\top}_j \|\mathfrak{X}^{(k)}_j\|^{-2}$ and $\mathfrak{X}^{(k)}_j = \{\mathbf{I}_n - \mathbf{H}_{\mathcal{S}^{(k)}}\}\mathfrak{X}_j$. Because, we are assuming that $a_{k+1} \notin \mathcal{T}$, we then must have

$$\begin{aligned}\Omega(k) &\ge \max_{j\in\mathcal{T}} \big\|\mathbf{H}^{(k)}_j\{\mathbf{I}_n - \mathbf{H}_{(\mathcal{S}^{(k)})}\}\mathfrak{Y}\big\|^2\\ &\ge \big\|\mathbf{H}^{(k)}_{\hat{j}}\{\mathbf{I}_n - \mathbf{H}_{(\mathcal{S}^{(k)})}\}\mathfrak{Y}\big\|^2,\end{aligned}$$

where $\hat{j} = \arg\max_{j\in\mathcal{T}} \|\mathbf{H}^{(k)}_j\{\mathbf{I}_n - \mathbf{H}_{(\mathcal{S}^{(k)})}\}\{\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})}\}\|^2$. Next, note that the right-hand side of the above inequality is no less than

$$\begin{aligned}&\ge \big\|\mathbf{H}^{(k)}_{\hat{j}}\{\mathbf{I}_n - \mathbf{H}_{(\mathcal{S}^{(k)})}\}\{\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})}\}\big\|^2 - \big\|\mathbf{H}^{(k)}_{\hat{j}}\{\mathbf{I}_n - \mathbf{H}_{(\mathcal{S}^{(k)})}\}\boldsymbol{\mathcal{E}}\big\|^2\\ &\ge \max_{j\in\mathcal{T}} \big\|\mathbf{H}^{(k)}_j\{\mathbf{I}_n - \mathbf{H}_{(\mathcal{S}^{(k)})}\}\{\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})}\}\big\|^2\\ &\quad - \max_{j\in\mathcal{T}} \big\|\mathbf{H}^{(k)}_j\{\mathbf{I}_n - \mathbf{H}_{(\mathcal{S}^{(k)})}\}\boldsymbol{\mathcal{E}}\big\|^2,\end{aligned} \tag{B.2}$$

where $\boldsymbol{\mathcal{E}} = (\varepsilon_1, \ldots, \varepsilon_n)^\top \in \mathbb{R}^n$. In what follows the two terms involved in (B.2) will be carefully evaluated separately.

*Step 1* [The 1st term in (B.2)]. For convenience, we define $\mathbf{Q}_{(\mathcal{S}^{(k)})} = \mathbf{I}_n - \mathbf{H}_{(\mathcal{S}^{(k)})}$. Then, the 1st term in (B.2) equals to

$$\begin{aligned}&\max_{j\in\mathcal{T}} \big\|\mathbf{H}^{(k)}_j\mathbf{Q}_{(\mathcal{S}^{(k)})}(\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})})\big\|^2\\ &\qquad = \max_{j\in\mathcal{T}}\Big\{\|\mathfrak{X}^{(k)}_j\|^{-2}\cdot|\mathfrak{X}^{(k)\top}_j\mathbf{Q}_{(\mathcal{S}^{(k)})}(\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})})|^2\Big\}.\end{aligned} \tag{B.3}$$

Note $\mathfrak{X}^{(k)\top}_j\mathbf{Q}_{(\mathcal{S}^{(k)})} = \mathfrak{X}^{\top}_j\mathbf{Q}_{(\mathcal{S}^{(k)})}$. Define $j^* = \arg\max_{j\in\mathcal{T}}|\mathfrak{X}^{(k)\top}_j \times \mathbf{Q}_{(\mathcal{S}^{(k)})}(\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})})|^2$. Thus, the right-hand size of (B.3) is no less than

$$\begin{aligned}&\ge \|\mathfrak{X}^{(k)}_{j^*}\|^{-2}\cdot|\mathfrak{X}^{(k)\top}_{j^*}\mathbf{Q}_{(\mathcal{S}^{(k)})}(\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})})|^2\\ &\ge \min_{j\in\mathcal{T}}\{\|\mathfrak{X}^{(k)}_j\|^{-2}\}\cdot|\mathfrak{X}^{(k)\top}_{j^*}\mathbf{Q}_{(\mathcal{S}^{(k)})}(\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})})|^2\\ &= \Big\{\max_{j\in\mathcal{T}}\|\mathfrak{X}^{(k)}_j\|^2\Big\}^{-1}\Big\{\max_{j\in\mathcal{T}}|\mathfrak{X}^{\top}_j\mathbf{Q}_{(\mathcal{S}^{(k)})}(\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})})|^2\Big\}\\ &\ge \Big\{\max_{j\in\mathcal{T}}\|\mathfrak{X}_j\|^2\Big\}^{-1}\Big\{\max_{j\in\mathcal{T}}|\mathfrak{X}^{\top}_j\mathbf{Q}_{(\mathcal{S}^{(k)})}(\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})})|^2\Big\},\end{aligned} \tag{B.4}$$

where (B.4) is due to the fact that $\|\mathfrak{X}_j\| \ge \|\mathfrak{X}^{(k)}_j\|$. On the other hand, note that

$$\begin{aligned}&\big\|\mathbf{Q}_{(\mathcal{S}^{(k)})}(\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})})\big\|^2\\ &= \boldsymbol{\beta}^\top_{(\mathcal{T})}\mathfrak{X}^\top_{(\mathcal{T})}\mathbf{Q}_{(\mathcal{S}^{(k)})}\{\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})}\}\\ &= \Big(\sum_{j\in\mathcal{T}}\beta_j\mathfrak{X}^\top_j\mathbf{Q}_{(\mathcal{S}^{(k)})}\{\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})}\}\Big)\\ &\le \Big(\sum_{j\in\mathcal{T}}\beta^2_j\Big)^{1/2}\Big\{\sum_{j\in\mathcal{T}}(\mathfrak{X}^\top_j\mathbf{Q}_{(\mathcal{S}^{(k)})}\{\mathfrak{X}_{\mathcal{T}}\boldsymbol{\beta}_{(\mathcal{T})}\})^2\Big\}^{1/2}\\ &\le \|\boldsymbol{\beta}\|\times\max_{j\in\mathcal{T}}|\mathfrak{X}^\top_j\mathbf{Q}_{(\mathcal{S}^{(k)})}\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})}|\times\sqrt{d_0}.\end{aligned} \tag{B.5}$$

Apply (B.5) back to (B.4) and also (B.3), and note that $\max_{j\in\mathcal{T}}\|\mathfrak{X}_j\|^2/n \le \tau_{\max}$ with probability tending to one, by conditions (C1)–(C2) and Lemma 1. Then, we have

$$\max_{j\in\mathcal{T}}\big\|\mathbf{H}^{(k)}_j\mathbf{Q}_{(\mathcal{S}^{(k)})}(\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})})\big\|^2 \ge \frac{\|\mathbf{Q}_{(\mathcal{S}^{(k)})}(\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})})\|^4}{n\tau_{\max}d_0\|\boldsymbol{\beta}\|^2}. \tag{B.6}$$

Defining $\boldsymbol{\zeta}_{(\mathcal{S}^{(k)})} = (\mathfrak{X}^\top_{(\mathcal{S}^{(k)})}\mathfrak{X}_{(\mathcal{S}^{(k)})})^{-1}\{\mathfrak{X}^\top_{(\mathcal{S}^{(k)})}\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})}\}$, we obtain

$$\big\|\mathbf{Q}_{(\mathcal{S}^{(k)})}(\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})})\big\|^2 = \big\|\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})} - \mathfrak{X}_{(\mathcal{S}^{(k)})}\boldsymbol{\zeta}_{(\mathcal{S}^{(k)})}\big\|^2.$$

Under the assumption that no additional relevant predictor has been identified by FR, we have $\mathcal{T} \not\subset \mathcal{S}^{(k)}$. Then, by conditions (C1)–(C3), and Lemma 1, we have

$$\big\|\mathbf{Q}_{(\mathcal{S}^{(k)})}(\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})})\big\|^2 \ge n\tau_{\min}\beta^2_{\min}$$

with probability tending to one. Applying this result back to (B.6) and also noting the technical condition (C3)–(C4), we find that

$$\begin{aligned}&\max_{j\in\mathcal{T}}\big\|\mathbf{H}^{(k)}_j\mathbf{Q}_{(\mathcal{S}^{(k)})}(\mathfrak{X}_{(\mathcal{T})}\boldsymbol{\beta}_{(\mathcal{T})})\big\|^2\\ &\quad\ge n\tau^{-1}_{\max}d_0^{-1}\|\boldsymbol{\beta}\|^{-2}\tau^2_{\min}\beta^4_{\min}\\ &\quad\ge \tau^{-1}_{\max}\nu^{-1}C^{-2}_{\boldsymbol{\beta}}\tau^2_{\min}\nu^4_{\boldsymbol{\beta}}n^{1-\xi_0-4\xi_{\min}}.\end{aligned} \tag{B.7}$$

*Step 2* [The 2nd term in (B.2)]. Note that $\mathfrak{X}^{(k)}_j = \mathfrak{X}_j - \mathbf{H}_{(\mathcal{S}^{(k)})}\mathfrak{X}_j = \mathfrak{X}_j - \mathfrak{X}_{(\mathcal{S}^{(k)})}\boldsymbol{\theta}_{j(\mathcal{S}^{(k)})}$, where $\boldsymbol{\theta}_{j(\mathcal{S}^{(k)})} = (\mathfrak{X}^\top_{(\mathcal{S}^{(k)})}\mathfrak{X}_{(\mathcal{S}^{(k)})})^{-1}(\mathfrak{X}^\top_{(\mathcal{S}^{(k)})} \times \mathfrak{X}_j)$. Consequently, we must have $\|\mathfrak{X}^{(k)}_j\|^2 \ge n\tau_{\min}$ by conditions (C1)–(C2) and Lemma 1. Moreover, recall that $\mathfrak{X}^{(k)}_j = (\mathbf{I}_n - \mathbf{H}_{(\mathcal{S}^{(k)})})\mathfrak{X}_j$. We then have

$$\begin{aligned}&\big\|\mathbf{H}^{(k)}_j\{\mathbf{I}_n - \mathbf{H}_{(\mathcal{S}^{(k)})}\}\boldsymbol{\mathcal{E}}\big\|^2\\ &= \|\mathfrak{X}^{(k)}_j\|^{-2}(\mathfrak{X}^\top_j\boldsymbol{\mathcal{E}} - \mathfrak{X}^\top_j\mathbf{H}_{(\mathcal{S}^{(k)})}\boldsymbol{\mathcal{E}})^2\\ &\le \tau^{-1}_{\min}n^{-1}(\mathfrak{X}^\top_j\boldsymbol{\mathcal{E}} - \mathfrak{X}^\top_j\mathbf{H}_{(\mathcal{S}^{(k)})}\boldsymbol{\mathcal{E}})^2\\ &= \tau^{-1}_{\min}n^{-1}(\mathfrak{X}^\top_j\mathbf{Q}_{(\mathcal{S}^{(k)})}\boldsymbol{\mathcal{E}})^2\\ &\le \tau^{-1}_{\min}n^{-1}\max_{j\in\mathcal{T}}\max_{|\mathcal{M}|\le m^*}(\mathfrak{X}^\top_j\mathbf{Q}_{(\mathcal{M})}\boldsymbol{\mathcal{E}})^2,\end{aligned} \tag{B.8}$$

where $m^* = K\nu n^{2\xi_0+4\xi_{\min}}$. Note that $\mathfrak{X}^\top_j\mathbf{Q}_{(\mathcal{M})}\boldsymbol{\mathcal{E}}$ is a normal random variable with mean 0 and variance given by $\|\mathbf{Q}_{(\mathcal{M})}\mathfrak{X}_j\|^2 \le \|\mathfrak{X}_j\|^2$. Thus, the right-hand side of (B.8) can be bounded further by

$$\le \tau^{-1}_{\min}n^{-1}\max_{j\in\mathcal{T}}\|\mathfrak{X}_j\|^2\cdot\max_{j\in\mathcal{T}}\max_{|\mathcal{M}|\le m^*}\chi^2_1,$$

where $\chi^2_1$ stands for a chi-square random variable with one degree of freedom. By conditions (C1)–(C2) and Lemma 1, we know that

$n^{-1}\max_{j\in\mathcal{T}}\|\mathfrak{X}_j\|^2 \le \tau_{\max}$ with probability tending to one. On the other hand, the total number of combinations for $j\in\mathcal{T}$ and $|\mathcal{M}|\le m^*$ is no more than $d^{m^*+2}$. Then by (C4), we have

$$\begin{aligned}\max_{j\in\mathcal{T}}\max_{|\mathcal{M}|=k}\chi_1^2 &\le 2(m^*+2)\log(d)\\ &\le 3K\nu n^{2\xi_0+4\xi_{\min}}\times\nu n^{\xi}\\ &= 3K\nu^2 n^{\xi+2\xi_0+4\xi_{\min}}\end{aligned}$$

with probability tending to one. Combining this result with (B.2) and (B.7), we find

$$\begin{aligned}n^{-1}\Omega(k) &\ge \{\tau_{\max}^{-1}\nu^{-1}C_{\boldsymbol{\beta}}^{-2}\tau_{\min}^2\nu_{\boldsymbol{\beta}}^4 n^{-\xi_0-4\xi_{\min}}\\ &\quad-\tau_{\min}^{-1}\tau_{\max}3K\nu^2\cdot n^{\xi+2\xi_0+4\xi_{\min}-1}\}\\ &= \tau_{\max}^{-1}\nu^{-1}C_{\boldsymbol{\beta}}^{-2}\tau_{\min}^2\nu_{\boldsymbol{\beta}}^4 n^{-\xi_0-4\xi_{\min}}\\ &\quad\times\{1-\tau_{\max}^2\nu^3C_{\boldsymbol{\beta}}^2\tau_{\min}^{-3}\nu_{\boldsymbol{\beta}}^{-4}\cdot 3Kn^{\xi+3\xi_0+8\xi_{\min}-1}\}\end{aligned}$$

uniformly for every $k\le Kn^{\xi_0+4\xi_{\min}}$. Recall $K=2\tau_{\max}\nu C_{\boldsymbol{\beta}}^2\tau_{\min}^{-2}\nu_{\boldsymbol{\beta}}^{-4}$, we then have

$$\begin{aligned}n^{-1}\|\mathfrak{Y}\|^2 &\ge n^{-1}\sum_{k=1}^{[K\cdot n^{\xi_0+4\xi_{\min}}]}\Omega(k)\\ &\ge 2\{1-\tau_{\max}^2\nu^3C_{\boldsymbol{\beta}}^2\tau_{\min}^{-3}\nu_{\boldsymbol{\beta}}^{-4}\cdot 3Kn^{\xi+3\xi_0+8\xi_{\min}-1}\}\\ &\to 2\end{aligned}\tag{B.9}$$

under the condition (C4). On the other hand, under the assumption $\mathrm{var}(Y_i)=1$ we have $n^{-1}\|\mathfrak{Y}\|^2\to_p 1$. Thus, it is impossible to have $\mathcal{S}^{(k)}\cup\mathcal{M}_t=\emptyset$ for every $1\le k\le Kn^{\xi_0+4\xi_{\min}}$, which implies that at least one relevant variable will be discovered within $Kn^{\xi_0+4\xi_{\min}}$ steps. This completes the proof.

## APPENDIX C: PROOF OF THEOREM 2

Define $k_{\min}=\min_{1\le k\le n}\{k:\mathcal{T}\subset\mathcal{S}^{(k)}\}$. By Theorem 1, we know that, with probability tending to one, $k_{\min}$ is well defined and satisfies $k_{\min}\le K\nu n^{2\xi_0+4\xi_{\min}}$. Thus, the theorem conclusion follows, if we can prove that $P(\hat m<k_{\min})\to 0$ as $n\to\infty$. To this end, it suffices to show that

$$P\Big(\min_{0\le k<k_{\min}}\{\mathrm{BIC}(\mathcal{S}^{(k)})-\mathrm{BIC}(\mathcal{S}^{(k+1)})\}>0\Big)\to 1.\tag{C.1}$$

To prove (C.1), note that

$$\begin{aligned}&\mathrm{BIC}(\mathcal{S}^{(k)})-\mathrm{BIC}(\mathcal{S}^{(k+1)})\\ &\quad=\log\Big(\frac{\hat\sigma^2_{(\mathcal{S}^{(k)})}}{\hat\sigma^2_{(\mathcal{S}^{(k+1)})}}\Big)-n^{-1}(\log n+2\log d)\\ &\quad\ge\log\Big(1+\frac{\hat\sigma^2_{(\mathcal{S}^{(k)})}-\hat\sigma^2_{(\mathcal{S}^{(k+1)})}}{\hat\sigma^2_{(\mathcal{S}^{(k+1)})}}\Big)-3n^{-1}\log d,\end{aligned}\tag{C.2}$$

under the assumption $d>n$. By definition, we have $\hat\sigma^2_{(\mathcal{S}^{(k+1)})}\le n^{-1}\|\mathfrak{Y}\|^2\to_p 1$. Then, with probability tending to one, the right-hand side of (C.2) is no less than

$$\begin{aligned}&\ge\log\Big(1+\frac{\hat\sigma^2_{(\mathcal{S}^{(k)})}-\hat\sigma^2_{(\mathcal{S}^{(k+1)})}}{2}\Big)-3n^{-1}\log d\\ &=\log\big[1+2^{-1}n^{-1}\{\Omega(k)-\Omega(k+1)\}\big]-3n^{-1}\log d,\end{aligned}\tag{C.3}$$

by the definition of $\Omega(k)$; see (B.1). Moreover, one can verify that $\log(1+x)\ge\min\{\log 2,0.5x\}$ for any $x>0$. Then, the right-hand side of (C.3) is no less than the following quantity

$$\begin{aligned}&\min\big[\log 2,4^{-1}n^{-1}\{\Omega(k)-\Omega(k+1)\}\big]-3n^{-1}\log d\\ &\qquad\ge\min(\log 2,5^{-1}K^{-1}n^{-\xi_0-4\xi_{\min}})-3n^{-1}\log d,\end{aligned}\tag{C.4}$$

according to (B.9). Note that the right-hand side (C.4) is independent of $k$, thus, it is a uniform lower bound for $\mathrm{BIC}(\mathcal{S}^{(k)})-\mathrm{BIC}(\mathcal{S}^{(k+1)})$ with $1\le k<k_{\min}$. Thus, it suffices to prove the right-hand side of (C.4) is positive with probability tending to one. To this end, we firstly note that $\log 2-3n^{-1}\log d\to 0$ under condition (C4). Thus, we can focus on

$$\begin{aligned}&5^{-1}K^{-1}n^{-\xi_0-4\xi_{\min}}-3n^{-1}\log d\\ &\quad\ge 5^{-1}K^{-1}n^{-\xi_0-4\xi_{\min}}-3\nu n^{\xi-1}\\ &\quad=5^{-1}K^{-1}n^{-\xi_0-4\xi_{\min}}(1-15\nu Kn^{\xi+\xi_0+4\xi_{\min}-1})>0\end{aligned}$$

with probability tending to one, according to (C4). This completes the proof.

*[Received September 2008. Revised April 2009.]*

## REFERENCES


Barron, A. R., and Cohen, A. (2008), "Approximation and Learning by Greedy Algorithms," *The Annals of Statistics*, 36, 64–94.

Bickel, P. J., and Levina, E. (2008), "Regularized Estimation of Large Covariance Matrices," *The Annals of Statistics*, 36, 199–227.

Breiman, L. (1995), "Better Subset Selection Using Nonnegative Garrote," *Technometrics*, 37, 373–384.

Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When $p$ Is Much Larger Than $n$," *The Annals of Statistics*, 35, 2313–2351.

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criterion for Model Selection With Large Model Spaces," *Biometrika*, 95, 759–771.

Donoho, D., and Stodden, V. (2006), "Breakdown Point of Model Selection When the Number of Variables Exceeds the Number of Observations," in *Proceedings of the International Joint Conference on Neural Networks*, Los Alamitos, CA: IEEE, pp. 1916–1921.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–489.

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360.

——— (2006), "Statistical Challenges With High Dimensionality: Feature Selection in Knowledge Discovery," in *Proceedings of the International Congress of Mathematicians*, Vol. III, eds. M. Sanz-Sole, J. Soria, J. L. Varona, and J. Verdera, Zurich: European Mathematical Society, pp. 595–622.

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultra-High Dimensional Feature Space" (with discussion), *Journal of the Royal Statistical Society, Ser. B*, 70, 849–911.

Fan, J., and Peng, H. (2004), "On Non-Concave Penalized Likelihood With Diverging Number of Parameters," *The Annals of Statistics*, 32, 928–961.

Fu, W. J. (1998), "Penalized Regression: The Bridge versus the LASSO," *Journal of Computational and Graphical Statistics*, 7, 397–416.

Huang, J., Horowitz, J., and Ma, S. (2008), "Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models," *The Annals of Statistics*, 36, 587–613.

Huang, J., Ma, S., and Zhang, C. H. (2007), "Adaptive LASSO for Sparse High Dimensional Regression," *Statistica Sinica*, 18, 1603–1618.

Hunter, D. R., and Li, R. (2005), "Variable Selection Using MM Algorithms," *The Annals of Statistics*, 33, 1617–1642.

Jia, J., and Yu, B. (2008), "On Model Selection Consistency of the Elastic Net When $p\gg n$," unpublished manuscript, UC-Berkeley, Dept. of Statistics.

Kong, E., and Xia, Y. (2007), "Variable Selection for Single-Index Model," *Biometrika*, 94, 217–229.

Leng, C., Lin, Y., and Wahba, G. (2006), "A Note on LASSO and Related Procedures in Model Selection," *Statistica Sinica*, 16, 1273–1284.

McCullagh, P., and Nelder, J. A. (1989), *Generalized Linear Models*, New York: Chapman & Hall.

Tibshirani, R. J. (1996), "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society, Ser. B*, 58, 267–288.

Yuan, M., and Lin, Y. (2007), "On the Nonnegative Garrote Estimator," *Journal of the Royal Statistical Society, Ser. B*, 69, 143–161.

Zhang, C. H., and Huang, J. (2008), "The Sparsity and Bias of the LASSO Selection in High-Dimensional Linear Regression," *The Annals of Statistics*, 36, 1567–1594.

Zhang, H. H., and Lu, W. (2007), "Adaptive LASSO for Cox's Proportional Hazard Model," *Biometrika*, 94, 691–703.

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of LASSO," *Journal of Machine Learning Research*, 7, 2541–2567.

Zou, H. (2006), "The Adaptive LASSO and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429.

Zou, H., and Hastie, T. (2005), "Regression Shrinkage and Selection via the Elastic Net With Application to Microarrays," *Journal of the Royal Statistical Society, Ser. B*, 67, 301–320.

Zou, H., and Li, R. (2008), "One-Step Sparse Estimates in Nonconcave Penalized Likelihood Models" (with discussion), *The Annals of Statistics*, 36, 1509–1533.

Zou, H., and Zhang, H. H. (2009), "On the Adaptive Elastic-Net With a Diverging Number of Parameters," *The Annals of Statistics*, 37, 1733–1751.