Biometrika (2018), **00**, 0, *pp.* 1–11
Printed in Great Britain
doi: 10.1093/biomet/asy032

When is the first spurious variable selected by sequential regression procedures?

BY WEIJIE J. SU

Department of Statistics, University of Pennsylvania, 472 John M. Huntsman Hall, 3730 Walnut Street, Philadelphia, Pennsylvania 19104, U.S.A.

suw@wharton.upenn.edu

SUMMARY

Applied statisticians use sequential regression procedures to rank explanatory variables and, in settings of low correlations between variables and strong true effect sizes, expect that variables at the top of this ranking are truly relevant to the response. In a regime of certain sparsity levels, however, we show that the lasso, forward stepwise regression, and least angle regression include the first spurious variable unexpectedly early. We derive a sharp prediction of the rank of the first spurious variable for these three procedures, demonstrating that it occurs earlier and earlier as the regression coefficients become denser. This phenomenon persists for statistically independent Gaussian random designs and arbitrarily large true effects. We gain insight by identifying the underlying cause and then introduce a simple visualization tool termed the double-ranking diagram to improve on these methods. We obtain the first result establishing the exact equivalence between the lasso and least angle regression in the early stages of solution paths beyond orthogonal designs. This equivalence implies that many important model selection results concerning the lasso can be carried over to least angle regression.

Some key words: False variable; Familywise error rate; Forward stepwise regression; Lasso; Least angle regression.

1. INTRODUCTION

Consider observing an n-dimensional response vector

$$y = X\beta + z,$$

where $X \in \mathbb{R}^{n\times p}$ is a design matrix, $\beta \in \mathbb{R}^p$ is a vector of regression coefficients, and $z \in \mathbb{R}^n$ is a noise term. To find explanatory variables that are associated with the response y, especially in the setting where $p > n$, forward stepwise regression, the lasso (Tibshirani, 1996) and least angle regression (Efron et al., 2004) are frequently used. These methods sequentially add or remove variables based on some criterion. A sequential method ranks explanatory variables according to when the variables enter the solution path. A routine practice for forming the final model is to select all variables ranked earlier than a certain cut-off and discard the rest.

A practitioner often wishes to understand where along the solution path regressors with zero regression coefficients, henceforth referred to as noise variables, start to enter the model. In particular, when is the first noise variable selected? A better understanding of this is desirable from at least two perspectives. First, the rank of the first noise variable sheds light on the difficulty

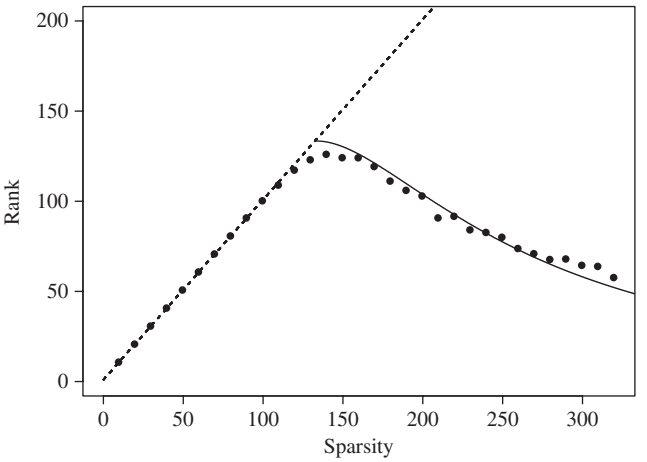

Fig. 1. Rank of the first spurious variable along the lasso path. The rank equals 1 plus the number of signal variables preceding the first spurious variable. Averages from 500 independent replicates are plotted as dots; the 45° dashed line is shown for comparison. The solid curve represents $\exp[\{(2n\log p)/k\}^{1/2} - n/(2k) + \log\{n/(2p\log p)\}]$ as a function of k, starting from $n/(2\log p)$.

of consistent model selection, offering guidelines for selecting important variables. If the rank is about the same size as the sparsity, i.e., the total number of nonzero regression coefficients, we could obtain a model that retains most of the important variables without leading to many false selections using a sequential method, whereas a small rank implies that signal variables, i.e., regressors with nonzero regression coefficients, and noise variables are interspersed early on in the solution path, so a false selection must occur long before all signal variables are selected. Second, the empirical performance of numerous tools for post-selection inference in linear regression is, to a large extent, contingent upon whether the first noise variable occurs early or not (Lockhart et al., 2014; G'Sell et al., 2016; Tibshirani et al., 2016). Insights into the occurrence of the first false variable would be valuable for improving these tools and developing new ones.

Despite an extensive body of work on these sequential methods, the literature remains relatively silent on the first false variable. Existing results address these questions in a limited setting, mostly characterizing under what conditions all the signal variables precede the first noise variable, corresponding to perfect support recovery or, put more simply, selection of the exactly correct model. These results guarantee perfect support recovery using a certain sequential method, given sufficiently strong effect sizes compared to the noise level and a form of local orthogonality of the design matrix; they have been established for the lasso (Zhao & Yu, 2006; Bickel et al., 2009; Wainwright, 2009) and forward stepwise regression (Tropp, 2004; Zhang, 2009; Cai & Wang, 2011).

Figure 1 illustrates when the first noise variable is selected by the lasso in simulations. The 2000×1800 design matrix X consists of independent $N(0, 1/2000)$ entries, the noise term z comprises independent standard normal variables, and the regression coefficients are $\beta_1 = \cdots = \beta_k = 100\sqrt{(2\log 1800)} = 387{\cdot}2$ and $\beta_j = 0$ for all $j > k$, with the sparsity k varying from 10 to 320. The true effect sizes are effectively infinitely strong, and the sample correlations between the regressors are small due to the independence. Figure 1 shows that, in the low sparsity regime, the sparsity/rank pairs lie close to the 45° line. This is equivalent to saying that all the signal variables are selected prior to any false variables, in agreement with theoretical results.

Once the sparsity exceeds 140, a phenomenon that is not explained by existing theory occurs: the average rank of the first noise variable becomes substantially smaller than the sparsity k, and it decreases as the sparsity increases. This demonstrates the impossibility of perfect support recovery in this nonextreme sparsity regime using the lasso, even under high signal-to-noise ratios and low correlations. As the signal β is amplified by setting more components to a large magnitude, one might anticipate that a sequential method such as the lasso tends to include more signal variables at the beginning, and thus one would imagine that the first noise variable would get selected later and later. Figure 1 falsifies this belief. A similar phenomenon was observed in Su et al. (2017), though without any justification.

We derive an analytical prediction that is asymptotically exact for the first noise variables for the lasso, forward stepwise regression, least angle regression, and potentially other sequential methods. Let T denote the rank of the first noise variable. Informally, the prediction states that, in the setting of strong effect sizes and statistically independent regressors as in Fig. 1, the three sequential procedures in the nonextreme sparsity regime all satisfy

$$\log T \approx \{(2n \log p)/k\}^{1/2} - n/(2k) + \log\{n/(2p \log p)\}. \tag{1}$$

The formal statement of this result is given in Theorem 2.

The prediction of T presented in Fig. 1 shows excellent agreement between the predicted and observed behaviours. To better appreciate this, note that the approximation to $\log T$ in (1) is smaller than $\log k$ once k exceeds $n/(2 \log p)$, suggesting the impossibility of perfect support recovery in this regime. This is consistent with Corollary 2 of Wainwright (2009). The prediction (1) implies more, however. To see this, write the right-hand side of (1) as

$$-\big[(\log p)^{1/2} - \{n/(2k)\}^{1/2}\big]^2 + \log\{n/(2 \log p)\},$$

which reveals that the predicted $\log T$ decreases as the sparsity $k \geqslant n/(2 \log p)$ increases. Put differently, the first noise variable is bound to occur earlier as the signal vector β gets denser, successfully predicting the phenomenon shown in Fig. 1. While problems in selecting the true model by the lasso have been empirically documented (Fan & Song, 2010), such analytical predictions have so far been unavailable in the literature.

The above result has several implications. First, once the underlying signals go beyond the very sparse regime, using sequential procedures will inevitably lead to a very small number of selections with familywise error rate control, the probability of selecting at least one noise variable, no matter how large the effect sizes are. If n and p are equal and large and $k = \epsilon p$ for some fixed $0 < \epsilon < 1$, the prediction asserts that the first false variable is included after no more than

$$\exp\big[\{1 + o(1)\}\{\sqrt{(2 \log p/\epsilon)} - 1/(2\epsilon) - \log(2 \log p)\}\big] = \exp\big[\{1 + o(1)\}\sqrt{(2 \log p/\epsilon)}\big]$$

steps. For a fixed ϵ, however, the predicted rank $\exp[\{1 + o(1)\}(2 \log p/\epsilon)^{1/2}]$ only accounts for a vanishing fraction of the $k = \epsilon p$ signal variables, as $(2 \log p/\epsilon)^{1/2} = o(\log p)$. In other words, the three sequential methods yield vanishing power if no noise variable may be included, even if $z = 0$. These negative results are derived under Gaussian designs with independent columns, which have vanishing sample correlations and satisfy conditions believed to be favourable for model selection, including restricted isometry properties (Candès & Tao, 2005) and restricted eigenvalue conditions (Bickel et al., 2009). Thus, the negative results are likely to carry over to a much broader class of design matrices. Simulations in § 3 demonstrate this in more general settings.

Another implication is that the three sequential regression methods seem to behave similarly in ranking variables, at least in the independent random design setting. Compared with forward stepwise regression, the lasso and least angle regression are considered less greedy because at each step they gradually blend in a new variable instead of adding it in a discontinuous manner (Efron et al., 2004). Forward stepwise regression selects the predictor having the largest absolute correlation with the residual vector and then takes a large step in the direction of the selected predictor, whereas the other methods proceed along a direction equiangular between the set of selected predictors. This distinction between the two strategies could lead to contrasting model selection performance. But this is not the case; the two strategies yield the same behaviour in terms of selecting the first noise variables in our setting. Theorem 3 is the first result establishing the exact equivalence between early solution paths of the lasso and least angle regression beyond orthogonal designs.

In the nonextreme sparsity regime, why do these distinct sequential methods select the first false variable so early? Loosely speaking, it is due to their greedy nature. Moreover, the equiangular strategy adopted by the lasso and least angle regression fails to alleviate greediness from the perspective of when the first noise variable gets selected. All three methods include a variable that roughly has the largest absolute inner product with the current residual. As the regression coefficients get denser, the solution at the beginning of the path is overwhelmingly biased and the residual vector absorbs many of the effects contributed by the nonzero components of β. As a result, some irrelevant variable would exhibit high correlations with the residual and hence be selected early.

2. Understanding the phenomenon

2·1. *Predicting the first spurious variable*

We consider a sequence of problems indexed by (k_l, n_l, p_l), where k_l, n_l and p_l are all assumed to grow to infinity as $l \to \infty$ in asymptotic statements. The subscript l is often omitted when clear from the context. Letters c_i and C_i in various settings denote positive constants that do not depend on the problem index l. Below we formalize our working hypothesis concerning the linear model $y = X\beta + z$.

Assumption 1. The design $X \in \mathbb{R}^{n\times p}$ has independent $N(0, 1/n)$ entries, and $z \in \mathbb{R}^n$ consists of independent $N(0, \sigma^2)$ errors; X and z are independent. The coefficient vector β has k fixed components equal to $M \neq 0$, with the rest all being zero. Last, we assume that $c_1 p/\log^{c_2} p \leqslant n \leqslant c_3 p$ and $c_4 n \leqslant k \leqslant \min(0{\cdot}99p, c_5 n \log^{0{\cdot}99} p)$ for arbitrary positive constants c_1, c_2, c_3, c_4 and c_5.

The assumption on (k, n, p) is satisfied in some popular examples, such as the linear sparsity framework where k/p and n/p converge to constants (Bayati & Montanari, 2012). Moreover, a number of cases leading to $k = o(p)$ satisfy Assumption 1, for example $n = c_1 p/\log^{c_2} p$ and $k = c_4 n$. Under this assumption, each column of X is approximately normalized, having about unit Euclidean norm. This random design is conventionally considered to be easy for model selection since it obeys restricted isometry properties (Candès & Tao, 2005) or restricted eigenvalue conditions (Bickel et al., 2009) with high probability. The nonrandom parameters $\sigma \geqslant 0$ and M both implicitly depend on the index l and are allowed to vary freely. The noiseless case $\sigma = 0$ is not excluded.

Throughout this paper, the intercept is not included and normalization is not applied to any columns of X. Let T denote the rank of the first noise variable, and let $o_{\mathrm{pr}}(1)$ denote a sequence of random variables converging to zero in probability.

THEOREM 1. *Under Assumption 1, the first spurious variable selected by each of forward stepwise regression, the lasso, and least angle regression satisfies*

$$\log T \leqslant \{1 + o_{\mathrm{pr}}(1)\}\big[\{2n(\log p)/k\}^{1/2} - n/(2k) + \log\{n/(2p\log p)\}\big].$$

If the signal magnitude M is not sufficiently large compared with σ, the logarithm of T would be much smaller than the upper bound, so the first false variable could arrive even earlier. This bound is sharp when M is large enough relative to σ.

THEOREM 2. *Under Assumption 1 and if $\sigma/M \to 0$, the three sequential methods satisfy*

$$\log T = \{1 + o_{\mathrm{pr}}(1)\}\big[\{2n(\log p)/k\}^{1/2} - n/(2k) + \log\{n/(2p\log p)\}\big]. \tag{2}$$

The proofs of both theorems, provided in the Supplementary Material, involve techniques that can be extended beyond the three sequential methods. The condition concerning the ratio between σ and M can be relaxed to $|M|/\sigma \gg \sqrt{(n/k)}$. Setting $k \approx n\log^{0{\cdot}99} p$ as in Assumption 1, for example, Theorem 2 follows if M/σ is bounded away from zero. An immediate consequence of this theorem is the following.

COROLLARY 1. *Under Assumption 1, the prediction (2) holds for all three of the methods in the noiseless case.*

In addition to explaining Fig. 1, Theorem 2 and Corollary 1 demonstrate that having an even stronger signal magnitude does not affect T much provided it exceeds a certain level.

Our theorems differ from results in the literature claiming a high probability of selecting the exactly correct model, mainly due to assuming different sparsity regimes of the regression coefficients β. We assume $c_4 n \leqslant k \leqslant \min(0{\cdot}99p, c_5 n\log^{0{\cdot}99} p)$ rather than a restrictive sparsity regime such as $k = O(n/\log p)$ or $k \ll n/\log p$. Under Assumption 1, it is unrealistic to expect perfect model selection using sequential methods: a corollary of Theorem 1 shows that the number of signal variables before the first false variable accounts for only an insignificant fraction of the total number of signal variables.

COROLLARY 2. *Under Assumption 1, each of the three methods satisfies $T/k \to 0$ in probability as $l \to \infty$.*

Consider the scenario where $k/p \to \epsilon$ and $n/p \to \delta$ for positive constants $\epsilon < 1$ and δ. Theorem 1 shows that, up to a vanishing fraction, the logarithm of T is no larger than $\{2\delta(\log p)/\epsilon\}^{1/2} - \delta/(2\epsilon) + \log\{\delta/(2\log p)\} = \{1 + o(1)\}\{2\delta(\log p)/\epsilon\}^{1/2}$. This expression for approximating $\log T$ yields $T \leqslant \exp[\{1 + o(1)\}\{2\delta(\log p)/\epsilon\}^{1/2}] \ll \epsilon p = k$, confirming Corollary 2 in this linear sparsity regime. We summarize this result as follows.

COROLLARY 3. *Under Assumption 1, if $k/p \to \epsilon$ and $n/p \to \delta$ for arbitrary positive constants $\epsilon < 1$ and δ, the three methods satisfy*

$$T \leqslant \exp\big[\{1 + o_{\mathrm{pr}}(1)\}\{2\delta(\log p)/\epsilon\}^{1/2}\big].$$

This linear sparsity regime was employed by Su et al. (2017), who studied limitations of the lasso for false discovery rate control. The techniques developed in that work are not applicable to studying the first noise variable, which is a much more delicate problem.

2·2. *Equivalence between the lasso and least angle regression*

In contrast to the other two methods, the lasso would drop a selected variable if its coefficient hits zero. This irregularity of the lasso path could lead to ambiguity in interpreting the rank T in Theorems 1 and 2. The next theorem rules out the possibility of such ambiguity.

THEOREM 3. *Assume that X has independent $N(0, 1/n)$ entries. Then, with probability at least $1 - p^{-2}$, no drop-out occurs before the first $\min\{\lceil c(n/\log p)^{1/2}\rceil, p\}$ variables along the lasso path are selected, where $\lceil x\rceil$ denotes the smallest integer greater than or equal to x and $c > 0$ is some universal constant.*

As seen from its proof in the Supplementary Material, the validity of Theorem 3 depends on the design matrix X only through its restricted isometry property. Thus, this result can seamlessly carry over to other matrix ensembles with an appropriate restricted isometry property constant, such as Bernoulli random matrices (Candès & Tao, 2005).

Assumption 1 implies that $\log[\min\{\lceil c(n/\log p)^{1/2}\rceil, p\}] \gg \{2n(\log p)/k\}^{1/2} - n/(2k)$. Consequently Theorem 3, together with Theorem 1, ensures that the first noise variable selected by the lasso is not preceded by any drop-out with probability approaching 1.

This provides new insights into the lasso path. The lasso is known to coincide exactly with least angle regression until the first time the lasso drops a selected variable (Efron et al., 2004; Tibshirani & Taylor, 2011). To the best of our knowledge, however, the question of where along the path the lasso and least angle regression first differ has not been addressed. By confirming the equivalence between them, Theorem 3 allows us to carry well-known results on lasso model selection over to least angle regression.

2·3. *Heuristics and insights*

In this section we give an informal derivation of Theorems 1 and 2. For the full proofs, see the Supplementary Material.

We focus on the noiseless case $z = 0$ in Assumption 1, which is the ideal scenario for model selection. Let $S = \{j : \beta_j \neq 0\}$ denote the support of the signals and let $\hat{\beta}$ denote an estimate given by any of the three methods somewhere along the solution path. Write $j_1 \notin S$ for the index outside the support having the largest inner product in magnitude with the residual $y - X\hat{\beta} = X(\beta - \hat{\beta})$, and write $j_2 \in S$ for the index on the support having the $(T-1)$th largest inner product in magnitude with the residual. Using arguments given in the Supplementary Material, we obtain that

$$X_{j_1}^{\mathrm{T}} X(\beta - \hat{\beta}) \approx M \left\{ \frac{2k \log(p-k)}{n} \right\}^{1/2}, \quad X_{j_2}^{\mathrm{T}} X(\beta - \hat{\beta}) \approx M + M \left\{ \frac{2k \log(k/T)}{n} \right\}^{1/2}, \quad (3)$$

where X^{T} denotes the transpose of X. As the sequential methods rank variables approximately according to the correlations with the residual, where in our case correlations are roughly equivalent to inner products since the columns of X are approximately normalized, from (3) we must have

$$M\{2k \log(p-k)/n\}^{1/2} \approx M + M\{2k \log(k/T)/n\}^{1/2}$$

at the point where the first false variable is just about to enter the model. In the linear sparsity regime $k/p \to \epsilon$ and $n/p \to \delta$, so $\log T \approx (2\delta \log p/\epsilon)^{1/2}$.

This suggests that an early spurious variable is mainly due to a large inner product $X_{j_1}^{\mathrm{T}} X(\beta - \hat{\beta})$, which would not be the case if $\hat{\beta}$ were a low-bias estimator of β. However, until a significant proportion of the variables has been selected, a solution $\hat{\beta}$ provided by a sequential method will be overwhelmingly biased. Another way to formalize this is by noting that the residual $X(\beta - \hat{\beta})$ still contains appreciable true effects, largely contributed by presently unselected variables. This bias acts as if it were noise, so some irrelevant variables happen to correlate highly with the residual vector, leading to false variables being selected early. This is not a matter of the signal-to-noise ratio; an increasing signal magnitude would enlarge the bias as well, and hence noise variables always occur early. Other examples of pseudo-noise caused by bias have been observed in previous work (Bayati & Montanari, 2012). This phenomenon does not appear in regimes of extreme sparsity (Wainwright, 2009).

3. Illustrations

3·1. *Numerical examples*

We present simulation experiments to illustrate the predictions given by Theorems 1 and 2. We conduct three studies examining the effects of design matrix shapes, signal magnitudes, and correlations between the columns of X on the first spurious variable. Two scenarios are considered for each study.

Study 1. In the first experiment the design X of size 1000×1000 has independent $N(0, 1/1000)$ entries, the signals are given by $\beta_j = 100$ for $j \leqslant k$ and $\beta_j = 0$ for $j \geqslant k + 1$, and each noise component z_i follows $N(0, 1)$ independently. In the second experiment the design X is changed to be of size 800×1200 and has independent Bernoulli entries, which take value $1/\sqrt{500}$ with probability $1/2$ and value $-1/\sqrt{500}$ otherwise, while all the other assumptions remain the same. The results are shown in Figs. 2(a) and (b).

Study 2. In both experiments, the 500×1000 design matrix X consists of independent $N(0, 1/500)$ entries and each noise component is independently distributed as $N(0, 1)$. For the first experiment, we set $\beta_j = M$ for $j = 1, \ldots, 80$ and $\beta_j = 0$ for $j = 81, \ldots, 1000$. For the second one, we set $\beta_j = M$ for $j = 1, \ldots, 40$, $\beta_j = M^2/\{10\sqrt{(2 \log p)}\}$ for $j = 41, \ldots, 80$ and $\beta_j = 0$ for $j = 81, \ldots, 1000$. The parameter M is varied from $0{\cdot}2\sqrt{(2 \log p)}$ to $10\sqrt{(2 \log p)}$. The two mixtures take the same value when $M = 10\sqrt{(2 \log p)} = 37{\cdot}17$. The results are shown in Figs. 2(c) and (d).

Study 3. This scenario uses β such that $\beta_j = 100\sqrt{(2 \log p)}$ for $j \leqslant 80$ and $\beta_j = 0$ otherwise. The noise z consists of independent standard normal variables. The 500×1000 design matrix X has each row independently drawn from $N(0, \Sigma)$. For the 1000×1000 covariance matrix Σ, the first experiment assumes that $\Sigma_{ij} = \rho/n$ if $i \neq j$ and $\Sigma_{jj} = 1/n$. In the second experiment, $\Sigma_{ij} = \rho^{|i-j|}/n$. The results are shown in Figs. 2(e) and (f).

The lasso and least angle regression closely match our predictions, yielding exactly the same ranks for the first noise variables. Forward stepwise regression exhibits larger departures from the theoretical predictions, mainly because of the slow convergence to the asymptotics, and also shows a decreasing rank once the sparsity exceeds a cut-off.

The first false variable occurs earlier as n decreases while p increases. In particular, the behaviours of the methods under Bernoulli random designs closely resemble those under Gaussian random designs. In Fig. 2(c), the rank of the first false variable increases as the signal magnitude M is amplified. While this increasing rank is expected, Fig. 2(d) shows that the rank drops after M exceeds a certain level. Given $M \geqslant 3{\cdot}4\sqrt{(2 \log p)} = 12{\cdot}64$, the lasso selects

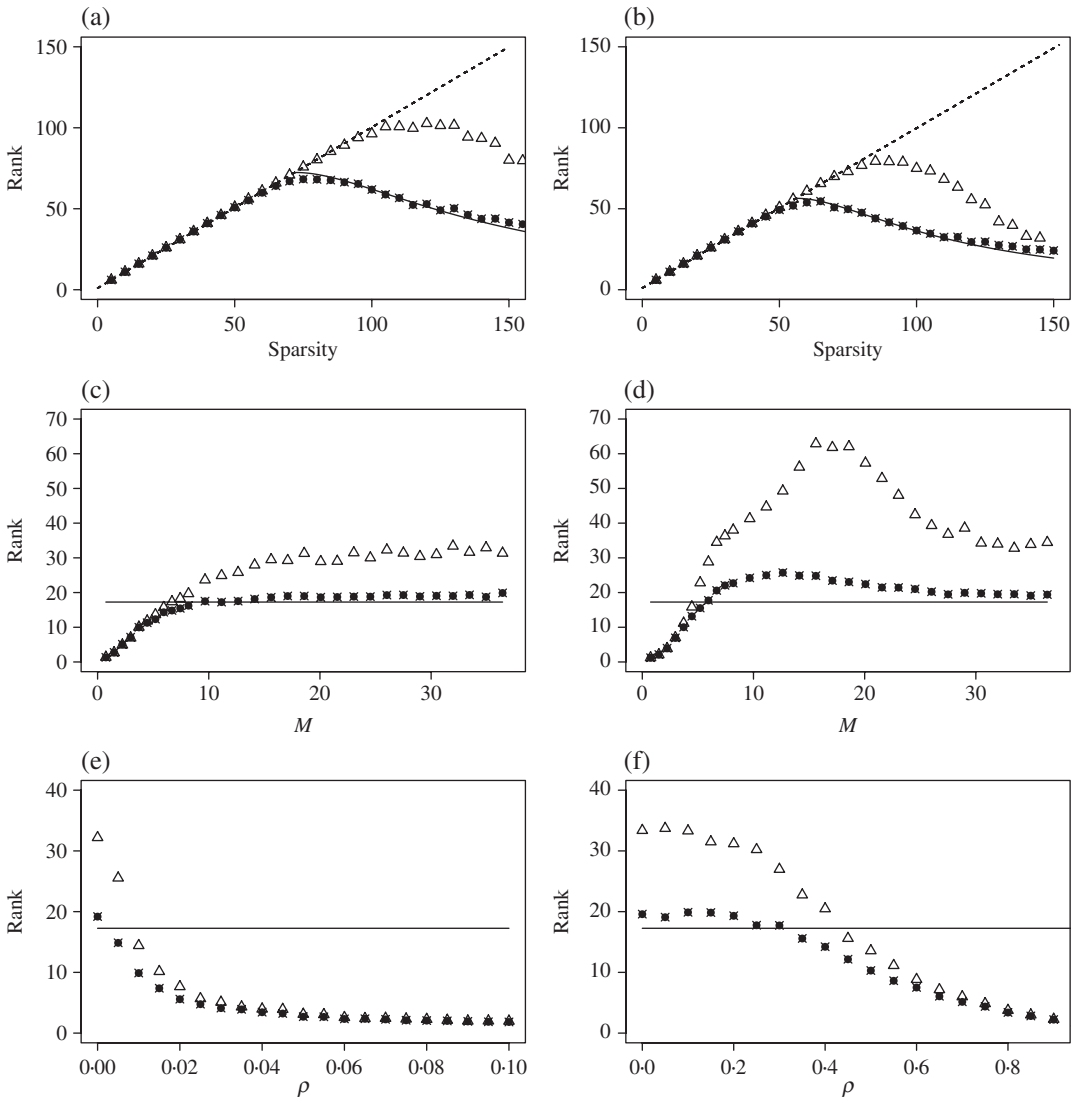

Fig. 2. Rank of the first spurious variable in the three studies described in § 3·1. Averaged over 500 replicates, ranks of forward stepwise regression, lasso and least angle regression are marked with triangles, dots and crosses, respectively; dots and crosses overlap exactly so they look like solid squares. The solid lines represent the predictions given by Theorem 2. For plots (c)–(f) the prediction is constant since $k = 80$ is fixed.

the first false variable earlier and earlier even though the sparsity is fixed and each signal gets strengthened, and the phenomenon is even more apparent for forward stepwise regression. Intuitively, this is because the effective sparsity in the case of a moderately large M is smaller than the nominal sparsity of 80. To see this, observe that the ratio of the signals of the first 40 components and the next 40 components is $M/(M^2/37{\cdot}17) = 37{\cdot}17/M$, which is noticeably larger than 1. Put another way, the first 40 components act as the main signals, so, loosely speaking, the effective sparsity is smaller than 80. In the presence of significant correlations between columns of X, Figs. 2(e) and (f) show that the problem of early false variables is further exacerbated.

Table 1. *Rank of the first noise predictor averaged over* 500 *runs, with standard errors in parentheses. The predictions for sparsity no larger than* $n/(2\log p) = 51{\cdot}6$ *are given simply as* $k+1$ *and otherwise are given by Theorems* 1 *and* 2*; the last row shows the predictions*

	Sparsity k						
	10	25	40	55	70	85	100
Lasso	10·4 (2·1)	15·4 (9·6)	10·3 (9·2)	6·8 (6·0)	5·5 (4·6)	4·4 (3·8)	3·7 (3·5)
Least angle regression	10·4 (2·1)	15·4 (9·6)	10·3 (9·2)	6·8 (6·0)	5·5 (4·6)	4·4 (3·8)	3·7 (3·5)
Forward stepwise	10·6 (1·9)	18·8 (10·0)	18·5 (15·9)	13·3 (16·0)	10·0 (11·7)	7·5 (8·2)	7·0 (7·6)
Gaussian designs	11·0	26·0	41·0	51·3	45·7	38·3	31·8

3·2. *HIV data*

We consider the HIV-1 data introduced in Rhee et al. (2006) to study the genetic basis of HIV-1 resistance to several drugs. Also used in a number of other works (Barber & Candès, 2015; G'Sell et al., 2016; Janson & Su, 2016), this dataset contains genotype information $X \in \mathbb{R}^{634\times 463}$ from 634 HIV-1-infected individuals across 463 locations after removing duplicate and missing values. The columns of X are standardized to have zero mean and unit Euclidean norm. The response y is synthetically generated by assigning an effect of $100\sqrt{(2\log p)}$ to each of k uniformly randomly chosen columns of X and setting the noise level σ to 1.

Table 1 reports the results averaged over 500 replicates. The three methods start to have a decreasing rank around $k = 25$, which is much smaller than $n/(2\log p) = 51{\cdot}6$. In addition, for each level of sparsity, the first spurious variable is included much earlier than the predictions say. This gap is not surprising given that the predictions are tailored to independent Gaussian designs, while the design X from the HIV-1 data has strongly correlated columns. To be more precise, about 4600 column pairs of X have correlations greater than 10%.

4. Visualizing early noise predictors

We introduce the double-ranking diagram to bring together the strengths of sequential methods and low-bias estimators such as the least-squares estimator $\hat\beta^{\mathrm{LS}}$. Figure 3 presents two examples: one is in the same setting as Fig. 1 except that X has size 200×180 and a fixed sparsity of $k = 50$, and the other is in the same setting as Table 1 with a fixed sparsity of $k = 60$. For each variable, the horizontal axis represents its rank by a sequential method, and the vertical axis represents its rank by a low-bias estimator. For example, the horizontal rank of the jth variable is given according to the magnitude of $|\hat\beta_j^{\mathrm{LS}}|/\sqrt{\{(X^{\mathrm{T}}X)^{-1}\}_{jj}}$: the larger this statistic is, the smaller the rank is. Equivalently, the variables can be ranked using the t-values.

The double-ranking diagram can help with the identification of early false variables for sequential regression methods. An important variable would presumably possess both a small horizontal rank and a small vertical rank, and so would appear in the bottom left corner of the diagram with a large probability. In light of this intuition, we screen out variables that are selected early by a sequential method but have unusually large vertical ranks, which in the case of least squares amount to small t-values or insignificant p-values. As seen from Fig. 3, the first five false variables in each instance have much larger vertical than horizontal ranks. These false variables are far from the signal variables in the diagram. To use this diagram, one can set some threshold for the vertical rank and select only variables that are below the threshold and in addition have significant horizontal ranks. On the other hand, in the low signal-to-noise ratio regime the diagram may not give a clear-cut separation between false and true predictors, and its use requires some caution.

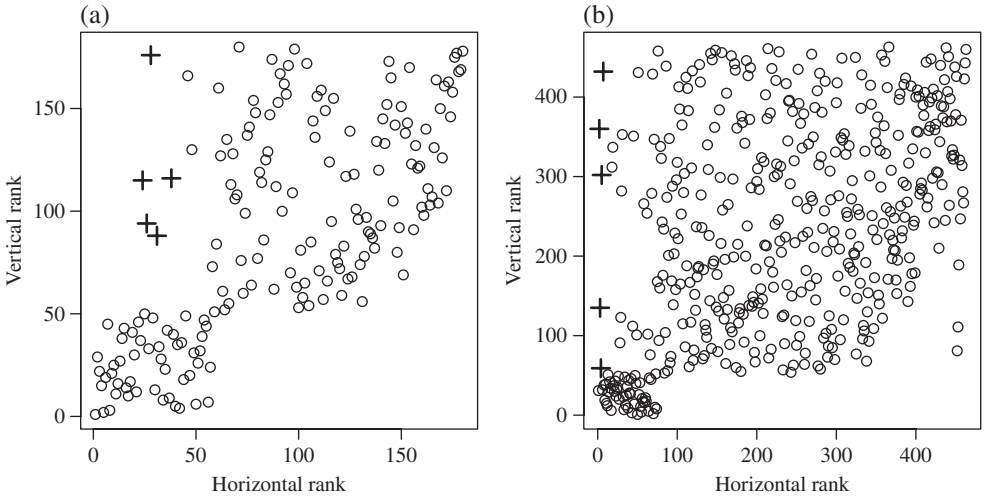

Fig. 3. Double-ranking diagrams: (a) in the same setting as Fig. 1; (b) in the same setting as the HIV data example. Vertical rankings are given by the least-squares estimators and horizontal rankings are given by least angle regression. The first five noise variables along the solution path of least angle regression are marked with crosses.

The following proposition states that the diagram can perfectly separate the first spurious variable from all the true variables.

PROPOSITION 1. *Under Assumption 1 and provided that* $n > \delta p$ *and* $M/\delta > 3\{2\delta \log p/(\delta - 1)\}^{1/2}$ *for some constant* $\delta > 1$*, the first noise variable in the double-ranking diagram has a greater vertical rank than all the true variables.*

This tool blends new and old ideas. Our discussion in § 2·3 demonstrates that, while sequential methods work well in very sparse settings, as the signals get denser, pseudo-noise can accumulate quickly and thus may dwarf some true signals, no matter how strong the corresponding coefficients are. The method of least squares favours dense signals, since the estimator variances stay the same as the sparsity of the signals increases. Variables with sufficiently strong effects can stand out using the least-squares estimator in the presence of highly correlated columns in the design matrix. This property of the least-squares estimator and its variants plays a pivotal role in a number of variable screening procedures (Wasserman & Roeder, 2009; Wang & Leng, 2016).

5. DISCUSSION

In the nonextreme sparsity regime, the common intuition that sequential regression procedures find a significant portion of all important variables before the first false variable deserves some scepticism. More caution is required when using these sequential methods, unless the true regression coefficients are known to be very sparse.

It is of interest to improve the predictions for forward stepwise regression and extend them to methods such as backward stepwise and forward-backward stepwise regression. The simulation studies imply that the lower bound $c_4 n$ on the sparsity k in Assumption 1 could be relaxed to $n/(2 \log p)$. In the high-dimensional setting where $p > n$, which low-bias estimator should we choose for the double-ranking diagram to yield the vertical ranking? Candidates include the lasso with a small penalty, ridge regression with a small penalty, and generalized least-squares estimators (Wang & Leng, 2016). It is also worth incorporating the strategies proposed by

Fan et al. (2015) and Fan & Zhou (2016) to investigate spurious discoveries. As seen from Table 1, the rank of the first noise variable has relatively large variation, and it is of practical relevance to characterize this variation.

Acknowledgement

This work was supported in part by the U.S. National Science Foundation. The author thanks Jianqing Fan, the editor, associate editor, and two referees for very helpful comments.

Supplementary material

Supplementary material available at *Biometrika* online contains the proofs of Theorems 1–3 and Proposition 1.

References

Barber, R. F. & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *Ann. Statist.* **43**, 2055–85.
Bayati, M. & Montanari, A. (2012). The LASSO risk for Gaussian matrices. *IEEE Trans. Info. Theory* **58**, 1997–2017.
Bickel, P. J., Ritov, Y. & Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37**, 1705–32.
Cai, T. T. & Wang, L. (2011). Orthogonal matching pursuit for sparse signal recovery with noise. *IEEE Trans. Info. Theory* **57**, 4680–8.
Candès, E. J. & Tao, T. (2005). Decoding by linear programming. *IEEE Trans. Info. Theory* **51**, 4203–15.
Efron, B., Hastie, T. J., Johnstone, I. M. & Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407–99.
Fan, J., Shao, Q.-M. & Zhou, W.-X. (2015). Are discoveries spurious? Distributions of maximum spurious correlations and their applications. *arXiv:* 1502.04237.
Fan, J. & Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Ann. Statist.* **38**, 3567–604.
Fan, J. & Zhou, W.-X. (2016). Guarding against spurious discoveries in high dimensions. *J. Mach. Learn. Res.* **17**, 1–34.
G'Sell, M. G., Wager, S., Chouldechova, A. & Tibshirani, R. (2016). Sequential selection procedures and false discovery rate control. *J. R. Statist. Soc.* B **78**, 423–44.
Janson, L. & Su, W. (2016). Familywise error rate control via knockoffs. *Electron. J. Statist.* **10**, 960–75.
Lockhart, R., Taylor, J., Tibshirani, R. J. & Tibshirani, R. (2014). A significance test for the lasso. *Ann. Statist.* **42**, 413–68.
Rhee, S.-Y., Taylor, J., Wadhera, G., Ben-Hur, A., Brutlag, D. L. & Shafer, R. W. (2006). Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proc. Nat. Acad. Sci.* **103**, 17355–60.
Su, W., Bogdan, M. & Candès, E. J. (2017). False discoveries occur early on the lasso path. *Ann. Statist.* **45**, 2133–50.
Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc.* B , 267–88.
Tibshirani, R. J. & Taylor, J. (2011). The solution path of the generalized lasso. *Ann. Statist.* **39**, 1335–71.
Tibshirani, R. J., Taylor, J., Lockhart, R. & Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *J. Am. Statist. Assoc.* **111**, 600–20.
Tropp, J. A. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Info. Theory* **50**, 2231–42.
Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Info. Theory* **55**, 2183–202.
Wang, X. & Leng, C. (2016). High dimensional ordinary least squares projection for screening variables. *J. R. Statist. Soc.* B **78**, 589–611.
Wasserman, L. & Roeder, K. (2009). High dimensional variable selection. *Ann. Statist.* **37**, 2178–201.
Zhang, T. (2009). On the consistency of feature selection using greedy least squares regression. *J. Mach. Learn. Res.* **10**, 555–68.
Zhao, P. & Yu, B. (2006). On model selection consistency of lasso. *J. Mach. Learn. Res.* **7**, 2541–63.

[*Received on* 7 *August* 2017. *Editorial decision on* 23 *March* 2018]