Available online at www.sciencedirect.com

ScienceDirect

Journal of Statistical Planning and Inference 136 (2006) 4349–4364

journal of statistical planning and inference

www.elsevier.com/locate/jspi

# Consistent variable selection in high dimensional regression via multiple testing

Florentina Bunea*, Marten H. Wegkamp, Anna Auguste

*Department of Statistics, Florida State University, Tallahassee, FL 32306-4330, USA*

Received 29 May 2004; accepted 29 March 2005
Available online 3 August 2005

**Abstract**

This paper connects consistent variable selection with multiple hypotheses testing procedures in the linear regression model $\boldsymbol{Y}=\boldsymbol{X}\boldsymbol{\beta}+\varepsilon$, where the dimension $p$ of the parameter $\boldsymbol{\beta}$ is allowed to grow with the sample size $n$. We view the variable selection problem as one of estimating the index set $I_0 \subseteq \{1, \ldots, p\}$ of the non-zero components of $\boldsymbol{\beta} \in \mathbb{R}^p$. Estimation of $I_0$ can be further reformulated in terms of testing the hypotheses $\beta_1 = 0, \ldots, \beta_p = 0$. We study here testing via the false discovery rate (FDR) and Bonferroni methods. We show that the set $\widehat{I} \subseteq \{1, \ldots, p\}$ consisting of the indices of rejected hypotheses $\beta_i = 0$ is a consistent estimator of $I_0$, under appropriate conditions on the design matrix $\mathbf{X}$ and the control values used in either procedure. This technique can handle situations where $p$ is large at a very low computational cost, as no exhaustive search over the space of the $2^p$ submodels is required.

*Keywords:* Bonferroni correction; False discovery rate; Multiple hypothesis testing; Consistent variable selection

## 1. Introduction

The false discovery rate (FDR) procedure has been developed in the context of multiple hypotheses testing by Benjamini and Hochberg (1995). Given a set of $p$ hypotheses, out of which an unknown number $p_0$ are true, the FDR method identifies the hypotheses to be rejected, while keeping the expected value of the ratio of the number of false rejections

[*] Corresponding author.

*E-mail addresses:* bunea@stat.fsu.edu (F. Bunea), wegkamp@stat.fsu.edu (M.H. Wegkamp), auguste@stat.fsu.edu (A. Auguste).

0378-3758/$ - see front matter © 2005 Published by Elsevier B.V.
doi:10.1016/j.jspi.2005.03.011

to the total number of rejections below $q$, a user specified control value. In addition, this technique can handle problems in which $p$ is very large at a very low computational cost. The span of its applications ranges from denoising in signal processing problems, see for instance Abramovich et al. (2000), to genetics and medicine, see for instance Storey (2002), Benjamini and Yekutieli (2001). Genovese and Wasserman (2004) discuss theoretical aspects of the procedure using a stochastic process approach.

In this paper we indicate how the FDR procedure can be used for variable selection in linear regression models and establish the consistency of selection. We assume that the data are generated from the model

$$\begin{aligned} &\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\ &\beta_j \neq 0, \ \ j \in I_0; \quad \beta_j = 0, \ \ j \in \{1, \ldots, p\} \backslash I_0, \end{aligned} \tag{1.1}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^{\mathrm{T}}$, $\boldsymbol{X}$ is a $n \times p$ design matrix with deterministic entries $x_{ij}$, $1 \leqslant i \leqslant n$, $1 \leqslant j \leqslant p$, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ is the unknown vector of regression coefficients. The number of predictors $x_j = (x_{1j}, \ldots, x_{nj})^{\mathrm{T}}$ considered, $p$, is allowed to grow with the sample size $n$. This means that as $n$ increases, the model is allowed to become more complex. In addition, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}}$ is a vector of independent, identically distributed errors $\varepsilon_i$ with

$$\mathbb{E}\varepsilon_i = 0, \quad \mathbb{E}\varepsilon_i^2 = \sigma^2, \quad \mathbb{E}|\varepsilon_i|^{4+\delta} < \infty \quad \text{for some } \delta > 0. \tag{1.2}$$

The consistent variable selection problem is equivalent with the problem of estimating consistently the unknown index set $I_0 \subseteq \{1, \ldots, p\} \overset{\text{def}}{=} I_p$ of the non-zero components of $\boldsymbol{\beta}$. This problem received considerable attention in the statistical literature. In particular, the bayesian information criterion (BIC) has been shown to lead to consistent estimators of $I_0$, see Hannan and Quinn (1979), Hannan (1980), Geweke and Meese (1981) for early references. Woodroofe (1982) and Haughton (1988) establish consistency in the context of exponential families and we refer to Bunea (2004) for a recent contribution in semiparametric regression.

The serious drawback of any model selection method based on a penalized criterion is of a computational nature, as a search through the space of all possible $2^p$ models may be needed. Cross-validation (Shao, 1993) provides an alternative, but again the leave $m$ out of $n$ strategy requires intensive computation. Zheng and Loh (1995), in the context of linear regression, suggested a two-stage procedure, where the first stage consists of ranking test statistics, and the second stage computing a penalized least squares estimator based on $p$ models only, which is a marked improvement over the other strategies, but may still be suboptimal computationally for $p$ large, which is the case of interest in this paper. Finally, Jiang and Liu (2004) study model selection based on parameter estimation in a more general setting. For instance, they allow for Poisson regression with random effects, Cox regression and graphical models. In the linear regression case, their method is intimately related to Zheng and Loh (1995). However $p$, the number of predictors, is not allowed to depend on the sample size $n$. This therefore creates the need for a computationally fast method that consistently estimates $I_0$ in this case. The approach we take here is based on multiple hypotheses testing. Note that the problem of estimating $I_0$ can be viewed as testing the null hypotheses

$$\begin{aligned}\mathbf{H_1}:&\quad \beta_1 = 0\\ &\quad\vdots\\ \mathbf{H_p}:&\quad \beta_p = 0.\end{aligned}$$

Any testing method which identifies hypotheses that can be rejected provides an estimator for $I_0$; see for example, Pötscher (1983), Bauer et al. (1988) for a consistent procedure consisting of individual tests of each of the parameters, when $p$ is fixed. We treat here the general case in which $p$ is allowed to grow with $n$, and show that, under appropriate conditions on the design matrix $\boldsymbol{X}$, adjustments of the FDR or Bonferoni procedures lead to consistent estimators of $I_0$. In addition, at the computational level, these methods only require fitting the full model. As such, the Bonferroni and FDR methods are faster than the other methods mentioned above, which is especially needed for $p$ large.

The rest of this article is structured as follows: Section 2.1 contains the description of the procedures, and Section 2.2 presents our theoretical results. Proofs of intermediate results are collected in the Appendix. The simulation study in Section 3 strongly supports our theoretical findings.

## 2. Consistent selection via thresholding *p*-values

We discuss two selection procedures based on multiple testing: the Bonferroni method and FDR procedure. Both procedures require a user specified level $q > 0$, which should be small (see Lemma 2.1).

Based on the full model (1.1), we start with computing the least squares estimates $\widehat{\beta}_i$, standard errors $\mathrm{se}(\widehat{\beta}_i)$, $t$-statistics $t_i = \widehat{\beta}_i/\mathrm{se}(\widehat{\beta}_i)$ and the $p$-values $\pi_i = 2\{1 - \Phi(|t_i|)\}$ for all $i = 1, \ldots, p$. (Here $\Phi$ is the standard normal distribution function.) The $t$-statistics $t_i$ and $p$-values $\pi_i$ correspond to the individual tests $\mathbf{H}_i : \beta_i = 0$ for $i = 1, \ldots, p$.

The Bonferroni method uses $\widehat{I} = \{i : \pi_i \leqslant q/p\}$ to estimate $I_0$. The FDR procedure, suggested by Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001), can be applied to variable selection in regression as follows:

- Order the $p$-values $\pi_{(1)} \leqslant \cdots \leqslant \pi_{(p)}$ and compute
$$k = \max\left\{ i : \pi_{(i)} \leqslant \frac{i}{p} \frac{q}{\sum_{j=1}^{p} j^{-1}} \right\}$$
and reject all $\mathbf{H}_{(i)} : \beta_{(i)} = 0,\ i = 1, \cdots, k$, where $\mathbf{H}_{(i)} : \beta_{(i)} = 0$ is the null hypothesis corresponding to the ordered $p$-value $\pi_{(i)}$. If no such $k$ exists, do not reject any hypothesis.
- Estimate $I_0$ by the set $\widehat{I}$ of indices corresponding to the first $k$ ordered $p$-values.

We turn now to studying the consistency of both estimators $\widehat{I}$ constructed above. These estimators may be different, but we use the same notation to keep the presentation focused. We first recall the theoretical properties of the FDR method that are relevant to this problem. Let $0 \leqslant R \leqslant p$ be the total number of rejected hypotheses, and let $0 \leqslant V \leqslant R$ be the number of falsely rejected hypotheses (i.e., reject whilst the null hypothesis is true). Benjamini and Yekutieli (2001), Theorem 1.3, showed that the procedure described in the previous subsection controls the false discovery rate at level $q$, that is,

$$\mathbb{E}Q \leqslant \frac{p - p_0}{p} q \leqslant q, \tag{2.1}$$

with

$$Q = \begin{cases} V/R & \text{if } R > 0, \\ 0 & \text{otherwise}, \end{cases} \tag{2.2}$$

where $p_0$ is the cardinality of $I_0$, and so $p - p_0$ is the number of true null hypotheses.

We call $\hat{I}$ consistent if $\lim_{n\to\infty} \mathbb{P}(\widehat{I} = I_0) = 1$. We can reformulate this in terms of the quantities $R$ and $V$ that are controlled by the procedure. Since $|I_0| = p_0$, the selection procedure will yield a consistent estimator $\widehat{I}$ of $I_0$ if and only if we have $p_0$ rejections ($R = p_0$), none of them erroneously ($V = 0$). Thus,

$$\mathbb{P}(\widehat{I} = I_0) = \mathbb{P}\{R = p_0, V = 0\}.$$

Proving consistency of $\hat{I}$ reduces then to showing

$$\mathbb{P}\{R = p_0, V = 0\} \to 1 \quad \text{as } n \to \infty.$$

In case $p_0 = 0$, we find that

$$\mathbb{P}\{R = p_0, V = 0\} = \mathbb{P}\{R = 0\}$$

and we need to show $\mathbb{P}\{R \neq p_0\} \to 0$. In the more interesting case where $p_0 \geqslant 1$, we need to show that both $\mathbb{P}\{R \neq p_0\}$ and $\mathbb{P}\{V \geqslant 1\}$ are asymptotically negligible.

**Lemma 2.1.** *Let $p_0 \geqslant 1$. For the Bonferroni method, we have*

$$\mathbb{P}\{V \geqslant 1\} \leqslant q. \tag{2.3}$$

*For the FDR method, we find*

$$\mathbb{P}\{V \geqslant 1\} \leqslant \mathbb{P}\{R \neq p_0\} + \frac{p_0(p - p_0)}{p} q. \tag{2.4}$$

**Proof.** Inequality (2.3) follows directly from the union bound $\mathbb{P}\{V \geqslant 1\} \leqslant p(q/p)$. It remains to prove (2.4). Note that

$$\begin{aligned} \mathbb{P}\{V \geqslant 1\} &\leqslant \mathbb{P}\{R \neq p_0\} + \mathbb{P}\{V \geqslant 1, R = p_0\} \\ &\leqslant \mathbb{P}\{R \neq p_0\} + \mathbb{P}\{Q \geqslant 1/p_0\} \\ &\leqslant \mathbb{P}\{R \neq p_0\} + p_0 \mathbb{E}Q \end{aligned}$$

by Markov's inequality. Theorem 1.3 in Benjamini and Yekutieli (2001) yields (2.1), which in turn implies (2.4). □

The previous result says that both procedures (Bonferroni and FDR) render consistent estimates $\hat{I}$ if we can show that $\mathbb{P}\{R \neq p_0\} \to 0$ and provided we choose $q \to 0$, as

$n \to \infty$. The next theorem (Theorem 2.5) establishes this under regularity assumptions on the design matrix, when the number of variables $p$ is allowed to tend to infinity with $n$, but is no larger than $\sqrt{n}$. We assume throughout that the (inverse) matrix

$$(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1} \overset{\text{def}}{=} M = (m_{ij})_{1 \leqslant i,j \leqslant p} \tag{2.5}$$

exists (for $n$ large enough). This means that $\mathrm{se}(\widehat{\beta}_i) = S\sqrt{m_{ii}}$, where $S^2 = \mathrm{RSS}/(n-p)$ is the usual estimate of $\sigma^2$ and RSS is the residual sum of squares. Let $H$ be the projection matrix onto the span of $\boldsymbol{X}$, i.e.,

$$\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}} \overset{\text{def}}{=} H = (h_{ij})_{1 \leqslant i,j \leqslant n}.$$

Furthermore, we impose the following assumptions that suppress the dependence on the sample size $n$ (of the quantities $m$, $p$, $q$ and $r$) to avoid notational clutter.

(A1) Assume that $p \leqslant \sqrt{n}/\log n$.
(A2) Define $m = \max_{1 \leqslant k \leqslant p} m_{kk}$. Assume that $m \to 0$, with $m \leqslant 1/\log n$.
(A3) Define $r = \max_{1 \leqslant k \leqslant n} h_{kk}$. Assume that $p^2 \cdot r \to 0$.

**Remark 2.2.** Condition (A2) is equivalent to $\max_{i \leqslant p} \mathrm{se}(\widehat{\beta}_i) \leqslant 1/\sqrt{\log n}$. A stronger condition is imposed by Jiang and Liu (2004). Bauer et al. (1988) and Zheng and Loh (1995) require that $\mathrm{se}(\widehat{\beta}_i) \to 0$ for all $i \leqslant p$.

The condition $r \to 0$ is standard for establishing asymptotic normality of $\widehat{\beta}_i$, see for instance Eicher (1965). Sen and Srivastava (1990) use $r < 0.2$ as a (very rough) rule of thumb. Condition (A3) strengthens this condition and it is needed for establishing a Berry–Esseen type bound (Lemma A.2 in the appendix) on the distribution of the regression estimates. When the errors $\varepsilon_i$ are normally distributed, the estimated coefficients $\widehat{\beta}_i$ have an exact normal distribution and, as a consequence, condition (A3) on $r$ becomes superfluous.

In view of Lemma 2.1, choosing the control parameter $q$ such that $q \to 0$ plays an important role in consistent variable selection. An additional condition that subsumes $q \to 0$ will be needed for our main Theorem 2.5.

$(C_q)$. Choose $q \to 0$ such that $q \geqslant \exp(-n)$ and $pq/\log p \to 0$.

**Remark 2.3.** For $p \to \infty$, the choice $q = \mathrm{O}(1/p)$ satisfies $(C_q)$. In practice, we suggest this as a rule of thumb for values of $p$ that are moderately large to large, relative to the sample size. For small values of $p$, relative to $n$, condition $(C_q)$ in connection with (A1) also offers a guideline: for $q = \mathrm{O}(1/\sqrt{n})$ we always have $q < \log p/p$. Section 3 contains a simulation study that explores various choices of this parameter.

Both selection procedures can be viewed as successively comparing the $p$-values

$$\pi_{(1)} \leqslant \cdots \leqslant \pi_{(p)}$$

with either a fixed threshold $q/p$ (Bonferroni) or a the variable threshold $iq/p\sum_{i=1}^{p} i^{-1}$ (FDR). The procedures stop when a certain $p$-value $\pi_{(k)}$ does not exceed the threshold. All hypotheses $\beta_{(j)} = 0$, $j = 1, \dots, k$, where $\beta_{(j)}$ corresponds to the ordered $p$-value $\pi_{(j)}$, will then be rejected. Since the $p$-values are all computed assuming that the null hypotheses are true, we note that the asymptotic distribution of $\pi_j$, $j \notin I_0$ is Uniform(0, 1), whereas for $j \in I_0$, we obtain a degenerate distribution, $\pi_j \to_P 0$; the asymptotic distributions are derived under model (1.1). To take this into account, we define the event

$$\mathscr{E}_n = \{(\pi_{(1)}, \dots, \pi_{(p_0)}) = (\pi_{j_1}, \dots, \pi_{j_{p_0}})\} \tag{2.6}$$

for $I_0 = \{j_1, \dots, j_{p_0}\}$.

**Lemma 2.4.** *Under assumptions* (A1)–(A3), *we have for both Bonferroni and FDR selection procedures that*

$$\lim_{n\to\infty} \mathbb{P}\{\mathscr{E}_n\} = 1. \tag{2.7}$$

**Proof.** We define the set $A_n$ by

$$A_n = \left\{ |S - \sigma| \leqslant \sqrt{\frac{\log n}{n}} \right\}. \tag{2.8}$$

Since Lemma A.1 in the appendix shows that $\lim_{n\to\infty} \mathbb{P}(A_n) = 1$, it suffices to show that $\lim_{n\to\infty} \mathbb{P}\{\mathscr{E}_n^c \cap A_n\} = 0$. Let $\delta = \{(\log n)/n\}^{1/2}$ and observe that for any $0 < \xi < 1$,

$$\begin{aligned} \mathbb{P}(\mathscr{E}_n^c \cap A_n) &\leqslant \sum_{j\in I_0} \sum_{i\notin I_0} \mathbb{P}(\{\pi_i < \pi_j\} \cap A_n) \\ &\leqslant \sum_{j\in I_0} \sum_{i\notin I_0} (\xi + \mathbb{P}(\{\pi_j \geqslant \xi\} \cap A_n) + \mathrm{O}(r + \delta)) \\ &\qquad \text{by Lemma A.3 in the appendix} \\ &= \mathrm{O}(p_0 p\{\xi + r + \delta + \sqrt{\log \xi} e^{-1/m}\}). \end{aligned}$$

Taking $\xi = \delta$ and invoking assumptions (A1)–(A3), we find that $\lim_{n\to\infty} \mathbb{P}(\mathscr{E}_n^c \cap A_n) = 0$.

**Theorem 2.5.** *Both Bonferroni and FDR procedures*, *under assumptions* (A1)–(A3) *and* $(C_q)$, *satisfy*

$$\lim_{n\to\infty} \mathbb{P}\{R \neq p_0\} = 0,$$

*and consequently they are consistent*:

$$\lim_{n\to\infty} \mathbb{P}(\hat{I} = I_0) = 1.$$

**Proof.** We first consider the FDR procedure. The event $\{R \neq p_0\}$ can be written in terms of the ordered $p$-values as follows:

$$\{R \neq p_0\} = \bigcup_{j=p_0+1}^{p} \left\{ \pi_{(j)} \leqslant q_p \frac{j}{p} \right\} \cup \left\{ \pi_{(p_0)} > q_p \frac{p_0}{p} \right\},$$

where we denoted $q/\sum_{i=1}^{p} i^{-1}$ by $q_p$. We then notice that we have

$$\begin{aligned}\mathbb{P}\{R \neq p_0\} \leqslant \mathbb{P}(A_n^c) + \mathbb{P}(\mathscr{E}_n^c \cap A_n) + \mathbb{P}\left(\left\{\pi_{(p_0)} > q_p \frac{p_0}{p}\right\} \cap \mathscr{E}_n \cap A_n\right) \\ + \sum_{j=p_0+1}^{p} \mathbb{P}\left(\{\pi_{(j)} \leqslant q_p \frac{j}{p}\} \cap \mathscr{E}_n \cap A_n\right). \qquad (2.9)\end{aligned}$$

In view of Lemmas A.1 and 2.4, it remains to show that the last two terms on the right in (2.9) converge to zero. We argue that

$$\begin{aligned}\sum_{j=p_0+1}^{p} \mathbb{P}\left(\left\{\pi_{(j)} \leqslant q_p \frac{j}{p}\right\} \cap \mathscr{E}_n \cap A_n\right) &\leqslant \sum_{j=p_0+1}^{p} \mathbb{P}(\{\pi_{(j)} \leqslant_p\} \cap \mathscr{E}_n \cap A_n) \\ &\leqslant \sum_{j \notin I_0} \mathbb{P}(\{\pi_j \leqslant q_p\} \cap A_n) \\ &= \mathrm{O}\left((p-p_0)\left\{\frac{q}{\log p} + r + \delta\right\}\right) \\ &= \mathrm{o}(1) \text{ as } n \to \infty\end{aligned}$$

by the choice $(C_q)$ and assumptions (A1) and (A3). Define

$$q_0 \overset{\mathrm{def}}{=} q_p \frac{p_0}{(2p)} = \frac{q p_0}{2p \sum_{i=1}^{p} i^{-1}}.$$

$$\begin{aligned}\mathbb{P}\left(\left\{\pi_{(p_0)} > q_p \frac{p_0}{p}\right\} \cap \mathscr{E}_n \cap A_n\right) &\leqslant p_0 \max_{j \in I_0} \mathbb{P}\left\{\pi_j \geqslant 2q_0 \cap A_n\right\} \\ &\leqslant p_0 \max_{j \in I_0} \mathbb{P}\{1 - \Phi(|T_j|) \geqslant q_0 \cap A_n\} \\ &= \mathrm{O}\left(p_0\left\{e^{-1/m}\sqrt{\log \frac{p \log p}{p_0 q}} + r + \delta\right\}\right) \\ &= \mathrm{o}(1) \text{ as } n \to \infty,\end{aligned}$$

by assumptions (A1)–(A3). This shows that $\mathbb{P}\{R \neq p_0\} \to 0$. Invoke Lemma 2.1 in connection with the choice of $q$ to conclude that the FDR procedure is consistent.

The consistency of the Bonferroni procedure can be proved in a similar way. The event $\{R \neq p_0\}$ can be written as

$$\{R \neq p_0\} = \bigcup_{j=p_0+1}^{p} \left\{\pi_{(j)} \leqslant \frac{q}{p}\right\} \cup \left\{\pi_{(p_0)} > \frac{q}{p}\right\},$$

and we find that

$$\begin{aligned}\sum_{j=p_0+1}^{p} \mathbb{P}\left(\left\{\pi_{(j)} \leqslant \frac{q}{p}\right\} \cap \mathscr{E}_n \cap A_n\right) &\leqslant \sum_{j\notin I_0} \mathbb{P}\left(\left\{\pi_j \leqslant \frac{q}{p}\right\} \cap A_n\right)\\ &= \mathrm{O}\left((p-p_0)\left\{\frac{q}{p}+r+\delta\right\}\right)\\ &= \mathrm{o}(1) \quad \text{as } n\to\infty\end{aligned}$$

by the choice $(C_q)$ and assumptions (A1) and (A3). Finally, we obtain

$$\begin{aligned}\mathbb{P}\left(\left\{\pi_{(p_0)} > \frac{q}{p}\right\} \cap \mathscr{E}_n \cap A_n\right) &\leqslant p_0 \max_{j\in I_0} \mathbb{P}\left(\left\{1-\Phi(|T_j|)\geqslant \frac{q}{p}\right\}, \cap A_n\right)\\ &= \mathrm{O}\left(p_0\left\{\mathrm{e}^{-1/m}\sqrt{\log\frac{p}{q}}+r+\delta\right\}\right)\\ &= \mathrm{o}(1) \text{ as } n\to\infty,\end{aligned}$$

by assumptions (A1)–(A3), which shows that the Bonferroni method is consistent. □

## 3. A simulation study

This section investigates the performance of our estimators constructed in Section 2 via a simulation study. We begin with comparing the FDR procedure with the one suggested by Zheng and Loh (1995). For comparison purposes, we considered the design of their simulations. We generated $p$ independent vectors $X_j^* \sim \mathrm{N}(0, I_n)$ and set the predictors $X_j = \sqrt{n}X_j^*/\|X_j^*\|$ for $j=1,\ldots,p$ ($\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^n$). The response variable $\boldsymbol{Y}$ is computed via $\boldsymbol{Y} = X_1+\cdots+X_{p_0}+\boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon}$ is a vector of independent standard Gaussian variables. We considered two instances of $p_0$, either 5 or 10. In each instance, we simulated samples of sizes $n=100, 500$ and 1000, respectively, from the corresponding linear model. For each combination $(n, p_0)$, we selected predictors out of a total of $p$ variables. We let $p$ to vary with the sample size as $p \doteq 10\times n^{\alpha}$, where $\alpha$ is one of 0.1, 0.25 or 0.45; the notation $a \doteq b$ means that $a$ equals the integer part of $b$. We note that in each case assumptions (A1) and (A2) are met, and the values of $p$ and $m$ are reported in the tables. Assumption (A3) is not needed as the errors are Gaussian, see Remark 2.2.

Our results are presented in Tables 1–3. The methods displayed as FDR1, FDR2, FDR3 and FDR4 in the simulation tables correspond to the FDR procedure described in Section 2, by choosing $q$ as 0.01, 0.05, 0.25 and 0.50, respectively. The BIC methods used here follow those of Zheng and Loh (1995), which we recall briefly for completeness. Their procedure starts by least squares estimation using the full model (including all $p$ predictors) and then obtaining the corresponding $p$-values $\pi_i$, as in Section 2. Let $X_{(1)},\ldots,X_{(p)}$ be the predictors corresponding to the ordered (in increasing order) $p$-values. For each $k=1,\ldots,p$, one computes the residual sum of squares $\mathrm{RSS}_k$ based on regression using the first $k$ predictors $X_{(1)},\ldots,X_{(k)}$ only. Finally, one selects the first $k^*$ predictors $X_{(1)},\ldots,X_{(k^*)}$,

Table 1
For each method, the proportion of exact selections is recorded in the row labeled Truth, the number of variables selected is recorded as Inclusions, the number of variables correctly included is recorded as Correct Inclusions and finally the average MSE is also displayed

| Results | Ideal | Procedures | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | FDR1<br>$q=0.01$ | FDR2<br>$q=0.05$ | FDR3<br>$q=0.25$ | FDR4<br>$q=0.50$ | BIC1<br>$c=0.5$ | BIC2<br>$c=1.0$ | BIC3<br>$c=2.0$ |
| $(p_0=5, n=100)$ | | | | | | | | |
| $m=0.013$, $p \doteq 10\times n^{0.1} \doteq 16$ | | | | | | | | |
| Truth | 1.000 | 0.996 | 0.938 | 0.728 | 0.562 | 0.270 | 0.718 | 0.966 |
| Inclusions | 5.000 | 5.004 | 5.064 | 5.340 | 5.674 | 6.414 | 5.352 | 5.034 |
| Correct inclusions | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 |
| A(MSE) | 0.052 | 0.053 | 0.058 | 0.072 | 0.084 | 0.107 | 0.073 | 0.056 |
| $m=0.018$, $p \doteq 10\times n^{0.25} \doteq 32$ | | | | | | | | |
| Truth | 1.000 | 0.982 | 0.940 | 0.752 | 0.624 | 0.122 | 0.508 | 0.932 |
| Inclusions | 5.000 | 5.020 | 5.074 | 5.342 | 5.640 | 8.486 | 5.838 | 5.082 |
| Correct inclusions | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 |
| A(MSE) | 0.054 | 0.056 | 0.060 | 0.076 | 0.089 | 0.185 | 0.105 | 0.062 |
| $(p_0=10, n=100)$ | | | | | | | | |
| $m=0.013$, $p \doteq 10\times n^{0.1} \doteq 16$ | | | | | | | | |
| Truth | 1.000 | 0.992 | 0.946 | 0.764 | 0.564 | 0.468 | 0.848 | 0.980 |
| Inclusions | 10.000 | 10.008 | 10.056 | 10.272 | 10.570 | 10.734 | 10.168 | 10.020 |
| Correct inclusions | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 |
| A(MSE) | 0.098 | 0.100 | 0.103 | 0.113 | 0.122 | 0.127 | 0.109 | 0.101 |
| $m=0.018$, $p \doteq 10\times n^{0.25} \doteq 32$ | | | | | | | | |
| Truth | 1.000 | 0.968 | 0.910 | 0.684 | 0.442 | 0.104 | 0.546 | 0.944 |
| Inclusions | 10.000 | 10.032 | 10.102 | 10.514 | 11.054 | 12.956 | 10.746 | 10.062 |
| Correct inclusions | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 |
| A(MSE) | 0.100 | 0.102 | 0.108 | 0.129 | 0.150 | 0.211 | 0.143 | 0.106 |

where

$$k^* = \arg \min_{1 \leqslant k \leqslant p} \{\mathrm{RSS}_k + ckS^2 \log n\},$$

for a user specified constant $c$ and $S^2 = \mathrm{RSS}_p/(n-p)$. The methods BIC1, BIC2 and BIC3 below are all performed in this way with constants $c = 0.5$, 1.0 and 2.0, respectively.

In all tables, the first column, labeled “Ideal”, is the benchmark. All the results reported here are over 500 replications. The first row, labeled “Truth” records the percentage of times we selected the true model. The second row, labeled “Inclusions”, records the number of variables, out of $p$, that are included. We reported the average number, over the 500 replications. The third row, “Correct Inclusions”, records the number of true variables that are included in the selected model. The mean squared error (MSE), averaged over simulations, is reported in the last row. The A(MSE) reported under “Ideal” has been computed by fitting the response (obtained from the true predictors) versus the true predictors, and it therefore serves as a benchmark for assessing the performance of the other methods.

Table 2
For each method, the proportion of exact selections is recorded in the row labeled Truth, the number of variables selected is recorded as Inclusions, the number of variables correctly included is recorded as Correct Inclusions and finally the average MSE is also displayed

| | | Procedures | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Results | Ideal | FDR1 | FDR2 | FDR3 | FDR4 | BIC1 | BIC2 | BIC3 |
| | | $q = 0.01$ | $q = 0.05$ | $q = 0.25$ | $q = 0.50$ | $c = 0.5$ | $c = 1.0$ | $c = 2.0$ |
| $(p_0 = 5, n = 500)$ | | | | | | | | |
| $m = 0.002$, $p \doteq 10 \times n^{0.1} \doteq 19$ | | | | | | | | |
| Truth | 1.000 | 0.986 | 0.940 | 0.704 | 0.516 | 0.314 | 0.812 | 0.992 |
| Inclusions | 5.000 | 5.014 | 5.062 | 5.366 | 5.724 | 6.112 | 5.370 | 5.204 |
| Correct inclusions | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 |
| E(MSE) | 0.010 | 0.010 | 0.011 | 0.015 | 0.018 | 0.021 | 0.013 | 0.010 |
| $m = 0.002$, $p \doteq 10 \times n^{0.25} \doteq 48$ | | | | | | | | |
| Truth | 1.000 | 0.982 | 0.940 | 0.750 | 0.526 | 0.072 | 0.616 | 0.978 |
| Inclusions | 5.000 | 5.018 | 5.060 | 5.318 | 5.698 | 8.150 | 5.534 | 5.022 |
| Correct inclusions | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 |
| A(MSE) | 0.010 | 0.010 | 0.011 | 0.015 | 0.020 | 0.040 | 0.018 | 0.010 |
| $m = 0.003$, $p \doteq 10 \times n^{0.45} \doteq 164$ | | | | | | | | |
| Truth | 1.000 | 0.996 | 0.952 | 0.776 | 0.606 | 0.010 | 0.416 | 0.948 |
| Inclusions | 5.000 | 5.004 | 5.050 | 5.256 | 5.566 | 15.144 | 6.260 | 5.060 |
| Correct inclusions | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 |
| A(MSE) | 0.010 | 0.010 | 0.011 | 0.014 | 0.018 | 0.102 | 0.030 | 0.011 |
| $(p_0 = 10, n = 500)$ | | | | | | | | |
| $m = 0.002$, $p \doteq 10 \times n^{0.1} \doteq 19$ | | | | | | | | |
| Truth | 1.000 | 0.984 | 0.918 | 0.686 | 0.436 | 0.454 | 0.884 | 0.998 |
| Inclusions | 10.000 | 10.016 | 10.084 | 10.392 | 10.814 | 10.738 | 10.126 | 10.002 |
| Correct inclusions | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 |
| A(MSE) | 0.019 | 0.020 | 0.021 | 0.024 | 0.027 | 0.026 | 0.021 | 0.019 |
| $m = 0.002$, $p \doteq 10 \times n^{0.25} \doteq 48$ | | | | | | | | |
| Truth | 1.000 | 0.978 | 0.904 | 0.588 | 0.378 | 0.070 | 0.626 | 0.984 |
| Inclusions | 10.000 | 10.022 | 10.106 | 10.554 | 11.138 | 12.956 | 10.462 | 10.016 |
| Correct inclusions | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 |
| A(MSE) | 0.020 | 0.021 | 0.022 | 0.028 | 0.034 | 0.048 | 0.027 | 0.021 |
| $m = 0.003$, $p \doteq 10 \times n^{0.45} \doteq 164$ | | | | | | | | |
| Truth | 1.000 | 0.988 | 0.926 | 0.658 | 0.420 | 0.014 | 0.476 | 0.962 |
| Inclusions | 10.000 | 10.012 | 10.084 | 10.484 | 11.028 | 20.422 | 11.140 | 10.042 |
| Correct inclusions | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 |
| A(MSE) | 0.019 | 0.020 | 0.021 | 0.026 | 0.033 | 0.111 | 0.037 | 0.021 |

For all combinations $p_0$, $p$ and $n$, the method FDR1 performs best amongst the other FDR methods, with very high percentages, 98–99%, of correct inclusions. It is followed closely in performance by FDR2. They correspond to the values of $q = 0.01$ and 0.05, respectively. These values are close to $1/p$, up to small multiplicative constants, for all $p$ considered. We opted for them as intermediate thresholds in order to study the progress or deterioration of

Table 3
For each method, the proportion of exact selections is recorded in the row labeled Truth, the number of variables selected is recorded as Inclusions, the number of variables correctly included is recorded as Correct Inclusions and finally the average MSE is also displayed.

| | | Procedures | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Results | Ideal | FDR1<br>$q = 0.01$ | FDR2<br>$q = 0.05$ | FDR3<br>$q = 0.25$ | FDR4<br>$q = 0.50$ | BIC1<br>$c = 0.5$ | BIC2<br>$c = 1.0$ | BIC3<br>$c = 2.0$ |
| $(p_0 = 5, n = 1000)$ | | | | | | | | |
| $m = 0.001$, $p \doteq 10 \times n^{0.1} \doteq 20$ | | | | | | | | |
| Truth | 1.000 | 0.990 | 0.950 | 0.742 | 0.536 | 0.400 | 0.880 | 1.000 |
| Inclusions | 5.000 | 5.010 | 5.054 | 5.324 | 5.678 | 5.914 | 5.124 | 5.000 |
| Correct inclusions | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 |
| A(MSE) | 0.005 | 0.005 | 0.005 | 0.007 | 0.009 | 0.010 | 0.006 | 0.005 |
| $m = 0.001$, $p \doteq 10 \times n^{0.25} \doteq 57$ | | | | | | | | |
| Truth | 1.000 | 0.984 | 0.942 | 0.724 | 0.552 | 0.032 | 0.646 | 0.992 |
| Inclusions | 5.000 | 5.016 | 5.062 | 5.344 | 5.664 | 8.144 | 5.452 | 5.008 |
| Correct inclusions | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 |
| A(MSE) | 0.005 | 0.005 | 0.006 | 0.008 | 0.010 | 0.021 | 0.009 | 0.005 |
| $m = 0.001$, $p \doteq 10 \times n^{0.45} \doteq 224$ | | | | | | | | |
| Truth | 1.000 | 0.990 | 0.944 | 0.790 | 0.650 | 0.006 | 0.428 | 0.966 |
| Inclusions | 5.000 | 5.010 | 5.056 | 5.250 | 5.498 | 16.720 | 6.130 | 5.036 |
| Correct inclusions | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 |
| A(MSE) | 0.005 | 0.005 | 0.006 | 0.007 | 0.010 | 0.063 | 0.015 | 0.006 |
| $(p_0 = 10, n = 1000)$ | | | | | | | | |
| $m = 0.001$, $p \doteq 10 \times n^{0.1} \doteq 20$ | | | | | | | | |
| Truth | 1.000 | 0.996 | 0.932 | 0.680 | 0.450 | 0.500 | 0.920 | 1.000 |
| Inclusions | 10.000 | 10.004 | 10.078 | 10.370 | 10.786 | 10.644 | 10.090 | 10.000 |
| Correct inclusions | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 |
| A(MSE) | 0.010 | 0.010 | 0.011 | 0.012 | 0.014 | 0.013 | 0.011 | 0.010 |
| $m = 0.001$, $p \doteq 10 \times n^{0.25} \doteq 57$ | | | | | | | | |
| Truth | 1.000 | 0.976 | 0.904 | 0.628 | 0.370 | 0.052 | 0.688 | 0.988 |
| Inclusions | 10.000 | 10.024 | 10.108 | 10.520 | 11.094 | 12.846 | 10.412 | 10.012 |
| Correct inclusions | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 |
| A(MSE) | 0.010 | 0.010 | 0.011 | 0.014 | 0.017 | 0.025 | 0.014 | 0.010 |
| $m = 0.001$, $p \doteq 10 \times n^{0.45} \doteq 224$ | | | | | | | | |
| Truth | 1.000 | 0.984 | 0.938 | 0.648 | 0.424 | 0.006 | 0.402 | 0.964 |
| Inclusions | 10.000 | 10.016 | 10.068 | 10.430 | 10.932 | 21.598 | 11.212 | 10.04 |
| Correct inclusions | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 |
| A(MSE) | 0.005 | 0.005 | 0.006 | 0.008 | 0.010 | 0.025 | 0.012 | 0.006 |

the method as $q$ increases. We observed the most drastic changes for $q \geqslant 0.1$, with marked decline in the percentage of perfect selection starting at $q = 0.25$ (FDR3) and continuing as $q$ increases, as recorded for $q = 0.5$ (FDR4). This is consistent with our theoretical considerations regarding the choice of the control parameter $q$. The methods BIC2 and BIC1 perform poorly, and are comparable with FDR3 and FDR4. This is an illustration of

Table 4
For each method, the proportion of exact selections is recorded in the row labeled Truth, the number of variables selected is recorded as Inclusions, the number of variables correctly included is recorded as Correct Inclusions and finally the average MSE is also displayed.

| | | $p$-value threshold | | | | | |
|---|---|---|---|---|---|---|---|
| Results | Ideal | 0.01 | 0.05 | 0.1 | $0.01/p$ | $0.05/p$ | $0.1/p$ |
| $(p_0 = 5, n = 500)$ | | | | | | | |
| $p \doteq 10 \times n^{0.1} \doteq 19$ | | | | | | | |
| Truth | 1.000 | 0.860 | 0.500 | 0.232 | 0.988 | 0.964 | 0.916 |
| Inclusions | 5.000 | 5.150 | 5.700 | 6.428 | 5.012 | 5.036 | 5.084 |
| Correct inclusions | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 |
| A(MSE) | 0.010 | 0.012 | 0.018 | 0.022 | 0.010 | 0.011 | 0.012 |
| $p \doteq 10 \times n^{0.25} \doteq 48$ | | | | | | | |
| Truth | 1.000 | 0.654 | 0.118 | 0.018 | 0.998 | 0.970 | 0938 |
| Inclusions | 5.000 | 5.420 | 7.110 | 9.292 | 5.002 | 5.030 | 5.064 |
| Correct inclusions | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 |
| A(MSE) | 0.010 | 0.016 | 0.032 | 0.046 | 0.010 | 0.011 | 0.011 |
| $p \doteq 10 \times n^{0.45} \doteq 164$ | | | | | | | |
| Truth | 1.000 | 0.240 | 0.002 | 0.000 | 0.986 | 0.944 | 0.912 |
| Inclusions | 5.000 | 6.552 | 12.752 | 20.688 | 5.014 | 5.056 | 5.092 |
| Correct inclusions | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 | 5.000 |
| A(MSE) | 0.010 | 0.029 | 0.079 | 0.126 | 0.010 | 0.011 | 0.011 |
| $(p_0 = 10, n = 500)$ | | | | | | | |
| $p \doteq 10 \times n^{0.1} \doteq 19$ | | | | | | | |
| Truth | 1.000 | 0.722 | 0.194 | 0.028 | 0.990 | 0.954 | 0.930 |
| Inclusions | 10.000 | 10.352 | 11.848 | 13.658 | 10.010 | 10.046 | 10.070 |
| Correct inclusions | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 |
| A(MSE) | 0.020 | 0.025 | 0.039 | 0.050 | 0.020 | 0.021 | 0.021 |
| $p \doteq 10 \times n^{0.25} \doteq 48$ | | | | | | | |
| Truth | 1.000 | 0.692 | 0.160 | 0.024 | 0.998 | 0.978 | 0.940 |
| Inclusions | 10.000 | 10.400 | 11.980 | 13.952 | 10.002 | 10.024 | 10.064 |
| Correct inclusions | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 |
| A(MSE) | 0.021 | 0.027 | 0.042 | 0.054 | 0.021 | 0.021 | 0.022 |
| $p \doteq 10 \times n^{0.45} \doteq 164$ | | | | | | | |
| Truth | 1.000 | 0.232 | 0.000 | 0.000 | 0.994 | 0.964 | 0.912 |
| Inclusions | 10.000 | 11.590 | 17.726 | 25.386 | 10.006 | 10.036 | 10.090 |
| Correct inclusions | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 | 10.000 |
| A(MSE) | 0.021 | 0.040 | 0.089 | 0.134 | 0.021 | 0.021 | 0.022 |

the importance of the choice of the constant $c$ in the BIC penalty. However, BIC3 shows excellent performance, comparable with FDR1 and FDR2. Thus, either of FDR or BIC leads to consistent selection, with the correct calibration of the parameter of the method, but the FDR method offers the advantage of increased computational speed. It is interesting to see that for all scenarios under consideration, and all methods, although there are many instances in which we do not have perfect selection, the true model is always included in the selected one. Hence, in all these cases we overestimate the model, but, with the occasional exception of BIC1, all other methods include, on average, less than one additional variable. In particular, all the FDR methods exhibit excellent behavior in this respect.

We also conducted variable selection, in the settings of Tables 1–3, respectively, by comparing each of the $p$-values with fixed thresholds. The results were very similar in all three scenarios, and we only report here, in Table 4, those obtained under the simulation design used for Table 2. Columns 2–4 correspond to comparing each $p$-value with $q = 0.01, 0.05$ and $0.1$, respectively, in which case consistency is no longer guaranteed, with substantial degradation as $p$ increases. The last three columns correspond to the Bonferroni method with $q = 0.01/p$, $0.05/p$ and $0.1/p$, respectively. The results support strongly the theoretical findings of Section 2, and we notice that the best performance is achieved for the Bonferroni method with $q = 0.01$, which is on par with the FDR1 ($q = 0.01$) method in Table 2.

## Acknowledgements

We are grateful to two referees for useful remarks that improved the quality of the paper.

## Appendix A.

**Lemma A.1.** *For the event $A_n$ defined in* (2.8), $\lim_{n\to\infty} \mathbb{P}(A_n) = 1$.

**Proof.** Set $\delta = \{(\log n)/n\}^{1/2}$ and let $I$ be the identity matrix in $\mathbb{R}^n$. Observe that for $\lambda > 0$ small enough,

$$\begin{aligned}\mathbb{P}(A_n^c) &\leqslant \mathbb{P}\{|S^2 - \sigma^2| \geqslant \sigma\delta\}\\ &\leqslant \frac{\mathbb{E}|\boldsymbol{\varepsilon}^{\mathrm{T}}(I-H)\boldsymbol{\varepsilon} - (n-p)\sigma^2|^{2\lambda}}{\left[(n-p)\sigma\delta\right]^{2\lambda}}\\ &\leqslant C(\lambda)\mathbb{E}|\varepsilon|^{4\lambda}\frac{[\mathrm{trace}(I-H)]^{\lambda}}{\left[(n-p)\delta\sigma\right]^{2\lambda}} \quad \text{by Rao and Kleffe (1988)}\\ &= C(\lambda,\sigma)[(n-p)\delta^2]^{-\lambda}\end{aligned}$$

which tends to zero as $n \to \infty$. □

**Lemma A.2.** *Set* $\tau_3 \overset{\mathrm{def}}{=} \mathbb{E}|\varepsilon_1/\sigma|^3$. *Let $G_{ni}$ be the distribution function of* $(\widehat{\beta}_i - \beta_i)/(\sigma\sqrt{m_{ii}})$. *Then*

$$\|G_{ni} - \Phi\|_\infty \leqslant 9\tau_3 \max_{1\leqslant k\leqslant n} \sqrt{h_{kk}}.$$

**Proof.** Write $U$ for the matrix with the eigenvectors of $X^{\mathrm{T}}X$ as its column vectors, and let $X^{\mathrm{T}}X$ be the diagonal matrix with the eigenvalues of $M$ as its diagonal elements.

Then

$$X^{\mathrm{T}}X = U\Lambda U^{\mathrm{T}}, \quad M = (X^{\mathrm{T}}X)^{-1} = U\Lambda^{-1}U^{\mathrm{T}}$$

are the eigenvalue decompositions of $X^{\mathrm{T}}X$ and $M$, respectively. Write $B = U\Lambda^{1/2}U^{\mathrm{T}}$, so that $X^{\mathrm{T}}X = B^2$ and $M = B^{-2}$. Let $e_i$ be the $i$th unit vector, and observe that

$$\frac{\widehat{\beta}_i - \beta_i}{\sigma\sqrt{m_{ii}}} = e_i^{\mathrm{T}}\sigma^{-1}B(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}).$$

Furthermore, write

$$B(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = U\Lambda^{-1/2}U^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{\varepsilon} \overset{\text{def}}{=} F\boldsymbol{\varepsilon},$$

and

$$e_i^{\mathrm{T}}B(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \overset{\text{def}}{=} \sum_{k=1}^{n} a_k\boldsymbol{\varepsilon}_k,$$

with $a_k = a_{nk} = \langle e_i, f_k\rangle$ and $f_k$ is the $k$th column vector of $F$. It is easily verified that $F^{\mathrm{T}}F = H$ and $FF^{\mathrm{T}} = I$, whence by Cauchy–Schwarz,

$$\max_k |a_k| \leqslant \max_k \|f_k\| \|e_i\| = \max_k \sqrt{h_{kk}},$$

and

$$\sum_{k=1}^{n} a_k^2 = \|F^{\mathrm{T}}e_i\|^2 = \|e_i\|^2 = 1.$$

Using the Berry–Esseen bound for sums of independent random variables (cf., e.g., Shorack, 2000, p. 259), we find

$$\begin{aligned}\|G_{ni} - \Phi\|_\infty &\leqslant 9\sum_{k=1}^{n} \mathbb{E}|a_k\varepsilon_k/\sigma|^3\\ &\leqslant 9\tau_3 \max_{1\leqslant k\leqslant n} |a_k|\\ &\leqslant 9\tau_3 \max_{1\leqslant k\leqslant n} \sqrt{h_{kk}}.\end{aligned}$$

The proof of the lemma is complete. □

**Lemma A.3.** *For $j \notin I_0$ and any $\xi > 0$,*

$$\mathbb{P}(\{\pi_j \leqslant \xi\} \cap A_n) = \xi + \mathrm{O}(\delta + r). \tag{A.1}$$

*For $j \in I_0$ and any $\xi > 0$,*

$$\mathbb{P}\left(\{\pi_j \geqslant \xi\} \cap A_n\right) = \mathrm{O}(e^{-m^{-1}}\sqrt{\log(1/\xi)} + r + \delta). \tag{A.2}$$

**Proof.** Set $\mu_j = \beta_j/\sqrt{\sigma^2 m_{jj}}$. For $j \notin I_0$, $\beta_j = 0$ and we find using Lemma A.2 that

$$\begin{aligned}\mathbb{P}(\{\pi_j \leqslant \xi\} \cap A_n) &= \mathbb{P}\left[\left\{\left|\frac{\widehat{\beta}_j - \beta_j}{\sqrt{m_{jj}}S}\right| \geqslant \Phi^{-1}\left(\frac{2-\xi}{2}\right)\right\} \cap A_n\right] \\ &\leqslant \mathbb{P}\left[\left|\frac{\widehat{\beta}_j - \beta_j}{\sqrt{m_{jj}}\sigma}\right| \geqslant \left(1 - \frac{\delta}{\sigma}\right)\Phi^{-1}\left(\frac{2-\xi}{2}\right)\right] \\ &= \xi + \mathrm{O}(\delta + r).\end{aligned}$$

On the other hand, for all $j \in I_0$, we have for all $0 < \xi < 1$

$$\begin{aligned}&\mathbb{P}\left(\{\pi_j \geqslant \xi\} \cap A_n\right) \\ &= \mathbb{P}\left[\left\{\left|\frac{\widehat{\beta}_j - \beta_j}{\sqrt{m_{jj}}S} + \frac{\beta_j}{\sqrt{m_{jj}}S}\right| \geqslant \Phi^{-1}\left(\frac{2-\xi}{2}\right)\right\} \cap A_n\right] \\ &\leqslant \Phi\left(\Phi^{-1}\left(\frac{2-\xi}{2}\right) - \mu_j\right) - \Phi\left(-\Phi^{-1}\left(\frac{2-\xi}{2}\right) - \mu_j\right) + \mathrm{O}(r + \delta) \\ &= \mathrm{O}(e^{-1/m}\sqrt{\log(1/\xi)} + r + \delta).\end{aligned}$$

We used in the last two lines the mean-value theorem, Lemma A.2 and the fact that

$$\min_{j \in I_0} \mu_j^2 = \min_{j \in I_0} \frac{\beta_j^2}{\sigma^2 m_{jj}} \geqslant C m^{-1}$$

for $n$ large enough and some finite constant $C > 0$. □

## References

Abramovich, F., Benjamini, Y., Donoho, D., Johnstone, I., 2000. Adapting to unknown sparsity by controlling the false discovery rate. Technical Report, Department of Statistics, Stanford University, Stanford. (available from http://www-stat.stanford.edu/∼imj)

Bauer, P., Pötscher, B.M., Hackl, P., 1988. Model selection by multiple test procedures. Statistics 19, 39–44.

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

Benjamini, Y., Yekutieli, D., 2001. The control of the false discovery rate in multiple testing under dependency. Ann. Statist. 29, 1165–1188.

Bunea, F., 2004. Consistent covariate selection and postmodel selection inference in semiparametric regression. Ann. Statist. 32, 898–927.

Eicher, F., 1965. Limit theorems for regression with unequal and dependent errors. In: Proceedings of Fifth Berkeley Symposium on Mathematical Statistics and Probability Statistics 1, pp. 59–82.

Genovese, C., Wasserman, L., 2004. A stochastic process approach to false discovery rates. Ann. Statist. 32, in press.

Geweke, J., Meese, R., 1981. Internat. Econom. Rev. 22 (1), 55–70.

Hannan, E.J., 1980. The estimation of the order of an ARMA process. Ann. Statist. 8, 1071–1081.

Hannan, E.J., Quinn, B.G., 1979. The determination of the order of an autoregression. J. Roy. Statist. Soc. B 41 (2), 190–195.

Haughton, D., 1988. On the choice of a model to fit data from an exponential. Ann. Statist. 16, 342–355.

Jiang, W., Liu, X., 2004. Consistent model selection based on parameter estimates. J. Statist. Plann. Inference 121, 265–283.

Pötscher, B.M., 1983. Order estimation in ARMA models by Lagrange multiplier tests. Ann. Statist. 11, 872–885.

Rao, C.R., Kleffe, J., 1988. Estimation of Variance Components and Applications. North-Holland, Amsterdam.

Sen, H., Srivastava, M., 1990. Regression Analysis. Springer, New York.

Shao, J., 1993. Linear model selection by cross-validation. J. Amer. Statist. Assoc. 88, 486–494.

Shorack, G.R., 2000. Probability for Statisticians. Springer, New York.

Storey, J., 2002. A direct approach to false discovery rates. J. Roy. Statist. Soc. B 64, 479–498.

Woodroofe, M., 1982. On model selection and the arcsine laws. Ann. Statist. 10, 1182–1194.

Zheng, X., Loh, W.L., 1995. Consistent variable selection in linear models. J. Amer. Statist. Assoc. 90, 151–156.