1-30-2012

A significance test for the lasso

Richard Lockhart
Simon Fraser University

Jonathan Taylor
Stanford University

Ryan J. Tibshirani
Carnegie Mellon University, ryantibs@cmu.edu

Robert Tibshirani
Stanford University

A significance test for the lasso

Richard Lockhart[1] Jonathan Taylor[2] Ryan J. Tibshirani[3]
Robert Tibshirani[2]

[1]Simon Fraser University, [2]Stanford University, [3]Carnegie-Mellon University

Abstract

In the sparse linear regression setting, we consider testing the significance of the predictor variable that enters the current lasso model, in the sequence of models visited along the lasso solution path. We propose a simple test statistic based on lasso fitted values, called the *covariance test statistic*, and show that when the true model is linear, this statistic has an Exp(1) asymptotic distribution under the null hypothesis (the null being that all truly active variables are contained in the current lasso model). Our proof of this result assumes some (reasonable) regularity conditions on the predictor matrix X, and covers the important high-dimensional case $p > n$.

Of course, for testing the significance of an additional variable between two nested linear models, one may use the usual chi-squared test, comparing the drop in residual sum of squares (RSS) to a χ^2_1 distribution. But when this additional variable is not fixed, but has been chosen adaptively or greedily, this test is no longer appropriate: adaptivity makes the drop in RSS stochastically much larger than χ^2_1 under the null hypothesis. Our analysis explicitly accounts for adaptivity, as it must, since the lasso builds an adaptive sequence of linear models as the tuning parameter λ decreases. In this analysis, shrinkage plays a key role: though additional variables are chosen adaptively, the coefficients of lasso active variables are shrunken due to the ℓ_1 penalty. Therefore the test statistic (which is based on lasso fitted values) is in a sense balanced by these two opposing properties—adaptivity and shrinkage—and its null distribution is tractable and asymptotically Exp(1).

Keywords: *lasso, least angle regression, significance test*

1 Introduction

Given an outcome vector $y \in \mathbb{R}^n$ and matrix $X \in \mathbb{R}^{n \times p}$ of predictor variables, we consider the usual linear regression setup:

$$y = X\beta^* + \sigma\epsilon, \tag{1}$$

where $\beta^* \in \mathbb{R}^p$ are unknown coefficients to be estimated, $\sigma^2 > 0$ is the marginal noise variance, and the components of the noise vector $\epsilon \in \mathbb{R}^n$ are i.i.d. $N(0,1)$. [1] We focus on the lasso estimator (Tibshirani 1996, Chen et al. 1998), defined as

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \; \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \tag{2}$$

where $\lambda \geq 0$ is a tuning parameter, controlling the degree of sparsity in the estimate $\hat{\beta}$. Here we assume that the columns of X are in general position in order to ensure uniqueness of the lasso solution [this is quite a weak condition, to be discussed again shortly; see also Tibshirani (2012)].

[1]If an intercept term is desired, then we can still assume a model of the form (1) after centering y and the columns of X. see Section 2.2 for discussion of this point.

There has been a considerable amount of recent work dedicated to the lasso problem, both in terms of computation and theory. A comprehensive summary of the literature in either category would be too long for our purposes here, so we instead give a short summary: for computational work, some relevant contributions are Friedman et al. (2007), Beck & Teboulle (2009), Friedman et al. (2010), Becker, Bobin & Candes (2011), Boyd et al. (2011), Becker, Candes & Grant (2011); and for theoretical work see, e.g., Greenshtein & Ritov (2004), Fuchs (2005), Donoho (2006), Candes & Tao (2006), Zhao & Yu (2006), Wainwright (2009), Candes & Plan (2009). Generally speaking, theory for the lasso is focused on bounding the estimation error $\|X\hat{\beta} - X\beta^*\|_2^2$ or $\|\hat{\beta} - \beta^*\|_2^2$, or ensuring exact recovery of the underlying model, $\mathrm{supp}(\hat{\beta}) = \mathrm{supp}(\beta^*)$ [with $\mathrm{supp}(\cdot)$ denoting the support function]; favorable results in both respects can be shown under the right assumptions on the generative model (1) and the predictor matrix X. Strong theoretical backing, as well as fast algorithms, have made the lasso a highly popular tool.

Yet, there are still major gaps in our understanding of the lasso as an estimation procedure. In many real applications of the lasso, a practitioner will undoubtedly seek some sort of inferential guarantees for his or her computed lasso model—but, generically, the usual constructs like p-values, confidence intervals, etc., do not exist for lasso estimates. There is a small but growing literature dedicated to inference for the lasso, and important progress has certainly been made, mostly through the use of resampling methods; we review this work in Section 2.5. The current paper focuses on a significance test for lasso models that does not employ resampling, but instead proposes a test statistic that has a simple and exact asymptotic null distribution.

Section 2 defines the problem that we are trying to solve, give the details of our proposal—the covariance test statistic. Section 3 considers an orthogonal predictor matrix X, in which case the statistic greatly simplifies. Here we derive its Exp(1) asymptotic distribution using relatively simple arguments from extreme value theory. Section 4 treats a general (nonorthogonal) X, and under some regularity conditions, derives an Exp(1) limiting distribution for the covariance test statistic, but through a different method of proof that relies on discrete-time Gaussian processes. Section 5 empirically verifies convergence of the null distribution to Exp(1) over a variety of problem setups. Up until this point we have assumed that the error variance σ^2 is known; in Section 6 we discuss the case of unknown σ^2. Section 7 gives some real data examples. Section 8 covers extensions to the elastic net, generalized linear models, and the Cox model for survival data. We conclude with a discussion in Section 9.

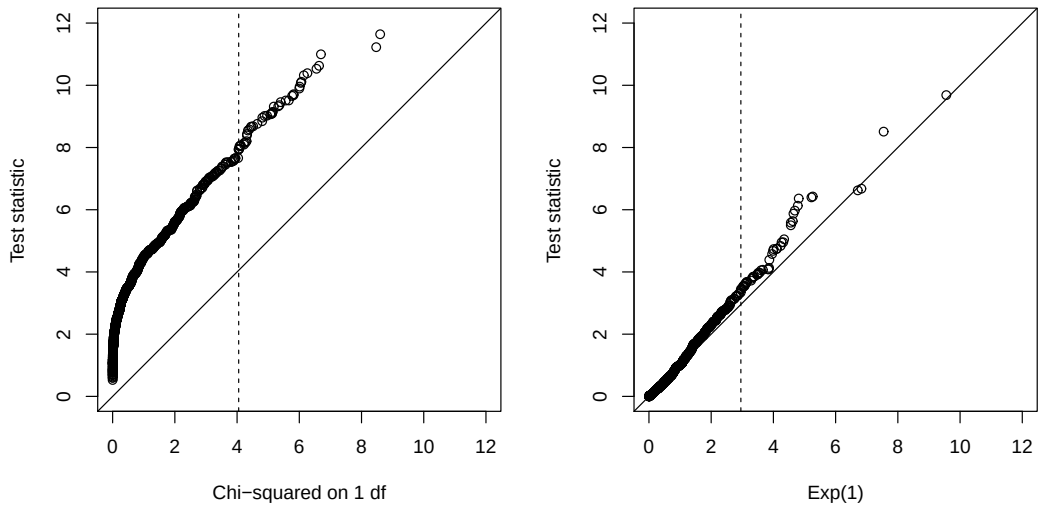
2 Significance testing in linear modeling

Classic theory for significance testing in linear regression operates on two fixed nested models. For example, if M and $M \cup \{j\}$ are fixed subsets of $\{1, \ldots p\}$, then to test the significance of the jth predictor in the model (with variables in) $M \cup \{j\}$, one naturally uses the chi-squared test, which computes the drop in residual sum of squares (RSS) from regression on $M \cup \{j\}$ and M,

$$R_j = (\mathrm{RSS}_M - \mathrm{RSS}_{M\cup\{j\}})/\sigma^2, \tag{3}$$

and compares this to a χ_1^2 distribution. (Here σ^2 is assumed to be known; when σ^2 is unknown, we use the sample variance in its place, which results in the F-test, equivalent to the t-test, for testing the significance of variable j.)

Often, however, one would like to run the same test for M and $M \cup \{j\}$ that are not fixed, but the outputs of an adaptive or greedy procedure. Unfortunately, adaptivity invalidates the use of a χ_1^2 null distribution for the statistic (3). As a simple example, consider forward stepwise regression: starting with an empty model $M = \emptyset$, we enter predictors one at a time, at each step choosing the predictor j that gives the largest drop in residual sum of squares. In other words, forward stepwise regression chooses j at each step in order to maximize R_j in (3), over all $j \notin M$. Since R_j follows a χ_1^2 distribution under the null hypothesis for each fixed j, the maximum possible R_j will clearly



(a) Forward stepwise (b) Lasso

Figure 1: *A simple example with* $n = 100$ *observations and* $p = 10$ *orthogonal predictors. All true regression coefficients are zero,* $\beta^* = 0$*. On the left is a quantile-quantile plot, constructed over 1000 simulations, of the standard chi-squared statistic* R_1 *in* (3)*, measuring the drop in residual sum of squares for the first predictor to enter in forward stepwise regression, versus the* χ^2_1 *distribution. The dashed vertical line marks the 95% quantile of the* χ^2_1 *distribution. The right panel shows a quantile-quantile plot of the covariance test statistic* T_1 *in* (5) *for the first predictor to enter in the lasso path, versus its asymptotic distribution* Exp(1)*. The covariance test explicitly accounts for the adaptive nature of lasso modeling, whereas the usual chi-squared test is not appropriate for adaptively selected models, e.g., those produced by forward stepwise regression.*

be stochastically larger than χ^2_1 under the null. Therefore, using a chi-squared test to evaluate the significance of a predictor entered by forward stepwise regression would be far too liberal (having type I error much larger than the nominal level). Figure 1(a) demonstrates this point by displaying the quantiles of R_1 in forward stepwise regression (the chi-squared statistic for the first predictor to enter) versus those of a χ^2_1 variate, in the fully null case (when $\beta^* = 0$). A test at level 0.05 for example, using the χ^2_1 cutoff of 3.84, would have actual type I error about 39%.

The failure of standard testing methodology when applied to forward stepwise regression is not an anomaly—in general, there is no simple way to carry out the significance tests designed for fixed linear models in an adaptive setting. Our aim is hence to provide a (new) significance test for the predictor variables chosen adaptively by the lasso, which we describe next.

2.1 The covariance test statistic

The test statistic that we propose here is constructed from the lasso solution path, i.e., the solution $\hat{\beta}(\lambda)$ in (2) a function of the tuning parameter $\lambda \in [0, \infty)$. The lasso path can be computed by the well-known LARS algorithm of Efron et al. (2004) [see also Osborne et al. (2000*a*), Osborne et al. (2000*b*)], which traces out the solution as λ decreases from ∞ to 0. Note that when $\text{rank}(X) < p$, there are possibly many lasso solutions at each λ and therefore many possible solution paths; we assume that the columns of X are in general position[2], implying that there is a unique lasso solution

[2]To be precise, we say that points $X_1, \ldots X_p \in \mathbb{R}^n$ are in *general position* provided that no k-dimensional affine subspace $L \subseteq \mathbb{R}^n$, $k < \min\{n, p\}$, contains more than $k + 1$ elements of $\{\pm X_1, \ldots \pm X_p\}$, excluding antipodal pairs. Equivalently: the affine span of any $k + 1$ points $s_1 X_{i_1}, \ldots s_{k+1} X_{i_{k+1}}$, for any signs $s_1, \ldots s_{k+1} \in \{-1, 1\}$, does not contain any element of the set $\{\pm X_i : i \neq i_1, \ldots i_{k+1}\}$.

at each $\lambda > 0$ and hence a unique path. The assumption that X has columns in general position is a very weak one [much weaker, e.g., than assuming that $\text{rank}(X) = p$]. For example, if the entries of X are drawn from a continuous probability distribution on $\mathbb{R}^{np}$, then the columns of X are almost surely in general position, and this is true regardless of the sizes of n and p. See Tibshirani (2012).

Before defining our statistic, we briefly review some properties of the lasso path.

- The path $\hat{\beta}(\lambda)$ is a continuous and piecewise linear function of λ, with knots (changes in slope) at values $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq 0$ (these knots depend on y, X).
- At $\lambda = \infty$, the solution $\hat{\beta}(\infty)$ has no active variables (i.e., all variables have zero coefficients); for decreasing λ, each knot λ_k marks the entry or removal of some variable from the current active set (i.e., its coefficient becomes nonzero or zero, respectively). Therefore the active set, and also the signs of active coefficients, remain constant in between knots.
- At any point λ in the path, the corresponding active set $A = \text{supp}(\hat{\beta}(\lambda))$ of the lasso solution indexes a linearly independent set of predictor variables, i.e., $\text{rank}(X_A) = |A|$, where we use X_A to denote the columns of X in A.
- For a matrix X satisfying the positive cone condition (a restrictive condition that covers, e.g., orthogonal matrices), there are no variables removed from the active set as λ decreases, and therefore the number of knots is $\min\{n, p\}$.

We can now precisely define the problem that we are trying to solve: at a given step in the lasso path (i.e., at a given knot), we consider testing the significance of the variable that enters the active set. To this end, we propose a test statistic defined at the kth step of the path.

First we define some needed quantities. Let A be the active set just before λ_k, and suppose that predictor j enters at λ_k. Denote by $\hat{\beta}(\lambda_{k+1})$ the solution at the next knot in the path λ_{k+1}, using predictors $\{A \cup j\}$. Finally, let $\tilde{\beta}_A(\lambda_{k+1})$ be the solution of the lasso problem using only the active predictors X_A, at $\lambda = \lambda_{k+1}$. To be perfectly explicit,

$$\tilde{\beta}_A(\lambda_{k+1}) = \underset{\beta_A \in \mathbb{R}^{|A|}}{\text{argmin}} \; \frac{1}{2}\|y - X_A\beta_A\|_2^2 + \lambda_{k+1}\|\beta_A\|_1. \tag{4}$$

We propose the *covariance test statistic* defined by

$$T_k = \Big(\langle y, X\hat{\beta}(\lambda_{k+1})\rangle - \langle y, X_A\tilde{\beta}_A(\lambda_{k+1})\rangle\Big)/\sigma^2. \tag{5}$$

Intuitively, the covariance statistic in (5) is a function of the difference between $X\hat{\beta}$ and $X_A\tilde{\beta}_A$, the fitted values given by incorporating the jth predictor into the current active set, and leaving it out, respectively. These fitted values are parametrized by λ, and so one may ask: at which value of λ should this difference be evaluated? Well, note first that $\tilde{\beta}_A(\lambda_k) = \hat{\beta}_A(\lambda_k)$, i.e., the solution of the reduced problem at λ_k is simply that of the full problem, restricted to the active set A (as verified by the KKT conditions). Clearly then, this means that we cannot evaluate the difference at $\lambda = \lambda_k$, as the jth variable has a zero coefficient upon entry at λ_k, and hence

$$X\hat{\beta}(\lambda_k) = X_A\hat{\beta}_A(\lambda_k) = X_A\tilde{\beta}_A(\lambda_k).$$

Indeed, the natural choice for the tuning parameter in (5) is $\lambda = \lambda_{k+1}$: this allows the jth coefficient to have its fullest effect on the fit $X\hat{\beta}$ before the entry of the next variable at λ_{k+1} (or possibly, the deletion of a variable from A at λ_{k+1}).

Secondly, one may also ask the particular choice of function of $X\hat{\beta}(\lambda_{k+1}) - X_A\tilde{\beta}_A(\lambda_{k+1})$. The covariance statistic in (5) uses an inner product of this difference with y, which can be roughly

thought of as an (uncentered) covariance, hence explaining its name.[3] At a high level, the larger the covariance of y with $X\hat{\beta}$ compared to that with $X_A\tilde{\beta}_A$, the more important the role of variable j in the proposed model $A \cup \{j\}$. There certainly may be other functions that would seem appropriate here, but the covariance form in (5) has a distinctive advantage: this statistic admits a simple and exact asymptotic null distribution, assuming normality of the errors in (1). The null hypothesis here is that the current lasso model contains all truly active variables, $A \supseteq \text{supp}(\beta^*)$, and in Sections 3 and 4, we show that under the null,

$$T_k \xrightarrow{d} \text{Exp}(1),$$

i.e., T_k is asymptotically distributed as a standard exponential random variable, given reasonable assumptions on X and the magnitudes of the nonzero true coefficients. In the above limit, we are considering both $n, p \to \infty$ (and in Section 4 we allow for the possibility $p > n$, the high-dimensional case).

We will also see the result $T_k \xrightarrow{d} \text{Exp}(1)$ applies to the first null predictor entered, after all signal variables (if there are any), have been entered. For the ℓth null variable entered, the distribution is $\text{Exp}(1/\ell)$. Since ℓ is unknown, we propose to always use the Exp(1) distribution for testing. Now $\text{Exp}(1/\ell) < \text{Exp}(1)$ for $\ell > 1$, so this yields a conservative test.

See Figure 1(b) for a quantile-quantile plot of T_1 versus an Exp(1) variate for the same fully null example ($\beta^* = 0$) used in Figure 1(a); this shows that the weak convergence to Exp(1) can be quite fast, as the quantiles are decently matched even for $p = 10$. Before proving this limiting distribution in Sections 3 (for an orthogonal X) and 4 (for a general X), we give an example of its application to real data, and discuss issues related to practical usage. We also derive useful alternative expressions for the statistic, discuss the connection to degrees of freedom, and review related work.

2.2 Prostate cancer data example and practical issues

We consider a training set of 67 observations and 8 predictors, the goal being to predict log of the PSA level of men who had surgery for prostate cancer. For more details see Hastie et al. (2008) and the references therein. Table 1 shows the results of forward stepwise regression and the lasso. Both methods entered the same predictors in the same order. The forward stepwise p-values are smaller than the lasso p-values, we would enter four predictors at level 0.05. The latter would enter only one or maybe two predictors. However we know that the forward stepwise p-values are accurate, as they are based on a null distribution that does not account for the adaptive choice of predictors.

Remark A. In the above example we implied that one might stop entering variables when the p-value rose above some threshold. More generally, our proposed test statistic and associated p-values could be used as the basis for multiple testing and false discovery rate control methods for this problem; we leave that to future work.

Remark B. Note that for a general X, a predictor may enter the active set more than one time, since it may leave the active set at some point, In this case we treat each entry as separate problems. Therefore, our test is specific to a step in the path, and not to a predictor at large;

Remark C. When we have an intercept in the model (1), we center y, and use the covariance test as is. Empirically the resulting distribution of T is close to Exp(1). Theoretically this creates dependence between components of ϵ. But the dependence is weak, and we don't study it.

Remark D. The covariance test is applied in a sequential manner, estimating p-values for each predictor as it enters. A more difficult problem is to test any of the nonzero predictors in a linear model fit by the lasso, at some arbitrary value of the tuning parameter λ. We discuss this problem briefly in Section 9.

[3] From its definition in (5), we get $T_k = \langle y - \mu, X\hat{\beta}(\lambda_{k+1})\rangle - \langle y - \mu, X_A\tilde{\beta}_A(\lambda_{k+1})\rangle + \langle \mu, X\hat{\beta}(\lambda_{k+1}) - X_A\tilde{\beta}_A(\lambda_{k+1})\rangle$ by expanding $y = y - \mu + \mu$, with $\mu = X\beta^*$ denoting the true mean. The first two terms are now really empirical covariances, and the last term is typically small. In fact, when X is orthogonal, it is not hard to see that this last term is exactly zero under the null hypothesis.

Table 1: *Forward stepwise and lasso applied to the prostate cancer data example. The error variance is estimated by $\hat{\sigma}^2$, the MSE of the full model. Forward stepwise regression p-values are based on comparing the drop in residual sum of squares (divided by $\hat{\sigma}^2$) to an $F(1, n-p)$ distribution (using χ^2_1 instead produced slight smaller p-values). The lasso p-values use a simple modification of the covariance test* (5) *for unknown variance, given in Section 6.*

Step	Predictor entered	Forward stepwise	Lasso
1	lcavol	0.000	0.000
2	lweight	0.000	0.052
3	svi	0.041	0.174
4	lbph	0.045	0.929
5	pgg45	0.226	0.353
6	age	0.191	0.650
7	lcp	0.065	0.051
8	gleason	0.883	0.978

2.3 Alternate expressions for the covariance statistic

Here we derive two alternate forms for the covariance statistic in (5). The first lends some insight into the role of shrinkage, and the second is helpful for the convergence results that we establish in Sections 3 and 4. We rely on some basic properties of lasso solutions; see, e.g., Tibshirani & Taylor (2012), Tibshirani (2012). To remind the reader, we are assuming that X has columns in general position.

For any fixed λ, if the lasso solution has active set $A = \mathrm{supp}(\hat{\beta}(\lambda))$ and signs $s_A = \mathrm{sign}(\hat{\beta}_A(\lambda))$, then it can be written explicitly (over active variables) as

$$\hat{\beta}_A(\lambda) = (X_A^T X_A)^{-1} X_A^T y - \lambda (X_A^T X_A)^{-1} s_A.$$

In the above expression, the first term $(X_A^T X_A)^{-1} X_A^T y$ simply gives the regression coefficients of y on the active variables X_A, and the second term $-\lambda(X_A^T X_A)^{-1} s_A$ can be thought of as a shrinkage term, shrinking the values of these coefficients towards zero. Further, the lasso fitted value at λ is

$$X\hat{\beta}(\lambda) = P_A y - \lambda (X_A^T)^+ s_A, \tag{6}$$

where $P_A = X_A(X_A^T X_A)^{-1} X_A^T$ denotes the projection onto the column space of X_A, and $(X_A^T)^+ = X_A(X_A^T X_A)^{-1}$ is the (Moore-Penrose) pseudo-inverse of X_A^T.

Using the representation (6) for the fitted values, we can derive our first alternate expression for the covariance statistic in (5). If A and s_A are the active set and signs just before the knot λ_k, and j is the variable added to the active set at λ_k, then by (6),

$$X\hat{\beta}(\lambda_{k+1}) = P_{A\cup\{j\}} y - \lambda_{k+1}(X_{A\cup\{j\}}^T)^+) s_{A\cup\{j\}},$$

where $s_{A\cup\{j\}}$ joins s_A and the sign of the jth coefficient, i.e., $s_{A\cup\{j\}} = \mathrm{sign}(\hat{\beta}_{A\cup\{j\}}(\lambda_{k+1}))$. Let us assume for the moment that the solution of reduced lasso problem (4) at λ_{k+1} has all variables active and $s_A = \mathrm{sign}(\tilde{\beta}_A(\lambda_{k+1}))$—remember, this holds for the reduced problem at λ_k, and we will return to this assumption shortly. Then, again by (6),

$$X_A\tilde{\beta}_A(\lambda_{k+1}) = P_A y - \lambda_{k+1}(X_A^T)^+ s_A,$$

and plugging the above two expressions into (5),

$$T_k = y^T(P_{A\cup\{j\}} - P_A)y/\sigma^2 - \lambda_{k+1}\Big((X_A^T)^+ s_A - (X_{A\cup\{j\}}^T)^+ s_{A\cup\{j\}}\Big)/\sigma^2. \tag{7}$$

Note that the first term above is $y^T(P_{A\cup\{j\}} - P_A)y/\sigma^2 = (\|y - P_A y\|_2^2 - \|y - P_{A\cup\{j\}}y\|_2^2)/\sigma^2$, which is exactly the chi-squared statistic for testing the significance of variable j, as in (3). Hence if A, j were fixed, then without the second term, T_k would have a χ_1^2 distribution under the null. But of course A, j are not fixed, and so much like we saw previously with forward stepwise regression, the first term in (7) will be generically larger than χ_1^2, because j is chosen adaptively based on its inner product with the current lasso residual vector. Interestingly, the second term in (7) adjusts for this adaptivity: with this term, which is composed of the shrinkage factors in the solutions of the two relevant lasso problems (on X and X_A), we prove in the coming sections that T_k has an asymptotic Exp(1) null distribution. Therefore, the presence of the second term restores the (asymptotic) mean of T_k to 1, which is what it would have been if A, j were fixed and the second term were missing. In short, adaptivity and shrinkage balance each other out.

This insight aside, the form (7) of the covariance statistic leads to a second representation that will be useful for the theoretical work in Sections 3 and 4. We call this the *knot form* of the covariance statistic, described in the next lemma.

Lemma 1. *Let A be the active set just before the kth step in the lasso path, i.e., $A = \mathrm{supp}(\hat{\beta}(\lambda_k))$, with λ_k being the kth knot. Also let s_A denote the signs of the active coefficients, $s_A = \mathrm{sign}(\hat{\beta}_A(\lambda_k))$, and j the predictor that enters the active set at λ_k. Then, assuming that*

$$s_A = \mathrm{sign}(\tilde{\beta}_A(\lambda_{k+1})), \tag{8}$$

or in other words, all coefficients are active in the reduced lasso problem (4) *at λ_{k+1} and have signs s_A, we have*

$$T_k = C(A, s_A, j) \cdot \lambda_k(\lambda_k - \lambda_{k+1})/\sigma^2, \tag{9}$$

where

$$C(A, s_A, j) = \|(X_{A\cup\{j\}}^T)^+ s_{A\cup\{j\}} - (X_A^T)^+ s_A\|_2^2.$$

The proof starts with expression (7), and arrives at (9) through simple algebraic manipulations. We defer it until Appendix A.1.

When does the condition (8) hold? This was a key assumption behind both of the forms (7) and (9) for the statistic. We first note that the solution $\tilde{\beta}_A$ of the reduced lasso problem has signs s_A at λ_k, so it will have the same signs s_A at λ_{k+1} provided that no variables are deleted from the active set in the solution path $\tilde{\beta}_A(\lambda)$ for $\lambda \in [\lambda_{k+1}, \lambda_k]$. Therefore, assumption (8) holds:

1. When X satisfies the positive cone condition (which includes X orthogonal), because no variables ever leave the active set in this case. In fact, for X orthogonal, it is straightforward to check that $C(A, s_A, j) = 1$, so $T_k = \lambda_k(\lambda_k - \lambda_{k+1})/\sigma^2$.

2. When $k = 1$ (we are testing the first variable to enter), as a variable cannot leave the active set right after it has entered. If $k = 1$ and X has unit norm columns, $\|X_i\|_2^2 = 1$ for $i = 1, \ldots p$, then we again have $C(A, s_A, j) = 1$ (note that $A = \emptyset$), so $T_1 = \lambda_1(\lambda_1 - \lambda_2)/\sigma^2$.

3. When $s_A = \mathrm{sign}((X_A)^+ y)$, i.e., s_A contains the signs of the least squares coefficients on X_A, because the same active set and signs cannot appear at two different knots in the lasso path (applied here to the reduced lasso problem on X_A).

The first and second scenarios are considered in Sections 3 and 4.1, respectively. The third scenario is actually somewhat general and occurs, e.g., when $s_A = \mathrm{sign}((X_A)^+ y) = \mathrm{sign}(\beta_A^*)$, both the lasso and least squares on X_A recover the signs of the true coefficients. Section 4.2 studies the general X and $k > 1$ case, wherein this third scenario is important.

2.4 Connection to degrees of freedom

There is a interesting connection between the covariance statistic in (5) and the degrees of freedom of a fitting procedure. In the regression setting (1), for an estimate $\hat{y}$ [which we think of as a fitting procedure $\hat{y} = \hat{y}(y)$], its degrees of freedom is typically defined (Efron 1986*a*) as

$$\mathrm{df}(\hat{y}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(\hat{y}_i, y_i). \tag{10}$$

In words, $\mathrm{df}(\hat{y})$ sums the covariances of each observation y_i with its fitted value $\hat{y}_i$. Hence the more adaptive a procedure, the higher this covariance, and the greater its degrees of freedom.

Using this definition, one can reason [and confirm by simulation, just as in Figure 1(a)] that with k predictors entered into the model, forward stepwise regression had used substantially more than k degrees of freedom. But something quite remarkable happens when we consider the lasso: for a model containing k nonzero coefficients, the degrees of freedom of the lasso fit is equal to k (either exactly or in expectation, depending on the assumptions) [Efron et al. (2004), Zou et al. (2007), Tibshirani & Taylor (2012)]. Why does this happen? Roughly speaking, it is the same adaptivity versus shrinkage phenomenon at play. [Recall our discussion in the last section following the expression (7) for the covariance statistic.] The lasso adaptively chooses the active predictors, which costs extra degrees of freedom; but it also shrinks the nonzero coefficients (relative to the usual least squares estimates), which decreases the degrees of freedom just the right amount, so that the total is simply k. The current work in this paper arose from our desire to find a statistic whose degrees of freedom was equal to one after each LARS step.

2.5 Related work

There is quite a bit of recent work that is related to the proposal of this paper. Wasserman & Roeder (2009) propose a procedure for variable selection and p-value estimation based on sample splitting, and this was extended by Meinhasuen et al. (2009). Meinshausen & Bühlmann (2010) propose Stability Selection, a generic method which controls the expected number of false positive selections. Zhang & Zhang (2011) and Bühlmann (2012) derive p-values for parameter components in lasso and ridge regression, based on stochastic upper bounds for the parameter estimates. Minnier et al. (2011) use perturbation resampling-based procedures to approximate the distribution of a general class of penalized parameter estimates. Berk et al. (2010) and Laber & Murphy (2011) propose methods for conservative statistical inference after model selection and classification, respectively. One big difference with the work here: we propose a simple statistic with an exact asymptotic null distribution and do not require any resampling or sample splitting. Unlike other approaches, our proposal exploits the special properties of the lasso, and does not work for other selection procedures.

3 An orthogonal predictor matrix X

We discuss the special case of an orthogonal predictor matrix X, i.e., one that satisfies $X^T X = I$. Even though the results here can be seen as special cases of those for a general X in Section 4, the arguments in the current orthogonal X case rely on relatively straightforward extreme value theory and are hence much simpler than their general X counterparts (which analyze the knots in the lasso path via Gaussian process theory). Furthermore, the Exp(1) limiting distribution for the covariance statistic translates here to some interesting and previously unknown (as far as we can tell) results on the order statistics of independent standard normals. For these reasons, we discuss the orthogonal case in detail.

As noted in the discussion following Lemma 1 (see the first point), we know that the covariance statistic for testing the entry of the variable at the kth step in the lasso path is

$$T_k = \lambda_k(\lambda_k - \lambda_{k+1})/\sigma^2.$$

Using orthogonality, we rewrite $\|y - X\beta\|_2^2 = \|X^T y - \beta\|_2^2 + C$ for a constant C (not depending on β) in the criterion in (2), and then we can see that the lasso solution at any given value of λ has the closed-form:

$$\hat{\beta}_j(\lambda) = S_\lambda(X_j^T y), \quad j = 1, \ldots p,$$

where $X_1, \ldots X_p$ are columns of X, and $S_\lambda : \mathbb{R} \to \mathbb{R}$ is the soft-thresholding function,

$$S_\lambda(x) = \begin{cases} x - \lambda & \text{if } x > \lambda \\ 0 & \text{if } -\lambda \leq x \leq \lambda \\ x + \lambda & \text{if } x < \lambda. \end{cases}$$

Letting $U_j = X_j^T y$, $j = 1, \ldots p$, the knots in the lasso path are simply the values of λ at which the coefficients become nonzero (i.e., cease to be thresholded),

$$\lambda_1 = |U_{(1)}|, \; \lambda_2 = |U_{(2)}|, \ldots \; \lambda_p = |U_{(p)}|,$$

where $|U_{(1)}| \geq |U_{(2)}| \geq \ldots \geq |U_{(p)}|$ are the order statistics of $|U_1|, \ldots |U_p|$ (somewhat of an abuse of notation). Therefore,

$$T_k = |U_{(k)}|(|U_{(k)}| - |U_{(k+1)}|)/\sigma^2.$$

Next, we study with the case $k = 1$, the test for the first predictor to enter the active set along the lasso path. We then examine the case $k > 1$, the test for subsequent predictors.

3.1 The first predictor to enter, $k = 1$

Consider the covariance test statistic for the first predictor to enter the active set, i.e., for $k = 1$,

$$T_1 = |U_{(1)}|(|U_{(1)}| - |U_{(2)}|)/\sigma^2.$$

We are interested in the distribution of T_1 under the null hypothesis; since we are testing the first step, this is

$$H_0 : \quad y \sim N(0, \sigma^2 I).$$

Under the null, $U_1, \ldots U_p$ are i.i.d., $U_j \sim N(0, \sigma^2)$, and so $|U_1|/\sigma, \ldots |U_p|/\sigma$ follow a χ_1 distribution (absolute value of a standard Gaussian). That T_1 has an asymptotic Exp(1) null distribution is now given by the next result.

Lemma 2. *Let $V_1 \geq V_2 \geq \ldots \geq V_p$ be the order statistics of an independent sample of χ_1 variates (i.e., they are the sorted absolute values of an independent sample of standard Gaussian variates). Then*

$$V_1(V_1 - V_2) \xrightarrow{d} \text{Exp}(1) \quad \textit{as } p \to \infty.$$

Proof. The χ_1 distribution has CDF

$$F(x) = (2\Phi(x) - 1)1(x > 0)$$

where Φ is the standard normal CDF. We first compute

$$\lim_{t \to \infty} \frac{F''(t)(1 - F(t))}{(F'(t))^2} = \lim_{t \to \infty} -\frac{t(1 - \Phi(t))}{\phi(t)} = -1,$$

the last equality using Mill's ratio. Then Theorem 2.2.1 in de Haan & Ferreira (2006) implies that, for constants $a_p = F^{-1}(1-1/p)$ and $b_p = pF'(a_p)$, the random variables $W_1 = b_p(V_1 - a_p)$ and $W_2 = b_p(V_2 - a_p)$ converge jointly in distribution,

$$(W_1, W_2) \xrightarrow{d} \big(-\log E_1, -\log(E_1 + E_2)\big),$$

where E_1, E_2 are independent standard exponentials. Now note that

$$V_1(V_1 - V_2) = (a_p + W_1/b_p)(W_1 - W_2)/b_p = \frac{a_p}{b_p}(W_1 - W_2) + \frac{W_1(W_1 - W_2)}{b_p}.$$

We claim that $a_p/b_p \to \infty$; this would give the desired result, as it would imply that first term above converges in distribution to $\log(E_2 + E_1) - \log(E_1)$, which is standard exponential, and the second term converges to zero, as $b_p \to \infty$. Writing a_p, b_p more explicitly, we see that $1 - 1/p = 2\Phi(a_p) - 1$, i.e., $1 - \Phi(a_p) = 1/(2p)$, and $b_p = 2p\phi(a_p)$. Using Mill's inequalities,

$$\frac{\phi(a_p)}{a_p}\left(1 - \frac{1}{a_p^2}\right) \le 1 - \Phi(a_p) \le \frac{\phi(a_p)}{a_p},$$

and multiplying by $2p$,

$$\frac{b_p}{a_p}\left(1 - \frac{1}{a_p^2}\right) \le 1 \le \frac{b_p}{a_p}.$$

Since $a_p \to \infty$, this means that $b_p/a_p \to 1$, completing the proof. □

We were unable to find this remarkably simple result elsewhere in the literature. An easy generalization is as follows.

Lemma 3. *If $V_1 \ge V_2 \ge \ldots \ge V_p$ are the order statistics of an independent sample of χ_1 variates, then for any fixed $k \ge 1$,*

$$\big(V_1(V_1 - V_2), V_2(V_2 - V_3), \ldots V_k(V_k - V_{k+1})\big) \xrightarrow{d} \big(\mathrm{Exp}(1), \mathrm{Exp}(1/2), \ldots \mathrm{Exp}(1/k)\big) \quad \textit{as } p \to \infty,$$

where the limiting distribution (on the right-hand side above) has independent components.

We leave the proof of Lemma 3 to Appendix A.2, since it follows from arguments very similar to those given for Lemma 2. Practically, Lemma 3 tells us that under the global null hypothesis $y \sim N(0, \sigma^2)$, comparing the covariance statistic T_k at the kth step of the lasso path to an Exp(1) distribution is increasingly conservative [at the first step, T_1 is asymptotically Exp(1), at the second step, T_2 is asymptotically Exp(1/3), at the third step, T_3 is asymptotically Exp(1/3), and so forth]. This progressive conservatism is favorable, if we place importance on parsimony in the fitted model: we are less and less likely to incur a false rejection of the null hypothesis as the size of the model grows. Moreover, we know that the test statistics $T_1, T_2, \ldots$ at successive steps are independent, and hence so are the corresponding p-values; from the point of view of multiple testing corrections, this is nearly an ideal scenario.

Of real interest is the distribution of T_k not under global null, but rather under the broader null hypothesis that all variables left out of the current model are truly inactive variables (i.e., they have zero coefficients in the true model). We study this in next section.

3.2 Subsequent predictors, $k > 1$

We suppose that exactly k_0 components of the true coefficient vector β^* are nonzero, and consider testing the entry of the predictor at the step $k = k_0 + 1$. We show that, under the null hypothesis that all truly active predictors are added to the model at steps $1, \ldots k_0$, the test statistic T_{k_0+1} is

asymptotically Exp(1); further, the test statistic T_{k_0+d} at a future step $k = k_0 + d$ is asymptotically Exp($1/d$).

The basic idea behind our argument is as follows: if we assume that the nonzero components of β^* are large enough in magnitude, then it is not hard to show (relying on orthogonality, here) that the truly active predictors are added to the model along the first k_0 steps of the lasso path, with probability tending to one. The test statistic at the (k_0+1)st step and beyond would therefore depend on the order statistics of $|U_i|$ for truly inactive variables i, subject to the constraint that the largest of these values is smaller than the smallest $|U_j|$ for truly active variables j. But with our strong signal assumption, i.e., that the nonzero entries of β^* are large in absolute value, this constraint has essentially no effect, and we are back to studying the order statistics from a χ_1 distribution, as in the last section. We make this idea precise below.

Theorem 1. *Assume that $X \in \mathbb{R}^{n\times p}$ is orthogonal, and that there are k_0 nonzero components in the true coefficient vector β^*. Let $A^* = \mathrm{supp}(\beta^*)$ be the true active set. Also assume that the smallest nonzero true coefficient is large compared to $\sqrt{2\log p}$,*

$$\min_{j\in A^*} |\beta_j^*| - \sqrt{2\log p} \to \infty \quad \textit{as } p \to \infty.$$

Let B denote the event that the first k_0 variables entering the model along the lasso path are those in A^*, i.e.,

$$B = \Big\{ \min_{j\in A^*} |U_j| > \max_{j\notin A^*} |U_j| \Big\},$$

where $U_j = X^T y$, $j = 1,\ldots p$. Then $\mathbb{P}(B) \to 1$ as $p \to \infty$, and for each fixed $d \geq 0$, we have

$$(T_{k_0+1}, T_{k_0+2}, \ldots T_{k_0+d}) \xrightarrow{d} \big(\mathrm{Exp}(1), \mathrm{Exp}(1/2), \ldots \mathrm{Exp}(1/d)\big) \quad \textit{as } p \to \infty.$$

The same convergence in distribution holds conditionally on B.

Proof. We first bound $\mathbb{P}(B)$. Let $\theta_p = \min_{i\in A^*} |\beta_i^*|$, and choose a_p such that

$$a_p - \sqrt{2\log p} \to \infty \quad \text{and} \quad \theta_p - a_p \to \infty.$$

Note that $U_j \sim N(\beta_j^*, \sigma^2)$, independently for $j = 1,\ldots p$. For $j \in A^*$,

$$\mathbb{P}(|U_j| \leq a_p) = \Phi\Big(\frac{a_p - \beta_i^*}{\sigma}\Big) - \Phi\Big(\frac{-a_p - \beta_i^*}{\sigma}\Big) \leq \Phi\Big(\frac{a_p - \theta_p}{\sigma}\Big) \to 0,$$

so

$$\mathbb{P}\Big(\min_{j\in A^*} |U_j| > a_p\Big) = \prod_{j\in A^*} \mathbb{P}(|U_j| > a_p) \to 1.$$

At the same time,

$$\mathbb{P}\Big(\max_{j\notin A^*} |U_j| \leq a_p\Big) = \Big(\Phi(a_p/\sigma) - \Phi(-a_p/\sigma)\Big)^{p-k_0} \to 1.$$

Therefore $\mathbb{P}(B) \to 1$. This in fact means that $\mathbb{P}(A) - \mathbb{P}(A|B) \to 0$ for any sequence of events A, so only the weak convergence of $(T_{k_0+1}, \ldots T_{k_0+d})$ remains to be proved. For this, we let $m = p - k_0$, and $V_1 \geq V_2 \geq \ldots \geq V_m$ denote the order statistics of the sample $|U_j|$, $j \notin A^*$ of independent χ_1 variates. Then, on the event B, we have

$$T_{k_0+i} = V_i(V_i - V_{i+1}), \quad \text{for } i = 1,\ldots d.$$

As $\mathbb{P}(B) \to 1$, we have in general

$$T_{k_0+i} = V_i(V_i - V_{i+1}) + o_{\mathbb{P}}(1), \quad \text{for } i = 1,\ldots d.$$

Hence we are essentially back in the setting of the last section, and the desired convergence result follows from the same arguments as in the proof of Lemma 3. ∎

4 A general predictor matrix X

In this section, we assume that the predictor matrix $X \in \mathbb{R}^{n \times p}$ has unit norm columns, $\|X_i\|_2 = 1$, $i = 1, \ldots p$, and that the columns $X_1, \ldots X_p \in \mathbb{R}^n$ are in general position, but otherwise X can be arbitrary. These are very weak assumptions. Our proposed covariance test statistic (5) is closely intertwined with the knots $\lambda_1 \geq \ldots \geq \lambda_r$ in the lasso path, as it was defined in terms of difference between fitted values at successive knots. Moreover, Lemma 1 showed that (provided there are no sign changes in the reduced lasso problem over $[\lambda_{k+1}, \lambda_k]$) this test statistic can be expressed even more explicitly in terms of the values of these knots. As was the case in the last section, this knot form is quite important for our analysis here.

Therefore, it is helpful to recall (Efron et al. 2004, Tibshirani 2012) the explicit formulae for the knots in the lasso path. If A denotes the active set and s_A denotes the signs of active coefficients at a knot λ_k,

$$A = \mathrm{supp}\big(\hat{\beta}(\lambda)\big), \quad s_A = \mathrm{sign}\big(\hat{\beta}_A(\lambda_k)\big),$$

then the next knot λ_{k+1} is given by

$$\lambda_{k+1} = \max\big\{\lambda_{k+1}^{\mathrm{join}}, \lambda_{k+1}^{\mathrm{leave}}\big\}, \tag{11}$$

where $\lambda_{k+1}^{\mathrm{join}}$ and $\lambda_{k+1}^{\mathrm{leave}}$ are the values of λ at which, if we were to decrease the tuning parameter from λ_k and continue along the current (linear) trajectory for the lasso coefficients, a variable would join and leave the active set A, respectively. These values are

$$\lambda_{k+1}^{\mathrm{join}} = \max_{j \notin A,\, s \in \{-1,1\}} \frac{X_j^T(I - P_A)y}{s - X_j^T(X_A^T)^+ s_A} \cdot 1\left\{\frac{X_j^T(I - P_A)y}{s - X_j^T(X_A^T)^+ s_A} \leq \lambda_k\right\}, \tag{12}$$

where P_A is the projection onto the column space of X_A, $P_A = X_A(X_A^T X_A)^{-1} X_A^T$, and $(X_A^T)^+$ is the pseudo-inverse $(X_A^T)^+ = (X_A^T X_A)^{-1} X_A^T$; and

$$\lambda_{k+1}^{\mathrm{leave}} = \max_{j \in A} \frac{[(X_A)^+ y]_j}{[(X_A^T X_A)^{-1} s_A]_j} \cdot 1\left\{\frac{[(X_A)^+ y]_j}{[(X_A^T X_A)^{-1} s_A]_j} \leq \lambda_k\right\}. \tag{13}$$

As we did in Section 3 with the orthogonal X case, we begin by studying the asymptotic distribution of the covariance statistic in the case $k = 1$ (i.e., the first model along the path), wherein the expressions for the next knot (11), (12), (13) greatly simplify. Following this, we give of sketch of the arguments for the more difficult case $k > 1$. For the sake of readability we defer the proofs and most technical details until the appendix.

4.1 The first predictor to enter, $k = 1$

As per our discussion following Lemma 1, we know that the first predictor to enter along the lasso path cannot leave at the next step, so assumption (8) holds, and the covariance statistic for testing the entry of the first variable is $T_1 = \lambda_1(\lambda_1 - \lambda_2)/\sigma^2$.

Now let $U_j = X_j^T y$, $j = 1, \ldots p$, and $R = X^T X$. With $\lambda_0 = \infty$, we have $A = \emptyset$, and trivially, no variables can leave the active set. The first knot is hence given by (12), which can be expressed as

$$\lambda_1 = \max_{j = 1, \ldots p,\, s \in \{-1,1\}} sU_j. \tag{14}$$

Letting j_1, s_1 be the first variable to enter and its sign (i.e., they achieve the maximum in the above expression), and recalling that j_1 cannot leave the active set immediately after it has entered, the second knot is again given by (12), written as

$$\lambda_2 = \max_{j \neq j_1,\, s \in \{-1,1\}} \frac{sU_j - sR_{j,j_1}U_{j_1}}{1 - ss_1R_{j,j_1}} \cdot 1\left\{\frac{sU_j - sR_{j,j_1}U_{j_1}}{1 - ss_1R_{j,j_1}} \leq s_1U_{j_1}\right\}.$$

The general position assumption on X implies that $|R_{j,j_1}| < 1$, and so $1 - ss_1 R_{j,j_1} > 0$, all $j \neq j_1$, $s \in \{-1, 1\}$. It is easy to show then that the indicator inside the maximum above can be dropped, and hence

$$\lambda_2 = \max_{j \neq j_1,\, s \in \{-1,1\}} \frac{sU_j - sR_{j,j_1}U_{j_1}}{1 - ss_1 R_{j,j_1}}. \tag{15}$$

Our goal now is to calculate the asymptotic distribution of $T_1 = \lambda_1(\lambda_1 - \lambda_2)/\sigma^2$, with λ_1 and λ_2 as above, under the null hypothesis; to be clear, since we are testing the significance of the first variable to enter along the lasso path, the null hypothesis is

$$H_0 : \;\; y \sim N(0, \sigma^2 I). \tag{16}$$

The strategy that we use here for the general X case—which differs from our extreme value theory approach for the orthogonal X case—is to treat the quantities inside the maxima in expressions (14), (15) for λ_1, λ_2 as discrete-time Gaussian processes. First, we consider the zero mean Gaussian process

$$g(j, s) = sU_j, \quad \text{for } j = 1, \ldots p, \; s \in \{-1, 1\}. \tag{17}$$

We can easily compute the covariance function of this process:

$$\mathbb{E}\big[g(j, s)g(j', s')\big] = ss' R_{j,j'} \sigma^2,$$

where the expectation is taken over the null distribution in (16). From (14), we know that the first knot is simply

$$\lambda_1 = \max_{j,\, s} \; g(j, s),$$

and from (15), the second knot is

$$\lambda_2 = \max_{j \neq j_1,\, s} \frac{g(j, s) - \mathbb{E}\big[g(j, s)g(j_1, s_1)\big]/\sigma^2 \cdot g(j_1, s_1)}{1 - \mathbb{E}\big[g(j, s)g(j_1, s_1)\big]/\sigma^2},$$

with j_1 and s_1 being the first variable to enter and its sign, i.e., $\lambda_1 = g(j_1, s_1)$. Hence in addition to (17), we consider the process

$$h^{(j_1,s_1)}(j, s) = \frac{g(j, s) - \mathbb{E}\big[g(j, s)g(j_1, s_1)\big]/\sigma^2 \cdot g(j_1, s_1)}{1 - \mathbb{E}\big[g(j, s)g(j_1, s_1)\big]/\sigma^2}. \tag{18}$$

An important property: for fixed j_1, s_1, the entire process $h^{(j_1,s_1)}(j, s)$ is independent of $g(j_1, s_1)$. This can be seen by verifying that

$$\mathbb{E}[h^{(j_1,s_1)}(j, s)g(j_1, s_1)] = 0,$$

and noting that $g(j_1, s_1)$ and $h^{(j_1,s_1)}(j, s)$, all $j \neq j_1$, $s \in \{-1, 1\}$, are jointly normal for fixed j_1, s_1.

Now define

$$M(j_1, s_1) = \max_{j \neq j_1,\, s} \; h^{(j_1,s_1)}(j, s), \tag{19}$$

and from the above we know that for fixed j_1, s_1, $M(j_1, s_1)$ is independent of $g(j_1, s_1)$. If j_1, s_1 are instead treated as random variables that maximize $g(j, s)$, then $\lambda_2 = M(j_1, s_1)$. Therefore, to study the distribution of $T = \lambda_1(\lambda_1 - \lambda_2)/\sigma^2$, we are interested in the random variable

$$g(j_1, s_1)\big(g(j_1, s_1) - M(j_1, s_1)\big)/\sigma^2,$$

on the event

$$\{g(j_1, s_1) \geq g(j, s) \text{ for all } j, s\}.$$

It turns out that this event, which concerns the argument maximizers of g, can be rewritten as an event concerning only the relative values of g and M [see Taylor et al. (2005) for the analogous result for continuous-time processes].

Lemma 4. *With g, M as defined in* (17), (18), (19), *we have*

$$\{g(j_1, s_1) \geq g(j, s) \text{ for all } j, s\} = \{g(j_1, s_1) \geq M(j_1, s_1)\}.$$

This is an important realization because the dual representation $\{g(j_1, s_1) \geq M(j_1, s_1)\}$ is more tractable, once we partition the space over the possible argument minimizers j_1, s_1, and use the fact that $M(j_1, s_1)$ is independent of $g(j_1, s_1)$ for fixed j_1, s_1. In this vein, we express the distribution of $T_1 = \lambda_1(\lambda_1 - \lambda_2)/\sigma^2$ in terms of the sum

$$\mathbb{P}(T_1 > t) = \sum_{j_1, s_1} \mathbb{P}\Big(g(j_1, s_1)\big(g(j_1, s_1) - M(j_1, s_1)\big)/\sigma^2 > t, \;\; g(j_1, s_1) \geq M(j_1, s_1)\Big).$$

The terms in the above sum can be simplified: dropping for notational convenience the dependence on j_1, s_1, we have

$$g(g - M)/\sigma^2 > t, \;\; g \geq M \quad \Leftrightarrow \quad g/\sigma > u(t, M/\sigma),$$

where $u(a, b) = (b + \sqrt{b^2 + 4a})/2$, which follows by simply solving for g in the quadratic equation $g(g - M)/\sigma^2 = t$. Therefore

$$\begin{aligned} \mathbb{P}(T_1 > t) &= \sum_{j_1, s_1} \mathbb{P}\Big(g(j_1, s_1)/\sigma > u\big(t, M(j_1, s_1)/\sigma\big)\Big) \\ &= \sum_{j_1, s_1} \int_0^\infty \bar{\Phi}\big(u(t, m/\sigma)\big) F_{M(j_1, s_1)}(dm), \end{aligned} \tag{20}$$

where $\bar{\Phi}$ is the standard normal survival function (i.e., $\bar{\Phi} = 1 - \Phi$, for Φ the standard normal CDF), $F_{M(j_1, s_1)}$ is the distribution of $M(j_1, s_1)$, and we have used the fact that $g(j_1, s_1)$ and $M(j_1, s_1)$ are independent for fixed j_1, s_1, and also $M(j_1, s_1) \geq 0$ on the event $\{g(j_1, s_1) \geq M(j_1, s_1)\}$. (The latter follows as Lemma 4 shows this event to be equivalent to j_1, s_1 being the argument maximizers of g, which means that $M(j_1, s_1) = \lambda_2 \geq 0$.) Continuing from (20), we can write the difference between $\mathbb{P}(T_1 > t)$ and the standard exponential tail, $\mathbb{P}(\text{Exp}(1) > t) = e^{-t}$, as

$$\big|\mathbb{P}(T_1 > t) - e^{-t}\big| = \bigg| \sum_{j_1, s_1} \int_0^\infty \bigg(\frac{\bar{\Phi}\big(u(t, m/\sigma)\big)}{\bar{\Phi}(m/\sigma)} - e^{-t} \bigg) \bar{\Phi}(m/\sigma) F_{M(j_1, s_1)}(dm) \bigg|, \tag{21}$$

where we used the fact that

$$\sum_{j_1, s_1} \int_0^\infty \bar{\Phi}\big(m/\sigma\big) F_{M(j_1, s_1)}(dm) = \sum_{j_1, s_1} \mathbb{P}\big(g(j_1, s_1) \geq M(j_1, s_1)\big) = 1.$$

We now examine the term inside the braces in (21), the difference between a ratio of normal survival functions and e^{-t}; our next lemma shows that this term vanishes as $m \to \infty$.

Lemma 5. *For any $t \geq 0$,*

$$\lim_{m \to \infty} \frac{\bar{\Phi}\big(u(t, m)\big)}{\bar{\Phi}(m)} = e^{-t}.$$

Hence, loosely speaking, if each $M(j_1, s_1) \to \infty$ fast enough as $p \to \infty$, then the right-hand side in (21) converges to zero, and T_1 converges weakly to Exp(1). This is made precise below.

Lemma 6. *Consider $M(j_1, s_1)$ defined in* (18), (19) *over $j_1 = 1, \ldots p$ and $s_1 \in \{-1, 1\}$. If for any fixed $m_0 > 0$*

$$\sum_{j_1, s_1} \mathbb{P}\big(M(j_1, s_1) \leq m_0\big) \to 0 \quad \text{as } p \to \infty, \tag{22}$$

then the right-hand side in (21) *converges to zero as $p \to \infty$, and so $\mathbb{P}(T_1 > t) \to e^{-t}$ for all $t \geq 0$.*

The assumption in (22) is written in terms of random variables whose distributions are induced by the steps along the lasso path; to make our assumptions more transparent, we show that (22) is implied by a conditional variance bound involving the predictor matrix X alone, and arrive at the main result of this section.

Theorem 2 (Exponential null distribution, general X, $k = 1$)**.** *Assume that $X \in \mathbb{R}^{n\times p}$ has unit norm columns in general position, and let $R = X^T X$. Assume also that there is some $\delta > 0$ such that for each $j = 1, \ldots p$, there exists a subset of indices $S \subseteq \{1, \ldots p\} \setminus \{j\}$ with*

$$1 - R_{i,S\setminus\{i\}}(R_{S\setminus\{i\},S\setminus\{i\}})^{-1}R_{S\setminus\{i\},i} \geq \delta \quad \textit{for all } i \in S, \tag{23}$$

and the size of S growing faster than $\log p$,

$$\frac{|S|}{\log p} \to \infty \quad \textit{as } p \to \infty. \tag{24}$$

Then $\mathbb{P}(T_1 > t) \to e^{-t}$ as $p \to \infty$ for all $t \geq 0$, under the null hypothesis in (16).

4.2 Subsequent predictors, $k > 1$

Here we give a sketch of the distribution theory for the statistic (5) for testing the significance of subsequent predictors, i.e., for $k > 1$. Precise statements will be made in future work.

We assume that there are k_0 large nonzero components of β^*, large enough that these variables are entered into the lasso model (and not deleted from the active set) over the first k_0 steps of the lasso path. We consider the distribution of the covariance statistic when a noise variable j_k is added at step $k = k_0 + 1$. Under the same assumptions, we can also show that the constant sign condition (8) holds (see the third point following Lemma 1), so we may use the knot form of the covariance statistic $T_k = C(A, s_A, j_k) \cdot \lambda_k(\lambda_k - \lambda_{k+1})$, with A and s_A being the active set and signs just before knot λ_k.

The general calculations for this case all have more or less the same form as they do in the last section. However, a main complication is that the indicator terms inside the maxima in (12), (13) do not vanish for general k as they did in (14), (15). Therefore, in the notation of the last section, we may define g and M analogously, but these two are no longer independent (for a fixed value of the variable j_k to enter and its sign s_k). We hence introduce a triplet of random variables M^+, M^-, M^0, defined carefully so that we can decompose the distribution of the test statistic as

$$\mathbb{P}(T_k > t) = \sum_{j_k,s_k} \mathbb{P}\Big(g(j_k,s_k)\big(g(j_k,s_k) - M(j_k,s_k)\big) > t,\ \ g(j_k,s_k) \geq M^+(j_k,s_k),\\ g(j_k,s_k) \leq M^-(j_k,s_k),\ \ 0 \geq M^0(j_k,s_k)\Big). \tag{25}$$

In the above expression, for each j_k, s_k, the random variable $g(j_k, s_k)$ is independent of the triplet $M^+(j_k, s_k), M^-(j_k, s_k), M^0(j_k, s_k)$. Furthermore, we have the constraint

$$\sum_{j_k,s_k} \mathbb{P}\Big(g(j_k,s_k) \geq M^+(j_k,s_k),\ \ g(j_k,s_k) \leq M^-(j_k,s_k),\ \ 0 \geq M^0(j_k,s_k)\Big) = 1,$$

which says that these events form a partition (up to a null set). A similar calculation to that given in the last section then shows that the right-hand side in (25), with $M(j_k, s_k)$ replaced by $M^+(j_k, s_k)$, converges to e^{-t} as $p \to \infty$, under the assumption that for an fixed m_0,

$$\sum_{j_k,s_k} \mathbb{P}\big(M^+(j_k,s_k) \leq m_0\big) \to 0.$$

Adding the condition $\sum_{j_k,s_k} \mathbb{P}(M^+(j_k, s_k) > M(j_k, s_k)) \to 0$, we get a (conservative) e^{-t} bound for $\mathbb{P}(T_k > t)$ in (25).

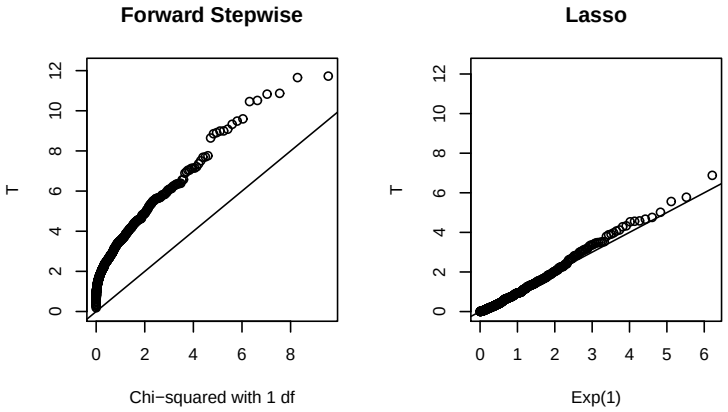


Figure 2: *Simulated data: quantile-quantile plots of the* RSS *test from forward stepwise regression (left) and covariance test from Lasso (right), for testing the (null) 4th predictor.*

5 Simulation of the null distribution

5.1 Orthonormal design

In this example we generated $N = 100$ observations each with $p = 10$ standard Gaussian features. The first three coefficients were equal to 3, and the rest zero. The error standard deviation was $\sigma = 0.5$ so that the first three predictors had strong effects and always enter first. Figure 2 shows the results for testing the 4th (null) predictor to enter. The panels show the drop in RSS test from forward stepwise regression and the covariance test, with σ^2 assumed known. We see that the Exp(1) distribution provides a good approximation for the distribution of the covariance statistic, while the χ^2_1 is a poor approximation for the RSS test. Figure 3 shows the results for entering the 5th, 6th and 7th predictors. The test will be conservative: with a nominal level 0.05, the actual type I errors are 0.01, 0.002, and 0.000 respectively. The solid line has slope 1, while the broken lines have slope $1/2, 1/3, 1/4$ respectively, as predicted by Theorem 1.

5.2 Simulations: general design matrix

In Table 2 we simulated null data, and the distribution of the covariance test statistic T for the first predictor to enter. We varied the numbers of predictors p, feature correlation ρ and form of the feature correlation matrix. In the first two correlation setups, the pairwise correlation between each pair of features was ρ, in the data and population, respectively. In the $AR(1)$ setup, the correlation between features j and k is $\rho^{|j-k|}$. Finally, in the block diagonal setup, the correlation matrix has two equal sized blocks, with population correlation ρ in each block. We see that the exponential (1) distribution is a reasonably good approximation throughout.

In Table 3 with the first k coefficients equal to 4.0, and the rest zero, for $k = 1, 2, 3$. The rest of the setup was the same as in Table 2, except that the sample size was fixed at 50. We computed the mean, variance and tail probability of the covariance statistic T_{k+1} for entering the next (null) $(k + 1)$st predictor, discarding those simulations in which a non-null predictor was chosen in the first k. (This occurred 1.7%, 4.0%, and 7.0% of the time, respectively) The Table shows that the exponential (1) distribution is again a reasonably good approximation.

In Figure 4 we estimate the power curves for forward stepwise and the lasso (the latter using

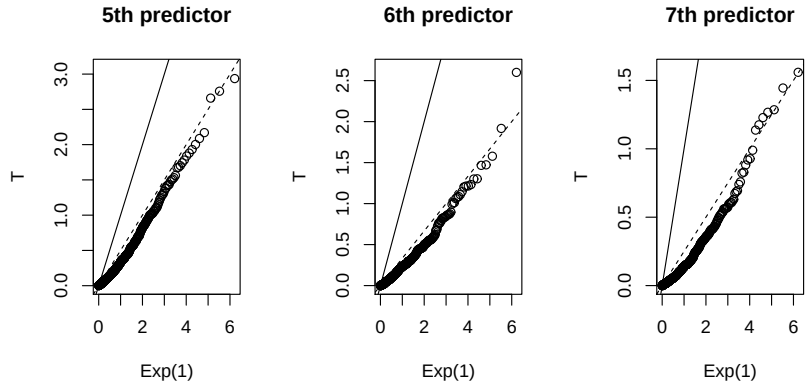


Figure 3: *Testing the (null) 5th, 6th and 7th predictors. The solid line has slope 1, while the broken lines have slope* $1/2, 1/3, 1/4$ *respectively, as predicted by Lemma 2.*

the covariance statistic), and find that they have similar power. Details are in the figure caption. However the cutpoints for forward stepwise, estimated here by simulation, are not typically available in practice,

6 The case of unknown σ

So far we have assumed that the error variance is known. In practice it will typically be unknown: in that case, we can estimate it and proceed by analogy to standard linear model theory. In particular, consider the numerator from (5)

$$W_k = \Big(\langle y, X\hat{\beta}(\lambda_{k+1})\rangle - \langle y, X_A\tilde{\beta}_A(\lambda_{k+1})\rangle\Big). \tag{26}$$

Consider first the case $N < p$. Then we can estimate σ^2 by the mean residual error $\hat{\sigma}^2 = \sum_{i=1}^{N}(y_i - \hat{\mu}_{\text{full}})^2/(N-p)$, with $\hat{\mu}_{\text{full}}$ the least squares fit for the full model. Then asymptotically

$$F_k = \frac{W_k}{\hat{\sigma}^2} \sim F_{2,N-p} \tag{27}$$

This follows because W_j is asymptotically Exp(1) which is the same as $\chi^2_2/2$, $(N-p)\cdot\hat{\sigma}^2$ is asymptotically χ^2_{N-p}, and the two are independent.

Here is a proof of the independence. Letting P_X project onto the column space of X, the lasso problem is unchanged is we solve it with $P_X y$ in place of y. The estimate for $\hat{\sigma}$ is a function of $(I-P_X)y$. It's clear that $P_X y$ and $(I-P_X)y$ are uncorrelated, and if we assume that y is normally distributed, then $P_X y$ and $(I-P_X)$ are independent. Therefore the lasso fit and estimate of σ are functions of independent quantities, hence independent. This is true for any λ.

The statistic that we propose involves both the lasso fit on X and on X_A. For the lasso problem on X_A, we can still replace y with $P_X y$, because $(I-P_X)X_A = 0$. Hence the same argument applies to the lasso fit on X_A, that is, it's independent of $\hat{\sigma}$.

As an example, consider one of the setups from Table 2, with $N = 100, p = 80$, AR(1) feature correlation $\rho^{|j-k|}$, and the model truly null. Consider testing the first predictor to enter. We chose $p \approx N$ to expose the difference between the σ known and unknown cases. Table 4 shows the results

Table 2: *Simulation results: null distribution for first predictor to enter, for different of numbers of predictors p, feature correlation ρ and form of the feature correlation matrix. Bottom left corner panel is missing because equal population correlation model isn't defined for $p > N$. There were 500 simulated datasets for each condition, and "se" is the Monte Carlo standard error.*

	$N = 100, p = 10$											
ρ	Equal data corr.			Equal pop'n corr.			AR(1)			Block diagonal		
	Mean	Var	Pr> q_{95}	Mean	Var	Pr> q_{95}	Mean	Var	Pr> q_{95}	Mean	Var	Pr> q_{95}
0	0.966	1.157	0.062	1.120	1.951	0.090	1.017	1.484	0.070	1.058	1.548	0.060
0.2	0.972	1.178	0.066	1.119	1.844	0.086	1.034	1.497	0.074	1.069	1.614	0.078
0.4	0.963	1.219	0.060	1.115	1.724	0.092	1.045	1.469	0.060	1.077	1.701	0.076
0.6	0.960	1.265	0.070	1.095	1.648	0.086	1.048	1.485	0.066	1.074	1.719	0.086
0.8	0.958	1.367	0.060	1.062	1.624	0.092	1.034	1.471	0.062	1.062	1.687	0.072
se	0.007	0.015	0.001	0.010	0.049	0.001	0.013	0.043	0.001	0.010	0.047	0.001
	$N = 100, p = 50$											
0	0.929	1.058	0.048	1.078	1.721	0.074	1.039	1.415	0.070	0.999	1.578	0.048
0.2	0.920	1.032	0.038	1.090	1.476	0.074	0.998	1.391	0.054	1.064	2.062	0.052
0.4	0.928	1.033	0.040	1.079	1.382	0.068	0.985	1.373	0.060	1.076	2.168	0.062
0.6	0.950	1.058	0.050	1.057	1.312	0.060	0.978	1.425	0.054	1.060	2.138	0.060
0.8	0.982	1.157	0.056	1.035	1.346	0.056	0.973	1.439	0.060	1.046	2.066	0.068
se	0.010	0.030	0.001	0.011	0.037	0.001	0.009	0.041	0.001	0.011	0.103	0.001
	$N = 100, p = 200$											
0				1.004	1.017	0.054	1.029	1.240	0.062	0.930	1.166	0.042
0.2				0.996	1.164	0.052	1.000	1.182	0.062	0.927	1.185	0.046
0.4				1.003	1.262	0.058	0.984	1.016	0.058	0.935	1.193	0.048
0.6				1.007	1.327	0.062	0.954	1.000	0.050	0.915	1.231	0.044
0.8				0.989	1.264	0.066	0.961	1.135	0.060	0.914	1.258	0.056
se				0.008	0.039	0.001	0.009	0.028	0.001	0.007	0.032	0.001

of 1000 simulations from each of $\rho = 0$ and $\rho = 0.8$. The F distribution provides a more accurate approximation than does Exp(1).

In Table 5, we have repeated the experiment of Table 2, except that σ^2 is unknown and is estimated from the data. This yields a F distribution for the covariance statistic. We derived the estimate of σ^2 from the mean residual error of the least squares fit to the full model when $N > p$, and when $p > N$, from the least squares fit to the support of the optimal model selected by cross-validation. The F approximation is reasonably good, although the variance of the observed statistic is sometimes inflated, especially for $p > N$.

The problem of estimation of σ^2 for the lasso when $p > N$ is a difficult one, and it is not clear that cross-validation is a good approach. See for example Fan et al. (2012) for a recent study of this issue.

7 Real data examples

7.1 Wine quality data

Table 6 shows the results for the red wine quality data taken from the UCI database. There are 11 predictors, and 1599 observations, which we split randomly into approximately equal-size training and test sets. The outcome is a wine quality rating, on a scale between 0 and 10.

The table shows the training set p-values from forward stepwise regression and the lasso. Forward stepwise enters 7 predictors at the 0.05 level, which the lasso enters only 3. In the top panel of Figure 5 we repeated this computation for 500 random training-test splits. In the bottom panel we show the corresponding test set error for the models of each size. There are three strong predictors, and then a few moderately strong ones, in general qualitative agreement with the lasso p-values in the

Table 3: *Simulation results: null distribution for second, third and fourth predictor to enter, for different feature correlation ρ and form of the feature correlation matrix. The number of features p was fixed at 50. There were 500 simulated datasets for each condition, and "se" is the Monte carlo standard error.*

	$N = 100, p = 10$											
ρ	Equal data corr.			Equal pop'n corr.			AR(1)			Block diagonal		
	Mean	Var	Pr> q_{95}	Mean	Var	Pr> q_{95}	Mean	Var	Pr> q_{95}	Mean	Var	Pr> q_{95}
	2nd predictor to enter											
0	0.933	1.091	0.048	1.105	1.628	0.078	1.023	1.146	0.064	1.039	1.579	0.060
0.2	0.940	1.051	0.046	1.039	1.554	0.082	1.017	1.175	0.060	1.062	2.015	0.062
0.4	0.952	1.126	0.056	1.016	1.548	0.084	0.984	1.230	0.056	1.042	2.137	0.066
0.6	0.938	1.129	0.064	0.997	1.518	0.079	0.964	1.247	0.056	1.018	1.798	0.068
0.8	0.818	0.945	0.039	0.815	0.958	0.044	0.914	1.172	0.062	0.822	0.966	0.037
se	0.010	0.024	0.002	0.011	0.036	0.002	0.010	0.030	0.002	0.015	0.087	0.002
	3rd predictor to enter											
0	0.927	1.051	0.046	1.119	1.724	0.094	0.996	1.108	0.072	1.072	1.800	0.064
0.2	0.928	1.088	0.044	1.070	1.590	0.080	0.996	1.113	0.050	1.043	2.029	0.060
0.4	0.918	1.160	0.050	1.042	1.532	0.085	1.008	1.198	0.058	1.024	2.125	0.066
0.6	0.897	1.104	0.048	0.994	1.371	0.077	1.012	1.324	0.058	0.945	1.568	0.054
0.8	0.719	0.633	0.020	0.781	0.929	0.042	1.031	1.324	0.068	0.771	0.823	0.038
se	0.011	0.034	0.002	0.014	0.049	0.003	0.009	0.022	0.002	0.013	0.073	0.002
	4th predictor to enter											
0	0.925	1.021	0.046	1.080	1.571	0.086	1.044	1.225	0.070	1.003	1.604	0.060
0.2	0.926	1.159	0.050	1.031	1.463	0.069	1.025	1.189	0.056	1.010	1.991	0.060
0.4	0.922	1.215	0.048	0.987	1.351	0.069	0.980	1.185	0.050	0.918	1.576	0.053
0.6	0.905	1.158	0.048	0.888	1.159	0.053	0.947	1.189	0.042	0.837	1.139	0.052
0.8	0.648	0.503	0.008	0.673	0.699	0.026	0.940	1.244	0.062	0.647	0.593	0.015
se	0.014	0.037	0.002	0.016	0.044	0.003	0.014	0.031	0.003	0.016	0.073	0.002

top panel.

7.2 HIV data

Rhee et al. (2003) study six nucleotide reverse transcriptase inhibitors (NRTIs) that are used to treat HIV-1. The target of these drugs can become resistant through mutation, and they compare a collection of models for predicting these drugs (log) susceptibility– a measure of drug resistance, based on the location of mutations. We focussed on the first drug (3TC), for which there are there are between $p = 217$ sites and and $N = 1057$ samples. To examine the behavior of the covariance test when $N < p$, We divided the data at random into training and test sets of size (150,907) fifty times. Figure 6 shows the results, in the same format as Figure 5. We used the model chosen by cross-validation to estimate σ^2. The covariance test for the lasso suggests that there are only one or two important predictors (in marked contrast to the forward stepwise test), and this is confirmed by the test error plot in the bottom panel.

8 Extensions

8.1 The elastic net

The (naive) elastic net criterion is

$$J_{\lambda,\gamma}(\beta) = \frac{1}{2}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j| + \frac{\gamma}{2}\sum_j \beta_j^2 \tag{28}$$

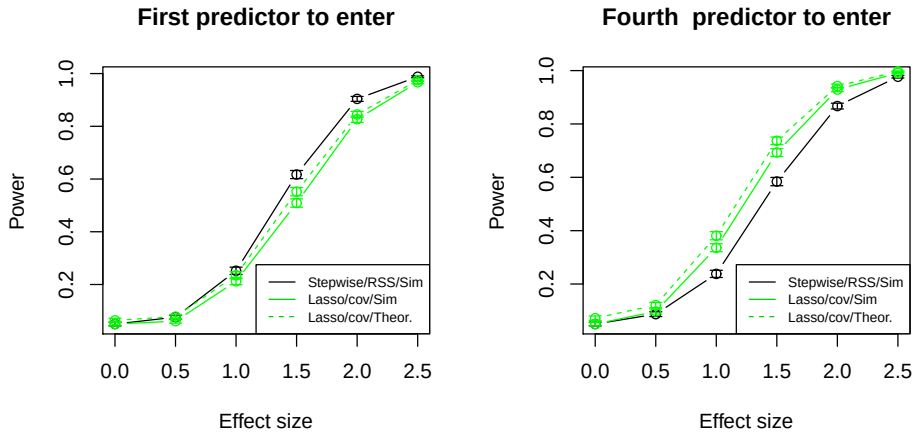


Figure 4: *Simulation of power with* $N = 100, p = 10$*, features i.i.d* $N(0,1)$ *,* $\sigma = 0.5$*. On left all true regression coefficients are zero, except for* β_1 *which is (Effect size)*$\cdot\sigma$*, and we examine the first step of the forward stepwise and lasso procedures. In the right, there are three large coefficients in addition to* β_1*, and we examine the 4th step after the three have been entered. Displayed are the power curves for forward stepwise using simulation to estimate the cutpoint yielding 5% type I error, and the same for lasso using both simulation-based and theoretical (Exp(1)) cutpoints.*

Table 4: *Comparison of Exp(1) and* $F_{2,N-p}$ *as approximations to the distribution of the covariance test statistic, when* σ *is estimated. Here* $N = 100, p = 20$*, the feature correlation is of the AR(1) form* $\rho^{|j-k|}$*, and the model is truly null. Results for* $\rho = 0.0$ *and* 0.8 *are shown.*

	Mean	Variance	.95 quantile	Prob$> q_{95}$
		$\rho = 0$		
observed	1.17	2.10	3.75	
Exp(1)	1.00	1.00	2.99	0.082
$F_{2,n-p}$	1.11	1.54	3.49	0.054
		$\rho = 0.8$		
observed	1.14	1.70	3.77	
Exp(1)	1.00	1.00	2.99	0.097
$F_{2,n-p}$	1.11	1.54	3.49	0.064

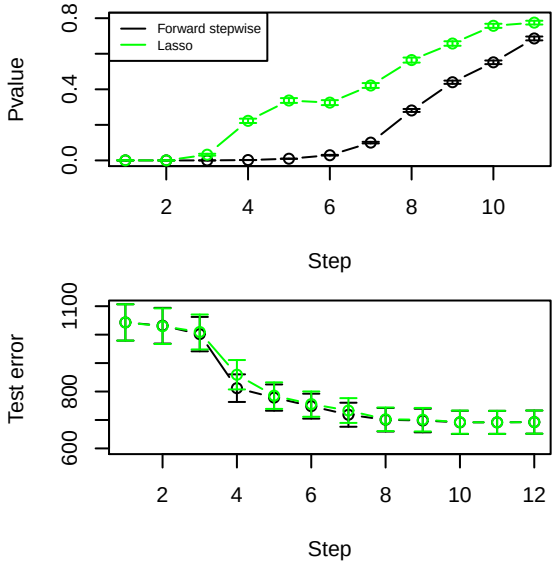


Figure 5: *Red wine data. The data were randomly divided 500 times into approximately equal sized training and set sets. Top row shows the training set p-values for forward stepwise regression and the lasso. The bottom panel show the test set error for the models of each size.*

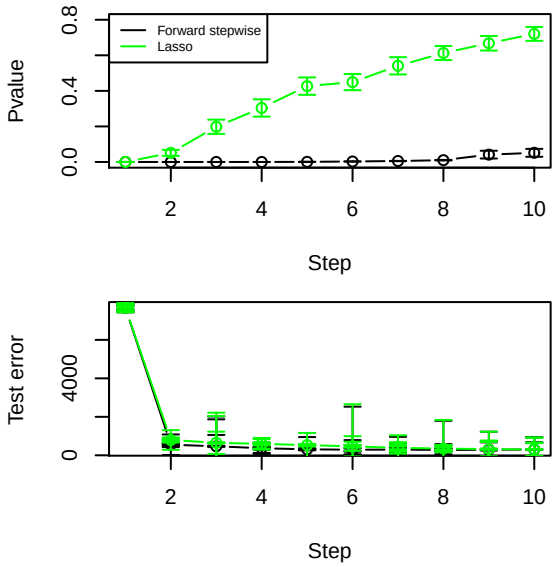


Figure 6: *HIV mutation data. The data were randomly divided 50 times into training and test sets of size (150,907). Top panel shows the training set p-values for forward stepwise regression and the lasso. The bottom panel shows the test set error for the models of each size.*

Table 5: *Same setup as is in Table 3, but σ^2 is unknown and is estimated from the data leading to an F distribution for the covariance statistic. The estimate of σ^2 is derived from the mean residual error from the least squares fit to the full model when $N > p$, and when $p > N$, from the least squares fit to the support of the optimal model selected by cross-validation.*

					$N = 100, p = 10$							
ρ	Equal data corr.			Equal pop'n corr.			AR(1)			Block diagonal		
	Mean	Var	Pr$> q_{95}$	Mean	Var	Pr$> q_{95}$	Mean	Var	Pr$> q_{95}$	Mean	Var	Pr$> q_{95}$
0	1.047	1.503	0.056	1.028	1.478	0.074	1.049	1.693	0.066	1.069	1.354	0.068
0.2	1.027	1.443	0.060	1.024	1.617	0.066	1.045	1.757	0.072	1.062	1.371	0.062
0.4	0.997	1.435	0.058	1.038	1.686	0.068	1.068	1.803	0.064	1.048	1.395	0.072
0.6	0.975	1.464	0.060	1.055	1.733	0.078	1.104	1.975	0.062	1.027	1.416	0.062
0.8	0.970	1.564	0.066	1.064	1.767	0.078	1.146	2.208	0.068	0.989	1.438	0.056
se	0.005	0.027	0.001	0.011	0.052	0.002	0.016	0.096	0.003	0.010	0.036	0.002
					$N = 100, p = 50$							
ρ	Equal data corr.			Equal pop'n corr.			AR(1)			Block diagonal		
0	0.993	1.354	0.046	1.140	1.939	0.080	1.097	1.558	0.062	1.038	1.713	0.052
0.2	0.988	1.295	0.050	1.160	1.663	0.082	1.055	1.552	0.058	1.112	2.400	0.052
0.4	0.992	1.285	0.054	1.153	1.556	0.066	1.042	1.573	0.050	1.128	2.594	0.056
0.6	1.016	1.338	0.054	1.136	1.508	0.068	1.039	1.642	0.060	1.115	2.568	0.060
0.8	1.049	1.477	0.062	1.117	1.575	0.062	1.039	1.685	0.066	1.105	2.437	0.068
se	0.011	0.039	0.001	0.012	0.042	0.003	0.011	0.047	0.002	0.012	0.123	0.002
					$N = 100, p = 200$							
ρ	Equal data corr.			Equal pop'n corr.			AR(1)			Block diagonal		
0				1.144	1.320	0.100	1.209	1.745	0.120	1.151	1.137	0.060
0.2				1.269	4.960	0.080	1.100	1.586	0.100	1.010	1.413	0.080
0.4				1.194	3.598	0.050	1.056	1.326	0.050	1.046	1.498	0.060
0.6				1.140	3.170	0.030	1.028	1.350	0.040	1.084	1.929	0.070
0.8				1.134	4.152	0.040	0.999	0.970	0.040	1.058	1.817	0.050
se				0.034	0.483	0.004	0.026	0.092	0.005	0.019	0.076	0.005

Let $X^* = \big(X, \sqrt{\gamma} \cdot I_p\big)$, $y^* = (y, 0_p)$, and let $\hat{\beta}^*$ solve the lasso problem with data (X^*, y^*). Then it is easy to show that the elastic net solutions $\hat{\beta}^{\text{en}}(\lambda, \gamma)$ are equal to $\hat{\beta}^*/\sqrt{1+\gamma_2}$. This shows that the paths are piecewise linear with breakpoints.

The covariance test statistic has the some form as it does for the lasso

$$T_k = \frac{1}{\sigma^2} \cdot \langle y, X\hat{\beta}^{\text{en}}(\lambda_{k+1}, \gamma)\rangle - \langle y, X_A\tilde{\beta}_A^{\text{en}}(\lambda_{k+1}, \gamma)\rangle. \tag{29}$$

Now in the orthogonal design case,

$$\hat{\beta}_k^{\text{en}}(\lambda, \gamma) = s(\hat{\beta}_k)\frac{(|\hat{\beta}_k| - \lambda)_+}{1+\gamma}, \tag{30}$$

$\hat{\beta}_k$ being the (univariate) least squares estimates. Therefore one can show that $T_k = |\hat{\beta}_{(k)}|(\hat{\beta}_{(k)}| - |\hat{\beta}_{(k+1)}|)/(1+\gamma)$. This leads to the approximation

$$(1+\gamma) \cdot T_k \sim \text{Exp}(1) \tag{31}$$

In Figure 7 we tried this approximation for the first predictor to enter, for orthogonal and correlated scenarios, and for three different values of γ. Here $n = 100, p = 10$ and the true model was null. It seems to work reasonably well in all cases.

Table 6: *Wine data: results of forward stepwise and lasso fits.*

	Forward stepwise		
	Predictor	RSSDrop	P-value
1	alcohol	263.672	0
2	volatile_acidity	111.857	0
3	sulphates	36.225	0
4	fixed_acidity	9.023	0.003
5	chlorides	10.066	0.002
6	total_sulfur_dioxide	6.115	0.014
7	pH	5.383	0.021
8	residual_sugar	1.946	0.163
9	free_sulfur_dioxide	1.933	0.165
10	citric_acid	0.503	0.478
11	density	0.451	0.502
	Lasso		
	Predictor	CovDrop	P-value
1	alcohol	53.702	0
2	volatile_acidity	37.156	0
3	sulphates	19.605	0
4	fixed_acidity	1.750	0.174
5	total_sulfur_dioxide	0.912	0.402
6	chlorides	1.299	0.273
7	pH	2.876	0.056
8	residual_sugar	0.806	0.447
9	free_sulfur_dioxide	1.225	0.294
10	density	0.002	0.998
11	citric_acid	0.502	0.606

8.2 Generalized linear models

Here we briefly discuss the extension of the covariance test to generalized linear models. Suppose that the outcome Y has an exponential family density

$$f(y|\beta) = \exp(\eta \cdot U(y) - A(\beta) + B(y)). \tag{32}$$

The natural parameter η is assumed to be linear function of the features: $\eta = X\beta$, and the mean $\mu = \mathbb{E}(Y)$ is related to η via a link function $\eta = g(\mu)$. Having fit this model to n observations by maximum likelihood, producing an n-vector of fitted values $\hat{\mu}$, we might define the degrees of freedom as

$$\mathrm{df}(\hat{\mu}) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{cov}(y_i, \hat{\eta}_i) \tag{33}$$

This is the implicit concept used by Efron (1986*b*) in his definition of the "optimism" of the training error.

In a similar way, we can define the covariance test in this setting, as follows. We add an ℓ_1 penalty $\lambda \sum |\beta_j|$ to the (negative) log-likelihood and for all values of λ maximize the resulting objective function to yield the estimate $\hat{\eta} = X\beta$. Unlike in the Gaussian case, the paths of the ℓ_1-penalized maximum likelihood estimates are not piecewise linear. However one can still determine numerically the knot values $\hat{\lambda}_k$ where the kth predictor enters. Then the covariance test for kth predictor to enter is defined as

$$T_k = \langle y, X\hat{\beta}(\lambda_{k+1})\rangle - \langle y, X_A\tilde{\beta}_A(\lambda_{k+1})\rangle, \tag{34}$$

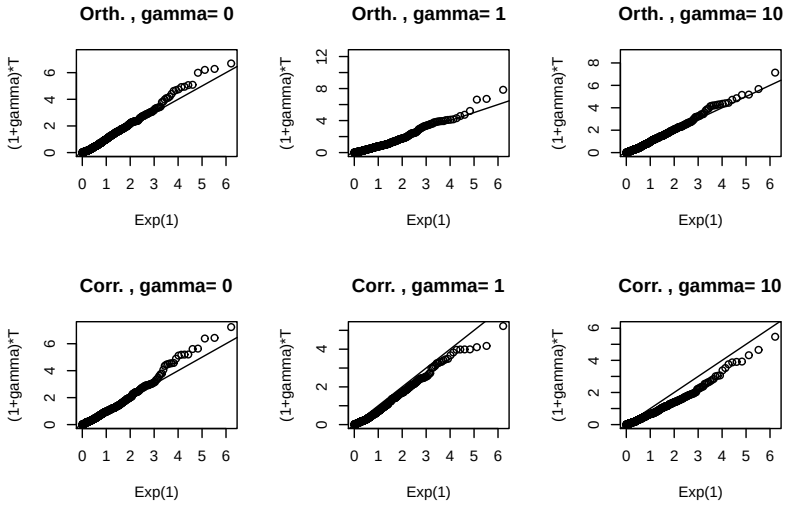


Figure 7: *Elastic net: approximation of the null distribution, for orthogonal and correlated designs, and three different values of the regularization parameter γ.*

where as before A is the active set of predictors at λ_j. By analogy to the Gaussian case, we hope that T_k has an asymptotic Exp(1) distribution, although we have not proven this.

As as example, we consider the linear logistic model for binary data. In this case, $\mu = \Pr(Y = 1|x)$ and $\eta = \log(\mu/(1-\mu))$. Figure 8 shows the results of a simulation comparing the null distribution of the covariance test statistic (34) to Exp(1). We used the `glmpath` R package (Park & Hastie 2007), which employs a predictor-corrector method to compute the regularization path for generalized linear models. The approximation looks quite good.

For general likelihood-based regression problems, let $\eta = X\beta$ and let $\ell(\eta)$ denote the log likelihood. We can view maximum likelihood estimation as an iteratively weighted least squares procedure using the response variable

$$z(\eta) = \eta + I_\eta^{-1} S_\eta \tag{35}$$

with $S_\eta = \partial\ell(\eta)/\partial\eta$, $I_\eta = \partial^2\ell(\eta)/\partial\eta\eta^T$. This applies for example, to the class of generalized linear models and Cox's proportional hazards model. We can analogously define the covariance test statistic as

$$T_k = [\langle S_0 I_0^{-1/2}, X\hat{\beta}(\lambda_{k+1})\rangle - \langle {S_0}^T I_0^{-1/2}, X_{\mathcal{A}}\tilde{\beta}_{\mathcal{A}}(\lambda_{k+1})\rangle]/2 \tag{36}$$

For the binomial model, this reduces to expression (34). In Figure 9 we computed this test for Cox's proportional hazards model, using a setup similar to that of Figure 8. The approximation looks reasonably accurate.

9 Discussion

We have proposed a simple "covariance" statistic for testing the significance of predictors as they are entered in a linear model, fit via the lasso. This statistic has an Exp(1) distribution asymptotically. This distribution accounts for the adaptive nature of the fitting, which is not true for the usual chi-squared or F test in adaptive least squares fitting, such as forward stepwise. As such, this result

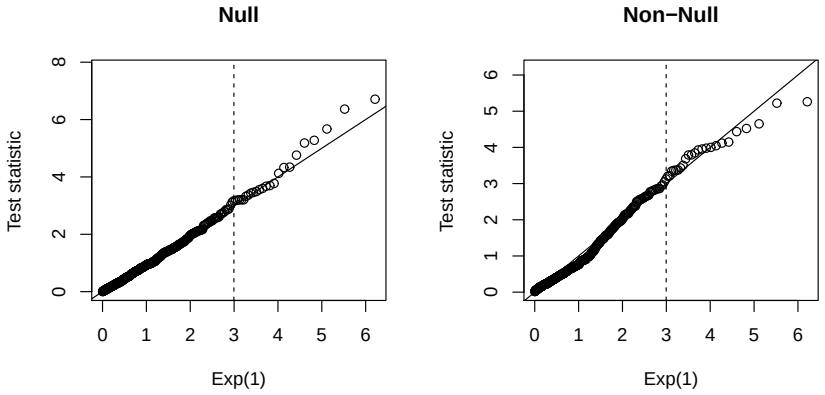


Figure 8: *Simulation for binary data, with* $n = 100, p = 10$*, features all i.i.d* $N(0, 1)$*. In left panel, all true regression coefficients are zero; in right, first coefficient is large, and the rest are zero. Shown are quantile-quantile plots of the covariance test statistic versus its asympototic distribution Exp(1).*

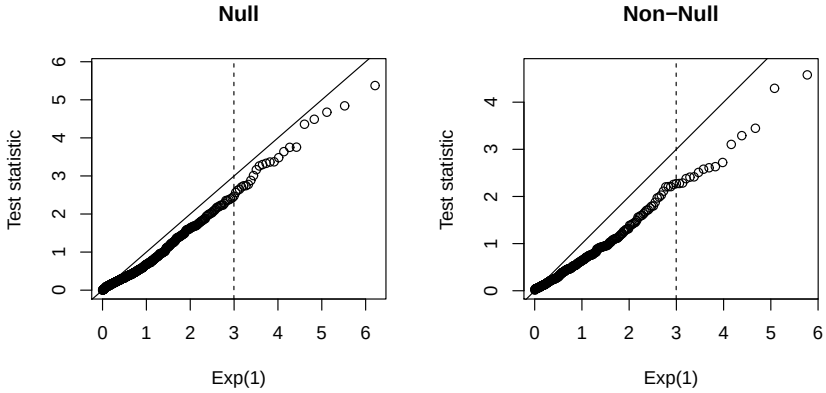


Figure 9: *Simulation for censored survival data, with* $n = 100, p = 10$*, features all i.i.d* $N(0, 1)$*. In left panel, all true regression coefficients are zero; in right, first coefficient is large, and the rest are zero. Shown are quantile-quantile plots of the covariance test statistic (36) versus its asympototic distribution Exp(1).*

is analogous to the degrees of freedom result for the lasso, which shows that the lasso with k non-zero parameters has used up k degrees of freedom.

There are many interesting directions in need of further work. Of particular importance is the generic lasso testing problem: given a lasso fit at an arbitrary value λ (not necessarily one of the knots in the LARS path), how do we carry out a significance test for any member of the active set at that λ? In more detail, suppose that we have a lasso fit $\hat{\beta}(\lambda)$ at some fixed λ, yielding a set of predictors $\mathcal{M}$ with nonzero coefficients. We want to test for the significance of each of the individual predictors $j \in \mathcal{M}$. For any predictor $j \in \mathcal{M}$, we consider use of a test statistic T_j of a similar form as the covariance test statistic defined earlier. Letting $\mathcal{M}(-j)$ denote the set $\mathcal{M}$ with the predictor j removed, we might define

$$T_j = \frac{1}{\sigma^2} \cdot \Big(\langle y, X\hat{\beta}(\lambda) \rangle - \langle y, X_{\mathcal{M}(-j)} \hat{\beta}_{\mathcal{M}(-j)}(\lambda) \rangle \Big). \tag{37}$$

Now in the orthogonal design case, one can show that

$$T_j = \Big[(|\hat{\beta}_{(j)}|(|\hat{\beta}_{(j)}| - |\hat{\beta}_{(j+1)}|) + |\hat{\beta}_{(j)}|(|\hat{\beta}_{(j+1)}| - \lambda) \Big] / \sigma^2. \tag{38}$$

Thus approximately, we might hope that

$$T_j' = T_j - \Big[|\hat{\beta}_{(j)}|(|\hat{\beta}_{(j+1)}| - \lambda) \Big] / \sigma^2 \sim \text{Exp}(1). \tag{39}$$

However the quality of this approximation is unclear. Overall, this problem seems more difficult than the sequential version, and is deserving of further study.

Extensions to other related fitting methods such as the group lasso and elastic net, and exploration of appropriate form the test statistic and its asymptotic distribution for generalized linear models and the proportional hazards model would be important.

An R package `covTest` for computing the covariance test, will be made freely available on the CRAN repository.

Acknowledgements. Richard Lockhart was supported from the Natural Sciences and Engineering Research Council of Canada; Jonathan Taylor was supported by NSF DMS 1208857 and AFOSR 113039; Robert Tibshirani was supported by NSF Grant DMS-9971405 and NIH grant N01-HV-28183.

A Appendix

A.1 Proof of Lemma 1

By continuity of the lasso solution path at λ_k,

$$P_A y - \lambda_k (X_A^T)^+ s_A = P_{A\cup\{j\}} y - \lambda_k (X_{A\cup\{j\}}^T)^+ s_{A\cup\{j\}},$$

and therefore

$$(P_{A\cup\{j\}} - P_A) y = \lambda_k \Big((X_{A\cup\{j\}}^T)^+ s_{A\cup\{j\}} - (X_A^T)^+ s_A \Big). \tag{40}$$

From this, we can obtain two identities: the first is

$$y^T (P_{A\cup\{j\}} - P_A) y = \lambda_k^2 \cdot \| (X_{A\cup\{j\}}^T)^+ s_{A\cup\{j\}} - (X_A^T)^+ s_A \|_2^2, \tag{41}$$

obtained by squaring both sides in (40) (more precisely, taking the inner product of the left-hand side with itself and the right-hand side with itself), and noting that $(P_{A\cup\{j\}} - P_A)^2 = P_{A\cup\{j\}} - P_A$; the second is

$$y^T \Big((X_{A\cup\{j\}}^T)^+ s_{A\cup\{j\}} - (X_A^T)^+ s_A \Big) = \lambda_k \cdot \| (X_{A\cup\{j\}}^T)^+ s_{A\cup\{j\}} - (X_A^T)^+ s_A \|_2^2, \tag{42}$$

obtained by taking the inner product of both sides in (40) with y, and then using (41). Plugging (41) and (42) in for the first and second terms in (7), respectively, then gives the result in (9). □

A.2 Proof of Lemma 3

Define $W_i = (V_i - a_p)/b_p$ for $i = 1, \ldots k+1$, with $a_p = F^{-1}(1-1/p)$ and $b_p = pF'(a_p)$, as in the proof of Lemma 2. Then Theorem 2.1.1 in de Haan & Ferreira (2006) shows that

$$(W_1, W_2, \ldots W_{k+1}) \xrightarrow{d} \big(-\log E_1, -\log(E_1+E_2), \ldots -\log(E_1+E_2+\ldots+E_{k+1})\big),$$

where $E_1, \ldots E_{k+1}$ are independent standard exponential variates. By the same arguments as those given for Lemma 2,

$$V_i(V_i - V_{i+1}) = W_i - W_{i+1} + o_{\mathbb{P}}(1), \quad \text{all } i = 1, \ldots k+1,$$

and so

$$\begin{aligned}
&\big(V_1(V_1-V_2), V_2(V_2-V_3), \ldots, V_k(V_k - V_{k+1})\big) \\
&\xrightarrow{d} \left(\log(E_1+E_2) - \log E_1,\ \log(E_1+E_2+E_3) - \log(E_1+E_2),\ \ldots\ \log\left(\sum_{i=1}^{k+1} E_i\right) - \log\left(\sum_{i=1}^{k} E_i\right)\right).
\end{aligned}$$

Now let

$$D_i = \frac{E_1 + \ldots + E_i}{E_1 + \ldots + E_{i+1}}, \quad \text{for } i = 1, \ldots k,$$

and $D_{k+1} = E_1 + \ldots + E_{k+1}$. A change of variables shows that $D_1, \ldots, D_{k+1}$ are independent. For each $i = 1, \ldots k$, the variable D_i is distributed as Beta$(i, 1)$—this is the distribution of the largest order statistic in a sample of size i from a uniform distribution. It is easily checked that $-\log D_i$ has distribution Exp$(1/i)$, and the result follows. □

A.3 Proof of Lemma 4

Note that

$$\begin{aligned}
&g(j_1, s_1) \geq g(j, s) \\
&\qquad \Leftrightarrow \quad \frac{g(j_1,s_1) - \mathbb{E}[g(j,s)g(j_1,s_1)]/\sigma^2 \cdot g(j_1,s_1)}{1 - \mathbb{E}[g(j,s)g(j_1,s_1)]/\sigma^2} \geq \frac{g(j,s) - \mathbb{E}[g(j,s)g(j_1,s_1)]/\sigma^2 \cdot g(j_1,s_1)}{1 - \mathbb{E}[g(j,s)g(j_1,s_1)]/\sigma^2} \\
&\qquad \Leftrightarrow \quad g(j_1, s_1) \geq h^{(j_1,s_1)}(j, s),
\end{aligned}$$

the first step following as $1 - \mathbb{E}[g(j,s)g(j_1,s_1)]/\sigma^2 = 1 - ss_1R_{j,j_1} > 0$, and the second step following from the definition of $h^{(j_1,s_1)}$. Taking the conjunction of these statements over all j, s [and using the definition of $M(j_1, s_1)$] gives the result. □

A.4 Proof of Lemma 5

By l'Hôpital's rule,

$$\lim_{m\to\infty} \frac{\bar{\Phi}\big(u(t,m)\big)}{\bar{\Phi}(m)} = \lim_{m\to\infty} \frac{\phi\big(u(t,m)\big)}{\phi(m)} \cdot \frac{\partial u(t,m)}{\partial m},$$

where ϕ is the standard normal density. First note that

$$\frac{\partial u(t,m)}{\partial m} = \frac{1}{2} + \frac{m}{2\sqrt{m^2+4t}} \to 1 \quad \text{as } m \to \infty.$$

Also, a straightforward calculation shows

$$\log\phi\big(u(t,m)\big) - \log\phi(m) = \frac{m^2}{2}\big(1 - \sqrt{1+4t/m^2}\big) - \frac{t}{2} \to -t \quad \text{as } m \to \infty,$$

where in the last step we used the fact that $(1-\sqrt{1+4t/m^2})/(2/m^2) \to -t/2$, again by l'Hôpital's rule. Therefore $\phi(u(t,m))/\phi(m) \to e^{-t}$, which completes the proof. □

A.5 Proof of Lemma 6

Fix $\epsilon > 0$, and choose m_0 large enough that

$$\left| \frac{\bar{\Phi}\big(u(t, m/\sigma^2)\big)}{\bar{\Phi}(m/\sigma^2)} - e^{-t} \right| \leq \epsilon \quad \text{for all } m \geq m_0.$$

Starting from (21),

$$\begin{aligned}
\mathbb{P}(T_1 > t) &\leq \sum_{j_1,s_1} \int_0^\infty \left| \frac{\bar{\Phi}\big(u(t, m/\sigma)\big)}{\bar{\Phi}(m/\sigma)} - e^{-t} \right| \bar{\Phi}(m/\sigma) F_{M(j_1,s_1)}(dm) \\
&\leq \epsilon \sum_{j_1,s_1} \int_{m_0}^\infty \bar{\Phi}(m/\sigma) F_{M(j_1,s_1)}(dm) + 2 \sum_{j_1,s_1} \int_0^{m_0} F_{M(j_1,s_1)}(dm) \\
&\leq \epsilon \sum_{j_1,s_1} \mathbb{P}\big(g(j_1,s_1) \geq M(j_1,s_1)\big) + 2 \sum_{j_1,s_1} \mathbb{P}\big(M(j_1,s_1) \leq m_0\big),
\end{aligned}$$

Above, the term multiplying ϵ is equal to 1, and the second term can be made arbitrarily small (say, less than ϵ) by taking p sufficiently large. □

A.6 Proof of Theorem 2

We will show that for fixed $m_0 > 0$ and any j_1, s_1,

$$\mathbb{P}\big(M(j_1,s_1) \leq m_0\big) \leq c^{|S|}, \tag{43}$$

where $c < 1$, and $S \subseteq \{1, \ldots p\} \setminus \{j_1\}$ is as in the theorem. This would imply that

$$\sum_{j_1,s_1} \mathbb{P}\big(M(j_1,s_1) \leq m_0\big) \leq 2p \cdot c^{|S|} \to 0 \quad \text{as } p \to \infty,$$

where we used that $|S|/\log p \to \infty$ by (24). The above sum tending to zero now implies the desired convergence result by Lemma 6, and hence it suffices to show (43). To this end, we use the bound

$$\mathbb{P}\big(M(j_1,s_1) \leq m_0\big) \leq \mathbb{P}(|V_j| \leq m_0,\, j \in S),$$

where V_j is defined as

$$V_j = \frac{U_j - R_{j,j_1} U_{j_1}}{1 + |R_{j,j_1}|}.$$

Let $r = |S|$, and without a loss of generality, write $S = \{1, \ldots r\}$. We show that

$$\mathbb{P}(|V_1| \leq m_0, \ldots |V_r| \leq m_0) \leq c^r, \tag{44}$$

for $c = \Phi(2m_0/\delta) - \Phi(-2m_0/\delta) < 1$, by induction. But before presenting this argument, we note a few important facts. First, the condition in (23) is really a statement about conditional variances:

$$\mathrm{Var}(U_i \,|\, U_\ell,\, \ell \in S \setminus \{i\}) = 1 - R_{i,S\setminus\{i\}} (R_{S\setminus\{i\},S\setminus\{i\}})^{-1} R_{S\setminus\{i\},i} \geq \delta \quad \text{for all } i \in S.$$

Second, since U_ℓ, $\ell \in S \setminus \{i\}$ are jointly normal, we have

$$\mathrm{Var}(U_i \,|\, U_\ell,\, \ell \in A) \geq \mathrm{Var}(U_i \,|\, U_\ell,\, \ell \in S \setminus \{i\}) \geq \delta^2 \tag{45}$$

for any subset $A \subseteq S \setminus \{i\}$ (including $A = \emptyset$), which can be verified using the conditional variance formula (law of total variance). Finally, the collection $V_1, \ldots V_r$ is independent of U_{j_1}, because these random variables are again jointly normal, and $\mathbb{E}[V_j U_{j_1}] = 0$ for all $j = 1, \ldots r$.

Now we give the inductive argument for (44). For the base case, note that $V_1 \sim N(0, \tau_1^2)$, where its variance is

$$\tau_1^2 = \mathrm{Var}(V_1) = \mathrm{Var}(V_1|U_{j_1}) = \mathrm{Var}(U_1)/(1 + |R_{j,j_1}|)^2 \geq \delta^2/4,$$

the second equality is due to the independence of V_1 and U_{j_1}, and the last inequality comes from the fact that conditioning can only decrease the variance, as stated above in (45). Hence

$$\mathbb{P}(|V_1| \leq m_0) = \Phi(m_0/\tau_1) - \Phi(-m_0/\tau_1) \leq \Phi(2m_0/\delta) - \Phi(-2m_0/\delta) = c.$$

Assume as the inductive hypothesis that $\mathbb{P}(|V_1| \leq m_0, \ldots |V_q| \leq m_0) \leq c^q$. Then

$$\mathbb{P}(|V_1| \leq m_0, \ldots |V_{q+1}| \leq m_0) = \mathbb{P}\big(|V_{q+1}| \leq m_0 \,\big|\, |V_1| \leq m_0, \ldots |V_q| \leq m_0\big) \cdot c^q,$$

We have, using the independence of $V_1, \ldots V_r$ and U_{j_1},

$$\begin{aligned} V_{q+1} \,|\, V_1, \ldots V_q &\stackrel{d}{=} V_{q+1} \,|\, V_1, \ldots V_q, U_{j_1} \\ &\stackrel{d}{=} V_{q+1} \,|\, U_1, \ldots U_q, U_{j_1} \\ &\stackrel{d}{=} N(0, \tau_{q+1}^2), \end{aligned}$$

where the variance is

$$\tau_{q+1}^2 = \mathrm{Var}(V_{q+1} \,|\, U_1, \ldots U_q, U_{j_1}) = \mathrm{Var}(U_{q+1} \,|\, U_1, \ldots U_q)/(1 + |R_{q+1,j_1}|)^2 \geq \delta^2/4,$$

and here we again used the fact that conditioning further can only reduce the variance, as in (45). Therefore

$$\mathbb{P}(|V_{q+1}| \leq m_0 \,|\, V_1, \ldots V_k) \leq \Phi(2m_0/\delta) - \Phi(-2m_0/\delta) = c,$$

and so

$$\mathbb{P}(|V_1| \leq m_0, \ldots |V_{q+1}| \leq m_0) \leq c \cdot c^q = c^{q+1}.$$

This completes the inductive proof, and verifies (43), as desired. □

References

Beck, A. & Teboulle, M. (2009), 'A fast iterative shrinkage-thresholding algorithm for linear inverse problems', *SIAM Journal on Imaging Sciences* **2**(1), 183–202.

Becker, S., Bobin, J. & Candes, E. J. (2011), 'NESTA: A fast and accurate first-order method for sparse recovery', *SIAM Journal on Imaging Sciences* **4**(1), 1–39.

Becker, S., Candes, E. J. & Grant, M. (2011), 'Templates for convex cone problems with applications to sparse signal recovery', *Mathematical Programming Computation* **3**(3), 165–218.

Berk, R., Brown, L. & Zhao, L. (2010), 'Statistical inference after model selection', *Journal of Quantitative Criminology* **26**, 217–236.

Boyd, S., Parikh, N., Chu, E., Peleato, B. & Eckstein, J. (2011), 'Distributed optimization and statistical learning via the alternative direction method of multipliers', *Foundations and Trends in Machine Learning* **3**(1), 1–122.

Bühlmann, P. (2012), Statistical significance in high-dimensional linear models.
URL: *http://arxiv.org/abs/1202.1377*

Candes, E. J. & Plan, Y. (2009), 'Near ideal model selection by ℓ_1 minimization', *Annals of Statistics* **37**(5), 2145–2177.

Candes, E. J. & Tao, T. (2006), 'Near optimal signal recovery from random projections: Universal encoding strategies?', *IEEE Transactions on Information Theory* **52**(12), 5406–5425.

Chen, S., Donoho, D. L. & Saunders, M. (1998), 'Atomic decomposition for basis pursuit', *SIAM Journal on Scientific Computing* **20**(1), 33–61.

de Haan, L. & Ferreira, A. (2006), *Extreme Value Theory: An Introduction*, Springer.

Donoho, D. L. (2006), 'Compressed sensing', *IEEE Transactions on Information Theory* **52**(12), 1289–1306.

Efron, B. (1986*a*), 'How biased is the apparent error rate of a prediction rule?', *Journal of the American Statistical Association: Theory and Methods* **81**(394), 461–470.

Efron, B. (1986*b*), 'How biased is the apparent error rate of a prediction rule?', *Journal of the American Statistical Association* **81**, 461–70.

Efron, B., Hastie, T., Johnstone, I. & Tibshirani, R. (2004), 'Least angle regression', *Annals of Statistics* **32**(2), 407–499.

Fan, J., Guo, S. & Hao, N. (2012), 'Variance estimation using refitted cross-validation in ultrahigh dimensional regression', *Journal of Royal Statistical Society* **75**, 37–65.

Friedman, J., Hastie, T., Hoefling, H. & Tibshirani, R. (2007), 'Pathwise coordinate optimization', *Annals of Applied Statistics* **1**(2), 302–332.

Friedman, J., Hastie, T. & Tibshirani, R. (2010), 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software* **33**(1), 1–22.

Fuchs, J. J. (2005), 'Recovery of exact sparse representations in the presense of bounded noise', *IEEE Transactions on Information Theory* **51**(10), 3601–3608.

Greenshtein, E. & Ritov, Y. (2004), 'Persistence in high-dimensional linear predictor selection and the virtue of overparametrization', *Bernoulli* **10**(6), 971–988.

Hastie, T., Tibshirani, R. & Friedman, J. (2008), *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, Springer, New York. Second edition.

Laber, E. & Murphy, S. (2011), 'Adaptive confidence intervals for the test error in classification (with discussion)', *J. Amer.Statist. Assoc,* **106**, 904–913.

Meinhasuen, N., Meier, L. & Bühlmann, P. (2009), 'P-values for high-dimensional regression', *Journal of the American Statistical Association* **104**, 1671–1681.

Meinshausen, N. & Bühlmann, P. (2010), 'Stability selection', *J. Royal. Stat. Soc. B* **72**, 417–473.

Minnier, J., Tian, L. & Cai, T. (2011), 'A perturbation method for inference on regularized regression estimates', *Journal of the American Statistical Association* **106**(496), 1371–1382.

Osborne, M., Presnell, B. & Turlach, B. (2000*a*), 'A new approach to variable selection in least squares problems', *IMA Journal of Numerical Analysis* **20**(3), 389–404.

Osborne, M., Presnell, B. & Turlach, B. (2000*b*), 'On the lasso and its dual', *Journal of Computational and Graphical Statistics* **9**(2), 319–337.

Park, M. Y. & Hastie, T. (2007), 'l_1-regularization path algorithm for generalized linear models', *Journal of the Royal Statistical Society: Series B* **69**(4), 659–677.

Rhee, S.-Y., Gonzales, M. J., Kantor, R., Betts, B. J., Ravela, J. & Shafer, R. W. (2003), 'Human immunodeficiency virus reverse transcriptase and pro- tease sequence database', *Nucleic Acids Research* **31**, 298–303.

Taylor, J., Takemura, A. & Adler, R. (2005), 'Validity of the expected Euler characteristic heuristic', *Annals of Probability* **33**(4), 1362–1296.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society: Series B* **58**(1), 267–288.

Tibshirani, R. J. (2012), The lasso problem and uniqueness. arXiv: 1206.0313.

Tibshirani, R. J. & Taylor, J. (2012), 'Degrees of freedom in lasso problems', *Annals of Statistics* **40**(2), 1198–1232.

Wainwright, M. J. (2009), 'Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso)', *IEEE Transactions on Information Theory* **55**(5), 2183–2202.

Wasserman, L. & Roeder, K. (2009), 'High-dimensional variable selection', *Journal of the American Statistical Association* **37**, 2178–2201.

Zhang, C.-H. & Zhang, S. (2011), Confidence intervals for low-dimensional parameters with high-dimensional data.

Zhao, P. & Yu, B. (2006), 'On model selection consistency of lasso', *Journal of Machine Learning Research* **7**, 2541–2564.

Zou, H., Hastie, T. & Tibshirani, R. (2007), 'On the "degrees of freedom" of the lasso', *Annals of Statistics* **35**(5), 2173–2192.