Stepwise Searching for Feature Variables in High-Dimensional Linear Regression

Hongzhi An[1], Da Huang[2], Qiwei Yao[3] and Cun-Hui Zhang[4]

[1]Institute of Applied Mathematics, Chinese Academy of Sciences, Beijing 100080, China

[2]Guanghua School of Management, Peking University
Beijing 100871, China

[3]Department of Statistics, London School of Economics, London, WC2A 2AE, UK

[4]Department of Statistics, Rutgers University, Piscataway, NJ08854-8019, USA

Abstract

We investigate the classical stepwise forward and backward search methods for selecting sparse models in the context of linear regression with the number of candidate variables p greater than the number of observations n. In the noiseless case, we give definite upper bounds for the number of forward search steps to recover all relevant variables, if each step of the forward search is approximately optimal in reduction of residual sum of squares, up to a fraction. These upper bounds for the number of steps are of the same order as the size of a true sparse model under mild conditions. In the presence of noise, traditional information criteria such as BIC and AIC are designed for $p < n$ and may fail spectacularly when p is greater than n. To overcome this difficulty, two information criteria BICP and BICC are proposed to serve as the stopping rules in the stepwise searches. The forward search with noise is proved to be approximately optimal with high probability, compared with the optimal forward search without noise, so that the upper bounds for the number of steps still apply. The proposed BICP is proved to stop the forward search as soon as it recovers all relevant variables and remove all extra variables in the backward deletion. This leads to the selection consistency of the estimated models. The proposed methods are illustrated in a simulation study which indicates that the new methods outperform a counterpart LASSO selector with a penalty parameter set at a fixed value.

Keyword: adjusted information criterion, backward deletion, forward addition, BICP, BICC, consistency, least squares, sparsity, stepwise regression, variable selection, sweep operation.

1 Introduction

Modern statistical applications often encounter the situation when a regression model is fitted with the number of candidate variables (i.e. regressors) p greater or far greater than the number of available observations n. Examples where such a scenario arises include radiology and biomedical imaging, gene expression studies, signal processing and even nonparametric estimation for curve or surface based on finite number of noisy observations. Without any assumptions on the structure of such a regression model, it is impossible to fit any practically useful models. One frequently used assumption is the so-called sparsity condition which assumes that the effective contribution to a dependent variable rests on a much small number of regressors than n. The challenge then is to find those 'true' regressors from a much larger number of candidate variables. This leads to a surging interest in new methods and theory for regression model selection with $p > n$.

In this paper we revisit the classical forward and backward stepwise regression methods for model selection and adapt them to the cases with $p > n$ or $p \gg n$. Forward stepwise regression is also know as matching pursuit (Mallat and Zhang, 1993) or greedy search. An and Gu (1985, 1987) showed that a forward addition followed by a backward deletion with the stopping rules defined by, for example, the Bayesian information criterion (BIC) leads to a consistent model selection when p is fixed. However the criteria such as BIC are designed for $p < n$. They may fail spectacularly when p is greater than or even close to n, leading to excessively overfitted models. We propose two new information criteria BICP and BICC in this paper. The BICP increases the penalty to overcome overfitting; see (5) below. The BICC controls the residuals in the sense that it will stop the search before the residuals diminish to 0 as the number of selected variables increases to n (Remark 1(i) in section 2.2 below). A simulation study shows that both methods work very well even when p is as ten times large as n.

Any attempt to develop the asymptotic theory in the setting of $p > n$ has to deal with the difficulties caused by the fact that the number of the candidate models also diverges to infinity. This unfortunately makes the approach of An and Gu (1985, 1987) inapplicable. We

take a radically different road by first considering approximately optimal forward search in the noiseless case which attains within a fraction the optimal reduction of the sum of residual squares. Under mild conditions on the design variables and the regression coefficients, we provide an upper bound for approximately optimal forward search steps to recover all nonzero coefficients. The upper bound is of the optimal order of the number of nonzero coefficients when the average and minimum of the squares of the nonzero coefficients are of the same order, and is no more than the square of the optimal order in general. We then prove that the cardinality of such approximate forward search strategies for the first k steps is of much smaller order than the conventional upper bound $\binom{p}{k}$. In the presence of noise, this entropy calculation leads to much smaller Bonferroni adjustments for the noise level, so that the forward search path lie within those deterministic collections of approximately optimal rules with high probability. Furthermore, we show that with high probability, the BICP criteria stops the forward addition search as soon as it recovers all nonzero regression coefficients and then ensures the removal of all variables with zero regression coefficients in backward deletion. Although we only deal with the stepwise search methods coupled with the BICP in this paper, our proofs also provide building blocks for investigating the theoretical properties of the other search procedures such as the ℓ_2 boosting (Bühlmann, 2006).

Regression with $p > n$ is a vibrant research area in statistics at present. Recent significant developments in the estimation of regression coefficients and prediction include Greenshtein and Ritov (2004), Candés and Tao (2007), Bunea, Tsybakov and Wegkamp (2007), van de Geer (2008), Zhang and Huang (2008), Meinshausen and Yu (2009), Bickel, Ritov and Tsybakov (2009), and Zhang (2009a). Important advances have also been made in selection consistency. The sign-consistency of the LASSO (Tibshirani, 1996; Chen and Donoho, 1994) was proved by Meinshausen and Buhlmann (2006), Tropp (2006), Zhao and Yu (2006) and Wainwright (2009a). However, due to the interference of the estimation bias, these results are obtained under quite strong conditions on the feature matrix and the minimum magnitude of the unknown nonzero regression coefficients. A number of approaches have been taken to achieve selection consistency by reducing the estimation bias, including concave penalized least squares (Frank and Friedman, 1993; Fan and Li, 2001; Fan and Peng, 2004; Zhang, 2008, 2010), adaptive LASSO (Zou, 2006; Huang,

Ma and Zhang, 2008), and correlation screening (SIS; Fan and Lv, 2008). Stepwise regression, widely used for feature selection, also aims nearly unbiased selection. In this direction, Zhang (2009b) provided sufficient conditions for the selection consistency of his FoBa stepwise search algorithm. Meanwhile, Bunea (2008), Wainwright (2009b) and Zhang (2007) proved that the minimum nonzero coefficients should be in the order of $\sqrt{(\log p)/n}$ in order to achieve the variable selection consistency in linear regression.

For penalization approaches, including LASSO, SCAD (Fan and Li 2001) and MCP (Zhang, 2008), selecting new variables were effectively carried out by a series of z-tests. Therefore one has to overcome the difficulties caused by the unknown variance of the error term in the regression model. In contrast, we employ a BIC-based approach which penalizes the logarithmic SSR, selecting new variables effectively by F-tests; see (32) below. Hence we do not require the knowledge of the variance of the error term or a computationally intensive method to choose the penalty parameter.

The asymptotic properties of a different extended BIC criterion for selecting sparse models have been investigated by Chen and Chen (2008). Under the assumption that the number of true regressors remains fixed while $\log p = O(\log n)$, they shows that with probability converging to 1 all the models with j regressors have the extended BIC values greater than that of the true model, where $j > 0$ is any finite integer (Theorem 1, Chen and Chen, 2008). Our approaches do not require the number of the regressors to be fixed, and can handle much larger p than $O(n^c)$ for a fixed constant $c > 0$.

The rest of the paper is organized as follows. The new methods and the associated algorithm are presented in section 2. It also contains a heuristic approach for the consistency of the stepwise forward addition. The numerical results are presented in section 3. A formal investigation into the consistency is presented in section 4. All the proofs are collected in the Appendix.

2 Methodology

2.1 Model

Consider linear regression model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} = (y_1, \cdots, y_n)'$ is an n-vector of random responses, $\mathbf{X} \equiv (x_{ij}) \equiv (\mathbf{x}_1, \cdots, \mathbf{x}_p)$ is an $n \times p$ design matrix, $\boldsymbol{\beta}$ is a p-vector of regression coefficients, $\boldsymbol{\varepsilon} = (\varepsilon_1, \cdots, \varepsilon_n)' \sim N(0, \sigma^2\mathbf{I}_n)$, where $\sigma^2 > 0$ is an unknown but fixed constant and $\mathbf{I}_n$ denotes the $n \times n$ identity matrix. We use $\mathbf{x}_1, \cdots, \mathbf{x}_p$ to denote the column vectors of $\mathbf{X}$. In the above model, we assume the regression coefficient vector $\boldsymbol{\beta} \equiv \boldsymbol{\beta}_n = (\beta_{n,1}, \cdots, \beta_{n,p})'$ varies with n, and furthermore, the number of coefficients p goes to infinity together with n. In fact p may be greater, or much greater than n. For such a highly under-determined regression model, it is necessary to impose some regularity condition, such as the sparsity, on $\boldsymbol{\beta}$. We assume that the contribution to the response $\mathbf{y}$ is from merely a much smaller number (than p) of $\mathbf{x}_i$. Let

$$\mathcal{I}_n = \{1 \le i \le p \,:\, \beta_{n,i} \neq 0\}, \tag{2}$$

and $d \equiv d_n = |\mathcal{I}_n|$, denoting the number of elements in $\mathcal{I}_n$. We assume that d is smaller or much smaller than p, although we allow $d \to \infty$ together with n and p (at a much slower rate; see (20) below).

2.2 Algorithm

We introduce some notation first. For any subset $\mathcal{J} \subset \{1, \cdots, p\}$, let $\mathbf{X}_\mathcal{J}$ denote the $n \times |\mathcal{J}|$ matrix consisting of the columns of $\mathbf{X}$ corresponding to the indices in $\mathcal{J}$, and $\boldsymbol{\beta}_\mathcal{J}$ the $|\mathcal{J}|$-vector consisting of the components $\boldsymbol{\beta}$ corresponding to the indices in $\mathcal{J}$. Put

$$\mathbf{P}_\mathcal{J} = \mathbf{X}_\mathcal{J}(\mathbf{X}_\mathcal{J}'\mathbf{X}_\mathcal{J})^{-}\mathbf{X}_\mathcal{J}',\ \mathbf{P}_\mathcal{J}^\perp = \mathbf{I}_n - \mathbf{P}_\mathcal{J},\ L_{\mathbf{u},\mathbf{v}}(\mathcal{J}) = \mathbf{u}'\mathbf{P}_\mathcal{J}^\perp\mathbf{v},\ \ \mathbf{u}, \mathbf{v} \in \mathcal{R}^n, \tag{3}$$

i.e. $\mathbf{P}_\mathcal{J}$ is the projection matrix onto the linear space spanned by the columns of $\mathbf{X}_\mathcal{J}$, and $L_{\mathbf{y},\mathbf{y}}(\mathcal{J})$ is the sum of the squared residuals resulted from the least squares fitting $\widehat{\mathbf{y}} = \mathbf{X}_\mathcal{J}\widehat{\boldsymbol{\beta}}_\mathcal{J} = \mathbf{P}_\mathcal{J}\mathbf{y}$.

The algorithm concerned is based on a combined use of the standard stepwise addition and deletion with some adjusted information criteria. The adjustment is necessary in order to ensure

that the algorithm works even when p is (much) greater than n. The searching consists of two stages: First we start with the optimal regression set with only one regressor $\mathcal{J}_1$. By adding one variable each time, we obtain an *optimum* regression set $\mathcal{J}_k$ with k regressors for $k = 2, 3, \cdots$. The newly added variable is selected such that the decrease in the sum of the squared residuals is maximized. Note that $\mathcal{J}_k$ is not necessarily the optimal regression set with k regressors. We adopt this stepwise addition searching for its computational efficiency. The forward search continues as long as the adjusted BIC value decreases. When the forward search stops at step $\widetilde{k}$, we set $\widehat{\mathcal{I}}_{n,1} \equiv \mathcal{J}_{\widetilde{k}}$ as an initial estimator for $\mathcal{I}_n$. The second stage starts with $\mathcal{J}^*_{\widetilde{k}} = \widehat{\mathcal{I}}_{n,1}$. We delete one variable each time; obtaining an *optimum* regression set $\mathcal{J}^*_k$ for $k = \widetilde{k} - 1, \widetilde{k} - 2, \cdots$. The variable deleted at each step is specified such that the increase in the sum of the squared residuals is minimized. The backward deletion continues as long as the adjusted BIC decreases. When the backward deletion stops at $\widehat{k}$, we set $\widehat{\mathcal{I}}_{n,2} \equiv \mathcal{J}^*_{\widehat{k}}$ as the final estimator for $\mathcal{I}_n$. Note that the searching in Stage II is among $\widetilde{k}$ (instead of p) variables only, and $\widetilde{k} \leq n$ with probability 1 even when $p >> n$. In practice it is often the case that $\widetilde{k}$ is much smaller than n. The computation involved is a standard stepwise regression problem which can be implemented in an efficient manner using the standard elimination algorithms; see Remark 1(ii) below.

Stage I – Forward addition:

1. Let $\mathcal{J}_1 = \{j_1\}$, where

$$j_1 = \arg\min_{1 \leq i \leq p} L_{\mathbf{y},\mathbf{y}}(\{i\}). \tag{4}$$

   Put

$$\mathrm{BICP}_1 = \log\{L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_1)/n\} + 2\log p/n.$$

2. Continue with $k = 2, 3, \cdots$, provided $\mathrm{BICP}_k < \mathrm{BICP}_{k-1}$, where

$$\mathrm{BICP}_k = \log\{L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_k)/n\} + \frac{2k}{n}\log p. \tag{5}$$

   In the above expression, $\mathcal{J}_k = \mathcal{J}_{k-1} \cup \{j_k\}$, and

$$\begin{aligned} j_k &= \arg\max_{i \notin \mathcal{J}_{k-1}} [L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_{k-1}) - L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_{k-1} \cup \{i\})] \\ &= \arg\max_{i \notin \mathcal{J}_{k-1}} L^2_{\mathbf{y},\mathbf{x}_i}(\mathcal{J}_{k-1})/L_{\mathbf{x}_i,\mathbf{x}_i}(\mathcal{J}_{k-1}). \end{aligned} \tag{6}$$

3. For $\text{BICP}_k \geq \text{BICP}_{k-1}$, let $\widetilde{k} = k-1$, and $\widehat{\mathcal{I}}_{n,1} = \mathcal{J}_{\widetilde{k}}$.

Stage II – Backward deletion:

4. Let $\text{BICP}^*_{\widetilde{k}} = \text{BICP}_{\widetilde{k}}$ and $\mathcal{J}^*_{\widetilde{k}} = \widehat{\mathcal{I}}_{n,1}$.

5. Continue with $k = \widetilde{k}-1$, $\widetilde{k}-2, \cdots$, provided $\text{BICP}^*_k \leq \text{BICP}^*_{k+1}$, where

$$\text{BICP}^*_k = \log\{L_{\mathbf{y},\mathbf{y}}(\mathcal{J}^*_k)/n\} + \frac{2k}{n}\log p. \tag{7}$$

In the above expression, $\mathcal{J}^*_k = \mathcal{J}^*_{k+1} \setminus \{j_k\}$, and

$$j_k = \arg\min_{i \in \mathcal{J}^*_{k+1}} [L_{\mathbf{y},\mathbf{y}}(\mathcal{J}^*_{k+1} \setminus \{i\}) - L_{\mathbf{y},\mathbf{y}}(\mathcal{J}^*_{k+1})].$$

6. For $\text{BICP}^*_k > \text{BICP}^*_{k+1}$, let $\widehat{k} = k+1$, and $\widehat{\mathcal{I}}_{n,2} = \mathcal{J}^*_{\widehat{k}}$.

Remark 1. (i) In the above algorithm, the criterion BICP (abbreviating for BIC modified for the cases with large p) is used, which replaces the penalty $\log n/n$ in the standard BIC by $2\log p/n$ and is designed for the cases $p \approx n$ or $p > n$. One alternative is to use the BICC (abbreviating for BIC with an added constant) defined as follows:

$$\text{BICC}_k = \log\{L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_k)/n) + c_0\} + \frac{k}{n}\log n, \tag{8}$$

where $c_0 > 0$ is a constant. Note that BICC uses exactly the same penalty term $\log n/n$ as in the standard BIC. The only difference is to insert a positive constant c_0 in the logarithmic function. This modification is necessary when p is greater than n. Note that for k sufficiently close to n, $L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_k)$ is very close to 0. Therefore

$$\log\{L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_{k-1})\} - \log\{L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_k)\} \approx \{L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_{k-1}) - L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_k)\}/L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_k)$$

may be very large even when the decrease in residual $L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_{k-1}) - L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_k)$ is negligible. Inserting c_0 overcomes this problem. In practical implementation, we may simply set c_0 equal to 0.2 times of the sample variance of y. Although the theoretical properties of the BICC for $p \geq n$ are unknown, our stimulation study shows that it outperforms the BICP. We also note that when p is fixed (as $n \to \infty$), the asymptotic properties of BIC established by An and Gu (1985, 1987) also apply to the BICC.

(ii) The stepwise addition and deletion may be implemented in terms of the so-called sweep operation. For the stepwise addition in Stage I, we set $\mathbf{L}^0 = (\mathbf{X}, \mathbf{y})'(\mathbf{X}, \mathbf{y}) \equiv (\ell^0_{i,j})$ which is a $(p+1) \times (p+1)$ matrix. Adding one variable, say, $\mathbf{x}_i$, in the k-th step corresponds to transfer $\mathbf{L}^{k-1} = (\ell^{k-1}_{i,j})$ to $\mathbf{L}^k = (\ell^k_{i,j})$ by the sweep operation:

$$\ell^k_{i,i} = 1/\ell^{k-1}_{i,i}, \qquad \ell^k_{j,m} = \ell^{k-1}_{j,m} - \ell^{k-1}_{i,m}\ell^{k-1}_{j,i}/\ell^{k-1}_{i,i} \quad \text{for } j \neq i \text{ and } m \neq i,$$
$$\ell^k_{i,j} = \ell^{k-1}_{i,j}/\ell^{k-1}_{i,i} \quad \text{and} \quad \ell^k_{j,i} = -\ell^{k-1}_{j,i}/\ell^{k-1}_{i,i} \quad \text{for } j \neq i.$$

Then

$$L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_{k-1}) - L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_{k-1} \cup \{i\}) = \left(\ell^{k-1}_{i,p+1}\right)^2/\ell^{k-1}_{i,i}, \qquad i \notin \mathcal{J}_{k-1}.$$

For the stepwise deletion in Stage II, the same sweep operation applies with the initial matrix $\mathbf{L}^0 = \mathbf{L}^{\widetilde{k}}$ obtained in Stage I. For $k = \widetilde{k} - 1, \widetilde{k} - 2, \cdots,$

$$L_{\mathbf{y},\mathbf{y}}(\mathcal{J}^*_{k+1} \setminus \{i\}) - L_{\mathbf{y},\mathbf{y}}(\mathcal{J}^*_{k+1}) = \left(\ell^{\widetilde{k}-k+1}_{i,p+1}\right)^2/\ell^{\widetilde{k}-k+1}_{i,i}, \qquad i \in \mathcal{J}^*_{k+1}.$$

(iii) It always holds that $\text{rank}(\mathbf{X}_{\mathcal{J}_{\widetilde{k}}}) = \widetilde{k}$. As $L_{\mathbf{x}_i,\mathbf{x}_i}(\mathcal{J}_{k-1}) = 0$ for any $\mathbf{x}_i$ in the linear space spanned by the columns of $\mathbf{X}_{\mathcal{J}_{k-1}}$, adding such an $\mathbf{x}_i$ to $\mathcal{J}_{k-1}$ will not reduce the sum of the squared residuals and, therefore, will only increase the BICP (or BICC) value. Hence the new entry to $\mathcal{J}_k$, for $k \leq \widetilde{k}$, must not be in the linear space spanned by the columns of $\mathbf{X}_{\mathcal{J}_{k-1}}$. Furthermore, $\widetilde{k} \leq n$ with probability 1.

(iv) In the forward search, if it happens that $L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_{k-1}) - L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_{k-1} \cup \{i\})$ is practically 0 for some $i \notin \mathcal{J}_{k-1}$, $\mathbf{x}_i$ should be excluded from the further search. This is more likely to happen when $p >> n$, and then the elimination would improve the computation efficiency.

(v) When $p \geq n$, the true mean $\boldsymbol{\mu} = \mathbf{X}_{\mathcal{I}_n}\boldsymbol{\beta}_{\mathcal{I}_n} = \mathbf{X}\boldsymbol{\beta}$ may be represented as linear combinations of any full-ranked $n \times n$ submatrix of $\mathbf{X}$. There is a possibility in theory, although unlikely in practice, that our forward stepwise addition estimator $\widehat{\mathcal{I}}_{n,1}$ misses the majority of the members in $\mathcal{I}_n$. In practice, we may start the forward search based on the subset of j regressors with the *optimal* fit, where $j \geq 1$ is a small integer. This should further reduce the small probability of the event that $\widehat{\mathcal{I}}_{n,1}$ ends as a non-sparse set.

2.3 Performance of the forward search: heuristics

Before presenting the formal asymptotic results in section 4 below, we first study the performance of the forward search.

Given $\mathcal{J}_{k-1}$, the objective of the k-th step of the forward search is to find an index $j = j_k$ with large $\mu_{j,k} \equiv \|\mathbf{P}_{\mathcal{J}_{k-1}\cup\{j\}}\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\boldsymbol{\mu}\|$. Theorem 1 below states that the forward search finds all j with $\beta_j \neq 0$ within $k_1^* \wedge k_2^*$ steps, if $\mu_{j_k,k}$ is within a γ fraction of the maximum $\mu_k^* \equiv \max_j \mu_{j_k,k}$, where $\gamma \in [0,1)$ is a constant, and k_1^* and k_2^* are positive integers satisfying

$$d_n \log\Big(\frac{e\|\boldsymbol{\mu}\|^2}{c_* d_n n\beta_*^2}\Big) + 1 + \log\sqrt{d_n/(2\pi)} \leq (1-\gamma)^2 \sum_{k=1}^{k_1^*} \psi_{k-1}, \tag{9}$$

and

$$\min\Big\{\frac{\|\boldsymbol{\mu}\|^2}{(1-\gamma)^2 n\beta_*^2}, \frac{d_n(d_n+1)}{2(1-\gamma)^2}\Big\} \leq \sum_{k=1}^{k_2^*} \psi_{k-1}^2. \tag{10}$$

In the above expressions, $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, $d_n = |\mathcal{I}_n|$ (see (2)), $\beta_* = \min\{j \in \mathcal{I}_n : |\beta_j|\}$, $0 < c_* \leq \psi_{k_1^* \wedge k_2^* - 1}$, $\phi_{\min}(\mathcal{J})$ denotes the minimum eigenvalue of $\mathbf{X}'_{\mathcal{J}}\mathbf{X}_{\mathcal{J}}/n$, and

$$\psi_k = \min\{\phi_{\min}(\mathcal{J} \cup \mathcal{I}_n) : |\mathcal{J}| \leq k\}. \tag{11}$$

Theorem 1 *Let $\mathcal{J}_k = \mathcal{J}_{k-1} \cup \{j_k\}$ be a sequence of models satisfying $\mu_{j_k,k} \geq (1-\gamma)\mu_k^*$. Then, $\mathcal{J}_k \supseteq \mathcal{I}_n$ for a certain $k \leq k^* = k_1^* \wedge k_2^*$.*

If ψ_k does not vanish too fast, Theorem 1 asserts the upper bound $k^* = O(d_n)$ when $\|\boldsymbol{\mu}\|^2 \asymp d_n n\beta_*^2$ and $k^* = O(d_n^2)$ in the worst case. In sparse recovery where $\mathbf{y} = \boldsymbol{\mu}$, the forward search attains the optimal μ_k^* in each step, so that $\gamma = 0$ in Theorem 1. For $\boldsymbol{\varepsilon} \neq 0$, Theorem 1 allows sub-optimal choices up to an error of $\gamma\mu_k^*$. To this end, the essential difficulty is to deal with the fact that the number of candidate models goes to infinity. Our idea is to find a series of collections $\mathscr{C}_k$ of deterministic models such that (i) the cardinality $|\mathscr{C}_k|$ diverges not too fast, and (ii) the forward search selects models in these collections, $\mathcal{J}_k \in \mathscr{C}_k$, with high probability. Note that the identification of those collections of deterministic models is purely for the purpose of our theoretical investigation, which is not required in the implementation of our stepwise search.

The natural choice of the $\mathscr{C}_k$ based on Theorem 1 is

$$\mathscr{C}_k = \Big\{\{j_1, \ldots, j_k\} : \mu_{j_k,k} \geq (1-\gamma)\mu_k^* > 0\Big\},\ k \geq 1, \tag{12}$$

where $\mu_{j,k} = \|\mathbf{P}_{\{j_1,\ldots,j_{k-1},j\}}\mathbf{P}^{\perp}_{\{j_1,\ldots,j_{k-1}\}}\boldsymbol{\mu}\|$ and $\mu^*_k = \max_{j\le p}\mu_{j,k}$. Since

$$1 - P\Big\{\mathcal{J}_k \in \mathscr{C}_k\ \forall \boldsymbol{\mu}^*_k > 0\Big\} \le \sum_{k=1}^{d^*_n} P\Big\{\mathcal{J}_{k-1} \in \mathscr{C}_{k-1},\ \mu_{j_k,k} < (1-\gamma)\mu^*_k\Big\}, \tag{13}$$

the probability calculation with step k only involves $|\mathscr{C}_{k-1}|(p-k+1)$ combinations of $\{\mathcal{J}_{k-1}, j_k\}$. This could be much smaller that the cardinality $\binom{p}{k}k!$ for the collection of all possible realizations of $\mathcal{J}_k$.

3 Numerical properties

We illustrate the proposed methods by two simulated examples. For BICC, we set $c_0 = 0.2s^2_y$, where s^2_y denotes the sample variance of y. For the comparison purpose, we also include three other model selection methods: a version of LASSO, the extended BIC (EBIC) of Chen and Chen (2008), and the greedy forward-backward search (FoBa) of Zhang (2009b).

A LASSO estimator is defined as the minimizer of the function

$$\frac{1}{2n}||\mathbf{y} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda\sum_{j=1}^{p}|\beta_{n,j}|. \tag{14}$$

To make it comparable with our BICP stepwise selector, we set $\lambda = \sigma\sqrt{2(\log p)/n}$ and standardize the data such that $||\mathbf{x}_j|| = \sqrt{n}$ for all j, where σ^2 is the true value of $\mathrm{Var}(\varepsilon_i)$, and is only known in simulation. It can be seen from (31) and Lemma 3 below, the BICP adds a new variable by performing an $F_{1,n-k-1}$ test at the threshold $(n-k-1)\{e^{2\log p/n}-1\} \approx 2\log p$, while the LASSO selects a new variable $\mathbf{x}_j$ by performing approximately a z-test $\chi^2_1 > n^2\lambda^2/\{||\mathbf{x}_j||^2\sigma^2\} = 2\log p$. As $F_{1,q} \approx \chi^2_1$ for large q, the two methods are approximately comparable.

The EBIC proposed by Chen and Chen (2008) represents an alternative extension of the classical BIC. Instead of BICP and BICC defined in (5) and (8) respectively, it uses the information criterion

$$\mathrm{EBIC}_k = \log\{L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_k)/n\} + k\log n/n + 2k\log p/n,$$

which adds an additional penalty term $2\log p/n$ in dealing with the case $p > n$.

Different from the method of forward addition followed by backward deletion, the FoBa advocated by Zhang (2009b) performs (as many as possible) backward deletions after each forward

Table 1: Simulation results for Example 1 with $n = 200$ and $\varepsilon_i \sim N(0,1)$: Means and standard deviations of $|\widehat{d} - d|$ and the relative error rate $\widehat{r}$.

			$\lvert\widehat{d} - d\rvert$		$\widehat{r}$	
p	d	Method	Mean	STD	Mean	STD
1000	10	BICC	0.0750	0.2641	0.0034	0.0120
		BICP	0.5700	1.5417	0.0210	0.0428
		EBIC	0.1350	0.8368	0.0048	0.0230
		FoBa+BICC	0.0000	0.0000	0.0000	0.0000
		FoBa+BICP	0.0900	0.3039	0.0041	0.0136
		FoBa+EBIC	0.0250	0.1565	0.0011	0.0071
		FoBa	0.1000	0.3170	0.0056	0.0181
		LASSO	4.5900	2.8181	0.1462	0.0600
1000	25	BICC	0.1900	0.4527	0.0036	0.0086
		BICP	1.3750	2.6646	0.0252	0.0623
		EBIC	8.4550	9.1859	0.9933	1.9555
		FoBa+BICC	0.0150	0.1578	0.0003	0.0034
		FoBa+BICP	0.5000	1.6442	0.0138	0.0922
		FoBa+EBIC	9.1600	9.3020	1.0680	1.9751
		FoBa	0.8350	1.1810	0.0186	0.0274
		LASSO	26.3350	8.5044	0.2510	0.0394
2000	10	BICC	0.1800	0.4456	0.0080	0.0195
		BICP	0.6750	1.7506	0.0244	0.0467
		EBIC	0.1550	0.8273	0.0057	0.0248
		FoBa+BICC	0.0000	0.0000	0.0000	0.0000
		FoBa+BICP	0.0300	0.1710	0.0014	0.0078
		FoBa+EBIC	0.0050	0.0707	0.0002	0.0032
		FoBa	0.1500	0.3717	0.0084	0.0210
		LASSO	7.0500	3.4185	0.1955	0.0583
2000	25	BICC	0.4300	0.7668	0.0080	0.0139
		BICP	2.4500	4.1283	0.1313	0.8888
		EBIC	14.5050	8.7293	2.1914	2.7699
		FoBa+BICC	0.0100	0.0997	0.0002	0.0021
		FoBa+BICP	1.0350	3.4530	0.1071	0.8903
		FoBa+EBIC	15.3800	8.4619	2.3482	2.8333
		FoBa	0.9150	1.2021	0.0206	0.0286
		LASSO	45.5750	11.8459	0.3211	0.0334

addition, and stops when no more variables can be added or deleted. More precisely, the FoBa

Table 2: Simulation results for Example 1 with $n = 800$ and $\varepsilon_i \sim N(0,1)$: Means and standard deviations of $|\widehat{d} - d|$ and the relative error rate $\widehat{r}$.

			$\lvert\widehat{d} - d\rvert$		$\widehat{r}$	
p	d	Method	Mean	STD	Mean	STD
10000	25	BICC	0.0850	0.2796	0.0016	0.0054
		BICP	0.2200	0.4719	0.0042	0.0089
		EBIC	0.0100	0.0997	0.0002	0.0019
		FoBa+BICC	0.0000	0.0000	0.0000	0.0000
		FoBa+BICP	0.0050	0.0707	0.0001	0.0016
		FoBa+EBIC	0.0000	0.0000	0.0000	0.0000
		FoBa	0.1150	0.3639	0.0024	0.0077
		LASSO	6.3750	3.0596	0.0980	0.0374
10000	40	BICC	0.0900	0.2869	0.0011	0.0035
		BICP	0.4050	0.9673	0.0048	0.0107
		EBIC	0.0200	0.1404	0.0002	0.0017
		FoBa+BICC	0.0000	0.0000	0.0000	0.0000
		FoBa+BICP	0.0600	0.2583	0.0007	0.0031
		FoBa+EBIC	0.0100	0.0997	0.0004	0.0012
		FoBa	0.2400	0.5037	0.0031	0.0066
		LASSO	20.0800	6.8350	0.1630	0.0366
20000	25	BICC	0.1000	0.3170	0.0019	0.0061
		BICP	0.2300	0.5083	0.0044	0.0096
		EBIC	0.0150	0.1219	0.0003	0.0023
		FoBa+BICC	0.0000	0.0000	0.0000	0.0000
		FoBa+BICP	0.0100	0.0997	0.0002	0.0019
		FoBa+EBIC	0.0000	0.0000	0.0000	0.0000
		FoBa	0.0900	0.3039	0.0019	0.0064
		LASSO	9.9350	4.8205	0.1360	0.0460
20000	40	BICC	0.1950	0.4335	0.0024	0.0053
		BICP	0.4950	0.9873	0.0058	0.0112
		EBIC	0.0150	0.1219	0.0002	0.0015
		FoBa+BICC	0.0000	0.0000	0.0000	0.0000
		FoBa+BICP	0.0200	0.1404	0.0002	0.0017
		FoBa+EBIC	0.0000	0.0000	0.0000	0.0000
		FoBa	0.1900	0.4299	0.0024	0.0056
		LASSO	32.8850	10.9816	0.2199	0.0385

adds a new variable to the selected model $\mathcal{J}$ if

$$\max_{i \notin \mathcal{J}} \left\{ L_{\mathbf{y},\mathbf{y}}(\mathcal{J}) - L_{\mathbf{y},\mathbf{y}}(\mathcal{J} \cup \{i\}) \right\} > \epsilon, \tag{15}$$

Table 3: Simulation results for Example 1 with $\varepsilon_i \sim N(0, \sigma^2)$: Means and standard deviations of $|\widehat{d} - d|$ and the relative error rate, where $n = 800$, $p = 10000$ and $d = 25$.

		$\lvert\widehat{d} - d\rvert$		$\widehat{r}$	
σ^2	Method	Mean	STD	Mean	STD
4	BICC	0.2800	0.5596	0.0063	0.0133
	BICP	0.2900	0.5266	0.0067	0.0123
	EBIC	0.4150	0.6893	0.0088	0.0150
	FoBa+BICC	0.1500	0.4103	0.0032	0.0087
	FoBa+BICC.1	0.6750	0.8620	0.0146	0.0192
	FoBa+BICC.2	2.1850	1.4286	0.0501	0.0358
	FoBa+BICP	0.1050	0.3530	0.0022	0.0075
	FoBa+EBIC	0.4150	0.6966	0.0089	0.0152
	FoBa	0.3550	0.6566	0.0076	0.0143
	LASSO	5.2300	2.6328	0.0909	0.0348
9	BICC	1.4850	1.4283	0.0569	0.0370
	BICP	1.7850	1.3520	0.0485	0.0358
	EBIC	3.5600	1.9430	0.0884	0.0553
	FoBa+BICC	1.4400	1.2345	0.0331	0.0294
	FoBa+BICC.1	2.4700	1.4245	0.0571	0.0363
	FoBa+BICC.2	4.2000	1.8045	0.1056	0.0543
	FoBa+BICP	1.9500	1.3846	0.0445	0.0341
	FoBa+EBIC	3.5850	1.9524	0.0888	0.0561
	FoBa	0.8950	0.8471	0.0272	0.0236
	LASSO	3.3850	2.2963	0.1262	0.0448
16	BICC	4.7800	3.6508	0.1648	0.0616
	BICP	4.8000	2.2551	0.1383	0.0753
	EBIC	7.7000	2.6581	0.2421	0.1272
	FoBa+BICC	3.5450	1.8343	0.0893	0.0526
	FoBa+BICC.1	4.7950	1.9780	0.1253	0.0633
	FoBa+BICC.2	6.5700	2.0728	0.1872	0.0813
	FoBa+BICP	5.1050	2.2268	0.1367	0.0766
	FoBa+EBIC	7.7500	2.6764	0.2440	0.1282
	FoBa	8.3100	3.2133	0.1564	0.0403
	LASSO	2.4900	1.9231	0.1887	0.0588
25	BICC	10.7500	6.0623	0.2811	0.0687
	BICP	8.0650	2.6939	0.2691	0.1269
	EBIC	11.3100	2.8574	0.4569	0.2153
	FoBa+BICC	6.1250	2.2054	0.1775	0.0839
	FoBa+BICC.1	7.5050	2.2751	0.2277	0.0988
	FoBa+BICC.2	8.9500	2.3782	0.2967	0.1238
	FoBa+BICP	8.3100	2.5389	0.2677	0.1266
	FoBa+EBIC	11.3700	2.8555	0.4611	0.2192
	FoBa	29.0550	6.1220	0.2998	0.0310
	LASSO	4.3400	2.7718	0.2734	0.0851

where $\epsilon > 0$ is a prescribed constant. After adding $j = \arg\max_{i \notin \mathcal{J}} \{L_{\mathbf{y},\mathbf{y}}(\mathcal{J}) - L_{\mathbf{y},\mathbf{y}}(\mathcal{J} \cup \{i\})\}$ to model $\mathcal{J}$, the FoBa deletes a variable if

$$\min_{i \in \mathcal{J}^\star} \frac{L_{\mathbf{y},\mathbf{y}}(\mathcal{J}^\star \backslash \{i\}) - L_{\mathbf{y},\mathbf{y}}(\mathcal{J}^\star)}{L_{\mathbf{y},\mathbf{y}}(\mathcal{J}) - L_{\mathbf{y},\mathbf{y}}(\mathcal{J}^\star)} < \nu, \tag{16}$$

where $\mathcal{J}^\star = \mathcal{J} \cup \{j\}$, and $\nu \in (0,1)$ is a prescribed constant. In our implementation below, we set $\epsilon = 9.766 \log(2p)/n$ and $\nu = 0.5$.

While the idea of deleting all the redundant variables after each addition is appealing and may improve the search, it is computationally more time-consuming than the algorithm presented in section 2.2 with one-way forward search followed by one-way backward deletion, although the difference is less substantial when p is large or very large, as then the computation of the initial matrix $\mathbf{L}^0$ (see Remark 1(ii)) contributes a major part of the computing time for those greedy algorithms.

Example 1. First we consider model (1) with all x_{ij} and ε_i independent $N(0,1)$. We let sample size $n = 200$ or 800. For $n = 200$, we set the number of regression variables $p = 1000$ or 2000, and the number of non-zero coefficients $d = 10$ or 25. For $n = 800$, we set $p = 10000$ or 20000, and $d = 25$ or 40. The non-zero regression coefficients are of the form $(-1)^u(b + |v|)$, where $b = 2.5\sqrt{2 \log p/n}$, u is a Bernoulli random variable with $P(u = 1) = P(u = 0) = 0.5$, and $v \sim N(0,1)$. For each setting, we replicate the simulation 200 times.

The simulation results are reported in Figure 1 which plots the selected numbers of regression variables from the 200 replications in the ascending order. For BICP, BICC and EBIC, both $\widetilde{d}$, the number of variables selected by the forward search, and $\widehat{d}$, the number of variables selected by the backward search, are plotted together. By the definitions, it holds that $\widetilde{d} \geq \widehat{d}$, though $\widetilde{d} = \widehat{d}$ in most the replications. Note both LASSO and FoBa only produce one estimated model.

Figure 1 indicates that the algorithm with the BICP works well in the sense that $\widehat{d} = d$ in most replications, especially when $n = 800$. Both BICC and FoBa provide better performance in every settings. When $n = 200$ and $d = 25$, EBIC tends to select too fewer covariates due to the heavier penalty (than BICP) when the sample size $n = 200$. But for large samples with $n = 800$, BICC, EBIC and FoBa perform about equally well. However the version of LASSO employed turned out not competitive with the other four methods, although it has the advantage for knowing σ which is not required by the other methods.

To have a fair comparisons on the different stopping rules used in the greedy searches, Figure 1 also includes the results from the using the FoBa algorithm but with both the stopping rules (15) and (16) replaced by BICP, BICC or EBIC. Now the FoBa coupled with BICC seems to further improve the performance. It is also clear that EBIC does not work as well as the other stopping rules for small sample size $n = 200$.

Tables 1 and 2 present the means and the standard deviations of the absolute difference $|\widehat{d} - d|$ in the 200 replications for the different estimation methods with $n = 200$ and 800 respectively. Also listed are the means and the standard deviations of a relative error of a fitted model defined as

$$\begin{aligned} \widehat{r} &= (\text{number of selected wrong variables} \\ &+ \text{number of unselected true variables})/(2\,\widehat{d}). \end{aligned}$$

Similar to the pattern in Figure 1, BICP, BICC and FoBa provided comparable performances while BICC is slightly better than FoBa, and BICP is slightly worse than FoBa. EBIC did not performed well when $n = 200$. The improvement from using BICC as the stopping rule instead of (15) and (16) in FoBa is also noticeable. However it is more striking that the LASSO estimation so defined is not competitive in comparison with all the other greedy search methods, in spite of its computational efficiency. Note that using $\lambda = \sigma\sqrt{(\log n)/n}$ in (14) leads to poorer estimates.

To examine the robustness with respect to the signal-to-noise ratio, we repeated the simulation with $\sigma^2 = \mathrm{Var}(\varepsilon_i) = 4, 9, 16$ and 25. To save the space, we only report the results from the setting $(n, p, d) = (800, 10000, 25)$; see Figure 2 and Table 3. As we would expect, the methods penalizing log(RSS) such as BICP and EBIC are robust against the increase of σ^2. The methods penalizing RSS directly such as FoBa are less so. The version of LASSO method appeared to work well. This is an artifact as we used $\lambda = \sigma\sqrt{2(\log p)/n}$ in (14) in the simulation with σ^2 being the true value of $\mathrm{Var}(\varepsilon_i)$. It is noticeable that BICC is less robust that BICP, as the choice of c_0 should depend on σ^2; see (8).

Example 2. We use the same setting as in Example 1 with the added dependence structure as follows: for any $1 \le k \le n$ and $1 \le i \ne j \le d$,

$$\mathrm{Corr}(X_{ki}, X_{kj}) = (-1)^{u_1}(0.5)^{|i-j|}, \qquad \mathrm{Corr}(X_{ki}, X_{k,i+d}) = (-1)^{u_2}\rho,$$

$$\mathrm{Corr}(X_{ki}, X_{k,i+2d}) = (-1)^{u_3}(1-\rho^2)^{1/2},$$

where ρ is drawn from the uniform distribution on the interval $[0.2, 0.8]$, and u_1, u_2 and u_3 are independent and are of the same distribution as u in Example 1. All x_{ki}, for $i > 3d$, are independent. We assume that the first d regression variables have the non-zero coefficients. The simulation results are depicted in Figure 3 which shows the similar pattern to that of Figure 1, although the performance is hampered slightly by the dependence among the regressors.

4 Main theoretical result

Theorem 2 below shows that with probability converging to 1, the estimator $\widehat{\mathcal{I}}_{n,1}$ from the forward addition alone contains the true model $\mathcal{I}_n$, and with a backward deletion following the forward addition, $\widehat{\mathcal{I}}_{n,2}$ is a consistent estimator for $\mathcal{I}_n$.

We introduce some notation first. Let $\phi_{\min}(\mathcal{J})$ and $\phi_{\max}(\mathcal{J})$ denote, respectively, the minimum and the maximum eigenvalues of $\mathbf{X}'_{\mathcal{J}}\mathbf{X}_{\mathcal{J}}/n$. Let ψ_k be as in (11) and define

$$\phi_k = \min_{|\mathcal{J}|=k} \phi_{\min}(\mathcal{J}), \quad \phi_k^* = \max_{|\mathcal{J}|=k} \phi_{\max}(\mathcal{J}), \tag{17}$$

$$m_k = \inf\left\{m : d_n\phi_m^*/m < (1-\gamma)^2\phi_{k-1}\psi_k\right\}. \tag{18}$$

We will prove in the Appendix that the cardinality of the collection $\mathscr{C}_k$ in (12) is bounded by $|\mathscr{C}_k| \le \prod_{j=1}^{k}(m_k - 1)$. Let $||\mathbf{x}_j||^2 = n$ for all $1 \le j \le p$ and $\{\mathcal{I}_n, \boldsymbol{\mu}, d_n, \beta_*, c_*, \gamma, k_1^*, k_2^*\}$ be as in (9) and (10) with a fixed $c_* > 0$. Some regularity conditions are now in order.

C1 (Conditions on d_n, p, β_* and $k^* = k_1^* \wedge k_2^*$)

$$\beta_* \ge M_0\sigma\sqrt{2(\log p)/n}, \quad \textstyle\sum_{k=1}^{k^*} \log m_k \ \le\ \eta_1 \log p, \tag{19}$$

with $M_0 \ge (\sqrt{1+\eta_1} + \sqrt{\eta_1})/(c_*\gamma)$ and $\eta_1 > 0$, and

$$d_n \le M_1 n/(2\log p),\ k^*(\log p)/n = O(1),\ (k^* + \log p)/n \to 0, \tag{20}$$

with $c_* M_0^2 M_1 < (c_* M_0 - \sqrt{\eta_1})^2/(1+\eta_0) - 1$.

C2 (Adjustment for BICP) The estimators $\widehat{\mathcal{I}}_{1,n}$ and $\widehat{\mathcal{I}}_{2,n}$ are defined as in section 2.2 but with BICP_k adjusted by a factor $(1+\eta_0)$ to

$$\text{BICP}_k = \log\{L_{\mathbf{y},\mathbf{y}}(\mathcal{J}_k)/n\} + 2k(1+\eta_0)(\log p)/n, \tag{21}$$

and BICP_k^* adjusted in the same manner.

C3 (Additional condition on k^* for backward deletion.)

$$\log(k^*!) \leq \eta_2 \log p \quad \text{with} \quad \eta_1 + \eta_2 \leq \min\{1+\eta_0, c_*(M_0/2)^2\}. \tag{22}$$

Theorem 2 *Suppose conditions C1 and C2 hold with $\eta_0 \geq \eta_1$. Then,*

$$P\Big\{\widehat{\mathcal{I}}_{n,1} \supset \mathcal{I}_n,\ \widetilde{k} < k^*,\ |\widehat{\sigma}_{n,1} - \sigma| \leq \epsilon\Big\} \to 1\ \forall\ \epsilon > 0, \tag{23}$$

where $\widehat{\sigma}^2_{n,1} = \|\mathbf{P}^{\perp}_{\widehat{\mathcal{I}}_{n,1}}\mathbf{y}\|^2/(n - |\widehat{\mathcal{I}}_{n,1}|)$. If in addition condition C3 hold, then

$$P\Big\{\widehat{\mathcal{I}}_{n,2} = \mathcal{I}_n\Big\} \to 1 \tag{24}$$

and the efficient estimation of $\boldsymbol{\mu}$, $\boldsymbol{\beta}$ and σ^2 is attained with

$$\widehat{\boldsymbol{\mu}} = \mathbf{P}_{\widehat{\mathcal{I}}_{n,2}}\mathbf{y}, \quad \widehat{\boldsymbol{\beta}} = (\mathbf{X}'_{\widehat{\mathcal{I}}_{n,2}}\mathbf{X}_{\widehat{\mathcal{I}}_{n,2}})^{-1}\mathbf{X}'_{\widehat{\mathcal{I}}_{n,2}}\mathbf{y}, \quad \widehat{\sigma}^2_n = \frac{\|(\mathbf{I}_n - \mathbf{P}_{\widehat{\mathcal{I}}_{n,2}})\mathbf{y}\|^2}{n - |\widehat{\mathcal{I}}_{n,2}|}. \tag{25}$$

Remark 2. (i) The sparse Riesz condition (Zhang and Huang, 2008) asserts

$$c_* \leq \phi_{d_n^*} \leq \phi^*_{d_n^*} \leq c^* \tag{26}$$

with fixed $0 < c_* < c^* < \infty$ and $d_n^* \to \infty$, which weakens the restricted isometry condition (Candés and Tao 2005) by allowing $c^* - 1 \neq 1 - c_*$. Such conditions are often used in the 'large p and and small n' literature. Random matrix theory provides (26) with $d_n^* \log(p/d_n^*) = a_0 n$ for fixed $\{c_*, c^*\}$ and certain small a_0, allowing $p \gg n$. Under (26), $k^* = k_1^* \leq d_n^*$ in (9) for the vector $\boldsymbol{\beta}$ when

$$d_n \log\Big(\frac{e\|\boldsymbol{\mu}\|^2}{c_* d_n n \beta_*^2}\Big) + 1 + \log\sqrt{d_n/(2\pi)} \leq c_*(d_n^* - d_n)$$

and for $k^* + d_n \leq d_n^*$,

$$m_k \leq d_n c^*/\{(1-\gamma)^2 c_*^2\} + 1.$$

Thus, (26) can be viewed as a crude sufficient condition to guarantee manageable growth of k^* and $|\mathscr{C}_k|$ in conditions *C1* and *C3*, with $\max_{k\le k^*} m_k = O(d_n)$ and $k^* = O(d_n)$ in the case of $\|\boldsymbol{\mu}\|^2 = O(d_n n\beta_*^2)$. Under such scenarios, the main sparsity requirement of Theorem 2 is $d_n \log d_n = O(\log p)$.

(ii) Condition *C1* requires that the non-zero regression coefficient in model (1) be at least of the order $O\big(\{(\log p)/n\}^{1/2}\big)$, the smallest order possible for consistent variable selection (Wainwright, 2009b; Zhang, 2007).

(iii) To prove the consistency in (23) and (24), we need to increase the penalty level in the BICP by a fraction η_0 as in condition *C2*. Meanwhile, conservative Bonferroni estimates of multiple testing errors are used in all stages of the proofs of our theorems. For highly sparse $\boldsymbol{\beta}$, we prove in Theorem 3 that the conclusions of Theorem 2 holds for the simple BICP (5). Thus, it is reasonable to use the slightly more aggressive (5) for both the forward and backward stopping rules. Our simulation results, not reported here, also conforms with this recommendation.

Theorem 2 is proved in the appendix by showing that with high probability, (a) the forward search stays within $\mathscr{J}_k \in \mathscr{C}_k$ before it finds all variables with $\beta_j \neq 0$, (b) BICP stops the forward search as soon as all the variables with $\beta_j \neq 0$ are found, and (c) the backward deletion deletes exactly all falsely discovered variables in the forward addition. The fact (b) is of independent interest, although it is not a consequence of Theorem 2.

For highly sparse $\boldsymbol{\beta}$ satisfying $d_n \log d_n \le (\log\log p)/2$, a modification of Theorem‘2 provides the selection consistency of the simpler and more explicit BICP in (5). We state this result in the theorem below. Note that when d_n is bounded, the conditions *C1*, *C3* and (27) below are easily fulfilled.

Theorem 3 *Suppose condition C1 holds with* $\eta_0 = 0$ *and*

$$\sum_{k=1}^{k^*} |\mathscr{C}_k| \ll \sqrt{\log p}. \tag{27}$$

Then, (23) holds for the BICP$_k$ in (5). If in addition condition C3 holds, then (24) and (25) hold for the BICP$_k$ in (5).

Although we focus in this paper on the BICP criterion, the methods developed for the proofs

can be used for investigating a number of related forward search procedures and also, for example, the ℓ_2- boosting of Bühlmann (2006).

5 Appendix

Here we prove Theorems 1, 2 and 3. We introduce technical lemmas and their proofs as needed. The proof of Theorem 1 uses the following lemma to derive lower bounds for the reduction of residual sum of squares in an approximate forward search.

Lemma 1 *Let ψ_k be as in (11), β_* in (9), $\mu_{j,k} = \|\mathbf{P}_{\mathcal{J}_{k-1}\cup\{j\}}\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\boldsymbol{\mu}\|$ with a certain $\mathcal{J}_{k-1}$ of size $|\mathcal{J}_{k-1}| = k-1$, and $\mu^*_k = \max_j \mu_{j,k}$. Then,*

$$|\mathcal{I}_n \setminus \mathcal{J}_{k-1}|(\mu^*_k)^2 \ge \psi_{k-1}\|\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\boldsymbol{\mu}\|^2 \ge \psi^2_{k-1} n \textstyle\sum_{j\in\mathcal{I}_n\setminus\mathcal{J}_{k-1}}\beta_j^2. \tag{28}$$

In particular, for $\mathcal{I}_n \setminus \mathcal{J}_{k-1} \neq \emptyset$, $\mu^*_k \ge \psi_{k-1}\sqrt{n}\beta_*$.

Proof of Lemma 1. Let $\mathcal{A} = \mathcal{I}_n \cup \mathcal{J}_{k-1}$ and $\mathcal{B} = \mathcal{I}_n \setminus \mathcal{J}_{k-1}$. By (11), $\phi_{\min}(\mathcal{A}) \ge \psi_{k-1}$. Since $\mu_{j,k} = 0$ for $j \in \mathcal{J}_{k-1}$ and $\|\mathbf{P}_{\mathcal{J}_{k-1}}\mathbf{x}_j\|^2 \le n$,

$$|\mathcal{B}|(\mu^*_k)^2 \ge \sum_{j\in\mathcal{B}} \frac{|\mathbf{x}_j'\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\boldsymbol{\mu}|^2}{n} = \boldsymbol{\mu}'\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}(\mathbf{X}_{\mathcal{A}}\mathbf{X}_{\mathcal{A}}'/n)\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\boldsymbol{\mu}.$$

This gives (28) since $\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\boldsymbol{\mu}$ is in the column space of $\mathbf{X}_{\mathcal{A}}$ and $\|\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\boldsymbol{\mu}\|^2 \ge \psi_{k-1}n\|\boldsymbol{\beta}_{\mathcal{B}}\|^2$ as in Lemma 1 of Zhang and Huang (2008). □

Proof of Theorem 1. Let $k_m = \min\{k : |\mathcal{J}_k \cap \mathcal{I}_n| = m\}$, $\mathcal{B}_m = \mathcal{I}_n \setminus \mathcal{J}_{k_m}$ and $\xi_m = \|\boldsymbol{\beta}_{\mathcal{B}_m}\|^2/|\mathcal{B}_m|$. Since $k_{d_n} \le k^*_2$ iff $\|\mathbf{P}^{\perp}_{\mathcal{J}_{d^*}}\boldsymbol{\mu}\| = 0$, we assume $\|\mathbf{P}^{\perp}_{\mathcal{J}_{d^*}}\boldsymbol{\mu}\| > 0$. Since $\|\mathbf{P}_{\mathcal{J}_k}\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\boldsymbol{\mu}\| \ge (1-\gamma)\mu^*_k$, Lemma 1 implies

$$\sum_{k=k_{j-1}+1}^{k_j\wedge k^*_2} \psi^2_{k-1} \le \sum_{k=k_{j-1}+1}^{k_j\wedge k^*_2} \frac{\|\mathbf{P}_{\mathcal{J}_k}\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\boldsymbol{\mu}\|^2}{(1-\gamma)^2 n\xi_{j-1}} \le \frac{\|\mathbf{P}_{\mathcal{J}_{k_j\wedge k^*_2}}\mathbf{P}^{\perp}_{\mathcal{J}_{k_{j-1}}}\boldsymbol{\mu}\|^2}{(1-\gamma)^2 n\xi_{j-1}}. \tag{29}$$

Since $\xi_{j-1} \ge \beta^2_*$ and $\|\mathbf{P}_{\mathcal{J}_{k_j}}\mathbf{P}^{\perp}_{\mathcal{J}_{k_{j-1}}}\boldsymbol{\mu}\|^2 \le \|\mathbf{X}_{\mathcal{B}_{j-1}}\boldsymbol{\beta}_{\mathcal{B}_{j-1}}\|^2 \le n\phi_{\max}(\mathcal{I}_n)\|\boldsymbol{\beta}_{\mathcal{B}_{j-1}}\|^2$,

$$\sum_{k=1}^{k_{d_n}\wedge k^*_2} \psi^2_{k-1} < \min\Big\{\frac{\|\boldsymbol{\mu}\|^2}{(1-\gamma)^2 n\beta^2_*}, \frac{\phi_{\max}(\mathcal{I}_n)d_n(d_n+1)}{2(1-\gamma)^2}\Big\} \le \sum_{k=1}^{k^*_2}\psi^2_{k-1}.$$

The inequality is strict due to $\|\mathbf{P}^{\perp}_{\mathcal{J}_{d^*}}\boldsymbol{\mu}\| > 0$. This gives $k_{d_n} \le k_2^*$. For k_1^*,

$$\begin{aligned}(1-\gamma)^2\sum_{k=1}^{k_{d_n}\wedge k_1^*}\psi_{k-1} &\le \sum_{j=1}^{d_n}\sum_{k=k_{j-1}\wedge k_1^*+1}^{k_j\wedge k_1^*}\frac{(m-j+1)\|\mathbf{P}_{\mathcal{J}_k}\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\boldsymbol{\mu}\|^2}{\max\{\|\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\boldsymbol{\mu}\|^2, nc_*\beta_*^2\}}\\ &\le \sum_{j=1}^{d_n}\int_{\|\mathbf{P}^{\perp}_{\mathcal{J}_{k_j\wedge k_1^*}}\boldsymbol{\mu}\|^2}^{\|\mathbf{P}^{\perp}_{\mathcal{J}_{k_{j-1}\wedge k_1^*}}\boldsymbol{\mu}\|^2}\frac{(m-j+1)dx}{\max\{x, nc_*\beta_*^2\}}\\ &= \sum_{j=1}^{d_n}\int_{\|\mathbf{P}^{\perp}_{\mathcal{J}_{k_j\wedge k_1^*}}\boldsymbol{\mu}\|^2}^{\|\mathbf{P}_{\mathcal{J}_{k_1^*}}\boldsymbol{\mu}\|^2}\frac{dx}{\max\{x, nc_*\beta_*^2\}}.\end{aligned}$$

Since $\|\mathbf{P}^{\perp}_{\mathcal{J}_{k_j\wedge k_1^*}}\boldsymbol{\mu}\|^2 \ge (d_n-j)nc_*\beta_*^2$ and $d_n! > d_n^{d_n+1/2}e^{-d_n}\sqrt{2\pi}$ by Stirling,

$$\begin{aligned}(1-\gamma)^2\sum_{k=1}^{k_{d_n}\wedge k_1^*}\psi_{k-1} &< \sum_{j=1}^{d_n}\int_{(d_n-j)c_*n\beta_*^2}^{\|\boldsymbol{\mu}\|^2}\frac{dx}{\max\{x, nc_*\beta_*^2\}}\\ &= d_n\log\Big(\frac{\|\boldsymbol{\mu}\|^2}{c_*n\beta_*^2}\Big)+1-\log((d_n-1)!)\\ &< d_n\log\Big(\frac{e\|\boldsymbol{\mu}\|^2}{c_*d_nn\beta_*^2}\Big)+1+\log\sqrt{d_n/(2\pi)}.\end{aligned}$$

This gives $k_{d_n} \le k_1^*$ by (9). □

The proof of Theorem 2 requires two lemmas. Lemma 2 provides upper bounds for the cardinality of the collection of models in (12), while Lemma 3 gives a sharp large deviation bound for the tail of the t-distribution.

Lemma 2 *For the $\{\mathscr{C}_k, m_k\}$ in (12) and (18), $|\mathscr{C}_k| \le \prod_{m=1}^{k}(m_k-1)$.*

Proof of Lemma 2. It suffices to prove $\#\{j : \mu_{j,k} \ge (1-\gamma)\mu_k^*\} \le m_k - 1$ for all k. For a fixed $\mathcal{J}_{k-1}$, let $\mathbf{v} = \mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\boldsymbol{\mu}$, $\mathbf{x}_j^{\perp} = \mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\mathbf{x}_j$ and $s_j = \mathrm{sgn}(\mathbf{x}_j'\mathbf{v})$. For any $\mathcal{A} \subset \{j : \mu_{j,k} \ge (1-\gamma)\mu_k^*\}$, (28) gives

$$\sum_{j\in\mathcal{A}}\frac{\mu_{j,k}}{\|\mathbf{v}\|} \ge |\mathcal{A}|(1-\gamma)\mu_k^*/\|\mathbf{v}\| \ge |\mathcal{A}|(1-\gamma)\sqrt{\psi_{k-1}/|\mathcal{I}_n\setminus\mathcal{J}_{k-1}|}.$$

On the other hand, for $|\mathcal{A}| \le m_k$, the upper bound in (17) gives

$$\sum_{j\in\mathcal{A}}\frac{s_j\mathbf{x}_j'\mathbf{v}}{\|\mathbf{v}\|} \le \Big\|\sum_{j\in\mathcal{A}}s_j\mathbf{x}_j\Big\| \le \Big(n\phi^*_{m_k}\sum_{j\in\mathcal{A}}s_j^2\Big)^2 = \sqrt{n\phi^*_{m_k}|\mathcal{A}|}.$$

Since $\mu_{j,k} = |\mathbf{x}_j'\mathbf{v}|/\|\mathbf{x}_j^\perp\| \le |\mathbf{x}_j'\mathbf{v}|/\sqrt{\phi_k n}$, the above two inequalities yield

$$|\mathcal{A}| \le m_k \;\Rightarrow\; |\mathcal{A}| \le \frac{\phi^*_{m_k}|\mathcal{I}_n \setminus \mathcal{J}_{k-1}|}{(1-\gamma)^2\phi_k\psi_{k-1}} < m_k$$

in view of the definition of m_k in (18). This gives $\#\{j : \mu_{j,k} \ge (1-\gamma)\mu^*_k\} < m_k$. □

Lemma 3 *Let T_m have the t-distribution with m degrees of freedom, or equivalently $T_m^2 \sim F_{1,m}$. Then, there exists $\epsilon_m \to 0$ such that for all $t > 0$*

$$P\Big\{T_m^2 > m(e^{2t^2/(m-1)} - 1)\Big\} \le \frac{1+\epsilon_m}{\sqrt{\pi}t}e^{-t^2}. \tag{30}$$

Proof of Lemma 3. Let $x = \sqrt{m(e^{2t^2/(m-1)} - 1)}$. Since T_m has the t-distribution,

$$\begin{aligned} P\Big\{T_m^2 > x^2\Big\} &= \frac{2\Gamma((m+1)/2)}{\Gamma(m/2)\sqrt{m\pi}}\int_x^\infty \Big(1+\frac{u^2}{m}\Big)^{-(m+1)/2}du \\ &\le \frac{2\Gamma((m+1)/2)}{x\Gamma(m/2)\sqrt{m\pi}}\int_x^\infty \Big(1+\frac{u^2}{m}\Big)^{-(m+1)/2}udu \\ &= \frac{2\Gamma((m+1)/2)m}{x\Gamma(m/2)\sqrt{m\pi}(m-1)}\Big(1+\frac{x^2}{m}\Big)^{-(m-1)/2}. \end{aligned}$$

Since $x \ge t\sqrt{2m/(m-1)}$,

$$P\Big\{T_m^2 > x^2\Big\} \le \frac{\sqrt{2}\Gamma((m+1)/2)}{\Gamma(m/2)\sqrt{m-1}}\frac{e^{-t^2}}{t\sqrt{\pi}} = (1+\epsilon_m)\frac{e^{-t^2}}{t\sqrt{\pi}},$$

where $\epsilon_m = \sqrt{2}\Gamma((m+1)/2)/\{\Gamma(m/2)\sqrt{m-1}\} - 1 \to 0$ as $m \to \infty$. □

Proof of Theorem 2. We prove $P\{\cap_{j=1}^5 \Omega_j\} \to 1$ in five steps, where

$$\begin{aligned} &\Omega_1 = \big\{\mathcal{J}_k \in \mathscr{C}_k \;\forall\; \|\mathbf{P}^\perp_{\mathcal{J}_{k-1}}\boldsymbol{\mu}\| > 0\big\},\; \Omega_2 = \big\{\|\mathbf{P}^\perp_{\mathcal{J}_k}\boldsymbol{\mu}\| = 0, k \ge \widetilde{k}\big\}, \\ &\Omega_3 = \big\{\|\mathbf{P}^\perp_{\mathcal{J}_k}\boldsymbol{\mu}\| > 0, k < \widetilde{k}\big\},\; \Omega_4 = \big\{|\widehat{\sigma}_{n,1} - \sigma| \le \epsilon\big\},\; \Omega_5 = \big\{\mathcal{I}_n = \widehat{\mathcal{I}}_{n,2}\big\}. \end{aligned}$$

In $\cap_{j=1}^5\Omega_j$, the forward search finds all variables with $\beta_j \ne 0$ reasonably quickly by Theorem 1, the stopping rule $\widetilde{k}$ does not stop the forward search too early, $\widetilde{k}$ stops the forward search immediately after finding all $\beta_j \ne 0$, the estimation error for σ is no greater than ϵ, and the backward deletion removes all extra variables. Let $\Phi(t)$ be the standard normal distribution function.

Step 1. We prove $P\{\Omega_1^c\} \to 0$. This does not involve the stopping rule for the forward search. Let $z_{j,k} = \|\mathbf{P}_{\mathcal{J}_{k-1}\cup\{j\}}\mathbf{P}^\perp_{\mathcal{J}_{k-1}}\boldsymbol{\varepsilon}\|$ and $j_k^* = \arg\max_j \mu_{j,k}$. Since $\|\mathbf{P}_{\mathcal{J}\cup\{j\}}\mathbf{P}^\perp_{\mathcal{J}}\mathbf{v}\|^2 = L^2_{\mathbf{v},\mathbf{x}_j}(\mathcal{J})/L_{x_j,x_j}(\mathcal{J})$,

(6) implies $\mu_k^* - \mu_{j_k,k} \le z_{j_k^*,k} + \max_j z_{j,k}$. Since $\|\mathbf{P}_{\mathcal{J}\cup\{j\}}\mathbf{P}_{\mathcal{J}}^{\perp}\boldsymbol{\varepsilon}\|/\sigma \sim |N(0,1)|$ for deterministic $\{\mathcal{J}, j\}$ and $\mu_k^* \ge \psi_{k-1}\sqrt{n}\beta_* \ge c_*\sqrt{n}\beta_* \ge c_* M_0\sigma\sqrt{2\log p}$ by Lemma 1 and (19),

$$\begin{aligned} & P\Big\{\mathcal{J}_k \not\subseteq \mathscr{C}_k, \mu_k^* > 0, \mathcal{J}_{k-1} \in \mathscr{C}_{k-1}\Big\} \\ \le\ & P\Big\{z_{j_k^*,k} + \max_j z_{j,k} \ge \gamma c_* M_0\sigma\sqrt{2\log p},\ \mathcal{J}_{k-1} \in \mathscr{C}_{k-1}\Big\} \\ \le\ & 2p|\mathscr{C}_{k-1}|\Phi\big(-\sqrt{(1+\eta_1)2\log p}\big) + 2|\mathscr{C}_{k-1}|\Phi\big(-\sqrt{\eta_1 2\log p}\big), \end{aligned}$$

due to $\gamma c_* M_0 \ge \sqrt{1+\eta_1}+\sqrt{\eta_1}$. Thus, since $|\mathscr{C}_{k-1}| \le \prod_{\ell=1}^{k-1}(m_\ell - 1)$ by Lemma 2 and $\sum_{k=1}^{k^*}\log m_k \le \eta_1 \log p$ by (19), (13) gives

$$\begin{aligned} P\Big\{\Omega_1^c\Big\} &\le \sum_{k=1}^{k^*} 2|\mathscr{C}_{k-1}|\Big\{p\Phi\big(-\sqrt{(1+\eta_1)2\log p}\big) + \Phi\big(-\sqrt{\eta_1 2\log p}\big)\Big\} \\ &\le e^{\eta_1\log p}\Big\{p\Phi\big(-\sqrt{(1+\eta_1)2\log p}\big) + \Phi\big(-\sqrt{\eta_1 2\log p}\big)\Big\} \to 0. \end{aligned}$$

Step 2. We prove $P\{\Omega_1 \cap \Omega_2^c\} \to 0$. Let $a_n = e^{2(1+\eta_0)(\log p)/n} - 1$. By (21)

$$\mathrm{BICP}_k < \mathrm{BICP}_{k-1} \Leftrightarrow \|\mathbf{P}_{\mathcal{J}_k}\mathbf{P}_{\mathcal{J}_{k-1}}^{\perp}\mathbf{y}\|^2/\|\mathbf{P}_{\mathcal{J}_{k-1}}^{\perp}\mathbf{y}\|^2 > a_n/(1+a_n). \tag{31}$$

Let $\epsilon_0 > 0$ be small and $\mathcal{A}_k = \mathcal{J}_k \cup \mathcal{I}_n$. Consider events

$$\begin{aligned} \Omega_{2,k} &= \Big\{\|\mathbf{P}_{\mathcal{J}_k}\mathbf{P}_{\mathcal{J}_{k-1}}^{\perp}\mathbf{y}\| \ge \mu_k^* - \sigma\sqrt{2\eta_1\log p} > 0, \\ &\qquad \|\mathbf{P}_{\mathcal{A}_{k-1}}\boldsymbol{\varepsilon}\| \le \epsilon_0\sigma\sqrt{n},\ |(\sigma^2 n)^{-1}\|\boldsymbol{\varepsilon}\|^2 - 1| \le \epsilon_0\Big\}. \end{aligned}$$

It follows from Lemma 1 that in the event $\Omega_{2,k}$,

$$\begin{aligned} \|\mathbf{P}_{\mathcal{J}_{k-1}}^{\perp}\mathbf{y}\|^2 &= \|\mathbf{P}_{\mathcal{A}_{k-1}}\mathbf{P}_{\mathcal{J}_{k-1}}^{\perp}\mathbf{y}\|^2 + \|\mathbf{P}_{\mathcal{A}_{k-1}}^{\perp}\boldsymbol{\varepsilon}\|^2 \\ &\le \|\mathbf{P}_{\mathcal{J}_{k-1}}^{\perp}\boldsymbol{\mu}\|^2 + 2\|\mathbf{P}_{\mathcal{J}_{k-1}}^{\perp}\boldsymbol{\mu}\| \times \|\mathbf{P}_{\mathcal{A}_{k-1}}\boldsymbol{\varepsilon}\| + \|\boldsymbol{\varepsilon}\|^2 \\ &\le (1+\epsilon_0)d_n(\mu_k^*)^2/\psi_{k-1} + (1+2\epsilon_0)\sigma^2 n. \end{aligned}$$

Since $\mu_k^* \ge c_* M_0\sigma\sqrt{2\log p}$ as in Step 1 with $c_* M_0 > \sqrt{\eta_1}$, we have in $\Omega_{2,k}$,

$$\begin{aligned} \frac{\|\mathbf{P}_{\mathcal{J}_k}\mathbf{P}_{\mathcal{J}_{k-1}}^{\perp}\mathbf{y}\|^2}{\|\mathbf{P}_{\mathcal{J}_{k-1}}^{\perp}\mathbf{y}\|^2} &\ge \frac{(\mu_k^* - \sigma\sqrt{2\eta_1\log p})^2}{(1+\epsilon_0)d_n(\mu_k^*)^2/\psi_{k-1} + (1+2\epsilon_0)\sigma^2 n} \\ &\ge \frac{(2\log p)(c_* M_0 - \sqrt{\eta_1})^2}{(1+\epsilon_0)d_n(c_* M_0)^2(2\log p)/c_* + (1+2\epsilon_0)n}. \end{aligned}$$

Since $\log p = o(n)$ by (20), $a_n/(1+a_n) \le (1+\epsilon_0)(1+\eta_0)(2\log p)/n$ for $k \le k^*$ and large n. Thus, for sufficiently small $\epsilon_0 > 0$ and in $\Omega_{2,k}$,

$$\frac{\|\mathbf{P}_{\mathcal{J}_k}\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\mathbf{y}\|^2}{\|\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\mathbf{y}\|^2 a_n/(1+a_n)} \ge \frac{(c_*M_0 - \sqrt{\eta_1})^2/\{(1+\eta_0)(1+\epsilon_0)\}}{(1+\epsilon_0)c_*M_0^2(2d_n\log p)/n + (1+2\epsilon_0)} > 1,$$

due to $c_*M_0^2(2d_n\log p)/n \le c_*M_0^2M_1 < (c_*M_0 - \sqrt{\eta_1})^2/(1+\eta_0) - 1$ by (20). Thus, $\Omega_2 \supseteq \cap_{k=1}^{k^*}\Omega_{2,k}$ and as in the probability calculation in Step 1,

$$\begin{aligned}
& P\{\Omega_1 \cap \Omega_2^c\} - P\big\{|(\sigma^2 n)^{-1}\|\boldsymbol{\varepsilon}\|^2 - 1| \le \epsilon_0\big\} \\
\le & \sum_{k=1}^{k^*} P\big\{\|\mathbf{P}_{\mathcal{J}_{k-1}\cup\{j_k^*\}}\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\boldsymbol{\varepsilon}\| > \sigma\sqrt{2\eta_1\log p},\ \mathcal{J}_{k-1} \in \mathscr{C}_{k-1}\big\} \\
& + \sum_{k=1}^{k^*} P\big\{\|\mathbf{P}_{\mathcal{A}_{k-1}}\boldsymbol{\varepsilon}\| > \epsilon_0\sigma\sqrt{n}, \mathcal{J}_{k-1} \in \mathscr{C}_{k-1}\big\} \\
\le & \sum_{k=1}^{k^*} |\mathscr{C}_{k-1}|\Big(\Phi\big(-\sqrt{2\eta_1\log p}\big) + P\big\{\chi^2_{d_n+k-1} > \epsilon_0^2 n\big\}\Big) \\
\le & \ p^{\eta_1}\Phi\big(-\sqrt{2\eta_1\log p}\big) + o(1) = o(1),
\end{aligned}$$

since $P\{\sqrt{\chi^2_d} > \sqrt{d} + t\} \le \Phi(-t)$ and $\epsilon_0\sqrt{n} - \sqrt{d_n + k^*} \gg \sqrt{2\eta_1\log p}$.

Step 3. We prove $P\{\Omega_1 \cap \Omega_2 \cap \Omega_3^c\} \to 0$. It follows from (31) that

$$\mathrm{BICP}_k < \mathrm{BICP}_{k-1} \Leftrightarrow \|\mathbf{P}_{\mathcal{J}_k}\mathbf{P}^{\perp}_{\mathcal{J}_{k-1}}\mathbf{y}\|^2/\|\mathbf{P}^{\perp}_{\mathcal{J}_k}\mathbf{y}\|^2 > a_n, \tag{32}$$

which is an $F_{1,n-k}$ test for the random model $\mathcal{J}_k$ against $\mathcal{J}_{k-1}$. Since $k(\log p)/n \le k^*(\log p)/n = O(1)$, by Lemma 3 the threshold $(n-k)a_n$ gives an $F_{1,n-k}$-test of size $O(1)p^{-1-\eta_0+(k+1)/n}/\sqrt{\log p} \lesssim p^{-1-\eta_0}/\sqrt{\log p}$. Since $\eta_0 \ge \eta_1$ in (19),

$$\begin{aligned}
P\{\Omega_1 \cap \Omega_2 \cap \Omega_3^c\} & \le \sum_{k=1}^{k^*} P\big\{\mu_k^* = 0, \mathrm{BICP}_k < \mathrm{BICP}_{k-1}, \mathcal{J}_{k-1} \in \mathscr{C}_{k-1}\big\} \\
& \lesssim \sum_{k=1}^{k^*} \frac{(p-k+1)|\mathscr{C}_{k-1}|}{p^{1+\eta_0}\sqrt{\log p}} = \frac{p^{\eta_1}}{p^{\eta_0}\sqrt{\log p}} = o(1).
\end{aligned} \tag{33}$$

Step 4. We prove $P\{\Omega_1 \cap \Omega_4^c\} \to 0$. Consider a fixed $\epsilon > 0$. In $\Omega_1 \cap \Omega_{2,k}$, $\widetilde{k} = k$ implies $\|\mathbf{P}_{\mathcal{I}_{n,1}}\boldsymbol{\varepsilon}\| = \|\mathbf{P}_{\mathcal{A}_{k-1}}\boldsymbol{\varepsilon}\| \le \epsilon_0\sigma\sqrt{n}$ and $|(\sigma^2 n)^{-1}\|\boldsymbol{\varepsilon}\|^2 - 1| \le \epsilon_0$. Since $|\widehat{\mathcal{I}}_{n,1}| \le k^* = o(n)$ and $\|\mathbf{P}^{\perp}_{\mathcal{I}_{n,1}}\mathbf{y}\|^2 = \|\boldsymbol{\varepsilon}\|^2 - \|\mathbf{P}_{\mathcal{I}_{n,1}}\boldsymbol{\varepsilon}\|^2$, for sufficiently small ϵ_0, $(\sigma-\epsilon)^2 \le \|\mathbf{P}_{\mathcal{I}_{n,1}} \perp \mathbf{y}\|^2/(n-\widetilde{k}) \le (\sigma+\epsilon)^2$. Thus, $P\{\Omega_1 \cap \Omega_4^c\} \le \sum_{k\le k^*} P\{\Omega_1 \cap \Omega_{2,k}^c\} \to 0$ by Step 2.

Step 5. We prove $P\{\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_5^c\} \to 0$. Let

$$\begin{aligned}\Omega_{5,k} &= \Big\{ \max_{j\in\mathcal{J}_k^*} \|\mathbf{P}_{\mathcal{J}_k^*}\mathbf{P}^{\perp}_{\mathcal{J}_k^*\setminus\{j\}}\boldsymbol{\varepsilon}\|^2/\sigma^2 < (\eta_1+\eta_2)2\log p, \\ &\qquad \max_{j\notin\mathcal{I}_n} \|\mathbf{P}_{\mathcal{J}_k^*}\mathbf{P}^{\perp}_{\mathcal{J}_k^*\setminus\{j\}}\boldsymbol{\varepsilon}\|^2/\|\mathbf{P}^{\perp}_{\mathcal{J}_k^*}\boldsymbol{\varepsilon}\|^2 \le a_n, \mathcal{J}_k^* \supset \mathcal{I}_n\Big\}. \end{aligned} \tag{34}$$

In the event $\mathcal{J}_k^* \supset \mathcal{I}_n$, $\min_{j\in\mathcal{I}_n} \|\mathbf{P}_{\mathcal{J}_k^*}\mathbf{P}^{\perp}_{\mathcal{J}_k^*\setminus\{j\}}\boldsymbol{\mu}\| \ge \beta_*\sqrt{c_*} \ge M_0\sigma\sqrt{c_*2(\log p)/n}$. Since $(c_*M_0/2)^2 \ge \eta_1+\eta_2$ and $\mathbf{P}^{\perp}_{\mathcal{J}}\mathbf{y} = \mathbf{P}^{\perp}_{\mathcal{J}}\boldsymbol{\varepsilon}$ for $\mathcal{J} \supseteq \mathcal{I}_n$, in the event $\Omega_{5,k}$,

$$\begin{aligned} & \min_{j\in\mathcal{I}_n} \|\mathbf{P}_{\mathcal{J}_k^*}\mathbf{P}^{\perp}_{\mathcal{J}_k^*\setminus\{j\}}\mathbf{y}\| - \max_{j\in\mathcal{J}_k^*\setminus\mathcal{I}_n} \|\mathbf{P}_{\mathcal{J}_k^*}\mathbf{P}^{\perp}_{\mathcal{J}_k^*\setminus\{j\}}\mathbf{y}\| \\ > \;& \beta_*\sqrt{c_*} - 2\sigma\sqrt{(\eta_1+\eta_2)2\log p} \ge 0, \end{aligned}$$

so that the backward deletion does not delete the elements of $\mathcal{I}_n$ for $k > d_n$. Moreover, since the forward addition and backward deletion are based on the same tests, by (32) and the second inequality in $\Omega_{5,k}$, the backward deletion does not stop in for $k > d_n$. Thus,

$$\begin{aligned} & P\big\{\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_5^c\big\} \\ \le \;& P\{\mathcal{J}_{d_n}^* = \mathcal{I}_n, \widetilde{k} < d_n\big\} + \sum_{k=d_n+1}^{k^*} P\big\{\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_{5,k}^c\big\}. \end{aligned}$$

In $\Omega_1\cap\Omega_2\cap\Omega_3$, $\widetilde{k} = k$ implies $\mathcal{I}_n \subset \widehat{\mathcal{I}}_{n,1} = \mathcal{J}_k$ and $\mathcal{J}_k \in \mathscr{C}_k$, so that the backward deletion involves at most N_{k-1}^* combinations of $\{j, \mathcal{J}_k^*\}$, where $N_k^* = (k^*!/k!)\sum_{k=1}^{k^*}|\mathscr{C}_k| \le p^{\eta_2+\eta_1}/k!$. Since the inequalities in (34) involve χ_1^2 and $F_{1,n-k}$ variables in random models,

$$\begin{aligned} & \sum_{k=d_n+1}^{k^*} P\big\{\Omega_1 \cap \Omega_2 \cap \Omega_3 \cap \Omega_{5,k}^c\big\}. \qquad (35) \\ \le \;& \sum_k \frac{p^{\eta_1+\eta_2}}{(k-1)!}\Big(P\{\chi_1^2 > 2(\eta_1+\eta_2)\log p\} + P\{F_{1,n-k} > (n-k)a_n\}\Big) \to 0. \end{aligned}$$

Since $\eta_1+\eta_2 \le 1+\eta_0$, the bound here for the tail of $F_{1,n-k}$ is identical to that of Step 3. The proof of $P\{\mathcal{J}_{d_n}^* = \mathcal{I}_n, \widetilde{k} < d_n\big\} \to 0$ is simpler than Step 2 and omitted. This completes the step. □

Proof of Theorem 3. For the BICP$_k$ in (5), $a_n = e^{2(\log p)/n} - 1$. Since $k^*(\log p)/n = O(1)$, the size of the test $F_{1,n-k} > (n-k)a_n$ is $O(p^{-1}(\log p)^{-1/2})$ by Lemma 3 uniformly for $k \le k^*$. Since the condition $\eta_0 \ge \eta_1$ is used only in (33), the proof of Theorem 2 is still valid when $\sum_{k=1}^{k^*}|\mathscr{C}_k| \ll \sqrt{\log p}$. □

References

[1] An, H. and Gu, L. (1985). On the selection of regression variables. *ACTA Mathematicae Applicatae Sinica*, **2**, 27-36.

[2] An, H. and Gu, L. (1987). Fast stepwise procedures of selection of variables by using AIC and BIC criteria. *ACTA Mathematicae Applicatae Sinica*, **5**, 60-67.

[3] Bickel, P., Ritov Y. and Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37** 1705-1732.

[4] Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.* **34** 559-583.

[5] Bunea, F. (2008). Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electronic Journal of Statistics*, **2** 11531194.

[6] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007). Sparsity oracle inequalities for the lasso. *Electron. J. Statist.* **1** 169-194 (electronic).

[7] Candés, E. and Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Inform. Theory* **51** 4203-4215.

[8] Candés, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Ann. Statist.* **35** 2313-2404.

[9] Chen, J. and Chen, Z. (2008). Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**, 759-771.

[10] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 13481360.

[11] Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *J. Roy. Stats. Soc.* **B**, **70**, 849-911.

[12] Fan, J. and Peng, H. (2004). On non-concave penalized likelihood with diverging number of parameters. *Ann. Stats.* **32** 928-961.

[13] Greenshtein E. and Ritov Y. (2004). Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli* **10** 971-988.

[14] Huang, J., Ma, S. and Zhang, C.-H. (2008). Adaptive LASSO for sparse high-dimensional regression models. *Statistica Sinica* **18** 1603-1618.

[15] Knight, K. and Fu, W. J. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356-1378.

[16] Mallows, C.L. (1973). Some comments on C_p. *Technometrics* **12** 661-675.

[17] Mallat, S. and Zhang, Z. (1993). Matching pursuits with time-frequency dictionaries. *IEEE Transa. Signal Processing* **41** 3397-3415.

[18] Meinshausen, N. and Bühlmann, P. (2006) High dimensional graphs and variable selection with the Lasso. *Ann. Statist.* **34** 1436-1462.

[19] Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.* **37** 246-270.

[20] Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461-464.

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

[22] Tropp, J.A. (2006). Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inform. Theory* **52** 1030-1051.

[23] van de Geer, S. (2008). High-dimensional generalized linear models and the Lasso. *Ann. Statist.* **36** 614-645.

[24] Wainwright, M.J. (2009a). Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Info. Theory* **55** 2183–2202.

[25] Wainwright, M.J. (2009b). Information-theoretic limitations on sparsity recovery in the high-dimensional and noisy setting. *IEEE Trans. Info. Theory* **55** 5728–5741.

[26] Zhang, C.-H. (2007). Information-theoretic optimality of variable selection with concave penalty. Technical Report 2007-008, Department of Statistics and Biostatistics, Rutgers University.

[27] Zhang, C.-H. (2008). Discussion: One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.* **36** 1553-1560.

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

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

[30] Zhang, T. (2009a). Some sharp performance bounds for least squares regression with L_1 regularization. *Ann. Statist.* **37** 2109-2144.

[31] Zhang, T. (2009b). Adaptive forward-backward greedy algorithm for sparse learning with linear models. In *NIPS 2008*, Koller, Schuurmans, Bengio and Bottou, Eds. pages 1921-1928.

[32] Zhao, P. and Yu, B. (2006). On model selection consistency of LASSO. *J. Machine Learning Research* **7** 2541-2567.

[33] Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101** 1418-1429.

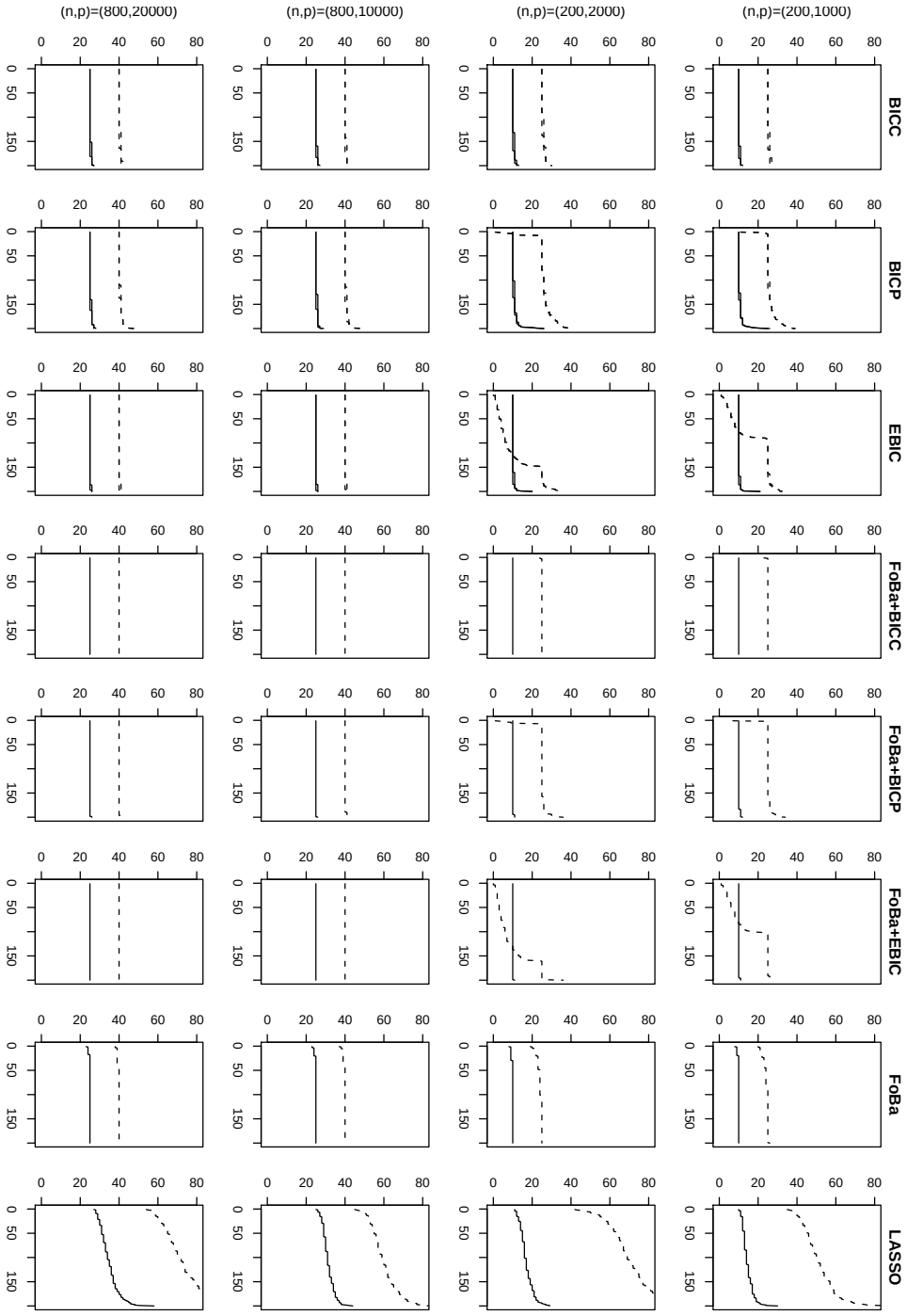

Figure 1: *Simulation results for Example 1 with* $\varepsilon_i \sim N(0,1)$. *Plots of the numbers of selected regression variables by the forward search and backward search in 200 replications. Two upper rows with* $n = 200$*:* $d = 10$ *(solid lines), and* $d = 25$ *(dashed lines). Two lower rows with* $n = 800$*:* $d = 25$ *(solid lines), and* $d = 40$ *(dashed lines).*

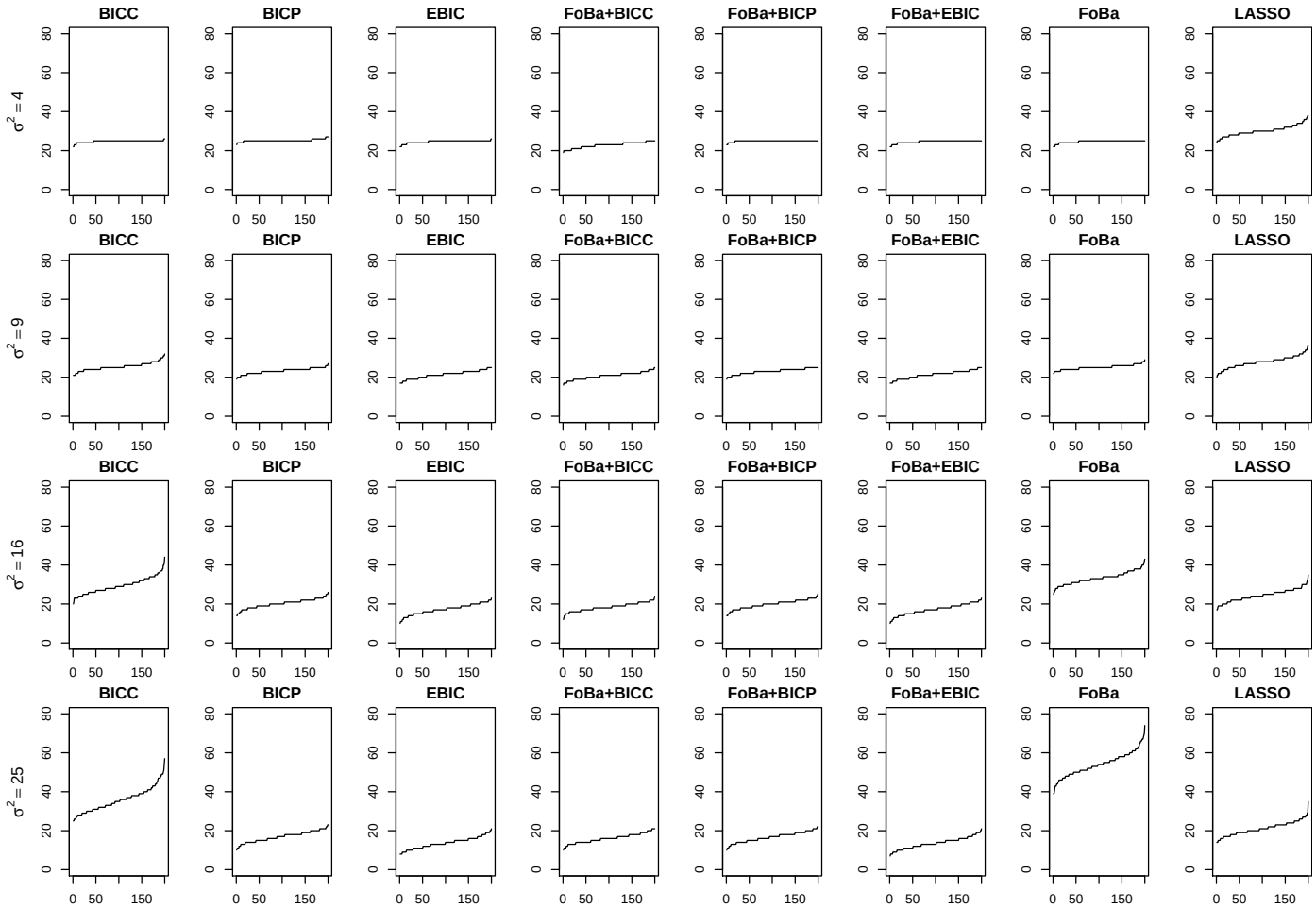

Figure 2: *Simulation results for Example 1 with* $\varepsilon_t \sim N(0, \sigma^2)$*: Plots of the numbers of selected regression variables in 200 replications, where* $(n, p, d) = (800, 10000, 25)$.

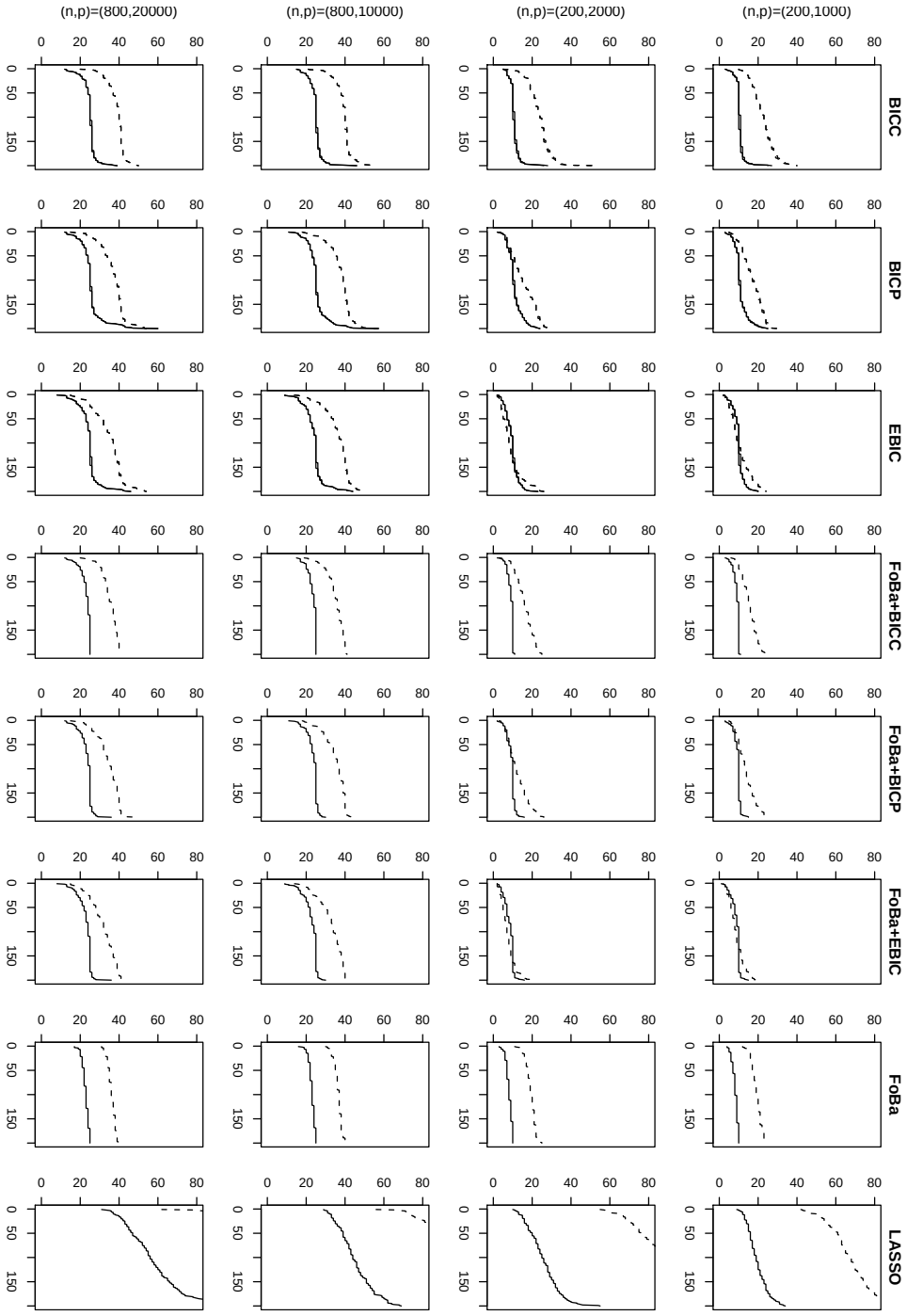

Figure 3: *Simulation results for Example 2: Plots of the numbers of selected regression variables by the forward search and backward search in 200 replications. Ten upper panels:* $n = 200$, $d = 10$ *(solid lines), and* $d = 25$ *(dashed lines). Ten lower panels:* $n = 800$, $d = 25$ *(solid lines), and* $d = 40$ *(dashed lines).*