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 for the special case of the first predictor to enter the model (i.e., testing for a single significant predictor variable against the global null) requires only weak assumptions on the predictor matrix X. On the other hand, our proof for a general step in the lasso path places further technical assumptions on X and the generative model, but still allows for the important high-dimensional case $p > n$, and does not necessarily require that the current lasso model achieves perfect recovery of the truly active variables.

Of course, for testing the significance of an additional variable between two nested linear models, one typically uses the 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, and 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, p-value, significance test*

1 Introduction

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

$$y = X\beta^* + \epsilon, \quad \epsilon \sim N(0, \sigma^2 I), \tag{1}$$

where $\beta^* \in \mathbb{R}^p$ are unknown coefficients to be estimated. [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 more details.] 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 level of sparsity in $\hat{\beta}$. 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)].

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, with most methods being based on resampling or data splitting; we review this work in Section 2.5. The current paper focuses on a significance test for lasso models that does not employ resampling or data splitting, but instead uses the full data set as given, and proposes a test statistic that has a simple and exact asymptotic null distribution.

Section 2 defines the problem that we are trying to solve, and gives 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

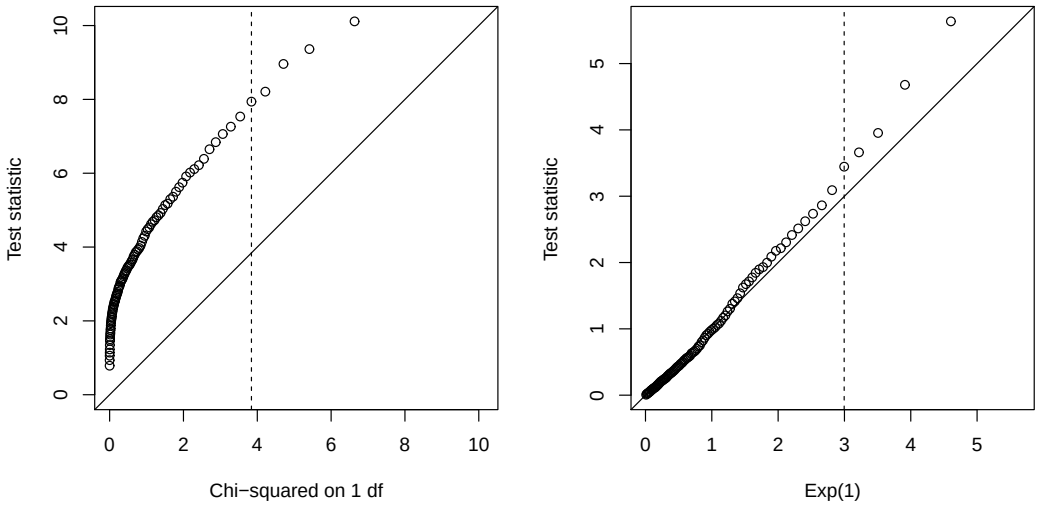
Classic theory for significance testing in linear regression operates on two fixed nested models. For example, if M and $M \cup \{j\}$ are fixed subsets of $\{1, \ldots p\}$, then to test the significance of the jth predictor in the model (with variables in) $M \cup \{j\}$, one naturally uses the chi-squared 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^2_1 distribution under the null hypothesis for each fixed j, the maximum possible R_j will clearly 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 the 5% level, for example, using the χ^2_1 cutoff of 3.84, would have an actual type I error of about 39%.



(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 null 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.*

The failure of standard testing methodology when applied to forward stepwise regression is not an anomaly—in general, there seems to be no direct way to carry out the significance tests designed for fixed linear models in an adaptive setting.[1] 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.

[1] It is important to mention that a simple application of sample splitting can yield proper p-values for an adaptive procedure like forward stepwise: e.g., run forward stepwise regression on one half of the observations to construct a sequence of models, and use the other half to evaluate significance via the usual chi-squared test. Some of the related work mentioned in Section 2.5 does essentially this, but with more sophisticated splitting schemes. Our proposal uses the entire data set as given, and we do not consider sample splitting or resampling techniques. Aside from adding a layer of complexity, the use of sample splitting can result in a loss of power in significance testing.

(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 possibly many solution paths; we assume that the columns of X are in general position[2], implying that there is a unique lasso solution 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 general X, the number of knots in the lasso path is bounded by 3^p (but in practice this bound is usually very loose). This bound comes from the following realization: if at some knot λ_k, the active set is $A = \text{supp}(\hat{\beta}(\lambda_k))$ and the signs of active coefficients are $s_A = \text{sign}(\hat{\beta}_A(\lambda_k))$, then the active set and signs cannot again be A and s_A at some other knot $\lambda_\ell \neq \lambda_k$. This in particular means that once a variable enters the active set, it cannot immediately leave the active set at the next step.
- 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 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

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

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,

[2]Points $X_1, \ldots X_p \in \mathbb{R}^n$ are said to be 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 $\{\pm X_i : i \neq i_1, \ldots i_{k+1}\}$.

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 about 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. In Sections 3 and 4, we show that under the null hypothesis that the current lasso model contains all truly active variables, $A \supseteq \text{supp}(\beta^*)$,

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

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 some cases, e.g., when we have a strict inclusion $A \supsetneq \text{supp}(\beta^*)$, the use of an Exp(1) null distribution is actually conservative, because the limiting distribution of T_k is stochastically smaller than Exp(1).] In the above limit, we are considering both $n, p \to \infty$; in Section 4 we allow for the possibility $p > n$, the high-dimensional case.

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, present a connection to degrees of freedom, review related work, and finally, discuss the null hypothesis in more detail.

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, and 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 inaccurate, as they are based on a null distribution that does not account for the adaptive choice of predictors. We now make several remarks.

Remark 1. The above example implicitly assumed that one might stop entering variables into the model when the computed p-value rose above some threshold. More generally, our proposed test

[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 slightly smaller p-values). The lasso p-values use a simple modification of the covariance test* (5) *for unknown variance, given in Section 6. All p-values are rounded to 3 decimal places.*

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

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 this to future work.

Remark 2. In the example, the lasso entered a predictor into the active set at each step. For a general X, however, a given predictor variable may enter the active set more than once along the lasso path, since it may leave the active set at some point. In this case we treat each entry as a separate problem. Our test is specific to a step in the path, and not to a predictor variable at large.

Remark 3. For the prostate cancer data set, it is important to include an intercept in the model. To accomodate this, we ran the lasso on centered y and column-centered X (which is equivalent to including an unpenalized intercept term in the lasso criterion), and then applied the covariance test (with the centered data). In general, centering y and the columns of X allows us to account for the effect of an intercept term, and still use a model of the form (1). From a theoretical perspective, this centering step creates a weak dependence between the components of the error vector $\epsilon \in \mathbb{R}^n$. If originally we assumed i.i.d. errors, $\epsilon_i \sim N(0, \sigma^2)$, then after centering y and the columns of X, our new errors are of the form $\tilde{\epsilon}_i = \epsilon_i - \bar{\epsilon}$, where $\bar{\epsilon} = \sum_{j=1}^n \epsilon_j/n$. It is easy see that these new errors are correlated:

$$\mathrm{Cov}(\tilde{\epsilon}_i, \tilde{\epsilon}_j) = -\sigma^2/n \quad \text{for } i \neq j.$$

One might imagine that such correlation would cause problems for our theory in Sections 3 and 4, which assumes i.i.d. normal errors in the model (1). However, a careful look at the arguments in these sections reveals that the only dependence on y is through $X^T y$, the inner products of y with the columns of X. Furthermore,

$$\mathrm{Cov}(X_i^T\tilde{\epsilon}, X_j^T\tilde{\epsilon}) = \sigma^2 X_i^T\Big(I - \frac{1}{n}\mathbb{1}\mathbb{1}^T\Big)X_j = \sigma^2 X_i^T X_j \quad \text{for all } i, j,$$

which is the same as it would have been without centering (here $\mathbb{1}\mathbb{1}^T$ is the matrix of all 1s, and we used that the columns of X are centered). Therefore, our arguments in Sections 3 and 4 apply equally well to centered data, and centering has no effect on the asymptotic distribution of T_k.

Remark 4. By design, the covariance test is applied in a sequential manner, estimating p-values for each predictor variable as it enters the model along the lasso path. A more difficult problem is to test the significance of any of the active predictors in a model fit by the lasso, at some arbitrary value of the tuning parameter λ. We discuss this problem briefly in Section 9.

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) pseudoinverse 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, with sign s upon entry, 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 $s_{A\cup\{j\}} = \mathrm{sign}(\hat{\beta}_{A\cup\{j\}}(\lambda_{k+1}))$. We can equivalently write $s_{A\cup\{j\}} = (s_A, s)$, the concatenation of s_A and the sign s of the jth coefficient when it entered (as no sign changes could have occurred inside of the interval $[\lambda_k, \lambda_{k+1}]$, by definition of the knots). 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),

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

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, $s_A = \mathrm{sign}(\hat{\beta}_A(\lambda_k))$, j be the predictor that enters the active set at λ_k, and s be its sign upon entry. 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*

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

where

$$C(A, s_A, j, s) = \|(X^T_{A\cup\{j\}})^+ s_{A\cup\{j\}} - (X^T_A)^+ s_A\|_2^2,$$

and $s_{A\cup\{j\}}$ is the concatenation of s_A and s.

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 $C(A, s_A, j, s) = 1$, 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 normed columns, $\|X_i\|_2 = 1$ for $i = 1, \ldots p$, then we again have $C(A, s_A, j, s) = 1$ (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 $s_A = \mathrm{sign}((X_A)^+ y) = \mathrm{sign}(\beta^*_A)$; in this case, both the lasso and least squares on X_A recover the signs of the true coefficients. Section 4.2 studies the general X and $k \geq 1$ case, wherein this third scenario is important.

2.4 Connection to degrees of freedom

There is an 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) as

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

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 fitting procedure, the higher this covariance, and the greater its degrees of freedom. The covariance test evaluates the significance of adding the jth predictor via a something loosely like a sample version of degrees of freedom, across two models: that fit on $A \cup \{j\}$, and that on A. This was more or less the inspiration for the current work.

Using the definition (10), 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.

2.5 Related work

There is quite a lot of recent work related to the proposal of this paper. Wasserman & Roeder (2009) propose a procedure for variable selection and p-value estimation in high-dimensional linear models based on sample splitting, and this idea was extended by Meinshausen et al. (2009). Meinshausen & Buhlmann (2010) propose a generic method using resampling called "stability selection", which controls the expected number of false positive variable selections. Minnier et al. (2011) use perturbation resampling-based procedures to approximate the distribution of a general class of penalized parameter estimates. One big difference with the work here: we propose a statistic that utilizes the data as given and does not employ any resampling or sample splitting.

Zhang & Zhang (2011) derive confidence intervals for contrasts of high-dimensional regression coefficients, by replacing the usual score vector with the residual from a relaxed projection (i.e., the residual from sparse linear regression). Buhlmann (2012) constructs p-values for coefficients in high-dimensional regression models, starting with ridge estimation and then employing a bias correction term that uses the lasso. Even more recently, van de Geer et al. (2013), Javanmard & Montanari (2013*a*), and Javanmard & Montanari (2013*b*) all present approaches for debiasing the lasso estimate based on estimates of the inverse covariance matrix of the predictors. (The latter work focuses on the special case of a predictor matrix X with i.i.d. Gaussian rows; the first two consider a general matrix X.) These debiased lasso estimates are asymptotically normal, which allows one to compute p-values both marginally for an individual coefficient, and simultaneously for a group of coefficients. All of the work mentioned in the present paragraph provides a way to make inferential statements about preconceived predictor variables of interest (or preconceived groups of interest); this is in contrast to our work, which instead deals directly with variables that have been adaptively selected by the lasso procedure. We discuss this next.

2.6 What precisely is the null hypothesis?

The referees of a preliminary version of this manuscript expressed some confusion with regard to the null distribution considered by the covariance test. Given a fixed number of steps $k \geq 1$ along the lasso path, the covariance test examines the set of variables A selected by the lasso before the kth step (i.e., A is the current active set not including the variable to be added at the kth step). In particular, the null distribution being tested is

$$H_0 : \ A \supseteq \text{supp}(\beta^*), \tag{11}$$

where β^* is the true underlying coefficient vector in the model (1). For $k = 1$, we have $A = \emptyset$ (no variables are selected before the first step), so this reduces to a test of the global null hypothesis: $\beta^* = 0$. For $k > 1$, the set A is random (it depends on y), and hence the null hypothesis in (11) is itself a random event. This makes the covariance test a *conditional hypothesis test* beyond the first step in the path, as the null hypothesis that it considers is indeed a function of the observed data. Statements about its null distribution must therefore be made conditional on the event that $A \supseteq \text{supp}(\beta^*)$, which is precisely what is done in Sections 3.2 and 4.2.

Compare the null hypothesis in (11) to a null hypothesis of the form

$$H_0 : \ S \cap \mathrm{supp}(\beta^*) = \emptyset, \tag{12}$$

where $S \subseteq \{1, \ldots p\}$ is a fixed subset. The latter hypothesis, in (12), describes the setup considered by Zhang & Zhang (2011), Buhlmann (2012), van de Geer et al. (2013), Javanmard & Montanari (2013*a*), and Javanmard & Montanari (2013*b*). At face value, the hypotheses (11) and (12) may appear similar [the test in (11) looks just like that in (12) with $S = \{1, \ldots p\} \setminus A$], but they are fundamentally very different. The difference is that the null hypothesis in (11) is random, whereas that in (12) is fixed; this makes the covariance test a conditional hypothesis test, while the tests constructed in all of the aforementioned work are traditional (unconditional) hypothesis tests. It should be made clear that the goal of our work and these works also differ. Our test examines an adaptive subset of variables A deemed interesting by the lasso procedure; for such a goal, it seems necessary to consider a random null hypothesis, as theory designed for tests of fixed hypotheses would not be valid here.[4] The main goal of Zhang & Zhang (2011), Buhlmann (2012), van de Geer et al. (2013), Javanmard & Montanari (2013*a*), and Javanmard & Montanari (2013*b*), it appears, is to construct a new set of variables, say $\tilde{A}$, based on testing the hypotheses in (12) with $S = \{j\}$ for $j = 1, \ldots p$. Though the construction of this new set $\tilde{A}$ may have started from a lasso estimate, it need not be true that $\tilde{A}$ matches the lasso active set A, and ultimately it is this new set $\tilde{A}$ (and inferential statements concerning $\tilde{A}$) that these authors consider the point of interest.

3 An orthogonal predictor matrix X

We examine 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 in the orthogonal case to a few interesting and previously unknown (as far as we can tell) results on the order statistics of independent standard χ_1 variates. For these reasons, we discuss the orthogonal X case in detail.

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

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

Again 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),

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

[4] In principle, fixed hypothesis tests can be used along with the appropriate correction for multiple comparisons in order to test a random null hypotheses. Aside from being conservative, it is unclear how to efficiently carry out such a procedure when the random null hypothesis consists of a group of coefficients (as opposed to a single one).

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 the special case $k = 1$, the test for the first predictor to enter the active set along the lasso path. We then examine the case $k \geq 1$, the test at a general step in the lasso path.

3.1 The first step, $k = 1$

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 predictor to enter, this is

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

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*

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

This lemma reveals a remarkably simple limiting distribution for the largest of independent χ_1 random variables times the gap between the largest two; we skip its proof, as it is a special case of the following generalization.

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

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

where the limiting distribution (on the right-hand side above) has independent components. To be perfectly clear, here and throughout we use $\text{Exp}(\alpha)$ *to denote the exponential distribution with scale parameter α (not rate parameter α), so that if $Z \sim \text{Exp}(\alpha)$, then $\mathbb{E}[Z] = \alpha$.*

Proof. The χ_1 distribution has CDF

$$F(x) = (2\Phi(x) - 1)1\{x \geq 0\},$$

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 Mills' ratio. Theorem 2.2.1 in de Haan & Ferreira (2006) then implies that, for constants $a_p = F^{-1}(1 - 1/p)$ and $b_p = pF'(a_p)$,

$$b_p(V_1 - a_p) \xrightarrow{d} -\log E_0,$$

where E_0 is a standard exponential variate, so $-\log E_0$ has the standard (or type I) extreme value distribution. Hence according to Theorem 3 in Weissman (1978), for any fixed $k \geq 1$, the random variables $W_0 = b_p(V_{k+1} - a_p)$ and $W_i = b_p(V_i - V_{i+1})$, $i = 1, \ldots, k$, converge jointly:

$$(W_0, W_1, W_2, \ldots, W_k) \xrightarrow{d} (-\log G_0, E_1/1, E_2/2, \ldots, E_k/k),$$

where $G_0, E_1, \ldots, E_k$ are independent, G_0 is Gamma distributed with scale parameter 1 and shape parameter k, and $E_1, \ldots, E_k$ are standard exponentials. Now note that

$$V_i(V_i - V_{i+1}) = \bigg(a_p + \frac{W_0}{b_p} + \sum_{j=i}^{k} \frac{W_j}{b_p}\bigg) \frac{W_i}{b_p}$$
$$= \frac{a_p}{b_p} W_i + \frac{1}{b_p^2} \bigg(W_0 + \sum_{j=i}^{k} W_j\bigg) W_i.$$

We claim that $a_p/b_p \to 1$; this would give the desired result as the second term converges to zero, using $b_p \to \infty$. 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 Mills' inequalities,

$$\frac{\phi(a_p)}{a_p} \frac{1}{1 + 1/a_p^2} \leq 1 - \Phi(a_p) \leq \frac{\phi(a_p)}{a_p},$$

and multiplying by $2p$,

$$\frac{b_p}{a_p} \frac{1}{1 + 1/a_p^2} \leq 1 \leq \frac{b_p}{a_p}.$$

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

Practically, Lemma 3 says that under the global null hypothesis $y \sim N(0, \sigma^2 I)$, 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(1/2), 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, $k \geq 1$ not under the global null hypothesis, but rather, under the weaker null hypothesis that all variables excluded from the current lasso model are truly inactive (i.e., they have zero coefficients in the true model). We study this in next section.

3.2 A general step, $k \geq 1$

We suppose that exactly k_0 components of the true coefficient vector β^* are nonzero, and consider testing the entry of the predictor at step $k = k_0 + 1$. Let $A^* = \mathrm{supp}(\beta^*)$ denote the true active set (so $k_0 = |A^*|$), and let B denote the event that all truly active variables are added at steps $1, \ldots k_0$,

$$B = \Big\{ \min_{j \in A^*} |U_j| > \max_{j \notin A^*} |U_j| \Big\}. \tag{13}$$

We show that under the null hypothesis (i.e., conditional on B), the test statistic T_{k_0+1} is asymptotically Exp(1), and 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. This is made precise below.

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

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

Let B denote the event in (13), namely, that the first k_0 variables entering the model along the lasso path are those in A^*. 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 study $\mathbb{P}(B)$. Let $\theta_p = \min_{i\in A^*} |\beta_i^*|$, and choose c_p such that

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

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

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

so

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

At the same time,

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

Therefore $\mathbb{P}(B) \to 1$. This in fact means that $\mathbb{P}(E|B) - \mathbb{P}(E) \to 0$ for any sequence of events E, 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 those for Lemma 3. □

4 A general predictor matrix X

In this section, we consider a general predictor matrix X, with columns in general position. Recall that 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 precise 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

$$\lambda_{k+1} = \max\big\{\lambda_{k+1}^{\mathrm{join}}, \lambda_{k+1}^{\mathrm{leave}}\big\}, \tag{14}$$

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[5]

$$\lambda_{k+1}^{\mathrm{join}} = \max_{j \notin A,\, s \in \{-1,1\}} \frac{X_j^T(I - P_A)y}{s - X_j^T(X_A^T)^+ s_A} \cdot 1\left\{\frac{X_j^T(I - P_A)y}{s - X_j^T(X_A^T)^+ s_A} < \lambda_k\right\}, \tag{15}$$

where recall $P_A = X_A(X_A^T X_A)^{-1} X_A^T$, and $(X_A^T)^+ = X_A(X_A^T X_A)^{-1}$; and

$$\lambda_{k+1}^{\mathrm{leave}} = \max_{j \in A} \frac{[(X_A)^+ y]_j}{[(X_A^T X_A)^{-1} s_A]_j} \cdot 1\left\{\frac{[(X_A)^+ y]_j}{[(X_A^T X_A)^{-1} s_A]_j} < \lambda_k\right\}. \tag{16}$$

As we did in Section 3 with the orthogonal X case, we begin by studying the asymptotic distribution of the covariance statistic in the special case $k = 1$ (i.e., the first model along the path), wherein the expressions for the next knot (14), (15), (16) greatly simplify. Following this, we study the more difficult case $k \geq 1$. For the sake of readability we defer the proofs and most technical details until the appendix.

4.1 The first step, $k = 1$

We assume here that X has unit normed columns: $\|X_i\|_2 = 1$, for $i = 1, \ldots p$; we do this mostly for simplicity of presentation, and the generalization to a matrix X whose columns are not unit normed is given in the next section (though the exponential limit is now a conservative upper bound). As per our discussion following Lemma 1 (see the second point), we know that the first predictor to enter the active set along the lasso path cannot leave at the next step, so the constant sign condition (8) holds, and by Lemma 1 the covariance statistic for testing the entry of the first variable can be written as

$$T_1 = \lambda_1(\lambda_1 - \lambda_2)/\sigma^2$$

(the leading factor C being equal to one since we assumed that X has unit normed columns). 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 (15), which can be expressed as

$$\lambda_1 = \max_{j=1,\ldots p,\, s \in \{-1,1\}} sU_j. \tag{17}$$

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 (15), written as

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

[5] In expressing the joining and leaving times in the forms (15) and (16), we are implicitly assuming that $\lambda_{k+1} < \lambda_k$, with strict inequality. Since X has columns in general position, this is true for (Lebesgue) almost every y, or in other words, with probability one taken over the normally distributed errors in (1).

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

$$\lambda_2 = \max_{j \neq j_1,\, s \in \{-1,1\}} \frac{sU_j - sR_{j,j_1}U_{j_1}}{1 - ss_1 R_{j,j_1}}. \tag{18}$$

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

$$H_0 : \; y \sim N(0, \sigma^2 I). \tag{19}$$

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 (17), (18) for λ_1, λ_2 as discrete-time Gaussian processes. First, we consider the zero mean Gaussian process

$$g(j, s) = sU_j \quad \text{for } j = 1, \ldots p, \; s \in \{-1, 1\}. \tag{20}$$

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 (19). From (17), we know that the first knot is simply

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

In addition to (20), we consider the process

$$h^{(j_1,s_1)}(j, s) = \frac{g(j, s) - ss_1 R_{j,j_1} g(j_1, s_1)}{1 - ss_1 R_{j,j_1}} \quad \text{for } j \neq j_1, \; s \in \{-1, 1\}. \tag{21}$$

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

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

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. Now define

$$M(j_1, s_1) = \max_{j \neq j_1,\, s} \; h^{(j_1,s_1)}(j, s), \tag{22}$$

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 $g(j, s)$ (the argument maximizers being almost surely unique), then from (18) we see that the second knot is $\lambda_2 = M(j_1, s_1)$. Therefore, to study the distribution of $T_1 = \lambda_1(\lambda_1 - \lambda_2)/\sigma^2$, 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

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

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* (20), (21), (22), *we have*

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

This is an important realization because the dual representation $\{g(j_1,s_1) > M(j_1,s_1)\}$ 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

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

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

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

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

$$\begin{aligned}\mathbb{P}(T_1 > t) &= \sum_{j_1,s_1} \mathbb{P}\Big(g(j_1,s_1)/\sigma > u\big(t, M(j_1,s_1)/\sigma\big)\Big) \\ &= \sum_{j_1,s_1} \int_0^\infty \bar{\Phi}\big(u(t,m/\sigma)\big) F_{M(j_1,s_1)}(dm), \end{aligned} \tag{23}$$

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, as well as $M(j_1,s_1) \geq 0$. Continuing from (23), we can write the difference between $\mathbb{P}(T_1 > t)$ and the standard exponential tail, $\mathbb{P}(\mathrm{Exp}(1) > t) = e^{-t}$, as

$$\big|\mathbb{P}(T_1 > t) - e^{-t}\big| = \bigg| \sum_{j_1,s_1} \int_0^\infty \bigg(\frac{\bar{\Phi}\big(u(t,m/\sigma)\big)}{\bar{\Phi}(m/\sigma)} - e^{-t} \bigg) \bar{\Phi}(m/\sigma) F_{M(j_1,s_1)}(dm) \bigg|, \tag{24}$$

where we used the fact that

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

We now examine the term inside the braces in (24), 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 $t \geq 0$,*

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

Hence, loosely speaking, if each $M(j_1,s_1) \to \infty$ fast enough as $p \to \infty$, then the right-hand side in (24) 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* (21), (22) *over $j_1 = 1, \ldots p$ and $s_1 \in \{-1,1\}$. If for any fixed $m_0 > 0$*

$$\sum_{j_1,s_1} \mathbb{P}\big(M(j_1,s_1) \leq m_0\big) \to 0 \quad \text{as } p \to \infty, \tag{25}$$

then the right-hand side in (24) *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 (25) 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 (25) is implied by a conditional variance bound involving the predictor matrix X alone, and arrive at the main result of this section.

Theorem 2. *Assume that $X \in \mathbb{R}^{n\times p}$ has unit normed 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*

$$1 - R_{i,S\setminus\{i\}}(R_{S\setminus\{i\},S\setminus\{i\}})^{-1}R_{S\setminus\{i\},i} \geq \delta^2 \quad \textit{for all } i \in S, \tag{26}$$

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

$$|S| \geq d_p, \quad \textit{where} \quad \frac{d_p}{\log p} \to \infty \quad \textit{as } p \to \infty. \tag{27}$$

The under the null distribution in (19) [i.e., y is drawn from the regression model (1) with $\beta^* = 0$], we have $\mathbb{P}(T_1 > t) \to e^{-t}$ as $p \to \infty$ for all $t \geq 0$.

Remark. Conditions (26) and (27) are sufficient to ensure (25), or in other words, that each $M(j_1, s_1)$ grows as in $\mathbb{P}(M(j_1, s_1) \leq m_0) = o(1/p)$, for any fixed m_0. While it is true that $\mathbb{E}[M(j_1, s_1)]$ will typically grow as p grows, some assumption is required so that $M(j_1, s_1)$ concentrates around its mean faster than standard Gaussian concentration results (such as the Borell-TIS inequality) imply.

Generally speaking, the assumptions (26) and (27) are not very strong. Stated differently, (26) is a lower bound on the variance of $U_i = X_i^T y$, conditional on $U_\ell = X_\ell^T y$ for all $\ell \in S \setminus \{i\}$. Hence for any j, we require the existence of a subset S not containing j such that the variables U_i, $i \in S$ are not too correlated, in the sense that the conditional variance of any one given all the others is bounded below. This subset S has to be larger in size than $\log p$, as made clear in (27). Note that, in fact, it suffices to find a total of two disjoint subsets S_1, S_2 with the properties (26) and (27), because then for any j, either one or the other will not contain j.

An example of a matrix X that does not satisfy (26) and (27) is one with fixed rank as p grows. (This, of course, would also not satisfy the general position assumption.) In this case, we would not be able to find a subset of the variables $U_i = X_i^T y$, $i = 1, \ldots p$ that is both linearly independent and has size larger than $r = \text{rank}(X)$, which violates the conditions. We note that in general, since $|S| \leq \text{rank}(X) \leq n$, and $|S|/\log p \to \infty$, conditions (26) and (27) require that $n/\log p \to \infty$.

4.2 A general step, $k \geq 1$

In this section, we no longer assume that X has unit normed columns (in any case, this provides no simplification in deriving the null distribution of the test statistic at a general step in the lasso path). Our arguments here have more or less the same form as they did in the last section, but overall the calculations are more complicated.

Fix an integer $k_0 \geq 0$, subset $A_0 \subseteq \{1, \ldots p\}$ containing the true active set $A_0 \supseteq A^* = \text{supp}(\beta^*)$, and sign vector $s_{A_0} \in \{-1, 1\}^{|A_0|}$. Consider the event

$$B = \Bigg\{ \text{The solution at step } k_0 \text{ in the lasso path has active set } A = A_0,$$
$$\text{signs } s_A = \text{sign}((X_{A_0})^+ y) = s_{A_0}, \text{ and the next two knots are given by}$$
$$\lambda_{k_0+1} = \max_{j \notin A\cup\{j_{k_0}\},\, s\in\{-1,1\}} \frac{X_j^T(I - P_A)y}{s - X_j^T(X_A^T)^+ s_A}, \quad \lambda_{k_0+2} = \lambda_{k_0+2}^{\text{join}} \Bigg\}. \tag{28}$$

We assume that $\mathbb{P}(B) \to 1$ as $p \to \infty$. In words, this is assuming that with probability approaching one: the lasso estimate at step k_0 in the path has support A_0 and signs s_{A_0}; the least squares estimate on A_0 has the same signs as this lasso estimate; the knots at steps k_0+1 and k_0+2 correspond to joining events; and in particular, the maximization defining the joining event at step k_0+1 can be taken to be unrestricted, i.e., without the indicators constraining the individual arguments to be $< \lambda_{k_0}$. Our goal is to characterize the asymptotic distribution of the covariance statistic T_k at the step $k = k_0 + 1$, under the null hypothesis (i.e., conditional on the event B). We will comment on the stringency of the assumption that $\mathbb{P}(B) \to 1$ following our main result in Theorem 3.

First note that on B, we have $s_A = \mathrm{sign}((X_A)^+ y)$, and as discussed in the third point following Lemma 1, this implies that the solution of the reduced problem (4) on X_A cannot incur any sign changes over the interval $[\lambda_k, \lambda_{k+1}]$. Hence we can apply Lemma 1 to write the covariance statistic on B as

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

where $C(A, s_A, j_k, s_k) = \|(X^T_{A\cup\{j_k\}})^+ s_{A\cup\{j_k\}} - (X_A^T)^+ s_A\|_2^2$, A and s_A are the active set and signs at step $k-1$, and j_k is the variable added to the active set at step k, with sign s_k. Now, analogous to our definition in the last section, we define the discrete-time Gaussian process

$$g^{(A,s_A)}(j,s) = \frac{X_j^T(I - P_A)y}{s - X_j^T(X_A^T)^+ s_A} \quad \text{for } j \notin A,\ s \in \{-1,1\}. \tag{29}$$

For any fixed A, s_A, the above process has mean zero provided that $A \supseteq A^*$. Additionally, for any such fixed A, s_A, we can compute its covariance function

$$\mathbb{E}\Big[g^{(A,s_A)}(j,s) g^{(A,s_A)}(j',s')\Big] = \frac{X_j^T(I-P_A)X_{j'}\sigma^2}{[s - X_j^T(X_A^T)^+ s_A][s' - X_{j'}^T(X_A^T)^+ s_A]}. \tag{30}$$

Note that on the event B, the kth knot in the lasso path is

$$\lambda_k = \max_{j\notin A,\, s\in\{-1,1\}} g^{(A,s_A)}(j,s).$$

For fixed j_k, s_k, we also consider the process

$$g^{(A\cup\{j_k\}, s_{A\cup\{j_k\}})}(j,s) = \frac{X_j^T(I - P_{A\cup\{j_k\}})y}{s - X_j^T(X^T_{A\cup\{j_k\}})^+ s_{A\cup\{j_k\}}} \quad \text{for } j \notin A\cup\{j_k\},\ s \in \{-1,1\} \tag{31}$$

(above, $s_{A\cup\{j_k\}}$ is the concatenation of s_A and s_k) and its achieved maximum value, subject to being less than the maximum of $g^{(A,s_A)}$,

$$M^{(A,s_A)}(j_k,s_k) = \max_{j\notin A\cup\{j_k\}\, s\in\{-1,1\}} g^{(A\cup\{j_k\}, s_{A\cup\{j_k\}})}(j,s) \cdot 1\bigg\{g^{(A\cup\{j_k\}, s_{A\cup\{j_k\}})}(j,s) < \max_{j\notin A,\, s\in\{-1,1\}} g^{(A,s_A)}(j,s)\bigg\}. \tag{32}$$

If j_k, s_k indeed maximize $g^{(A,s_A)}$, i.e., they correspond to the variable added to the active set at λ_k and its sign (note that these are almost surely unique), then on B, we have $\lambda_{k+1} = M^{(A,s_A)}(j_k,s_k)$. To study the distribution of T_k on B, we are therefore interested in the random variable

$$C(A,s_A,j_k,s_k) \cdot g^{(A,s_A)}(j_k,s_k)\Big(g^{(A,s_A)}(j_k,s_k) - M^{(A,s_A)}(j_k,s_k)\Big)/\sigma^2,$$

on the event

$$E(j_k,s_k) = \Big\{g^{(A,s_A)}(j_k,s_k) > g^{(A,s_A)}(j,s) \text{ for all } (j,s) \neq (j_k,s_k)\Big\}. \tag{33}$$

Equivalently, we may write

$$\mathbb{P}(\{T_k > t\} \cap B) = \sum_{j_k, s_k} \mathbb{P}\bigg(\Big\{ C(A, s_A, j_k, s_k) \cdot g^{(A,s_A)}(j_k, s_k) \cdot \Big(g^{(A,s_A)}(j_k, s_k) - M^{(A,s_A)}(j_k, s_k) \Big) / \sigma^2 > t \Big\} \cap E(j_k, s_k) \bigg).$$

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

$$\mathbb{P}(T_k > t) = \sum_{j_k, s_k} \mathbb{P}\bigg(\Big\{ C(A_0, s_{A_0}, j_k, s_k) \cdot g^{(A_0,s_{A_0})}(j_k, s_k) \cdot \Big(g^{(A_0,s_{A_0})}(j_k, s_k) - M^{(A_0,s_{A_0})}(j_k, s_k) \Big) / \sigma^2 > t \Big\} \cap E(j_k, s_k) \bigg) + o(1), \quad (34)$$

where we have replaced all instances of A and s_A on the right-hand side above with the fixed subset A_0 and sign vector s_{A_0}. This is a helpful simplification, because in what follows we may now take $A = A_0$ and $s_A = s_{A_0}$ as fixed, and consider the distribution of the random processes $g^{(A_0,s_{A_0})}$ and $M^{(A_0,s_{A_0})}$. With $A = A_0$ and $s_A = s_{A_0}$ fixed, we drop the notational dependence on them and write these processes as g and M. We also write the scaling factor $C(A_0, s_{A_0}, j_k, s_k)$ as $C(j_k, s_k)$.

The setup in (34) looks very much like the one in the last section [and to draw an even sharper parallel, the scaling factor $C(j_k, s_k)$ is actually equal to one over the variance of $g(j_k, s_k)$, meaning that $\sqrt{C(j_k, s_k)} \cdot g(j_k, s_k)$ is standard normal for fixed j_k, s_k, a fact that we will use later in the proof of Lemma 8]. However, a major complication is that $g(j_k, s_k)$ and $M(j_k, s_k)$ are no longer independent for fixed j_k, s_k. Next, we derive a dual representation for the event (33) (analogous to Lemma 4 in the last section), introducing a triplet of random variables M^+, M^-, M^0—it turns out that g is independent of this triplet, for fixed j_k, s_k.

Lemma 7. *Let g be as defined in (29) (with A, s_A fixed at A_0, s_{A_0}). Let $\Sigma_{j,j'}$ denote the covariance function of g [short form for the expression in (30)].*[6] *Define*

$$S^+(j,s) = \Big\{ (j', s') : j' \notin A \cup \{j\},\ \frac{\Sigma_{j,j'}}{\Sigma_{jj}} < 1 \Big\},\quad M^+(j,s) = \max_{(j',s') \in S^+(j,s)} \frac{g(j', s') - (\Sigma_{j,j'}/\Sigma_{jj}) g(j,s)}{1 - \Sigma_{j,j'}/\Sigma_{jj}}, \quad (35)$$

$$S^-(j,s) = \Big\{ (j', s') : j' \notin A \cup \{j\},\ \frac{\Sigma_{j,j'}}{\Sigma_{jj}} > 1 \Big\},\quad M^-(j,s) = \min_{(j',s') \in S^-(j,s)} \frac{g(j', s') - (\Sigma_{j,j'}/\Sigma_{jj}) g(j,s)}{1 - \Sigma_{j,j'}/\Sigma_{jj}}, \quad (36)$$

$$S^0(j,s) = \Big\{ (j', s') : j' \notin A \cup \{j\},\ \frac{\Sigma_{j,j'}}{\Sigma_{jj}} = 1 \Big\},\quad M^0(j,s) = \max_{(j',s') \in S^0(j,s)} g(j', s') - (\Sigma_{j,j'}/\Sigma_{jj}) g(j,s). \quad (37)$$

Then the event $E(j_k, s_k)$ in (33), that j_k, s_k maximize g, can be written as an intersection of events involving M^+, M^-, M^0:

$$\Big\{ g(j_k, s_k) > g(j,s) \text{ for all } (j,s) \neq (j_k, s_k) \Big\} = \Big\{ g(j_k, s_k) > 0 \Big\} \cap \Big\{ g(j_k, s_k) > M^+(j_k, s_k) \Big\} \cap \Big\{ g(j_k, s_k) < M^-(j_k, s_k) \Big\} \cap \Big\{ 0 > M^0(j_k, s_k) \Big\}. \quad (38)$$

[6] To be perfectly clear, here $\Sigma_{j,j'}$ actually depends on s, s', but our notation suppresses this dependence for brevity.

As a result of Lemma 7, continuing from (34), we can decompose the tail probability of T_k as

$$\mathbb{P}(T_k > t) = \sum_{j_k, s_k} \mathbb{P}\Big(C(j_k, s_k) \cdot g(j_k, s_k)\big(g(j_k, s_k) - M(j_k, s_k)\big)/\sigma^2 > t,\ \ g(j_k, s_k) > 0,$$
$$g(j_k, s_k) > M^+(j_k, s_k),\ \ g(j_k, s_k) < M^-(j_k, s_k),\ \ 0 > M^0(j_k, s_k)\Big) + o(1). \quad (39)$$

A key point here is that, for fixed j_k, s_k, the triplet $M^+(j_k, s_k), M^-(j_k, s_k), M^0(j_k, s_k)$ is independent of $g(j_k, s_k)$, which is true because

$$\mathbb{E}\Big[g(j_k, s_k)\Big(g(j, s) - (\Sigma_{j_k,j}/\Sigma_{j_k,j_k})g(j_k, s_k)\Big)\Big] = 0,$$

and $g(j_k, s_k)$, along with $g(j, s) - (\Sigma_{j_k,j}/\Sigma_{j_k,j_k})g(j_k, s_k)$, for all j, s, form a jointly Gaussian collection of random variables. If we were to now replace M by M^+ in the first line of (39), and define a modified statistic $\tilde{T}_k$ via its tail probability,

$$\mathbb{P}(\tilde{T}_k > t) = \sum_{j_k, s_k} \mathbb{P}\Big(C(j_k, s_k) \cdot g(j_k, s_k)\big(g(j_k, s_k) - M^+(j_k, s_k)\big)/\sigma^2 > t,\ \ g(j_k, s_k) > 0,$$
$$g(j_k, s_k) > M^+(j_k, s_k),\ \ g(j_k, s_k) < M^-(j_k, s_k),\ \ 0 > M^0(j_k, s_k)\Big), \quad (40)$$

then arguments similar to those in the second half of Section 4.1 give a (conservative) exponential limit for $\mathbb{P}(\tilde{T}_k > t)$.

Lemma 8. *Consider g as defined in* (29) *(with A, s_A fixed at A_0, s_{A_0}), and M^+, M^-, M^0 as defined in* (35), (36), (37). *Assume that for any fixed m_0,*

$$\sum_{j_k, s_k} \mathbb{P}\Big(M^+(j_k, s_k) \leq m_0/\sqrt{C(j_k, s_k)}\Big) \to 0 \quad \text{as } p \to \infty, \quad (41)$$

Then the modified statistic $\tilde{T}_k$ in (40) *satisfies $\lim_{p\to\infty} \mathbb{P}(\tilde{T}_k > t) \leq e^{-t}$, for all $t \geq 0$.*

Of course, deriving the limiting distribution of $\tilde{T}_k$ was not the goal, and it remains to relate $\mathbb{P}(\tilde{T}_k > t)$ to $\mathbb{P}(T_k > t)$. A fortuitous calculation shows that the two seemingly different quantities M^+ and M—the former of which is defined as the maximum of particular functionals of g, and the latter concerned with the joining event at step $k+1$—admit a very simple relationship: $M^+(j_k, s_k) \leq M(j_k, s_k)$ for the maximizing j_k, s_k. We use this to bound the tail of T_k.

Lemma 9. *Consider g, M as defined in* (29), (31), (32) *(with A, s_A fixed at A_0, s_{A_0}), and consider M^+ as defined in* (36). *Then for any fixed j_k, s_k, on the event $E(j_k, s_k)$ in* (33), *we have*

$$M^+(j_k, s_k) \leq M(j_k, s_k).$$

Hence if we assume as in Lemma 8 the condition (41), *then $\lim_{p\to\infty} \mathbb{P}(T_k > t) \leq e^{-t}$ for all $t \geq 0$.*

Though Lemma 9 establishes a (conservative) exponential limit for the covariance statistic T_k, it does so by enforcing assumption (41), which is phrased in terms of the tail distribution of a random process defined at the kth step in the lasso path. We translate this into an explicit condition on the covariance structure in (30), to make the stated assumptions for exponential convergence more concrete.

Theorem 3. *Assume that $X \in \mathbb{R}^{n\times p}$ has columns in general position, and $y \in \mathbb{R}^n$ is drawn from the normal regression model* (1). *Assume that for a fixed integer $k_0 \geq 0$, subset $A_0 \subseteq \{1, \ldots p\}$ with*

$A_0 \supseteq A^* = \mathrm{supp}(\beta^*)$, and sign vector $s_{A_0} \in \{-1,1\}^{|A_0|}$, the event B in (28) satisfies $\mathbb{P}(B) \to 1$ as $p \to \infty$. Assume that there exists a constant $0 < \eta \leq 1$ such that

$$\|(X_{A_0})^+ X_j\|_1 \leq 1 - \eta \quad \textit{for all } j \notin A_0. \tag{42}$$

Define the matrix R by

$$R_{ij} = X_i^T (I - P_{A_0}) X_j, \quad \textit{for } i, j \notin A_0.$$

Assume that the diagonal elements in R are all of the same order, i.e., $R_{ii}/R_{jj} \leq C$ for all i, j and some constant $C > 0$. Finally assume that, for each fixed $j \notin A_0$, there is a set $S \subseteq \{1, \ldots p\} \setminus (A_0 \cup \{j\})$ such that for all $i \in S$,

$$\big[R_{ii} - R_{i,S\setminus\{i\}} (R_{S\setminus\{i\},S\setminus\{i\}})^{-1} R_{S\setminus\{i\},i}\big] / R_{ii} \geq \delta^2, \tag{43}$$

$$|R_{ij}|/R_{jj} < \eta/(2-\eta), \tag{44}$$

$$\|(X_{A_0\cup\{j\}})^+ X_i\|_1 < 1, \tag{45}$$

where $\delta > 0$ is a constant (not depending on j), and the size of S grows faster than $\log p$,

$$|S| \geq d_p, \quad \textit{where} \quad \frac{d_p}{\log p} \to \infty \quad \textit{as } p \to \infty. \tag{46}$$

Then at step $k = k_0 + 1$, we have $\lim_{p\to\infty} \mathbb{P}(T_k > t) \leq e^{-t}$ for all $t \geq 0$. The same result holds for the tail of T_k conditional on B.

Remark 1. If X has unit normed columns, then by taking $k_0 = 0$ (and accordingly, $A_0 = \emptyset$, $s_{A_0} = \emptyset$) in Theorem 3, we essentially recover the result of Theorem 2. To see this, note that with $k_0 = 0$ (and $A_0, s_{A_0} = \emptyset$), we have $\mathbb{P}(B) = 1$ for all finite p (recall the arguments given at the beginning of Section 4.1). Also, condition (42) trivially holds with $\eta = 1$ because $A_0 = \emptyset$. Next, the matrix R defined in the theorem reduces to $R = X^T X$, again because $A_0 = \emptyset$; note that R has all diagonal elements equal to one, because X has unit normed columns. Hence (43) is the same as condition (26) in Theorem 2. Finally, conditions (44) and (45) both reduce to $|R_{ij}| < 1$, which always holds as X has columns in general position. Therefore, when $k_0 = 0$, Theorem 3 imposes the same conditions as Theorem 2, and gives essentially the same result—we say "essentially" here is because the former gives a conservative exponential limit for T_1, while the latter gives an exact exponential limit.

Remark 2. If X is orthogonal, then for any A_0, conditions (42) and (43)–(46) are trivially satisfied [for the latter set of conditions, we can take, e.g., $S = \{1, \ldots p\} \setminus (A_0 \cup \{j\})$]. With an additional condition on the strength of the true nonzero coefficients, we can assure that $\mathbb{P}(B) \to 1$ as $p \to \infty$ with $A_0 = A^*$, $s_{A_0} = \mathrm{sign}(\beta^*_{A_0})$, and $k_0 = |A_0|$, and hence prove a conservative exponential limit for T_k; note that this is precisely what is done in Theorem 1 (except that in this case, the exponential limit is proven to be exact).

Remark 3. Defining $U_i = X_i^T (I - P_{A_0}) y$ for $i \notin A_0$, the condition (43) is a lower bound on ratio of the conditional variance of U_i given U_ℓ, $\ell \notin S$, to the unconditional variance of U_i. Loosely speaking, conditions (43), (44), and (45) can all be interpreted as requiring, for any $j \notin A_0$, the existence of a subset S not containing j (and disjoint from A_0) such that the variables U_i, $i \in S$ are not very correlated. This subset has to be large in size compared to $\log p$, by (46). An implicit consequence of (43)–(46), as argued in the remark following Theorem 2, is that $n/\log p \to \infty$.

Remark 4. Some readers will likely recognize condition (42) as that of *mutual incoherence* or *strong irrepresentability*, commonly used in the lasso literature on exact support recovery [see, e.g., Wainwright (2009), Zhao & Yu (2006)]. This condition, in addition to a lower bound on the magnitudes of the true coefficients, is sufficient for the lasso solution to recover the true active set A^* with probability tending to one, at a carefully chosen value of λ. It is important to point out that we do

not place any requirements on the magnitudes of the true nonzero coefficients; instead, we assume directly that the lasso converges (with probability approaching one) to some fixed model defined by A_0, s_{A_0} at the (k_0)th step in the path. Here A_0 is large enough that it contains the true support, $A_0 \supseteq A^*$, and the signs s_{A_0} are arbitrary—they may or may not match the signs of the true coefficients over A_0. In a setting in which the nonzero coefficients in β^* are well-separated from zero, a condition quite similar to the irrepresentable condition can be used to show that the lasso converges to the model with support $A_0 = A^*$ and signs $s_{A_0} = \mathrm{sign}(\beta^*_{A_0})$, at step $k_0 = |A_0|$ of the path. Our result extends beyond this case, and allows for situations in which the lasso model converges to a possibly larger set of "screened" variables A_0, and fixed signs s_{A_0}.

Remark 5. In fact, one can modify the above arguments to account for the case that A_0 does not contain the entire set A^* of truly nonzero coefficients, but rather, only the "strong" coefficients. While "strong" is rather vague, a more precise way of stating this is to assume that β^* has nonzero coefficients both large and small in magnitude, and with A_0 corresponding to the set of large coefficients, we assume that the (left-out) small coefficients must be small enough that the mean of the process g in (29) (with $A = A_0$ and $s_A = s_{A_0}$) grows much faster than M^+. The details, though not the main ideas, of the arguments would change, and the result would still be a conservative exponential limit for the covariance statistic T_k at step $k = k_0 + 1$. We may pursue this extension in future work.

5 Simulation of the null distribution

We investigate the null distribution of the covariance statistic through simulations, starting with an orthogonal predictor matrix X, and then considering more general forms of X.

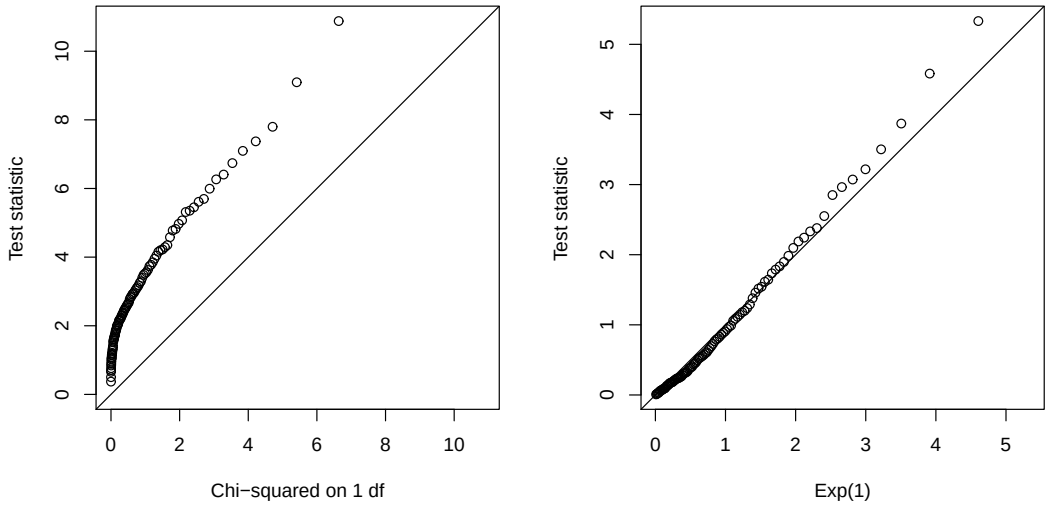
5.1 Orthogonal predictor matrix

Similar to our example from the start of Section 2, we generated $n = 100$ observations with $p = 10$ orthogonal predictors. The true coefficient vector β^* contained 3 nonzero components equal to 6, and the rest zero. The error variance was $\sigma^2 = 1$, so that the truly active predictors had strong effects and always entered the model first, with both forward stepwise and the lasso. Figure 2 shows the results for testing the 4th (truly inactive) predictor to enter, averaged over 500 simulations; the left panel shows the chi-squared test (drop in RSS) applied at the 4th step in forward stepwise regression, and the right panel shows the covariance test applied at the 4th step of the lasso path. We see that the Exp(1) distribution provides a good finite-sample approximation for the distribution of the covariance statistic, while χ^2_1 is a poor approximation for the drop in RSS.

Figure 3 shows the results for testing the 5th, 6th, and 7th predictors to enter the lasso model. An Exp(1)-based test will now be conservative: at a nominal 5% level, the actual type I errors are about 1%, 0.2%, and 0.0%, respectively. The solid line has slope 1, and the broken lines have slopes $1/2, 1/3, 1/4$, as predicted by Theorem 1.

5.2 General predictor matrix

In Table 2, we simulated null data (i.e., $\beta^* = 0$), and examined the distribution of the covariance test statistic T_1 for the first predictor to enter. We varied the numbers of predictors p, correlation parameter ρ, and structure of the predictor correlation matrix. In the first two correlation setups, the correlation between each pair of predictors was ρ, in the data and population, respectively. In the $AR(1)$ setup, the correlation between predictors j and j' is $\rho^{|j-j'|}$. Finally, in the block diagonal setup, the correlation matrix has two equal sized blocks, with population correlation ρ in each block. We computed the mean, variance, and tail probability of the covariance statistic T_1 over



(a) Forward stepwise (b) Lasso

Figure 2: *An example with $n = 100$ and $p = 10$ orthogonal predictors, and the true coefficient vector having 3 nonzero, large components. Shown are quantile-quantile plots for the drop in* RSS *test applied to forward stepwise regression at the 4th step and the covariance test for the lasso path at the 4th step.*

500 simulated data sets for each setup. We see that the Exp(1) distribution is a reasonably good approximation throughout.

In Table 3, the setup was the same as in Table 2, except that we set the first k coefficients of the true coefficient vector equal to 4, and the rest zero, for $k = 1, 2, 3$. The dimensions were also fixed at $n = 100$ and $p = 50$. We computed the mean, variance, and tail probability of the covariance statistic T_{k+1} for entering the next (truly inactive) $(k+1)$st predictor, discarding those simulations in which a truly inactive predictor was selected in the first k steps. (This occurred 1.7%, 4.0%, and 7.0% of the time, respectively.) Again we see that the Exp(1) approximation is reasonably accurate throughout.

In Figure 4, we estimate the power curves for significance testing via the drop in RSS test for forward stepwise regression, and the covariance test for the lasso. In the former we use simulation-derived cutpoints, and in the latter we use the theoretically-based Exp(1) cutpoints, to control the type I error at the 5% level. We find that the tests have similar power, though the cutpoints for forward stepwise would not be typically available in practice. For more details see the figure caption.

6 The case of unknown σ^2

Up until now we have assumed that the error variance σ^2 is known; in practice it will typically be unknown. In this case, provided that $n > p$, we can easily estimate it and proceed by analogy to standard linear model theory. In particular, we can estimate σ^2 by the mean squared residual error $\hat{\sigma}^2 = \|y - X\hat{\beta}^{\mathrm{LS}}\|_2^2/(n-p)$, with $\hat{\beta}^{\mathrm{LS}}$ being the regression coefficients from y on X (i.e., the full model). Plugging this estimate into the covariance statistic in (5) yields a new statistic F_k, that has an asymptotic F-distribution under the null:

$$F_k = \frac{\big\langle y, X\hat{\beta}(\lambda_{k+1})\big\rangle - \big\langle y, X_A\tilde{\beta}_A(\lambda_{k+1})\big\rangle}{\hat{\sigma}^2} \xrightarrow{d} F_{2,n-p}. \tag{47}$$

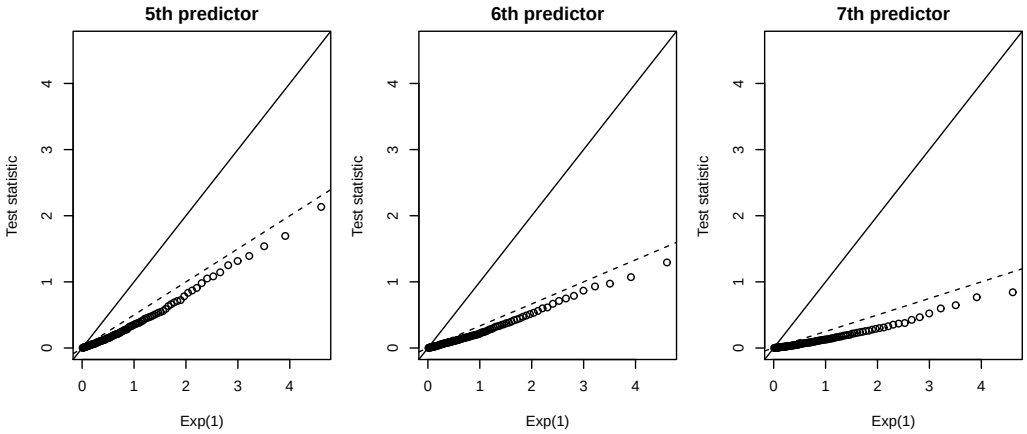


Figure 3: *The same setup as in Figure 2, but here we show the covariance test at the 5th, 6th, and 7th steps along the lasso path, from left to right, respectively. The solid line has slope 1, while the broken lines have slopes* $1/2, 1/3, 1/4$*, as predicted by Theorem 1.*

This follows because $F_k = T_k/(\hat{\sigma}^2/\sigma^2)$, the numerator T_k being asymptotically $\text{Exp}(1) = \chi^2_2/2$, the denominator $\hat{\sigma}^2/\sigma^2$ being asymptotically $\chi^2_{n-p}/(n-p)$, and we claim that the two are independent. Why? Note that the lasso solution path is unchanged if we replace y by $P_X y$, so the lasso fitted values in T_k are functions of $P_X y$; meanwhile, $\hat{\sigma}^2$ is a function of $(I - P_X)y$. The quantities $P_X y$ and $(I - P_X)y$ are uncorrelated and hence independent (recalling normality of y), so T_k and $\hat{\sigma}^2$ are functions of independent quantities, and therefore independent.

As an example, consider one of the setups from Table 2, with $n = 100$, $p = 80$, and predictor correlation of the $AR(1)$ form $\rho^{|j-j'|}$. The true model is null, and we test the first predictor to enter along the lasso path. (We choose n, p of roughly equal sizes here to expose the differences between the σ^2 known and unknown cases.) Table 4 shows the results of 1000 simulations from each of the $\rho = 0$ and $\rho = 0.8$ scenarios. We see that with σ^2 estimated, the $F_{2,n-p}$ distribution provides a more accurate finite-sample approximation than does Exp(1).

When $p \geq n$, estimation of σ^2 is not nearly as straightforward; one idea is to estimate σ^2 from the least squares fit on the support of the model selected by cross-validation. One would then hope that the resulting statistic, with this plug-in estimate of σ^2, is close in distribution to $F_{2,n-r}$ under the null, where r is the size of the model chosen by cross-validation. This is by analogy to the low-dimensional $n > p$ case in (47), but is not supported by rigorous theory. Simulations (withheld for brevity) show that this approximation is not too far off, but that the variance of the observed statistic is sometimes inflated compared that of an $F_{2,n-r}$ distribution (this unaccounted variability is likely due to the model selection process via cross-validation). Other authors have argued that using cross-validation to estimate σ^2 when $p \gg n$ is not necessarily a good approach, as it can be anti-conservative; see, e.g., Fan et al. (2012), Sun & Zhang (2012) for alternative techniques. In future work, we will address the important issue of estimating σ^2 in the context of the covariance statistic, when $p \geq n$.

7 Real data examples

We demonstrate the use of covariance test with some real data examples. As mentioned previously, in any serious application of significance testing over many variables (many steps of the lasso path), we would need to consider the issue of multiple comparisons, which we do not here. This is a topic

$n = 100$, $p = 10$												
ρ	Equal data corr			Equal pop'n corr			$AR(1)$			Block diagonal		
	Mean	Var	Tail pr	Mean	Var	Tail pr	Mean	Var	Tail pr	Mean	Var	Tail pr
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

Table 2: *Simulation results for the first predictor to enter for a global null true model. We vary the number of predictors p, correlation parameter ρ, and structure of the predictor correlation matrix. Shown are the mean, variance, and tail probability $\mathbb{P}(T_1 > q_{.95})$ of the covariance statistic T_1, where $q_{.95}$ is the* 95% *quantile of the* Exp(1) *distribution, computed over 500 simulated data sets for each setup. Standard errors are given by "se". (The panel in the bottom left corner is missing because the equal data correlation setup is not defined for $p > n$.)*

for future work.

7.1 Wine data

Table 5 shows the results for the wine quality data taken from the UCI database. There are $p = 11$ predictors, and $n = 1599$ observations, which we split randomly into approximately equal-sized 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 (with the chi-squared test) and the lasso (with the covariance test). Forward stepwise enters 6 predictors at the 0.05 level, while the lasso enters only 3.

In the left panel of Figure 5, we repeated this p-value computation over 500 random splits into training test sets. The right panel shows the corresponding test set prediction error for the models of each size. The lasso test error decreases sharply once the 3rd predictor is added, but then somewhat flattens out from the 4th predictor onwards; this is in general qualitative agreement with the lasso p-values in the left panel, the first 3 being very small, and the 4th p-value being about 0.2. This also echoes the well-known difference between hypothesis testing and minimizing prediction error. For example, the C_p statistic stops entering variables when the p-value is larger than about 0.16.

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 the (log) susceptibility of the drugs, a measure of drug resistance, based on the location of mutations. We focused on the first drug (3TC), for which there are $p = 217$

$k=1$ and 2nd predictor to enter												
ρ	Equal data corr			Equal pop'n corr			$AR(1)$			Block diagonal		
	Mean	Var	Tail pr	Mean	Var	Tail pr	Mean	Var	Tail pr	Mean	Var	Tail pr
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
$k=2$ and 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
$k=3$ and 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

Table 3: *Simulation results for the $(k+1)$st predictor to enter for a model with k truly nonzero coefficients, across $k = 1, 2, 3$. The rest of the setup is the same as in Table 2 except that the dimensions were fixed at $n = 100$ and $p = 50$. The values are conditional on the event that the k truly active variables enter in the first k steps.*

sites and $n = 1057$ samples. To examine the behavior of the covariance test in the $p > n$ setting, we divided the data at random into training and test sets of size 150 and 907, respectively, a total of 50 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 chi-squared test for forward stepwise), and this is confirmed by the test error plot in the right panel.

8 Extensions

We discuss some extensions of the covariance statistic, beyond significance testing for the lasso. The proposals here are supported by simulations [in terms of having an Exp(1) null distribution], but we do offer any theory. This may be a direction for future work.

8.1 The elastic net

The elastic net estimate (Zou & Hastie 2005) is defined as

$$\hat{\beta}^{\text{en}} = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \; \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1 + \frac{\gamma}{2}\|\beta\|_2^2, \tag{48}$$

where $\gamma \geq 0$ is a second tuning parameter. It is not hard to see that this can actually be cast as a lasso estimate with predictor matrix $\tilde{X} = \begin{bmatrix} X \\ \sqrt{\gamma} I \end{bmatrix} \in \mathbb{R}^{(n+p)\times p}$, and outcome $\tilde{y} = (y, 0) \in \mathbb{R}^{n+p}$. This shows that, for a fixed γ, the elastic net solution path is piecewise linear over λ, with each knot marking the entry (or deletion) of a variable from the active set. We therefore define the covariance

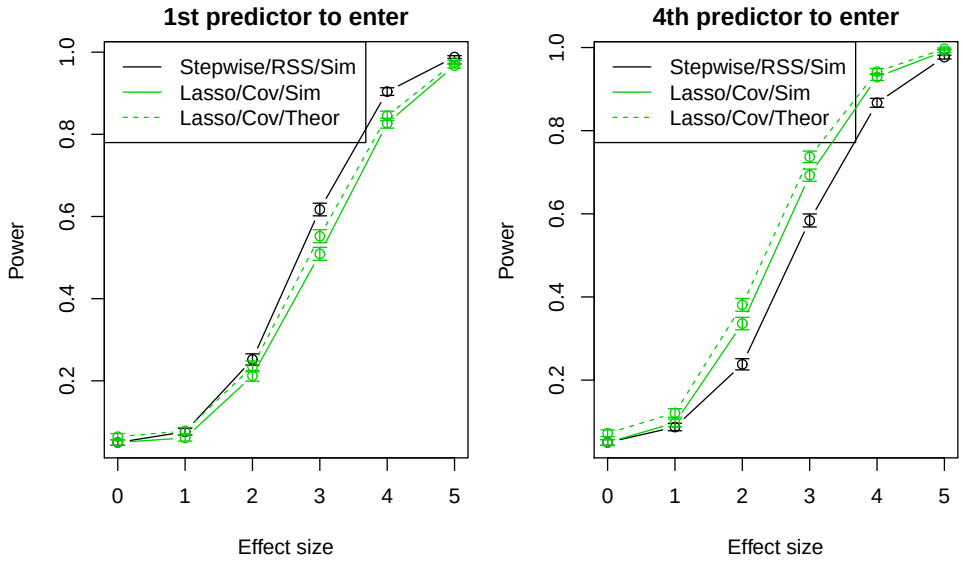


Figure 4: *Estimated power curves for significance tests using forward stepwise regression and the drop in RSS statistic, as well as the lasso and the covariance statistic. The results are averaged over 1000 simulations with* $n = 100$ *and* $p = 10$ *predictors drawn i.i.d. from* $N(0, 1)$*, and* $\sigma^2 = 1$*. On the left, there is one truly nonzero regression coefficient, and we varied its magnitude (the effect size parameter on the x-axis). We examined the first step of the forward stepwise and lasso procedures. On the right, in addition to a nonzero coefficient with varying effect size (on the x-axis), there are 3 additional large coefficients in the true model. We examined the 4th step in forward stepwise and the lasso, after the 3 strong variables have been entered. For the power curves in both panels, we use simulation-based cutpoints for forward stepwise to control the type I error at the 5% level; for the lasso we do the same, but also display the results for the theoretically-based [*Exp(1)*] cutpoint. Note that in practice, simulation-based cutpoints would not typically to be available.*

statistic in the same manner as we did for the lasso; fixing γ, to test the predictor entering at the kth step (knot λ_k) in the elastic net path, we consider the statistic

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

where as before, λ_{k+1} is next knot in the path, A is the active set of predictors just before λ_k, and $\tilde{\beta}_A^{\text{en}}$ is the elastic net solution using only the predictors X_A. The precise expression for the elastic net solution in (48), for a given active set and signs, is the same as it is for the lasso (see Section 2.3), but with $(X_A^T X_A)^{-1}$ replaced by $(X_A^T X_A + \gamma I)^{-1}$. This generally creates a complication for the theory in Sections 3 and 4. But in the orthogonal X case, we have $(X_A^T X_A + \gamma I)^{-1} = I/(1+\gamma)$ and so

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

with $U_j = X_j^T y$, $j = 1, \ldots p$. This means that for an orthogonal X, under the null,

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

and one is tempted to use this approximation beyond the orthogonal setting as well. In Figure 7, we evaluated the distribution of $(1+\gamma)T_1$ (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 be reasonably close to Exp(1) in all cases.

$\rho = 0$				
	Mean	Variance	95% quantile	Tail prob
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

Table 4: *Comparison of* Exp(1), $F_{2,N-p}$, *and the observed (empirical) null distribution of the covariance statistic, when* σ^2 *has been estimated. We examined 1000 simulated data sets with* $n = 100$, $p = 80$, *and the correlation between predictors* j *and* j' *equal to* $\rho^{|j-j'|}$. *We are testing the first step of the lasso path, and the true model is the global null. Results are shown for* $\rho = 0.0$ *and* 0.8. *The third column shows the tail probability* $\mathbb{P}(T_1 > q_{0.05})$ *computed over the 1000 simulations, where* $q_{.95}$ *is the* 95% *quantile from the appropriate distribution (either* Exp(1) *or* $F_{2,n-p}$*).*

Forward stepwise				Lasso			
Step	Predictor	RSS test	p-value	Step	Predictor	Cov test	p-value
1	alcohol	315.216	0.000	1	alcohol	79.388	0.000
2	volatile_acidity	137.412	0.000	2	volatile_acidity	77.956	0.000
3	sulphates	18.571	0.000	3	sulphates	10.085	0.000
4	chlorides	10.607	0.001	4	chlorides	1.757	0.173
5	pH	4.400	0.036	5	total_sulfur_dioxide	0.622	0.537
6	total_sulfur_dioxide	3.392	0.066	6	pH	2.590	0.076
7	residual_sugar	0.607	0.436	7	residual_sugar	0.318	0.728
8	citric_acid	0.878	0.349	8	citric_acid	0.516	0.597
9	density	0.288	0.592	9	density	0.184	0.832
10	fixed_acidity	0.116	0.733	10	free_sulfur_dioxide	0.000	1.000
11	free_sulfur_dioxide	0.000	0.997	11	fixed_acidity	0.114	0.892

Table 5: *Wine data: forward stepwise and lasso p-values. The values are rounded to 3 decimal places. For the lasso, we only show p-values for the steps in which a predictor entered the model and stayed in the model for the remainder of the path (i.e., if a predictor entered the model at a step but then later left, we do not show this step—we only show the step corresponding to its last entry point).*

8.2 Generalized linear models and the Cox model

Consider the estimate from an ℓ_1-penalized generalized linear model:

$$\hat{\beta}^{\text{glm}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} - \sum_{i=1}^{n} \log f(y_i; x_i, \beta) + \lambda \|\beta\|_1, \tag{49}$$

where $f(y_i; x_i, \beta)$ is an exponential family density, a function of the predictor measurements $x_i \in \mathbb{R}^p$ and parameter $\beta \in \mathbb{R}^p$. Note that the usual lasso estimate in (2) is a special case of this form when f is the Gaussian density with known variance σ^2. The natural parameter in (49) is $\eta_i = x_i^T \beta$, for $i = 1, \ldots n$, related to the mean of y_i via a link function $g(\mathbb{E}[y_i|x_i]) = \eta_i$.

Having solved (49) with $\lambda = 0$ (i.e., this is simply maximum likelihood), producing a vector of

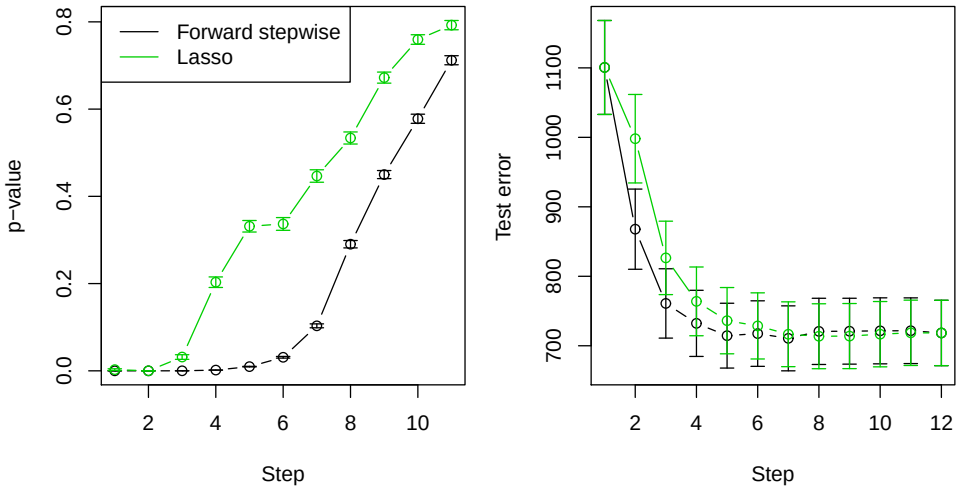


Figure 5: *Wine data: the data were randomly divided 500 times into roughly equal-sized training and test sets. The left panel shows the training set p-values for forward stepwise regression and the lasso. The right panel show the test set error for the corresponding models of each size.*

fitted values $\hat{\eta} = X\hat{\beta}^{\text{glm}} \in \mathbb{R}^n$, we might define degrees of freedom as[7]

$$\text{df}(\hat{\eta}) = \sum_{i=1}^{n} \text{Cov}(y_i, \hat{\eta}_i). \tag{50}$$

This is the implicit concept used by Efron (1986) in his definition of the "optimism" of the training error. The same idea could be used to define degrees of freedom for the penalized estimate in (49) for any $\lambda > 0$, and this motivates the definition of the covariance statistic, as follows. If the tuning parameter value $\lambda = \lambda_k$ marks the entry of a new predictor into the active set A, then we define the covariance statistic

$$T_k = \big\langle y, X\hat{\beta}^{\text{glm}}(\lambda_{k+1}) \big\rangle - \big\langle y, X_A \tilde{\beta}_A^{\text{glm}}(\lambda_{k+1}) \big\rangle, \tag{51}$$

where λ_{k+1} is the next value of the tuning parameter at which the model changes (a variable enters or leaves the active set), and $\tilde{\beta}_A^{\text{glm}}$ is the estimate from the penalized generalized linear model (49) using only predictors in A. Unlike in the Gaussian case, the solution path in (49) is not generally piecewise linear over λ, and there is not an algorithm to deliver the exact the values of λ at which variables enter the model (we still refer to these as knots in the path). However, one can numerically approximate these knot values; e.g., see Park & Hastie (2007). By analogy to the Gaussian case, we would hope that T_k has an asymptotic Exp(1) distribution under the null. Though we have not rigorously investigated this conjecture, simulations seem to support it.

As as example, consider the logistic regression model for binary data. Now $\eta_i = \log(\mu_i/(1-\mu_i))$, with $\mu_i = \mathbb{P}(y_i = 1|x_i)$. Figure 8 shows the simulation results from comparing the null distribution of the covariance test statistic in (51) to Exp(1). Here we used the `glmpath` package in R (Park & Hastie 2007) to compute an approximate solution path and locations of knots. The null distribution of the test statistic looks fairly close to Exp(1).

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

[7]Note that in the Gaussian case, this definition is actually σ^2 times the usual notion of degrees of freedom; hence in the presence of a scale parameter, we would divide the right-hand side in the definition (50) by this scale parameter, and we would do the same for the covariance statistic as defined in (50).

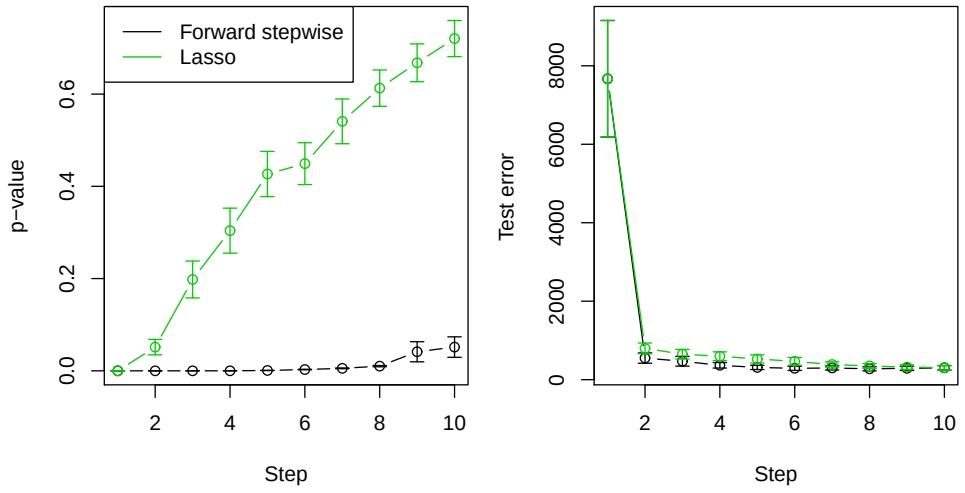


Figure 6: *HIV data: the data were randomly divided 50 times into training and test sets of size 150 and 907, respectively. The left panel shows the training set p-values for forward stepwise regression and the lasso. The right panel shows the test set error for the corresponding models of each size.*

the outcome variable

$$z(\eta) = \eta + I_\eta^{-1} S_\eta \tag{52}$$

where $S_\eta = \nabla \ell(\eta)$, and $I_\eta = \nabla^2 \ell(\eta)$. This applies, e.g., to the class of generalized linear models and Cox's proportional hazards model. For the general ℓ_1-penalized estimator

$$\hat{\beta}^{\text{lik}} = \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \; -\ell(X\beta) + \lambda \|\beta\|_1, \tag{53}$$

we can analogously define the covariance test statistic at a knot λ_k, marking the entry of a predictor into the active set A, as

$$T_k = \Big(\big\langle I_0^{-1/2} S_0, X \hat{\beta}^{\text{lik}}(\lambda_{k+1}) \big\rangle - \big\langle I_0^{-1/2} S_0, X_A \tilde{\beta}_A^{\text{lik}}(\lambda_{k+1}) \big\rangle \Big) / 2, \tag{54}$$

with λ_{k+1} being the next knot in the path (at which a variable is added or deleted from the active set), and $\tilde{\beta}_A^{\text{lik}}$ the solution of the general penalized likelihood problem (53) with predictor matrix X_A. For the binomial model, the statistic (54) reduces to expression (51). In Figure 9, we computed this statistic for Cox's proportional hazards model, using a similar setup to that in Figure 8. The Exp(1) approximation for its null distribution looks reasonably accurate.

9 Discussion

We proposed a simple *covariance statistic* for testing the significance of predictor variables as they enter the active set, along the lasso solution path. We showed that the distribution of this statistic is asymptotically Exp(1), under the null hypothesis that all truly active predictors are contained in the current active set. (See Theorems 1, 2, and 3; the conditions required for this convergence result vary depending on the step k along the path that we are considering, and the covariance structure of the predictor matrix X; the Exp(1) limiting distribution is in some cases a conservative upper bound under the null.) Such a result accounts for the adaptive nature of the lasso procedure, which

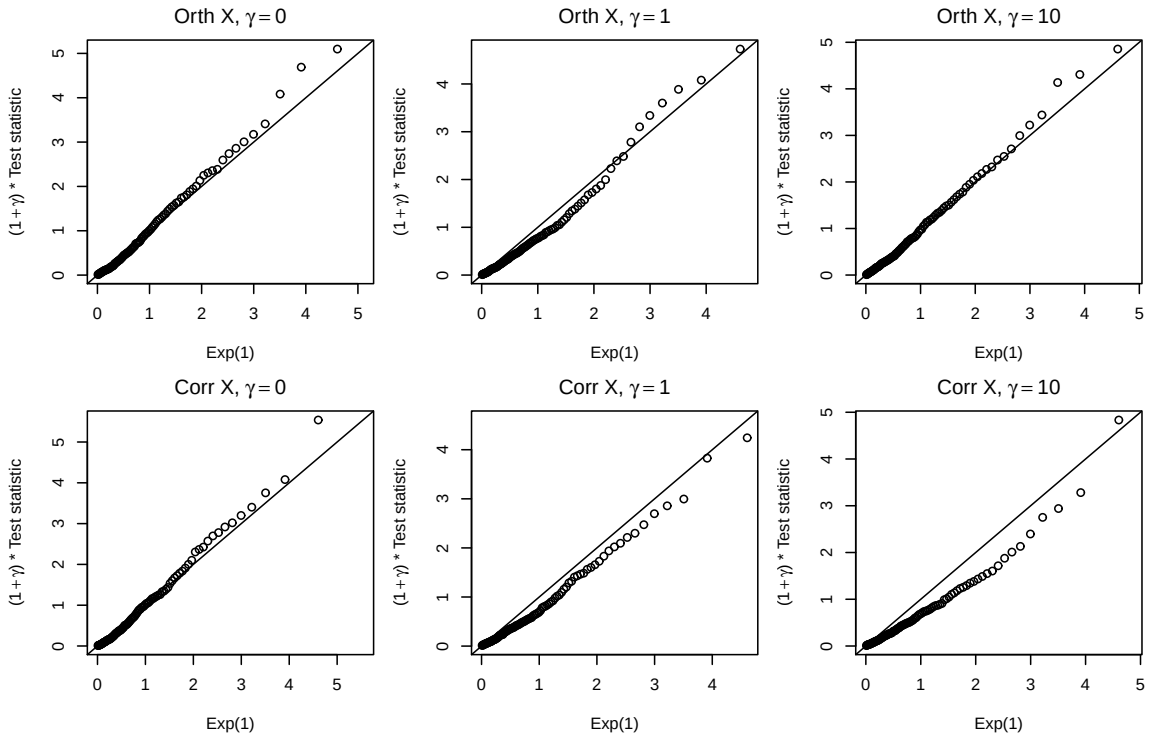


Figure 7: *Elastic net: an example with* $n = 100$ *and* $p = 10$*, for orthogonal and correlated predictors (having pairwise population correlation 0.5), and three different values of the ridge penalty parameter* γ*.*

is not true for the usual chi-squared test (or F-test) applied to, e.g., forward stepwise regression. An R package `covTest` for computing the covariance test will be made freely available on the CRAN repository.

We feel that our work has shed light not only on the lasso path (as given by LARS), but also, at a high level, on forward stepwise regression. Both the lasso and forward stepwise start by entering the predictor variable most correlated with the outcome (thinking of standardized predictors), but the two differ in what they do next. Forward stepwise is greedy, and once it enters this first variable, it proceeds to fit the first coefficient fully, ignoring the effects of other predictors. The lasso, on the other hand, increases (or decreases) the coefficient of the first variable only as long as its correlation with the residual is larger than that of the inactive predictors. Subsequent steps follow similarly. Intuitively, it seems that forward stepwise regression inflates coefficients unfairly, while the lasso takes more appropriately sized steps. This intuition is confirmed in one sense by looking at degrees of freedom (recall Section 2.4). The covariance test and its simple asymptotic null distribution reveal another way in which the step sizes used by the lasso are "just right".

The problem of assessing significance in an adaptive linear model fit by the lasso is a difficult one, and what we have presented in this paper by no means a complete solution. We describe some current work and ideas for future projects below.

- *Significance test for generic lasso models.* A natural direction to consider is the generic lasso testing problem: given a lasso model computed at some fixed value of λ, how do we carry out a significance test for each predictor in the active set? Work on this is in progress.

- *Nonasymptotic null distributions.* A geometric characterization of the first knot in the lasso

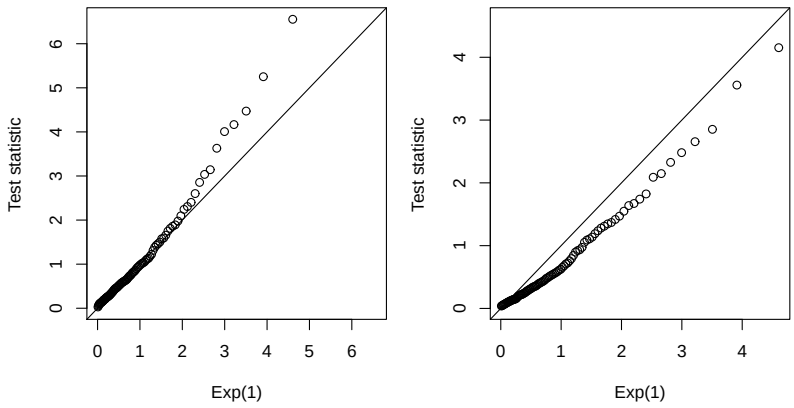


Figure 8: *Lasso logistic regression: an example with* $n = 100$ *and* $p = 10$ *predictors, i.i.d. from* $N(0, 1)$. *In the left panel, all true coefficients are zero; on the right, the first coefficient is large, and the rest are zero. Shown are quantile-quantile plots of the covariance tet statistic (at the first and second steps, respectively), generated over 500 data sets, versus its conjectured asymptotic distribution,* Exp(1).

path provides an alternative test for the global null hypothesis, $\beta^* = 0$. When all predictors have unit norm, $\|X_i\|_2 = 1$, for $i = 1, \ldots p$, this test has the form

$$\frac{1 - \Phi(\lambda_1/\sigma)}{1 - \Phi(\lambda_2/\sigma)} \sim \text{Unif}(0, 1).$$

Remarkably, this above result is exact (nonasymptotic), valid for any n and p, requiring (essentially) only Gaussianity of the errors, and no real assumptions about the matrix X. For most reasonably behaved predictor matrices X, the Exp(1) approximation agrees closely with this test. Details are in Taylor et al. (2013). Work to extend this formula to subsequent steps along the solution path, i.e., to test a hypothesis beyond the global null, is underway.

- *Generalizations to other penalties and models.* The manuscript of Taylor et al. (2013) applies to a regularized regression setting with a general seminorm penalty, and derives explicit results for the group lasso and nuclear norm penalties (in addition to the lasso penalty). The nuclear norm result yields a test for principal components and matrix completion. The recent work of Grazier G'Sell, Taylor & Tibshirani (2013) studies the covariance test for graphical models, based on a sparse estimate of the inverse covariance matrix.

- *Sequential procedures with false discovery rate control.* It is also interesting to consider how the sequence of covariance test p-values can be used to construct a sequential test with good power properties, and a guaranteed bound on its false discovery rate. A number of such approaches are proposed in Grazier G'Sell, Wager, Chouldechova & Tibshirani (2013).

- *Proper p-values for forward stepwise.* Perhaps surprisingly, a test analogous to the covariance test can be used in forward stepwise regression, to construct valid p-values for this greedy procedure. This work is in progress.

- Other related problems include: estimation of σ^2 when $p \geq n$, in the context of the covariance test; power calculations and confidence interval estimation; theory for linear models having strong and weak signals (large and small true coefficients); theory for the elastic net, generalized linear models, and the Cox model.

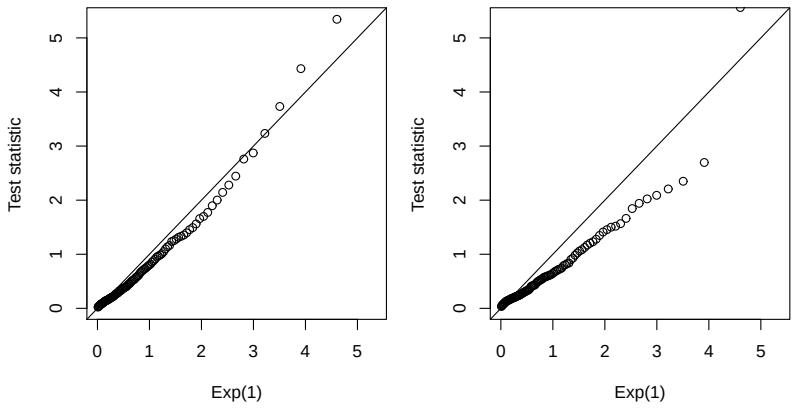


Figure 9: *Lasso Cox model estimate: the basic setup is the same as in Figure 8 (n, p, the distribution of the predictors X, the true coefficient vector—on the left, entirely zero, and on the right, one large coefficient). Shown are quantile-quantile plots of the covariance test statistic (at the first and second steps, respectively), generated over 500 data sets, versus the* Exp(1) *distribution.*

As is clear from the above discussion, the covariance test work has created much excitement and activity among our close collaborators and students. It is our hope that the current paper will also broadly stimulate other researchers' interest in this area, and that at some point, the joint efforts of the community will yield a full set of inferential tools for the lasso and other commonly used adaptive procedures.

Acknowledgements

We thank Jacob Bien, Trevor Hastie, Fred Huffer, and Larry Wasserman for helpful comments. Richard Lockhart was supported by the Natural Sciences and Engineering Research Council of Canada; Jonathan Taylor was supported by NSF grant DMS 1208857 and AFOSR grant 113039; Ryan Tibshirani was supported by NSF grant DMS-1309174; 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

$$(P_{A\cup\{j\}} - P_A) y = \lambda_k \Big((X_{A\cup\{j\}}^T)^+ s_{A\cup\{j\}} - (X_A^T)^+ s_A \Big). \tag{55}$$

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

$$y^T (P_{A\cup\{j\}} - P_A) y = \lambda_k^2 \cdot \| (X_{A\cup\{j\}}^T)^+ s_{A\cup\{j\}} - (X_A^T)^+ s_A \|_2^2, \tag{56}$$

obtained by squaring both sides in (55) (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

$$y^T \Big((X_{A\cup\{j\}}^T)^+ s_{A\cup\{j\}} - (X_A^T)^+ s_A \Big) = \lambda_k \cdot \| (X_{A\cup\{j\}}^T)^+ s_{A\cup\{j\}} - (X_A^T)^+ s_A \|_2^2, \tag{57}$$

obtained by taking the inner product of both sides in (55) with y, and then using (56). Plugging (56) and (57) in for the first and second terms in (7), respectively, then gives the result in (9). $\square$

A.2 Proof of Lemma 4

Note that

$$\begin{aligned} g(j_1,s_1) > g(j,s) \quad &\Leftrightarrow \quad \frac{g(j_1,s_1) - ss_1 R_{j,j_1} g(j_1,s_1)}{1 - ss_1 R_{j,j_1}} > \frac{g(j,s) - ss_1 R_{j,j_1} g(j_1,s_1)}{1 - ss_1 R_{j,j_1}} \\ &\Leftrightarrow \quad g(j_1,s_1) > h^{(j_1,s_1)}(j,s), \end{aligned}$$

the first step following since $1 - ss_1 R_{j,j_1} > 0$, and the second step following from the definition of $h^{(j_1,s_1)}$. The intersection of the right-hand side above, over all $(j,s) \neq (j_1,s_1)$, is equivalent to

$$g(j_1,s_1) > g(j_1,-s_1), \;\; g(j_1,s_1) > M(j_1,s_1).$$

But the former inequality is the same as $g(j_1,s_1) > 0$, because $g(j_1,s_1)$ and $g(j_1,-s_1)$ have opposite signs. Further, the inequality $g(j_1,s_1) > 0$ is redundant, as $M(j_1,s_1) \geq 0$. This gives the result. $\square$

A.3 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

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

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. $\square$

A.4 Proof of Lemma 6

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

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

Starting from (24),

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

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. $\square$

A.5 Proof of Theorem 2

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

$$\mathbb{P}\big(M(j_1, s_1) \le m_0\big) \le c^{|S|}, \tag{58}$$

where $S \subseteq \{1, \dots p\} \setminus \{j_1\}$ is as in the theorem for $j = j_1$, with size $|S| \ge d_p$, and $c < 1$ is a constant (not depending on j_1). This would imply that

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

where we used the fact that $d_p / \log p \to \infty$ by (27). The above sum tending to zero now implies the desired convergence result by Lemma 6, and hence it suffices to show (58). To this end, consider

$$\begin{aligned} M(j_1, s_1) &= \max_{j \neq j_1,\, s} \frac{sU_j - sR_{j,j_1}U_{j_1}}{1 - ss_1R_{j,j_1}} \\ &\ge \max_{j \neq j_1} \frac{|U_j - R_{j,j_1}U_{j_1}|}{1 + |R_{j,j_1}|} \\ &\ge \max_{j \in S} \frac{|U_j - R_{j,j_1}U_{j_1}|}{2}, \end{aligned}$$

where in both inequalities above we used the fact that $|R_{j,j_1}| < 1$. We can therefore use the bound

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

where we define $V_j = (U_j - R_{j,j_1}U_{j_1})/2$ for $j \in S$. Let $r = |S|$, and without a loss of generality, let $S = \{1, \dots r\}$. We will show that

$$\mathbb{P}(|V_1| \le m_0, \dots |V_r| \le m_0) \le c^r, \tag{59}$$

for $c = \Phi(2m_0/(\sigma\delta)) - \Phi(-2m_0/(\sigma\delta)) < 1$, by induction; this would complete the proof, as it would imply (58). Before presenting this argument, we note a few important facts. First, the condition in (26) is really a statement about conditional variances:

$$\mathrm{Var}(U_i \,|\, U_\ell,\, \ell \in S \setminus \{i\}) = \sigma^2 \cdot \Big[1 - R_{i,S\setminus\{i\}}(R_{S\setminus\{i\},S\setminus\{i\}})^{-1}R_{S\setminus\{i\},i}\Big] \ge \sigma^2\delta^2 \quad \text{for all } i \in S,$$

where recall that $U_j = X_j^T y$, $j = 1, \dots p$. Second, since $U_1, \dots U_r$ are jointly normal, we have

$$\mathrm{Var}(U_i \,|\, U_\ell,\, \ell \in S') \ge \mathrm{Var}(U_i \,|\, U_\ell,\, \ell \in S \setminus \{i\}) \ge \sigma^2\delta^2 \quad \text{for any } S' \subseteq S \setminus \{i\}, \text{ and } i \in S, \tag{60}$$

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

Now we give the inductive argument for (59). For the base case, note that $V_1 \sim N(0, \tau_1^2)$, where its variance is

$$\tau_1^2 = \mathrm{Var}(V_1) = \mathrm{Var}(V_1 | U_{j_1}) = \mathrm{Var}(U_1)/4 \ge \sigma^2\delta^2/4,$$

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 (60). Hence

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

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

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

We have, using the independence of $V_1, \dots V_{q+1}$ and U_{j_1},

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

where the variance is

$$\tau^2_{q+1} = \mathrm{Var}(V_{q+1} \,|\, U_1, \dots U_q, U_{j_1}) = \mathrm{Var}(U_{q+1} \,|\, U_1, \dots U_q)/4 \geq \sigma^2\delta^2/4,$$

and here we again used the fact that conditioning further can only reduce the variance, as in (60). Therefore

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

and so

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

completing the inductive step. □

A.6 Proof of Lemma 7

Notice that

$$g(j_k, s_k) > g(j, s) \quad \Leftrightarrow \quad g(j_k, s_k)(1 - \Sigma_{j,j'}/\Sigma_{jj}) > g(j, s) - (\Sigma_{j,j'}/\Sigma_{jj})g(j_k, s_k).$$

We now handle division by $1 - \Sigma_{j,j'}/\Sigma_{jj}$ in three cases:

- if $1 - \Sigma_{j,j'}/\Sigma_{jj} > 0$, then

$$g(j_k, s_k) > g(j, s) \;\; \Leftrightarrow \;\; g(j_k, s_k) > \frac{g(j, s) - (\Sigma_{j,j'}/\Sigma_{jj})g(j_k, s_k)}{1 - \Sigma_{j,j'}/\Sigma_{jj}};$$

- if $1 - \Sigma_{j,j'}/\Sigma_{jj} < 0$, then

$$g(j_k, s_k) > g(j, s) \;\; \Leftrightarrow \;\; g(j_k, s_k) < \frac{g(j, s) - (\Sigma_{j,j'}/\Sigma_{jj})g(j_k, s_k)}{1 - \Sigma_{j,j'}/\Sigma_{jj}};$$

- if $1 - \Sigma_{j,j'}/\Sigma_{jj} = 0$, then

$$g(j_k, s_k) > g(j, s) \;\; \Leftrightarrow \;\; 0 > g(j, s) - (\Sigma_{j,j'}/\Sigma_{jj})g(j_k, s_k).$$

Using this breakdown, we see that the statement $g(j_k, s_k) > g(j, s)$ for all $(j, s) \neq (j_k, s_k)$ is then equivalent to

$$g(j_k, s_k) > g(j_k, -s_k), \;\; g(j_k, s_k) > M^+(j_k, s_k), \;\; g(j_k, s_k) < M^-(j_k, s_k), \;\; 0 > M^0(j_k, s_k).$$

Noting that $g(j_k, s_k)$ and $g(j_k, -s_k)$ must have opposite signs, the above is equivalent to

$$g(j_k, s_k) > 0, \;\; g(j_k, s_k) > M^+(j_k, s_k), \;\; g(j_k, s_k) < M^-(j_k, s_k), \;\; 0 > M^0(j_k, s_k).$$

which gives the result in the lemma. □

A.7 Proof of Lemma 8

Define $\sigma_k = \sigma/\sqrt{C(j_k, s_k)}$ and $u(a,b) = (b+\sqrt{b^2+4a})/2$. Exactly as before (dropping for simplicity the notational dependence of g, M^+ on j_k, s_k),

$$g(g - M^+)/\sigma_k^2 > t,\ \ g > M^+ \quad \Leftrightarrow \quad g/\sigma_k > u(t, M^+/\sigma_k).$$

Therefore we can rewrite (40) as

$$\mathbb{P}(\tilde{T}_k > t) = \sum_{j_k, s_k} \mathbb{P}\Big(g(j_k,s_k)/\sigma_k > u\big(t, M^+(j_k,s_k)/\sigma_k\big),\ \ g(j_k,s_k) < M^-(j_k,s_k),\ \ 0 > M^0(j_k,s_k)\Big).$$

Note that we have dropped the inequality $g(j_k, s_k) > 0$ from each term, as it is implied by the first inequality $g(j_k,s_k)/\sigma_k > u(t, M^+(j_k,s_k)/\sigma_k) \geq 0$. We can upper bound the right-hand side above by replacing $g(j_k,s_k) < M^-(j_k,s_k)$ with

$$g(j_k,s_k) < M^-(j_k,s_k) + u\big(t\sigma_k^2, M^+(j_k,s_k)\big) - M^+(j_k,s_k),$$

because $u(a,b) \geq b$ for all $a \geq 0$ and b. Furthermore, Lemma 10 (Appendix A.10) shows that indeed $\sigma_k^2 = \sigma^2/C(j_k,s_k) = \mathrm{Var}(g(j_k,s_k))$ for fixed j_k, s_k, and hence $g(j_k,s_k)/\sigma_k$ is standard normal for fixed j_k, s_k. Therefore

$$\mathbb{P}(\tilde{T}_k > t) \leq \sum_{j_k,s_k} \int \Big[\Phi\big(m^-/\sigma_k + u(t, m^+/\sigma_k) - m^+/\sigma_k\big) - \Phi\big(u(t,m^+/\sigma_k)\big)\Big] \cdot G_{j_k,s_k}(dm^+, dm^-, dm^0), \quad (61)$$

where

$$G_{j_k,s_k}(dm^+, dm^-, dm^0) = 1\{m^+ < m^-,\ m^0 < 0\} \cdot F_{M^+(j_k,s_k), M^-(j_k,s_k), M^0(j_k,s_k)}(dm^+, dm^-, dm^0),$$

with $F_{M^+(j_k,s_k),M^-(j_k,s_k),M^0(j_k,s_k)}$ the joint distribution of $M^+(j_k,s_k), M^-(j_k,s_k), M^0(j_k,s_k)$, and we used the fact that g is independent of M^+, M^-, M^0 for fixed j_k, s_k. From (61),

$$\mathbb{P}(\tilde{T}_k > t) - e^{-t} \leq \sum_{j_k,s_k} \int \left(\frac{\Phi\big(m^-/\sigma_k + u(t,m^+/\sigma_k) - m^+/\sigma_k\big) - \Phi\big(u(t,m^+/\sigma_k)\big)}{\Phi(m^-/\sigma_k) - \Phi(m^+/\sigma_k)} - e^{-t}\right) \cdot \Big[\Phi(m^-/\sigma_k) - \Phi(m^+/\sigma_k)\Big] \cdot G_{j_k,s_k}(dm^+, dm^-, dm^0), \quad (62)$$

where we here used the fact that

$$\begin{aligned}
&\sum_{j_k,s_k} \int \Big[\Phi(m^-/\sigma_k) - \Phi(m^+/\sigma_k)\Big] G_{j_k,s_k}(dm^+, dm^-, dm^0) \\
&\quad = \sum_{j_k,s_k} \mathbb{P}\Big(g(j_k,s_k) > M^+(j_k,s_k),\ \ g(j_k,s_k) < M^-(j_k,s_k),\ \ 0 > M^0(j_k,s_k)\Big) \\
&\quad \geq \sum_{j_k,s_k} \mathbb{P}\Big(g(j_k,s_k) > 0,\ \ g(j_k,s_k) > M^+(j_k,s_k),\ \ g(j_k,s_k) < M^-(j_k,s_k),\ \ 0 > M^0(j_k,s_k)\Big) \\
&\quad = 1,
\end{aligned}$$

the last equality following by Lemma 7 (i.e., each term in the last sum is exactly the probability of j_k, s_k maximizing g). We show in Lemma 11 (Appendix A.11) that

$$\lim_{m^+ \to \infty} \frac{\Phi\big(m^- + u(t,m^+) - m^+\big) - \Phi\big(u(t,m^+)\big)}{\Phi(m^-) - \Phi(m^+)} \leq e^{-t},$$

provided that $m^- > m^+$. Hence fix $\epsilon > 0$, and choose m_0 sufficiently large, so that for each k,

$$\frac{\Phi\big(m^-/\sigma_k + u(t, m^+/\sigma_k) - m^+/\sigma_k\big) - \Phi\big(u(t, m^+/\sigma_k)\big)}{\Phi(m^-/\sigma_k) - \Phi(m^+/\sigma_k)} - e^{-t} \leq \epsilon,$$
$$\text{for all } m^-/\sigma_k > m^+/\sigma_k \geq m_0.$$

Working from (62),

$$\mathbb{P}(\tilde{T}_k > t) - e^{-t} \leq \epsilon \sum_{j_k, s_k} \int_{m^+/\sigma_k \geq m_0} \Big[\Phi(m^-/\sigma_k) - \Phi(m^+/\sigma_k)\Big] G_{j_k, s_k}(dm^+, dm^-, dm^0) + \sum_{j_k, s_k} \int_{m^+/\sigma_k \leq m_0} G_{j_k, s_k}(dm^+, dm^-, dm^0).$$

Note that the first term on the right-hand side above is $\leq \epsilon$, and the second term is bounded by $\sum_{j_k, s_k} \mathbb{P}(M^+(j_k, s_k) \leq m_0 \sigma_k)$, which by assumption can be made arbitrarily small (smaller than, say, ϵ) by taking p large enough. □

A.8 Proof of Lemma 9

For now, we reintroduce the notational dependence of the process g on A, s_A, as this will be important. We show in Lemma 12 (Appendix A.12) that for any fixed j_k, s_k, j, s,

$$\frac{g^{(A,s_A)}(j,s) - (\Sigma_{j_k,j}/\Sigma_{j_k,j_k}) g^{(A,s_A)}(j_k, s_k)}{1 - \Sigma_{j_k,j}/\Sigma_{j_k,j_k}} = g^{(A \cup \{j_k\}, s_{A \cup \{j_k\}})}(j,s),$$

where $\Sigma_{j_k,j} = \mathbb{E}[g^{(A,s_A)}(j_k, s_k), g^{(A,s_A)}(j,s)]$, as given in (30), and as usual, $s_{A\cup\{j_k\}}$ denotes the concatenation of s_A and s_k. According to its definition in (35), therefore,

$$M^+(j_k, s_k) = \max_{(j,s) \in S^+(j_k, s_k)} g^{(A \cup \{j_k\}, s_{A \cup \{j_k\}})}(j,s),$$

and hence on the event $E(j_k, s_k)$, since we have $g^{(A,s_A)}(j_k, s_k) > M^+(j_k, s_k)$,

$$\begin{aligned} M^+(j_k, s_k) &= \max_{(j,s) \in S^+(j_k, s_k)} g^{(A \cup \{j_k\}, s_{A \cup \{j_k\}})}(j,s) \cdot 1\Big\{g^{(A \cup \{j_k\}, s_{A \cup \{j_k\}})}(j,s) < g^{(A,s_A)}(j_k, s_k)\Big\} \\ &\leq \max_{j \notin A \cup \{j_k\},\, s} g^{(A \cup \{j_k\}, s_{A \cup \{j_k\}})}(j,s) \cdot 1\Big\{g^{(A \cup \{j_k\}, s_{A \cup \{j_k\}})}(j,s) < g^{(A,s_A)}(j_k, s_k)\Big\} \\ &= M(j_k, s_k). \end{aligned}$$

This means that (now we return to writing $g^{(A,s_A)}$ as g, for brevity)

$$\sum_{j_k, s_k} \mathbb{P}\Big(\Big\{C(j_k, s_k) \cdot g(j_k, s_k)\big(g(j_k, s_k) - M(j_k, s_k)\big)/\sigma^2 > t\Big\} \cap E(j_k, s_k)\Big) \leq \sum_{j_k, s_k} \mathbb{P}\Big(\Big\{C(j_k, s_k) \cdot g(j_k, s_k)\big(g(j_k, s_k) - M^+(j_k, s_k)\big)/\sigma^2 > t\Big\} \cap E(j_k, s_k)\Big),$$

and so $\lim_{p\to\infty} \mathbb{P}(T_k > t) \leq \lim_{p\to\infty} \mathbb{P}(\tilde{T}_k > t) \leq e^{-t}$, the desired conclusion. □

A.9 Proof of Theorem 3

Since we are assuming that $\mathbb{P}(B) \to 1$, we know that $\mathbb{P}(T_k > t|B) - \mathbb{P}(T_k > t) \to 0$, so we only need to consider the marginal limiting distribution of T_k. We write $A = A_0$ and $s_A = s_{A_0}$. The general idea here is similar to that used in the proof of Theorem 2. Fixing m_0 and j_k, s_k, we will show that

$$\mathbb{P}\big(M^+(j_k, s_k) \leq m_0\sigma_k\big) \leq c^{|S|}, \tag{63}$$

where $S \subseteq \{1, \ldots p\} \setminus (A \cup \{j_k\})$ is as in the statement of the theorem for $j = j_k$, with size $|S| \geq d_p$, and $c < 1$ is a constant (not depending on j_k). Also, as in the proof of Lemma 8, we abbreviated $\sigma_k = \sigma/\sqrt{C(j_k, s_k)}$. This bound would imply that

$$\sum_{j_k, s_k} \mathbb{P}\big(M^+(j_k, s_k) \leq m_0\sigma_k\big) \leq 2p \cdot c^{d_p} \to 0 \quad \text{as } p \to \infty,$$

since $d_p/\log p \to 0$. The above sum converging to zero is precisely the condition required by Lemma 9, which then gives the desired (conservative) exponential limit for T_k. Hence it is suffices to show (63). For this, we start by recalling the definition of M^+ in (35):

$$M^+(j_k, s_k) = \max_{(j,s) \in S^+(j_k, s_k)} \frac{g(j,s) - (\Sigma_{j_k,j}/\Sigma_{j_k,j_k}) g(j_k, s_k)}{1 - \Sigma_{j_k,j}/\Sigma_{j_k,j_k}},$$
$$\text{where} \quad S^+(j_k, s_k) = \Big\{(j,s) : j \notin A \cup \{j_k\}, \ \frac{\Sigma_{j_k,j}}{\Sigma_{j_k,j_k}} < 1\Big\}.$$

Here we write $\Sigma_{j_k,j} = \mathbb{E}[g(j_k, s_k)g(j,s)]$; note that $\Sigma_{j_k,j_k} = \sigma_k^2$ (as shown in Lemma 10). First we show that the conditions of the theorem actually imply that $S^+(j_k, s_k) \supseteq S \times \{-1, 1\}^{|S|}$. This is true because for $j \in S$ and any $s \in \{-1, 1\}$, we have $|R_{j,j_k}|/R_{j_k,j_k} < \eta/(2-\eta)$ by (44), and

$$\begin{aligned} |R_{j,j_k}|/R_{j_k,j_k} < \eta/(2-\eta) \quad &\Rightarrow \quad \left| \frac{R_{j,j_k}}{R_{j_k,j_k}} \cdot \frac{s_k - X_{j_k}^T (X_A^T)^+ s_A}{s - X_j^T (X_A^T)^+ s_A} \right| < 1 \\ &\Rightarrow \quad \Sigma_{j_k,j}/\Sigma_{j_k,j_k} < 1. \end{aligned}$$

The first implication uses the assumption (42), as $|s_k - X_{j_k}^T (X_A^T)^+ s_A| \leq 1 + \|(X_A)^+ X_{j_k}\|_1 \leq 2 - \eta$ and $|s - X_j^T (X_A^T)^+ s_A| \geq 1 - \|(X_A)^+ X_{j_k}\|_1 \geq \eta$, and the second simply follows from the definition of $\Sigma_{j_k,j}$ and Σ_{j_k,j_k}. Therefore

$$M^+(j_k, s_k) \geq \max_{j \in S, \, s} \frac{g(j,s) - (\Sigma_{j_k,j}/\Sigma_{j_k,j_k}) g(j_k, s_k)}{1 - \Sigma_{j_k,j}/\Sigma_{j_k,j_k}}.$$

Let $U_j = X_j^T (I - P_A) y$ and $\theta_{j_k,j} = R_{j_k,j}/R_{j_k,j_k}$ for $j \in S$. By the arguments given in the proof of Lemma 12, we can rewrite the right-hand side above, yielding

$$\begin{aligned} M^+(j_k, s_k) &\geq \max_{j \in S, \, s} \frac{U_j - \theta_{j_k,j} U_{j_k}}{s - X_j^T (X_{A \cup \{j_k\}}^T)^+ s_{A \cup \{j_k\}}} \\ &= \max_{j \in S, \, s} \frac{s(U_j - \theta_{j_k,j} U_{j_k})}{1 - s X_j^T (X_{A \cup \{j_k\}}^T)^+ s_{A \cup \{j_k\}}} \\ &\geq \max_{j \in S} \frac{|U_j - \theta_{j_k,j} U_{j_k}|}{1 + |X_j^T (X_{A \cup \{j_k\}}^T)^+ s_{A \cup \{j_k\}}|} \\ &\geq \max_{j \in S} \frac{|U_j - \theta_{j_k,j} U_{j_k}|}{2}, \end{aligned}$$

where the last two inequalities above follow as $|X_j^T(X_{A\cup\{j_k\}})^+ s_{A\cup\{j_k\}}| < 1$ for all $j \in S$, which itself follows from the assumption that $\|(X_{A\cup\{j_k\}})^+ X_j\|_\infty < 1$ for all $j \in S$, in (45). Hence

$$\mathbb{P}\big(M^+(j_k, s_k) \le m_0\sigma_k\big) \le \mathbb{P}(|V_j| \le m_0\sigma_k,\, j \in S),$$

where $V_j = (U_j - \theta_{j_k,j}U_{j_k})/2$. Writing without a loss of generality $r = |S|$ and $S = \{1, \dots r\}$, it now remains to show that

$$\mathbb{P}(|V_1| \le m_0\sigma_k, \dots |V_r| \le m_0\sigma_k) \le c^r. \tag{64}$$

Similar to the arguments in the proof of Theorem 2, we will show (64) by induction, for the constant $c = \Phi(2m_0\sqrt{C}/(\delta\eta)) - \Phi(-2m_0\sqrt{C}/(\delta\eta)) < 1$. Before this, it is helpful to discuss three important facts. First, we note that (43) is actually a lower bound on the ratio of conditional to unconditional variances:

$$\mathrm{Var}(U_i \,|\, U_\ell,\, \ell \in S\setminus\{i\})/\mathrm{Var}(U_i) = \Big[R_{ii} - R_{i,S\setminus\{i\}}(R_{S\setminus\{i\},S\setminus\{i\}})^{-1}R_{S\setminus\{i\},i}\Big]/R_{ii} \ge \delta^2 \quad \text{for all } i \in S.$$

Second, conditioning on a smaller set of variables can only increase the conditional variance:

$$\mathrm{Var}(U_i \,|\, U_\ell,\, \ell \in S') \ge \mathrm{Var}(U_i \,|\, U_\ell,\, \ell \in S\setminus\{i\}) \ge \delta^2\sigma^2 R_{ii} \quad \text{for any } S' \subseteq S\setminus\{i\}, \text{ and } i \in S,$$

which holds as $U_1, \dots U_r$ are jointly normal. Third, and lastly, the collection $V_1, \dots V_r$ is independent of U_{j_k}, since these variables are all jointly normal, and it is easily verified that $\mathbb{E}[V_jU_{j_k}] = 0$ for each $j = 1, \dots r$.

We give the the inductive argument for (64). For the base case, we have $V_1 \sim N(0, \tau_1^2)$, where

$$\tau_1^2 = \mathrm{Var}(V_1) = \mathrm{Var}(V_1|U_{j_k}) = \mathrm{Var}(U_1)/4 \ge \delta^2\sigma^2 R_{11}/4.$$

Above, in the second equality, we used that V_1 and U_{j_k} are independent, and in the last inequality, that conditioning on fewer variables (here, none) only increases the variance. This means that

$$\mathbb{P}(|V_1| \le m_0\sigma_k) \le \mathbb{P}\Big(|Z| \le 2m_0\sigma_k/\big(\delta\sigma\sqrt{R_{11}}\big)\Big) \le \mathbb{P}\big(|Z| \le 2m_0\sqrt{C}/(\delta\eta)\big) = c,$$

where Z is standard normal; note that in the last inequality above, we applied the upper bound

$$\frac{\sigma_k^2}{\sigma^2 R_{11}} = \frac{\Sigma_{j_k,j_k}}{\sigma^2 R_{11}} = \frac{R_{j_k,j_k}}{R_{11}} \cdot \frac{1}{\big[s_k - X_{j_k}^T(X_A^T)^+ s_A\big]^2} \le \frac{C}{\eta^2}.$$

Now, for the inductive hypothesis, assume that $\mathbb{P}(|V_1| \le m_0\sigma_k, \dots |V_q| \le m_0\sigma_k) \le c^q$. Consider

$$\mathbb{P}(|V_1| \le m_0\sigma_k, \dots |V_{q+1}| \le m_0\sigma_k) = \mathbb{P}\big(|V_{q+1}| \le m_0\sigma_k \,\big|\, |V_1| \le m_0\sigma_k, \dots |V_q| \le m_0\sigma_k\big) \cdot c^q.$$

Using the independence of $V_1, \dots V_{q+1}$ and U_{j_k},

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

The variance τ_{q+1}^2 is

$$\tau_{q+1}^2 = \mathrm{Var}(V_{q+1} \,|\, U_1, \dots U_q, U_{j_k}) = \mathrm{Var}(U_{q+1} \,|\, U_1, \dots U_q)/4 \ge \delta^2\sigma^2 R_{q+1,q+1}/4,$$

where we again used the fact that conditioning on a smaller set of variables only makes the variance larger. Finally,

$$\mathbb{P}(|V_{q+1}| \le m_0\sigma_k \,|\, V_1, \dots V_q) \le \mathbb{P}\Big(|Z| \le 2m_0\sigma_k/\big(\delta\sigma\sqrt{R_{q+1,q+1}}\big)\Big) \le \mathbb{P}\big(|Z| \le 2m_0\sqrt{C}/(\delta\eta)\big) = c,$$

where we used $\sigma_k^2/(\sigma^2 R_{q+1,q+1}) \le C/\eta^2$ as above, and so

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

This completes the inductive proof. □

A.10 Statement and proof of Lemma 10

Lemma 10. *For any fixed A, s_A, and any $j \notin A$, $s \in \{-1, 1\}$, we have*

$$\mathrm{Var}\big(g(j,s)\big) = \frac{X_j^T(I - P_A)X_j^T\sigma^2}{[s - X_j^T(X_A^T)^+ s_A]^2} = \frac{\sigma^2}{\|(X_{A\cup\{j\}}^T)^+ s_{A\cup\{j\}} - (X_A^T)^+ s_A\|_2^2},$$

where $s_{A\cup\{j\}}$ denotes the concatenation of s_A and s.

Proof. We will show that

$$\frac{[s - X_j^T(X_A^T)^+ s_A]^2}{X_j^T(I - P_A)X_j^T} = \|(X_{A\cup\{j\}}^T)^+ s_{A\cup\{j\}} - (X_A^T)^+ s_A\|_2^2. \tag{65}$$

The right-hand side above, after a straightforward calculation, is shown to be equal to

$$s_{A\cup\{j\}}^T(X_{A\cup\{j\}}^T X_{A\cup\{j\}})^{-1} s_{A\cup\{j\}} - s_A(X_A^T X_A)^{-1} s_A. \tag{66}$$

Now let $z = (X_{A\cup\{j\}}^T X_{A\cup\{j\}})^{-1} s_{A\cup\{j\}}$. In block form,

$$\begin{bmatrix} X_A^T X_A & X_A^T X_j \\ X_j^T X_A & X_j^T X_j \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} s_A \\ s \end{bmatrix}. \tag{67}$$

Solving for z_1 in the first row yields

$$z_1 = (X_A^T X_A)^{-1} s_A - (X_A)^+ X_j z_2,$$

and therefore (66) is equal to

$$s_A^T z_1 + s z_2 - s_A^T(X_A^T X_A)^{-1} s_A = [s - s_A^T(X_A)^+ X_j] z_2. \tag{68}$$

Solving for z_2 in the second row of (67) gives

$$z_2 = \frac{s - s_A^T(X_A)^+ X_j}{X_j^T(I - P_A)X_j}.$$

Plugging this value into (68) produces the left-hand side in (65), completing the proof. □

A.11 Statement and proof of Lemma 11

Lemma 11. *If $v = v(m)$ satisfies $v > m$, then for any $t \geq 0$,*

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

Proof. First note, using a Taylor series expansion of $\sqrt{1 + 4t/m^3}$, that for sufficiently large m,

$$u(t,m) \geq m + \frac{t}{m} - \frac{t^2}{m^3}. \tag{69}$$

Also, a simple calculation shows that $\partial(u(t,m) - m)/\partial m \leq 0$ for all m, so that

$$u(t,w) - w \leq u(t,m) - m \quad \text{for all } w \geq m. \tag{70}$$

Now consider

$$\begin{aligned}
\Phi\big(v + u(t,m) - m\big) - \Phi\big(u(t,m)\big) &= \int_{u(t,m)}^{v+u(t,m)-m} \frac{e^{-z^2/2}}{\sqrt{2\pi}}\, dz \\
&= \int_m^v \frac{e^{-(w+u(t,m)-m)^2/2}}{\sqrt{2\pi}}\, dw \\
&\leq \int_m^v \frac{e^{-u(t,m)^2/2}}{\sqrt{2\pi}}\, dw \\
&\leq \int_m^v \frac{e^{-(w+t/m-t^2/m^3)^2/2}}{\sqrt{2\pi}}\, dw,
\end{aligned}$$

where the first inequality follows from (70), and the second from (69) (assuming m is large enough). Continuing from the last upper bound,

$$\int_m^v \frac{e^{-(w+t/m-t^2/m^3)^2/2}}{\sqrt{2\pi}}\, dw = e^{-t} \int_m^v \frac{e^{-w^2/2}}{\sqrt{2\pi}} f(w,t)\, dw,$$

where

$$f(w,t) = \exp\Big(\frac{t^2}{2w^2} + \frac{t^3}{w^4} - \frac{t^4}{2w^6}\Big).$$

Therefore, we have

$$\frac{\Phi\big(v + u(t,m) - m\big) - \Phi\big(u(t,m)\big)}{\Phi(v) - \Phi(m)} - e^{-t} \leq \left(\frac{\int_m^v \frac{e^{-w^2/2}}{\sqrt{2\pi}} f(w,t)\, dw}{\int_m^v \frac{e^{-w^2/2}}{\sqrt{2\pi}}\, dw} - 1 \right) \cdot e^{-t}. \tag{71}$$

It is clear that $f(w,t) \to 1$ as $w \to \infty$. Fixing ϵ, choose m_0 large enough so that for all $w \geq m_0$, we have $|f(w,t) - 1| \leq \epsilon$. Then the term multiplying e^{-t} on the right-hand side in (71), for $m \geq m_0$, is

$$\frac{\int_m^v \frac{e^{-w^2/2}}{\sqrt{2\pi}} f(w,t)\, dw}{\int_m^v \frac{e^{-w^2/2}}{\sqrt{2\pi}}\, dw} - 1 \leq \frac{\int_m^v \frac{e^{-w^2/2}}{\sqrt{2\pi}} \big|f(w,t) - 1\big|\, dw}{\int_m^v \frac{e^{-w^2/2}}{\sqrt{2\pi}}\, dw} \leq \epsilon,$$

which shows that the right-hand side in (71) is $\leq \epsilon \cdot e^{-t} \leq \epsilon$, and completes the proof. ∎

A.12 Statement and proof of Lemma 12

Lemma 12. *For any fixed j_k, s_k, j, s (and fixed A, s_A), we have*

$$\frac{g^{(A,s_A)}(j,s) - (\Sigma_{j_k,j}/\Sigma_{j_k,j_k}) g^{(A,s_A)}(j_k,s_k)}{1 - \Sigma_{j_k,j}/\Sigma_{j_k,j_k}} = g^{(A\cup\{j_k\}, s_{A\cup\{j_k\}})}(j,s), \tag{72}$$

where $\Sigma_{j_k,j}$ denotes the covariance between $g^{(A,s_A)}(j_k,s_k)$ and $g^{(A,s_A)}(j,s)$,

$$\Sigma_{j_k,j} = \frac{X_{j_k}^T(I-P_A)X_j\sigma^2}{[s_k - s_A^T(X_A)^+X_{j_k}][s - s_A^T(X_A)^+X_j]}.$$

Proof. Simple manipulations of the left-hand side in (72) yield the expression

$$\frac{X_j^T(I-P_A)y - \theta_{j_k,j}\cdot X_{j_k}^T(I-P_A)y}{s - s_A^T(X_A)^+X_j - \theta_{j_k,j}\cdot[s_k - s_A^T(X_A)^+X_{j_k}]}, \tag{73}$$

where $\theta_{j_k,j} = X_{j_k}^T(I-P_A)X_j/(X_{j_k}^T(I-P_A)X_{j_k})$. Now it remains to show that (73) is equal to

$$\frac{X_j^T(I-P_{A\cup\{j_k\}})y}{s - s_{A\cup\{j_k\}}^T(X_{A\cup\{j_k\}})^+X_j}. \tag{74}$$

We show individually that the numerators and denominators in (73) and (74) are equal. First the denominators: starting with (73), notice that

$$s - s_A^T(X_A)^+X_j - \theta_{j_k,j}[s_k - s_A^T(X_A)^+X_{j_k}] = s - s_{A\cup\{j_k\}}^T\begin{bmatrix} (X_A)^+(X_j - \theta_{j_k,j}X_{j_k}) \\ \theta_{j_k,j}\end{bmatrix}. \tag{75}$$

By the well-known formula for partial regression coefficients,

$$\theta_{j_k,j} = \frac{X_{j_k}^T(I-P_A)X_j}{X_{j_k}^T(I-P_A)X_{j_k}} = \big[(X_{A\cup\{j_k\}})^+X_j\big]_{j_k},$$

i.e., $\theta_{j_k,j}$ is the (j_k)th coefficient in the regression of X_j on $X_{A\cup\{j_k\}}$. Hence to show that (75) is equal to the denominator in (74), we need to show that $(X_A)^+(X_j - \theta_{j_k,j}X_{j_k})$ gives the coefficients in A in the regression of X_j on $X_{A\cup\{j_k\}}$. This follows by simply noting that the coefficients $(X_{A\cup\{j_k\}})^+X_j = (\theta_{A,j}, \theta_{j_k,j})$ satisfy the equation

$$X_A\theta_{A,j} + X_{j_k}\theta_{j_k,j} = P_{A\cup\{j_k\}}X_j,$$

and so solving for $\theta_{A,j}$,

$$\theta_{A,j} = (X_A)^+(P_{A\cup\{j_k\}}X_j - \theta_{j_k,j}X_{j_k}) = (X_A)^+(X_j - \theta_{j_k,j}X_{j_k}).$$

Now for the numerators: again beginning with (73), its numerator is

$$y^T(I-P_A)(X_j - \theta_{j_k,j}X_{j_k}), \tag{76}$$

and by essentially the same argument as above, we have

$$P_A(X_j - \theta_{j_k,j}X_{j_k}) = P_{A\cup\{j_k\}}X_j,$$

therefore (76) matches the numerator in (74). □

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.

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.

Buhlmann, P. (2012), Statistical significance in high-dimensional linear models. arXiv: 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, New York.

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

Efron, B. (1986), '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., 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: Series B* **74**(1), 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.

Grazier G'Sell, M., Taylor, J. & Tibshirani, R. (2013), Adaptive testing for the graphical lasso. arXiv: 1307.4765.

Grazier G'Sell, M., Wager, S., Chouldechova, A. & Tibshirani, R. (2013), False discovery rate control for sequential selection procedures, with application to the lasso. arXiv: 1309.5352.

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.

Javanmard, A. & Montanari, A. (2013*a*), Confidence intervals and hypothesis testing for high-dimensional regression. arXiv: 1306.3171.

Javanmard, A. & Montanari, A. (2013*b*), Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory. arXiv: 1301.4240.

Meinshausen, N. & Buhlmann, P. (2010), 'Stability selection', *Journal of the Royal Statistical Society: Series B* **72**(4), 417–473.

Meinshausen, N., Meier, L. & Buhlmann, P. (2009), 'p-values for high-dimensional regression', *Journal of the American Statistical Association* **104**(488), 1671–1681.

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 protease sequence database', *Nucleic Acids Research* **31**(1), 298–303.

Sun, T. & Zhang, C.-H. (2012), 'Scaled sparse linear regression', *Biometrika* **99**(4), 879–898.

Taylor, J., Loftus, J. & Tibshirani, R. J. (2013), Tests in adaptive regression via the Kac-Rice formula. arXiv: 1308.3020.

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.

van de Geer, S., Buhlmann, P. & Ritov, Y. (2013), On asymptotically optimal confidence regions and tests for high-dimensional models. arXiv: 1303.0518.

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', *Annals of Statistics* **37**(5), 2178–2201.

Weissman, I. (1978), 'Estimation of parameters and larger quantiles based on the k largest observations', *Journal of the American Statistical Association* **73**(364), 812–815.

Zhang, C.-H. & Zhang, S. (2011), Confidence intervals for low-dimensional parameters with high-dimensional data. arXiv: 1110.2563.

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

Zou, H. & Hastie, T. (2005), 'Regularization and variable selection via the elastic net', *Journal of the Royal Statistical Society: Series B* **67**(2), 301–320.

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