# Exact Post-selection Inference for Forward Stepwise and Least Angle Regression

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

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

**Abstract**

In this paper we propose new inference tools for forward stepwise and least angle regression. We first present a general scheme to perform valid inference after any selection event that can be characterized as the observation vector $y$ falling into some polyhedral set. This framework then allows us to derive conditional (post-selection) hypothesis tests at any step of the forward stepwise and least angle regression procedures. We derive an exact null distribution for our proposed test statistics in finite samples, yielding p-values with exact type I error control. The tests can also be inverted to produce confidence intervals for appropriate underlying regression parameters. Application of this framework to general likelihood-based regression models (e.g., generalized linear models and the Cox model) is also discussed.

Keywords: *forward stepwise regression, least angle regression, lasso, p-value, confidence interval, generalized linear model, Cox model*

## 1 Introduction

We consider observations $y \in \mathbb{R}^n$ drawn from a Gaussian model

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

Given a fixed matrix $X \in \mathbb{R}^{n \times p}$ of predictor variables, our focus is to provide inferential tools for methods that perform variable selection and estimation in an adaptive linear regression of $y$ on $X$. Unlike much of the related literature on adaptive linear modeling, we *do not assume that the true model is itself linear*, i.e., we do not assume that $\theta = X\beta^*$ for a vector of true coefficients $\beta^* \in \mathbb{R}^p$. As for adaptive regression methods, we mainly study forward stepwise regression (FS) and least angle regression (LAR), but we stress that our approach is not limited to these cases. First, our results for LAR also apply to the lasso path, up to the first variable deletion encountered (since these two partial paths are equivalent). Second, and more importantly, our results for both the FS and LAR cases follow from a more general theory that describes how to carry out exact inference after any *polyhedral selection* mechanism—this is a selection event that can be characterized by $y \in \mathcal{P}$ for a polyhedron $\mathcal{P} \subseteq \mathbb{R}^n$. Both the FS and LAR procedures are of this type.

To motivate the basic problem and illustrate our proposed solutions, we examine a data set of 67 observations and 8 variables, where the outcome is the log PSA level of men who had surgery for prostate cancer. The same data set was used to motivate the covariance test in Lockhart et al. (2014).[1] The first two numeric columns of Table 1 show the p-values for regression coefficients of

[1]The results for FS and the covariance test differ slightly from those that appear in Lockhart et al. (2014). Here we use a version of FS that selects variables to maximize the drop in residual sum of squares at each step; Lockhart et al. (2014) used a version that adds variables based on the maximal absolute correlation with the residual. Also, we use an Exp(1) limit for the covariance test; Lockhart et al. (2014) used an F-distribution to account for the unknown variance.

variables that enter the model, across steps of FS. The first column shows the results of applying naive, ordinary t-tests to compute the significance of these regression coefficients. We see that four variables are apparently significant at level 0.05, but this is suspect, as the p-values do not account for the greedy selection of variables that is inherent in the FS method. The second column shows our new selection-adjusted p-values that do properly account for the greediness: these are conditional on the active model at each step, and now just two variables are significant at the 0.05 level.

| | FS, naive | FS, adjusted | | Covariance test | Spacing test |
|---|---|---|---|---|---|
| lcavol | 0.000 | 0.000 | lcavol | 0.000 | 0.000 |
| lweight | 0.000 | 0.012 | lweight | 0.044 | 0.100 |
| svi | 0.047 | 0.849 | svi | 0.165 | 0.269 |
| lbph | 0.047 | 0.337 | lbph | 0.929 | 0.166 |
| pgg45 | 0.234 | 0.847 | pgg45 | 0.346 | 0.032 |
| lcp | 0.083 | 0.546 | age | 0.648 | 0.837 |
| age | 0.137 | 0.118 | lcp | 0.043 | 0.116 |
| gleason | 0.883 | 0.311 | gleason | 0.978 | 0.284 |

Table 1: *Prostate data example: naive and selection-adjusted forward stepwise sequential tests, and the covariance test and spacing test for LAR.*

Columns three and four show of Table 1 show analogous results for the LAR algorithm applied to these data. The covariance statistic (Lockhart et al. 2014), reviewed in the Section 7, measures the improvement in the fit due to adding a predictor at each step of LAR, and the third column shows the p-value from its Exp(1) asymptotic null distribution. The *spacing test*—this is the name we give to our new framework applied to LAR—produces the results in the rightmost column. We note that the spacing test assumes far less than the covariance test. In fact, our selection-adjusted p-values for both FS and LAR do not require assumptions about the predictors $X$, or about the true model being linear. They also use a null distribution that is correct in finite samples, rather than asymptotically, under Gaussian errors in (1). Later, we establish an asymptotic equivalence between the spacings and covariance test. Given this connection, one might expect the spacing and covariance test p-values in Table 1 to be closer. This discrepancy can be explained by the difference in alternatives considered: in the table, the adjusted FS and spacing p-values are based on a two-sided alternative, which we maintain as the default in this paper; meanwhile, the covariance test is actually equivalent to a one-sided version of the spacing test. Hence some of the smaller covariance test p-values are only about half the size of their spacing test counterparts.

The last three tests in Table 1 (the adjusted FS test, the covariance test, and the spacing test), all deem the first two predictors significant at the 0.10 level, and then the p-values rise. A highly important and nontrivial problem is to figure out how to combine the sequential p-values like these to build a stopping rule (here a model selection rule) with say a guaranteed false discovery rate. This is, of course, a pertinent question that a data analyst would ask when faced with p-values such as those in Table 1. We do not address this issue in the current paper, but we refer the reader to Grazier G'Sell et al. (2013) for a general method that can leverage p-values like the ones we establish in this work.

Here is an outline for this paper. In Section 2, we give an overview of our general approach to hypothesis testing after polyhedral selection, and describe how it applies to procedures such as FS and LAR. Sections 3, 4, and 5 then fill out the details for the general polyhedral, FS, and LAR cases, respectively. In the latter problem, our newly derived spacing test, which was featured in Table 1, has a remarkably simple form. Section 6 covers empirical examples, and Section 7 draws connections between the spacing and covariance tests. We finish with a discussion in Section 8, where we also present an extension of the FS test to likelihood-based regression model.

# 2 Summary of results

## 2.1 Hypothesis testing after selection

We now summarize our conditional testing framework that led to the results in the prostate data example, beginning briefly with the general problem setting that we consider. This general problem is to test the hypothesis

$$H_0: \ v^T\theta = 0, \tag{2}$$

conditional on having observed $y \in \mathcal{P}$, where $\mathcal{P}$ is a given polyhedral set (represented as an intersection of halfspaces), and $v = v(\mathcal{P})$ is a given vector (allowed to depend on $\mathcal{P}$). We derive a test statistic $T(y, \mathcal{P}, v)$ with the property that

$$T(y, \mathcal{P}, v) \overset{\mathbb{P}_0}{\sim} \text{Unif}(0, 1), \tag{3}$$

where $\mathbb{P}_0(\cdot) = \mathbb{P}_{v^T\theta=0}(\cdot \,|\, y \in \mathcal{P})$, the probability measure under $v^T\theta = 0$, conditional on $y \in \mathcal{P}$. The statistic in (3) is exactly uniform under the null measure, for any finite $n$ and $p$. This statement assumes nothing about the polyhedron $\mathcal{P}$ (aside from having an explicit representation for it as an intersection of halfspaces), and requires only Gaussian errors in the model (1). As it has a uniform null distribution, the test statistic in (3) serves as its own p-value, and so hereafter we will refer to it in both ways (test statistic and p-value).

Specifically, we can think of $y \in \mathcal{P}$ as describing a cumulative selection event for FS, up to some given number of steps $k$. Later in Section 4, we show that if $A_k = [j_1, \ldots j_k]$ denotes the FS active list after $k$ steps (so that FS selects variables $j_1, \ldots j_k$, in this order), and $s_{A_k} = [s_1, \ldots s_k]$ denotes the signs of corresponding fitted coefficients, then

$$\mathcal{P} = \Big\{ y : \hat{A}_k(y) = A_k, \, \hat{s}_{A_k}(y) = s_{A_k} \Big\}, \tag{4}$$

is a polyhedral set. Above, $\hat{A}_k(y)$ and $\hat{s}_{A_k}(y)$ denote the active list and signs after $k$ steps of FS, as functions of $y$. Now consider $v_k = (X^+_{A_k})^T e_k$, where $X_{A_k}$ is the subset of the predictor matrix whose columns are indexed by $A_k$, $(X_{A_k})^+ = (X^T_{A_k} X_{A_k})^{-1} X^T_{A_k}$ is the pseudoinverse of $X_{A_k}$, and $e_k$ is the $k$th standard basis vector. With the choice $v = v_k$, the null hypothesis $H_0$ in (2) is

$$H_0: \ e_k^T X^+_{A_k} \theta = 0. \tag{5}$$

which tests whether the coefficient of the latest selected variable, in the regression of $\theta$ onto $X_{A_k}$, is equal to zero. Said differently, this tests whether the contribution of the latest selected variable is significant, in the context of a projected linear model of $\theta$ onto the active variables in $A_k$. Our framework yields a test statistic (3) for this hypothesis, whose distribution we study conditional on the event $y \in \mathcal{P}$, i.e., conditional on the fact that FS selected the variables $A_k$ with signs $s_{A_k}$ over $k$ steps. To be explicit, this test statistic $T_k$ satisfies

$$\mathbb{P}_{e_k^T X^+_{A_k} \theta = 0}\Big( T_k \leq \alpha \,\Big|\, \hat{A}_k(y) = A_k, \, \hat{s}_{A_k}(y) = s_{A_k} \Big) = \alpha, \tag{6}$$

for all $0 \leq \alpha \leq 1$. The conditioning in (6) is important because it properly accounts for the adaptive (greedy) nature of FS in building the list $A_k$ in the first place. In loose terms, we are measuring the magnitude of the linear function $e_k^T X^+_{A_k} y$—not among all $y$ marginally—but among the observation vectors $y$ that would have resulted in FS selecting the variables $A_k$ (and signs $s_{A_k}$) after $k$ steps. The second column of Table 1 is showing p-values $T_k$ for tests of the hypotheses (5), across steps $k = 1, \ldots 8$ of the FS algorithm.

We emphasize that the p-values in (6) are exactly uniform under the null, in finite samples. This is true without placing any real restrictions on $X$ (aside from a general position assumption), and notably, without assuming linearity of the underlying model (i.e., without assuming $\theta = X\beta^*$). Finally, though we described the FS case here, basically the same story holds for LAR. We develop the theory the FS and LAR procedures in Sections 4 and 5, respectively.

## 2.2 Selection intervals

A strength of our framework is that our test statistics can be inverted to make coverage statements about arbitrary linear contrasts of $\theta$. In particular, consider the sequential hypothesis tests in (5) across steps $k$ of the FS procedure (the same results apply to LAR). By inverting our test statistic in (6), we derive a *selection interval* $I_k$ satisfying

$$\mathbb{P}\Big(e_k^T X_{A_k}^+ \theta \in I_k \,\Big|\, \hat{A}_k(y) = A_k,\, \hat{s}_{A_k}(y) = s_{A_k}\Big) = 1 - \alpha. \tag{7}$$

In words, the random interval $I_k$ traps with probability $1 - \alpha$ the coefficient of the latest selected variable, in a regression model that projects $\theta$ onto $X_{A_k}$, conditional on FS having selected variables $A_k$ with signs $s_{A_k}$, after $k$ steps of the algorithm. Since (7) is true conditional on $y \in \mathcal{P}$, we can marginalize this statement to yield

$$\mathbb{P}\Big(e_k^T X_{\hat{A}_k}^+ \theta \in I_k\Big) = 1 - \alpha. \tag{8}$$

Recall that $\hat{A}_k$ denotes the random active list after $k$ FS steps. From (8) we see that the selection interval $I_k$ covers with probability $1 - \alpha$ the regression coefficient of the variable selected at step $k$, in the projection of $\theta$ onto $X_{\hat{A}_k}$. This differs from a traditional confidence interval in that the selection interval is covering a *moving target*, as the identity of the $k$th selected variable is itself random (depending on $y$).

We have seen that selection intervals can be interpreted conditionally (7) or unconditionally (8). The former interpretation is more aligned with the spirit of post-selection inference, as it describes coverage for a projected regression coefficient of a selected variable, conditional on the output of our selection procedure. But the latter interpretation is also interesting, and in a way, cleaner. From the unconditional point of view, we can roughly think of the selection interval $I_k$ as covering the coefficient of the "$k$th most important variable" as deemed by the sequential regression procedure at hand (e.g., FS or LAR). See Figure 1 for 90% selection intervals at each step of FS regression on the prostate data set.

## 2.3 Conditional or unconditional?

The discussion of selection intervals in the last subsection raised a question that could be asked in general of our conditional hypothesis testing framework: should significance results be interpreted conditionally or unconditionally? Consider again the sequential hypotheses (5) over steps $k$ of the FS algorithm, for concreteness. The statement in (6) certifies the correctness of the $k$th constructed p-value, conditional on the current list of active variables and their signs. We can marginalize over all possible pairs $A_k, s_{A_k}$ to yield the unconditional statement

$$\mathbb{P}\Big(T_k \leq \alpha \,\Big|\, e_k^T X_{\hat{A}_k}^+ \theta = 0\Big) = \alpha, \tag{9}$$

where $\hat{A}_k$ is the random active list after $k$ steps of FS. Hence, as before, we have a second possible interpretation for the statistic (p-value) $T_k$: under the null measure, in which the latest selected variable has a zero projected regression coefficient, the statistic $T_k$ is uniformly distributed. (This is different from (6) in that the identity of the latest selected variable is random, because we are not conditoning the observed active list $A_k$.) In fact, any amount of coarsening of the conditioning set in (6) is also valid. For example, by marginalizing over all possible sign lists $s_{A_k}$ associated with $A_k$, we obtain

$$\mathbb{P}_{e_k^T X_{A_k}^+ \theta = 0}\Big(T_k \leq \alpha \,\Big|\, \hat{A}_k(y) = A_k\Big) = \alpha,$$

so that we are only conditioning on the observed active lists and not the observed signs.

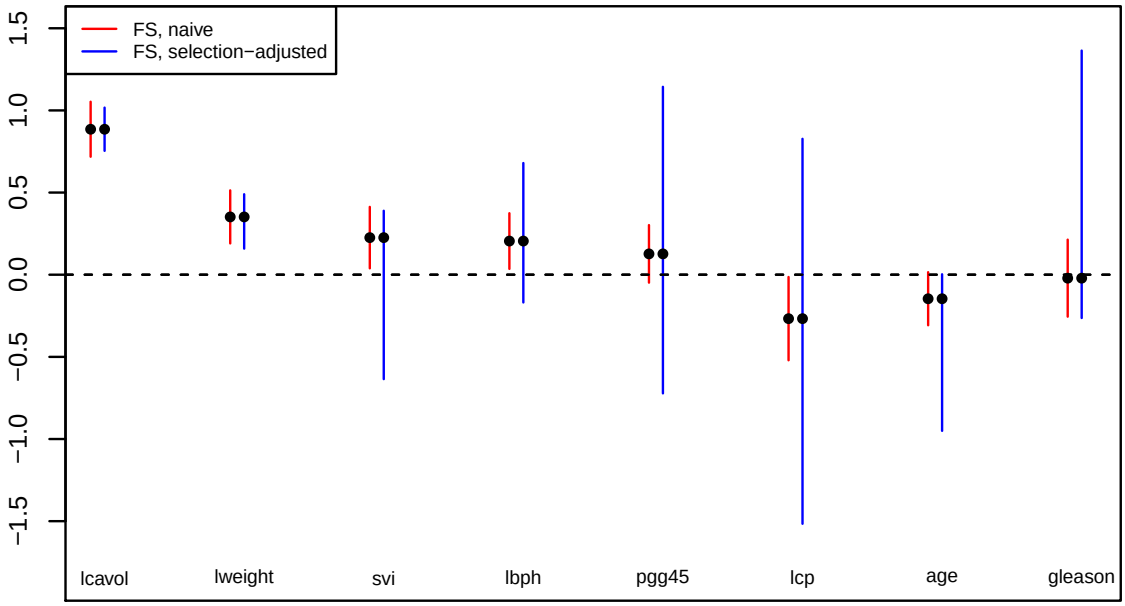


Figure 1: *Prostate data example: naive confidence intervals, and adjusted confidence intervals (selection intervals) with FS regression.*

## 2.4 Related work

There is much recent work on inference for adaptively fitted regression models. One class of techniques by Wasserman & Roeder (2009), Meinshausen & Buhlmann (2010), Minnier et al. (2011) is based on sample-splitting or resampling methods. Another class of approaches by Zhang & Zhang (2011), Buhlmann (2012), van de Geer et al. (2013), Javanmard & Montanari (2013*a*,*b*) is based on "denoising" a regularized regression estimator, like the lasso. Berk et al. (2013) carry out valid post-selection inference ("PoSI") by considering all possible model selection procedures that could have produced the given submodel. As the authors state, the inferences are generally conservative for particular selection procedures, but have the advantage that do not depend on the correctness the selected submodel. This same advantage is shared by the tests we propose here.

In a sister paper, Lee et al. (2013) use an approach that closely relates to the core idea in this paper, on polyhedral selection, to construct p-values and intervals for lasso coefficients at a fixed value of the regularization parameter $\lambda$ (instead of a fixed number of steps $k$ along the lasso or LAR path).

## 2.5 Notation

Here is the notation used in the coming sections. For a matrix $M \in \mathbb{R}^{n\times p}$ and list $S = [s_1, \ldots s_p] \subseteq [1, \ldots p]$, we write $M_S \in \mathbb{R}^{n\times|S|}$ for the submatrix formed by extracting the corresponding columns of $M$ (in the specified order). Similarly for a vector $x \in \mathbb{R}^p$, we write $x_S$ to denote the relevant subvector. We write $(M^TM)^+$ for the (Moore-Penrose) pseudoinverse of the square matrix $M^TM$, and $M^+ = (M^TM)^+M^T$ for the pseudoinverse of the rectangular matrix $M$. Lastly, we use $P_L$ for the projection operator onto a linear space $L$.

# 3 Conditional Gaussian inference after polyhedral selection

In this section we give an interesting result on Gaussian contrasts after polyhedral selection, which provides a basis for the methods proposed in this paper. Let

$$y \sim N(\theta, \Sigma),$$

where $\theta \in \mathbb{R}^n$ is unknown but $\Sigma \in \mathbb{R}^{n \times n}$ is known. This generalizes our original setup in (1) (in that we allow for a general error covariance structure). Consider a polyhedron

$$\mathcal{P} = \{y : \Gamma y \geq u\},$$

where $\Gamma \in \mathbb{R}^{m \times n}$, $u \in \mathbb{R}^m$ are fixed, and the above inequality is meant to be interpreted componentwise. Recall that we call $y \in \mathcal{P}$ a polyhedral selection event. Our goal is to make inferences about $v^T\theta$ conditional on $y \in \mathcal{P}$. Indeed $v = v(\mathcal{P})$ is allowed to be defined in terms of $\mathcal{P}$. The next result provides a key alternate representation for $\mathcal{P}$.

**Lemma 1 (Polyhedral selection as truncation).** *For any $\Sigma, v$ such that $v^T\Sigma v \neq 0$,*

$$\Gamma y \geq u \iff \mathcal{V}^{\mathrm{lo}}(y) \leq v^T y \leq \mathcal{V}^{\mathrm{up}}(y),\ \mathcal{V}^0(y) \leq 0, \tag{10}$$

*where*

$$\rho = \frac{\Gamma\Sigma v}{v^T\Sigma v} \tag{11}$$

$$\mathcal{V}^{\mathrm{lo}}(y) = \max_{j:\rho_j > 0} \frac{u_j - (\Gamma y)_j + \rho_j v^T y}{\rho_j} \tag{12}$$

$$\mathcal{V}^{\mathrm{up}}(y) = \min_{j:\rho_j < 0} \frac{u_j - (\Gamma y)_j + \rho_j v^T y}{\rho_j} \tag{13}$$

$$\mathcal{V}^0(y) = \max_{j:\rho_j = 0} u_j - (\Gamma y)_j. \tag{14}$$

*Moreover, the triplet $(\mathcal{V}^{\mathrm{lo}}, \mathcal{V}^{\mathrm{up}}, \mathcal{V}^0)(y)$ is independent of $v^T y$.*

*Remark.* The result in (10), with $\mathcal{V}^{\mathrm{lo}}, \mathcal{V}^{\mathrm{up}}, \mathcal{V}^0$ defined as in (11)–(14), is a deterministic result that holds for all $y$. Only the last independence result depends on normality of $y$.

See Figure 2 for a geometric illustration of this lemma. Intuitively, the result can be explained as follows, assuming for simplicity (and without a loss of generality) that $\Sigma = I$. We decompose $y = P_v y + P_{v^\perp} y$, where $P_v y = vv^T/\|v\|_2$ is the projection of $y$ along $v$, and $P_{v^\perp} y = y - P_v y$ is the projection onto the orthocomplement of $v$. Accordingly, we view $y$ as a deviation from $P_{v^\perp} y$, of an amount $v^T y$, along the line determined by $v$. The quantities $\mathcal{V}^{\mathrm{lo}}$ and $\mathcal{V}^{\mathrm{up}}$ describe how far we can deviate on either side of $P_{v^\perp} y$, before $y$ leaves the polyhedron. This gives rise to the inequality $\mathcal{V}^{\mathrm{lo}} \leq v^T y \leq \mathcal{V}^{\mathrm{up}}$. Some faces of the polyhedron, however, may be perfectly aligned with $v$ (i.e., their normal vectors may be orthogonal to $v$), and $\mathcal{V}^0$ accounts for this by checking that $y$ lies on the correct side of these faces. The proof, which is simply mathematical translation of this description, is given in Appendix A.1.

From Lemma 1, the distribution of any linear function $v^T y$, conditional on the selection $\Gamma y \geq u$, can be written as the conditional distribution

$$v^T y \,\big|\, \mathcal{V}^{\mathrm{lo}}(y) \leq v^T y \leq \mathcal{V}^{\mathrm{up}}(y),\ \mathcal{V}^0(y) \leq 0. \tag{15}$$

Since $v^T y$ has a Gaussian distribution, the above is a truncated Gaussian distribution (with random truncation limits). A simple transform reveals a pivotal statistic, which will be critical for inference about $v^T\theta$.

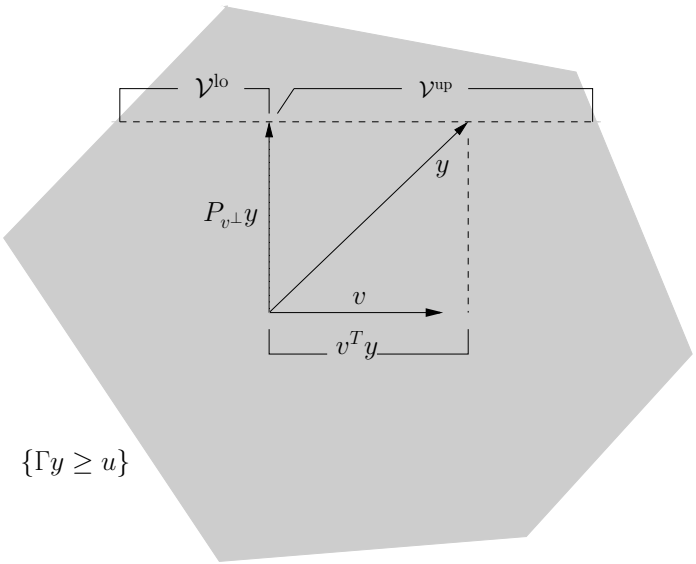


Figure 2: *Geometry of polyhedral selection as truncation. For simplicity we assume that $\Sigma = I$ (otherwise we can always standardize as appropriate). The shaded gray area is the polyhedral set $\{y : \Gamma y \geq u\}$. By breaking up $y$ into its projection onto $v$ and its projection onto the orthogonal complement of $v$, we see that $\Gamma y \geq u$ holds if and only if $v^T y$ does not deviate too far from $P_{v^\perp} y$, hence trapping it in between bounds $\mathcal{V}^{\text{lo}}, \mathcal{V}^{\text{up}}$. Furthermore, these bounds $\mathcal{V}^{\text{lo}}, \mathcal{V}^{\text{up}}$ are functions of $P_{v^\perp} y$ alone, so under normality, they are independent of $v^T y$.*

**Lemma 2** (**Pivotal statistic after polyhedral selection**). *Let $\Phi(x)$ denote the standard normal cumulative distribution function (CDF), and let $F^{[a,b]}_{\mu,\sigma^2}$ denote the CDF of a $N(\mu, \sigma^2)$ random variable truncated to lie in $[a, b]$, i.e.,*

$$F^{[a,b]}_{\mu,\sigma^2}(x) = \frac{\Phi((x-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}.$$

*For $v^T \Sigma v \neq 0$, the statistic $F^{[\mathcal{V}^{\text{lo}}, \mathcal{V}^{\text{up}}]}_{v^T\theta, v^T\Sigma v}(v^T y)$ is a pivotal quantity conditional on $\Gamma y \geq u$:*

$$\mathbb{P}\Big(F^{[\mathcal{V}^{\text{lo}}, \mathcal{V}^{\text{up}}]}_{v^T\theta, v^T\Sigma v}(v^T y) \leq \alpha \,\Big|\, \Gamma y \geq u\Big) = \alpha, \tag{16}$$

*for any $0 \leq \alpha \leq 1$, where $\mathcal{V}^{\text{lo}}$, $\mathcal{V}^{\text{up}}$ are as defined in (12), (13).*

See Appendix A.2 for the proof of Lemma 2 (it is very simple, and essentially just relies on the independence of $\mathcal{V}^{\text{lo}}, \mathcal{V}^{\text{up}}, \mathcal{V}^0$ and $v^T y$). The pivotal statistic in the lemma leads to valid conditional p-values for testing the null hypothesis $H_0 : v^T\theta = 0$, and correspondingly, conditional confidence intervals for $v^T\theta$. We divide our presentation into two parts, on one-sided and two-sided inference.

## 3.1 One-sided conditional inference

The result below is a direct consequence of the pivot in Lemma 2.

**Lemma 3** (**One-sided conditional inference after polyhedral selection**). *Given $v^T\Sigma v \neq 0$, suppose that we are interested in testing*

$$H_0 : \; v^T\theta = 0 \quad \textit{against} \quad H_1 : \; v^T\theta > 0.$$

*Define the test statistic*

$$T = 1 - F^{[\mathcal{V}^{\text{lo}},\mathcal{V}^{\text{up}}]}_{0,v^T\Sigma v}(v^T y), \tag{17}$$

*where we use the notation of Lemma 2 for the truncated normal CDF. Then $T$ is a valid p-value for $H_0$, conditional on $\Gamma y \geq u$:*

$$\mathbb{P}_{v^T\theta=0}(T \leq \alpha \,|\, \Gamma y \geq u) = \alpha, \tag{18}$$

*for any $0 \leq \alpha \leq 1$. Further, define $\delta_\alpha$ to satisfy*

$$1 - F^{[\mathcal{V}^{\text{lo}},\mathcal{V}^{\text{up}}]}_{\delta_\alpha,v^T\Sigma v}(v^T y) = \alpha. \tag{19}$$

*Then $I = [\delta_\alpha, \infty)$ is a valid one-sided confidence interval for $v^T\theta$, conditional on $\Gamma y \geq u$:*

$$\mathbb{P}(v^T\theta \geq \delta_\alpha \,|\, \Gamma y \geq u) = 1 - \alpha. \tag{20}$$

Note that by defining our test statistic in terms of the conditional survival function, as in (17), we are implicitly aligning ourselves to have power against the one-sided alternative $H_1 : v^T\theta > 0$. This is because the truncated normal survival function $1 - F^{[a,b]}_{\mu,\sigma^2}(x)$, evaluated at any fixed point $x$, is monotone increasing in $\mu$. The same fact (monotonicity of the survival function in $\mu$) validates the coverage of the confidence interval in (19), (20). Appendix A.3 covers the proof of Lemma 3.

## 3.2 Two-sided conditional inference

For a two-sided alternative, we use a simple modification of the one-sided test in Lemma 3.

**Lemma 4** (**Two-sided conditional inference after polyhedral selection**). *Given $v^T\Sigma v \neq 0$, suppose that we are interested in testing*

$$H_0 : \ v^T\theta = 0 \quad \textit{against} \quad H_1 : \ v^T\theta \neq 0.$$

*Define the test statistic*

$$T = 2 \cdot \min\Big\{F^{[\mathcal{V}^{\text{lo}},\mathcal{V}^{\text{up}}]}_{0,v^T\Sigma v}(v^T y),\ 1 - F^{[\mathcal{V}^{\text{lo}},\mathcal{V}^{\text{up}}]}_{0,v^T\Sigma v}(v^T y)\Big\}, \tag{21}$$

*where we use the notation of Lemma 2 for the truncated normal CDF. Then $T$ is a valid p-value for $H_0$, conditional on $\Gamma y \geq u$:*

$$\mathbb{P}_{v^T\theta=0}(T \leq \alpha \,|\, \Gamma y \geq u) = \alpha, \tag{22}$$

*for any $0 \leq \alpha \leq 1$. Further, define $\delta_{\alpha/2}, \delta_{1-\alpha/2}$ satisfy*

$$1 - F^{[\mathcal{V}^{\text{lo}},\mathcal{V}^{\text{up}}]}_{\delta_{\alpha/2},v^T\Sigma v}(v^T y) = \alpha/2 \tag{23}$$

$$1 - F^{[\mathcal{V}^{\text{lo}},\mathcal{V}^{\text{up}}]}_{\delta_{1-\alpha/2},v^T\Sigma v}(v^T y) = 1 - \alpha/2. \tag{24}$$

*Then*

$$\mathbb{P}(\delta_{\alpha/2} \leq v^T\theta \leq \delta_{1-\alpha/2} \,|\, \Gamma y \geq u) = 1 - \alpha. \tag{25}$$

The test statistic in (21), defined in terms of the minimum of the truncated normal CDF and survival function, has power against the two-sided alternative $H_1 : v^T\theta \neq 0$. The proof of its null distribution in (22) follows from the simple fact that if $U$ is a standard uniform random variable, then so is $2 \cdot \min\{U, 1-U\}$. The construction of the confidence interval in (23), (24), (25) again uses the monotonicity of the truncated normal survival function in the underlying mean parameter. See Appendix A.3.

The FS and spacing p-values in Table 1 and intervals in Figure 1 were all computed using the two-sided framework of Lemma 4. Unless specified otherwise, we will stick to the two-sided tests and two-sided intervals for the rest of this paper.

### 3.3 A simple example with orthogonal regression

Lemmas 1–4 provide tools for inference about any linear function $v^T\theta$, conditional on the selection event $\Gamma y \geq u$. To help understand their results, we present a simple example with $n = 60$ and the number of variables ranging over $p = 2, 5, 10, 20, 50$. The setup: orthogonal predictors $X$ (satisfying $X^TX = I$), and observations $y_i \sim N(0, 1)$ drawn independently for $i = 1, \ldots n$ (representing the global null case). We found the predictor $j_1$ having the largest absolute inner product with $y$; note that this is the variable chosen by FS in its first step. To simplify matters, we conditioned on data realizations such that the inner product $X_{j_1}^T y$ was positive.

The FS setting will be discussed in detail in the next section, but for now, it is straightforward to see that the set of observation vectors $y$ for which the first step of FS chooses $j_1$, with a positive attained inner product, forms a polyhedral set:

$$\mathcal{P} = \{y : X_{j_1}^T y \geq \pm X_j^T y, \text{ for all } j \neq j_1\},$$

so that we can apply our polyhedral selection theory in Lemma 1. With $v = X_{j_1}$, and $v^T y$ being the attained maximal inner product, the density of the truncated normal random variable in (15) is plotted in Figure 3. We use a different colored line for each number of predictors $p$. It is not hard to check that in this setting, we have $\mathcal{V}^{\text{up}} = \mathcal{V}^0 = \infty$, so that the truncation is effectively only on the lower end. It is also not hard to check that the lower bound $\mathcal{V}^{\text{lo}}$ is exactly equal to the maximum of $|X_j^T y|$ over $j \neq j_1$, i.e., the second largest absolute inner product among the set, and so $X_{j_1}^T y \geq \mathcal{V}^{\text{lo}}$ is clearly equivalent to $y \in \mathcal{P}$. As $p$ grows larger, the second largest absolute inner product $\mathcal{V}^{\text{lo}}$ also grows, and the truncation becomes steeper, as we can see in the figure.

Moreover, Lemmas 3 and 4 can be interpreted in terms of the observed values of $v^T y$, denoted by colored dots in Figure 3, and the underlying conditional densities, denoted by the colored lines. To compute a p-value as in (17), (18), we simply read off the mass in the truncated density to the right of the observed value $v^T y$. To compute a confidence interval as in (19), (20), we shift around the truncated density, away from its anchor point (mean of the untruncated density) at $v^T\theta = 0$, and we collect all points $v^T\theta$ for which the p-value is larger than the nominal level. The two-sided inference tools from Lemma 4 can be viewed similarly, but use both tails of the density.

## 4 Selection-adjusted forward stepwise tests

We apply the tools of Section 3 to the particular case of selection in regression using the forward stepwise or FS approach. Recall that in FS, we repeatedly add the predictor to the current active model that most improves the fit. After each addition, the active coefficients are recomputed by least squares regression on the active predictors. This process ends when all predictors are in the model, or when the residual error is zero. Slightly different versions of FS result from different ways of defining the notion of the predictor that "most improves the fit"—specifically, we use the largest drop in residual error (when a candidate predictor is added to the current active model) to choose the predictor at each step. However, other common versions of FS will also fit into our polyhedral framework.

Our assumptions on $X$ are minimal: we only require that $X$ have columns in general position. This means that for any $k < \min\{n, p\}$, any columns $X_{j_1}, \ldots X_{j_k}$, and any signs $\sigma_1, \ldots \sigma_k \in \{-1, 1\}$, the linear span of $\sigma_1 X_{j_1}, \ldots \sigma_k X_{j_k}$ does not contain any of the remaining columns, up to a sign flip (i.e., does not contain any of $\pm X_j$, $j \neq j_1, \ldots j_k$). One can check that this implies that the selected variable and sign pair, at each step of FS, is unique. (In other words, there is no loss of generality in the use of nonstrict inequalities in the next subsection.) The general position assumption is not at all stringent; e.g., as argued in Tibshirani (2013), if the columns of $X$ are drawn according to a continuous probability distribution, then they are in general position almost surely.

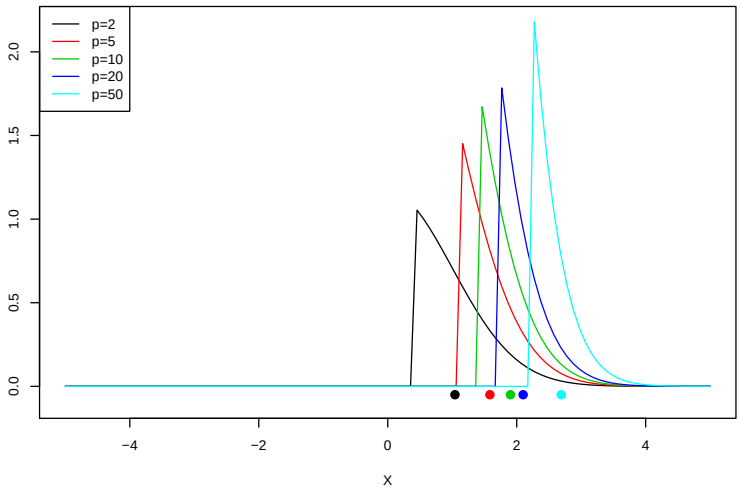


Figure 3: *Truncated Gaussian densities* (15)*, when the conditioning set is specified by the first step of FS, in problems with* $n = 60$ *observations and* $p = 2, 5, 10, 20, 25$ *predictors. In each problem, the variables are orthonormal, and the true model is null. Also,* $v$ *is chosen so that* $v^T y$ *is the maximum absolute inner product between a predictor and* $y$*, and this value is denoted by a colored dot along the bottom. (All curves and dots were averaged over 100 simulations from the described model setup to stabilize the results.)*

## 4.1 Details of the polyhedral set

After $k$ FS steps, let $A_k = [j_1, \ldots j_k]$ denote the active list of variables, selected in this order, and $s_{A_k} = [s_1, \ldots s_k]$ denote their signs upon entering. That is, at each step $k$, the variable $j_k$ and sign $s_k$ satisfy

$$\begin{aligned} \mathrm{RSS}\big(y, X_{[j_1,\ldots j_{k-1},j_k]}\big) \le \mathrm{RSS}\big(y, X_{[j_1,\ldots j_{k-1},j]}\big) \quad &\text{for all } j \neq j_1, \ldots j_k, \text{ and} \\ s_k = \mathrm{sign}\big(e_k^T (X_{[j_1,\ldots j_k]})^+ y\big), & \end{aligned}$$

where $\mathrm{RSS}(y, X_S)$ denotes the residual sum of squares from regressing $y$ onto $X_S$, for some list of variables $S$. We show that the set of all observation vectors $y$ that result in a FS active list $A_k$ and sign list $s_{A_k}$ over $k$ steps

$$\mathcal{P} = \Big\{ y : \hat{A}_k(y) = A_k,\ \hat{s}_{A_k}(y) = s_{A_k} \Big\} \tag{26}$$

is indeed a polyhedron of the form $\mathcal{P} = \{y : \Gamma y \ge 0\}$. Above, $\hat{A}_k(y)$ denotes the random FS active list at $y$, and $\hat{s}_{A_k}(y)$ the random sign list, at step $k$.

The proof proceeds by induction. The case when $k = 1$ can be seen directly by inspection, as $j_1$ and $s_1$ are the variable and sign to be chosen by FS if and only if

$$\begin{aligned} \Big\| \Big( I - X_{j_1} X_{j_1}^T / \|X_{j_1}\|_2^2 \Big) y \Big\|_2^2 \le \Big\| \Big( I - X_j X_j^T / \|X_j\|_2^2 \Big) y \Big\|_2^2 \quad &\text{for all } j \neq j_1, \text{ and} \\ s_1 = \mathrm{sign}(X_{j_1}^T y), & \end{aligned}$$

which is equivalent to

$$s_1 X_{j_1}^T y / \|X_{j_1}\|_2 \ge \pm X_j^T y / \|X_j\|_2 \quad \text{for all } j \neq j_1.$$

Hence the matrix $\Gamma$ begins with $2(p-1)$ rows of the form $s_1 X_{j_1}/\|X_{j_1}\|_2 \pm X_j/\|X_j\|_2$, for $j \neq j_1$. Now assume the statement is true for $k-1$ steps. At step $k$, the optimality conditions for $j_k, s_k$ can

be expressed as

$$\left\|\left(I - \tilde{X}_{j_k}\tilde{X}_{j_k}^T/\|\tilde{X}_{j_k}\|_2^2\right)r\right\|_2^2 \leq \left\|\left(I - \tilde{X}_j\tilde{X}_j^T/\|\tilde{X}_j\|_2^2\right)r\right\|_2^2 \quad \text{for all } j \neq j_1, \ldots j_k, \text{ and}$$
$$s_k = \text{sign}(\tilde{X}_{j_k}^T r),$$

where $\tilde{X}_j$ denotes the residual from regressing $X_j$ onto $X_{A_{k-1}}$, and $r$ the residual from regressing $y$ onto $X_{A_{k-1}}$. As in the $k = 1$ case, the above is equivalent to

$$s_k\tilde{X}_{j_k}^T r/\|\tilde{X}_{j_k}\|_2 \geq \pm\tilde{X}_j^T r/\|\tilde{X}_j\|_2 \quad \text{for all } j \neq j_1, \ldots j_k.$$

or

$$s_k X_{j_k}^T P_{A_{k-1}}^\perp y/\|P_{A_{k-1}}^\perp X_{j_k}\|_2 \geq \pm X_j^T P_{A_{k-1}}^\perp y/\|P_{A_{k-1}}^\perp X_j\|_2 \quad \text{for all } j \neq j_1, \ldots j_k,$$

with $P_{A_{k-1}}^\perp$ denoting the projection orthogonal to the column space of $X_{A_{k-1}}$. Therefore we append $2(p-k)$ rows to $\Gamma$, of the form $P_{A_{k-1}}^\perp(s_k X_{j_k}/\|P_{A_{k-1}}^\perp X_{j_k}\|_2 \pm X_j/\|P_{A_{k-1}}^\perp X_j\|_2)$, for $j \neq j_1, \ldots j_k$.

In summary, after $k$ steps, the polyhedral representation for the FS selection event (26) corresponds to a matrix $\Gamma$ with precisely $2pk - k(k+1)$ rows.[2]

## 4.2 Details of the tests and intervals

Given some number of steps $k$, after we have formed the $\Gamma$ matrix, computing conditional $p$-values and intervals for FS is straightforward. Consider testing a generic hypothesis $H_0 : v^T\theta = 0$ where $v$ is arbitrary. First we compute, as prescribed by Lemma 1, the quantities

$$\mathcal{V}^{\text{lo}} = \max_{j:(\Gamma v)_j > 0} -(\Gamma P_{v^\perp} y)_j/(\Gamma v)_j \cdot \|v\|_2^2 \tag{27}$$

$$\mathcal{V}^{\text{up}} = \min_{j:(\Gamma v)_j < 0} -(\Gamma P_{v^\perp} y)_j/(\Gamma v)_j \cdot \|v\|_2^2. \tag{28}$$

Then we define

$$R_k = 1 - F_{0,\sigma^2\|v\|_2^2}^{[\mathcal{V}^{\text{lo}},\mathcal{V}^{\text{up}}]}(v^T y) = \frac{\Phi\left(\frac{\mathcal{V}^{\text{up}}}{\sigma\|v\|_2}\right) - \Phi\left(\frac{v^T y}{\sigma\|v\|_2}\right)}{\Phi\left(\frac{\mathcal{V}^{\text{up}}}{\sigma\|v\|_2}\right) - \Phi\left(\frac{\mathcal{V}^{\text{lo}}}{\sigma\|v\|_2}\right)}, \tag{29}$$

and the test statistic

$$T_k = 2 \cdot \min\{R_k, 1 - R_k\}. \tag{30}$$

By Lemma 4, this serves as a valid p-value, conditional on the selection. That is,

$$\mathbb{P}_{v^T\theta=0}\Big(T_k \leq \alpha \,\Big|\, \hat{A}_k(y) = A_k,\, \hat{s}_{A_k}(y) = s_{A_k}\Big) = \alpha, \tag{31}$$

for any $0 \leq \alpha \leq 1$. Also according to Lemma 4, a conditional confidence interval is derived by first computing $\delta_\alpha$, $\delta_{1-\alpha/2}$ that satisfy

$$1 - F_{\delta_{\alpha/2},\sigma^2\|v\|_2^2}^{[\mathcal{V}^{\text{lo}},\mathcal{V}^{\text{up}}]}(v^T y) = \alpha/2 \tag{32}$$

$$1 - F_{\delta_{1-\alpha/2},\sigma^2\|v\|_2^2}^{[\mathcal{V}^{\text{lo}},\mathcal{V}^{\text{up}}]}(v^T y) = 1 - \alpha/2, \tag{33}$$

and then $I_k = [\delta_{\alpha/2}, \delta_{1-\alpha/2}]$ has the proper conditional coverage, in that

$$\mathbb{P}\Big(v^T\theta \in I_k \,\Big|\, \hat{A}_k(y) = A_k,\, \hat{s}_{A_k}(y) = s_{A_k}\Big) = 1 - \alpha. \tag{34}$$

[2]Note: we have been implicitly assuming so far that $k < p$. If $k = p$ (so that necessarily $p \leq n$), then we must add an "extra" row to $\Gamma$, this row being $P_{A_{p-1}}^\perp s_p X_{j_p}$, which encodes the sign constraint $s_p X_{j_p}^T P_{A_{p-1}}^\perp y \geq 0$. For $k < p$, this constraint is implicitly encoded due to the constraints of the form $s_k X_{j_k}^T P_{A_{k-1}}^\perp y \geq \pm a$ for some $a$.

The case when $v_k = (X^+_{A_k})^T e_k$, and the null hypothesis is $H_0 : e_k^T X^+_{A_k} \theta = 0$, is of particular interest, as discussed in the introduction. Now, we are testing whether the coefficient of the latest selected variable, in the population regression of $\theta$ on $X_{A_k}$, is equal to zero. For this case, the details of the p-values and intervals follow exactly as above with the appropriate substitution $v = v_k$. It may be interesting to point out that here $v_k^T y = e_k^T X^+_{A_k} y$ is the fitted regression component of the latest selected variable as it enters the model, which we may also write as $v_k^T y = \hat{\beta}_{j_k}$. Therefore, our proposed selection-adjusted FS test compares the magnitude of the latest computed regression coefficient against the tails of a truncated normal distribution, and clearly, the larger $|\hat{\beta}_{j_k}|$ is, the more significant the result.

### 4.3 A model with intercept

Often, FS is run by first starting with an intercept term in the model, and then sequentially adding predictors. Our selection theory can accomodate this case. It is easiest to simply consider centering $y$ and the columns of $X$, which is equivalent to including an intercept term in the regression. After centering, the covariance matrix of $y$ is $\Sigma = \sigma^2(I - \mathbb{1}\mathbb{1}^T/n)$, where $\mathbb{1}$ is the vector of all 1s. This is fine, because the polyhedral theory from Section 3 applies to Gaussian random variables with an arbitrary (but known) covariance. With the centered $y$ and $X$, the construction of the polyhedral set ($\Gamma$ matrix) carries over just as described in Section 4.1. The conditional tests and intervals also carry over as in Section 4.2, except with the general contrast vector $v$ replaced by its own centered version. Note that when $v$ in the column space of $X$, e.g., when $v_k = (X^+_{A_k})^T e_k$, no change at all is needed.

## 5 Selection-adjusted least angle regression tests

The least angle regression or LAR algorithm (Efron et al. 2004) is an iterative method, like forward stepwise, that produces a sequence of nested regression models. As before, we keep a list of active variables and signs across steps of the algorithm. Also as before, we assume that $X$ has columns in general position, which ensures uniqueness of the selected variable and sign at each step (Tibshirani 2013). Here is a concise description of the LAR steps. For step $k = 1$, we set the active variable and sign list to be $A = [j_1]$ and $s_{A_1} = [s_1]$, where $j_1, s_1$ satisfy

$$(j_1, s_1) = \underset{j=1,\ldots p,\, s\in\{-1,1\}}{\operatorname{argmax}} \; s X_j^T y. \tag{35}$$

This is the same selection as made by FS at the first step, provided that $X$ has columns with unit norm. We also record the first knot

$$\lambda_1 = s_1 X_{j_1}^T y. \tag{36}$$

For a general step $k > 1$, we form the list $A_k$ by appending $j_k$ to $A_{k-1}$, and form $s_{A_k}$ by appending $s_k$ to $s_{A_k}$, where $j_k, s_k$ satisfy

$$(j_k, s_k) = \underset{j\notin A_{k-1},\, s\in\{-1,1\}}{\operatorname{argmax}} \; \frac{X_j^T P^\perp_{A_{k-1}} y}{s - X_j^T (X^+_{A_{k-1}})^T s_{A_{k-1}}} \cdot 1\left\{ \frac{X_j^T P^\perp_{A_{k-1}} y}{s - X_j^T (X^+_{A_{k-1}})^T s_{A_{k-1}}} \le \lambda_{k-1} \right\}. \tag{37}$$

Above, $P^\perp_{A_{k-1}}$ is the projection orthogonal to the column space of $X_{A_{k-1}}$, $1\{\cdot\}$ denotes the indicator function, and $\lambda_{k-1}$ is the knot value from step $k-1$. We also record the $k$th knot

$$\lambda_k = \frac{X_{j_k}^T P^\perp_{A_{k-1}} y}{s_k - X_{j_k}^T (X^+_{A_{k-1}})^T s_{A_{k-1}}}. \tag{38}$$

The algorithm terminates after the $k$-step model if $k = p$, or if $\lambda_{k+1} < 0$.

LAR is often viewed as "less greedy" than FS. Though we have not described it this way, LAR maintains an estimate of the regression coefficients, at each step. It moves these coefficients in the direction of the least squares estimate on the current active variables, but instead of fitting a full least squares estimate, it halts when another variable has equal correlation with the residual as do the active variables. It then adds this variable to the active set, and repeats this procedure. LAR is also intimiately tied to the lasso: by introducing a step that deletes variables from the active set when their coefficients pass through zero, the modified LAR algorithm traces out the lasso solution path exactly. All of this is covered in elegant detail in Efron et al. (2004).

Over the next subsections, we show that the set of observation vectors $y$ that result in a given sequence of active variable and sign lists,

$$\mathcal{P} = \Big\{y : \hat{A}_k(y) = A_k,\, \hat{s}_{A_k}(y) = s_{A_k}\Big\}, \tag{39}$$

is polyhedral. In the above, $\hat{A}_k(y)$ and $\hat{s}_{A_k}(y)$ denote the random LAR active variables and signs at $y$, after $k$ steps. We provide more than one characterization for the polyhedron $\mathcal{P}$, leading to a concise characterization that holds in special cases. Note that, given a polyhedral representation of the form $\mathcal{P} = \{y : \Gamma y \geq 0\}$, we can compute proper p-values and intervals conditional on the LAR selection event $\mathcal{P}$, just as with FS in Section 4.2; for LAR, they simply rely on a different matrix $\Gamma$. That is, $\mathcal{V}^{\mathrm{lo}}, \mathcal{V}^{\mathrm{up}}$ are still computed precisely as in (27), (28), conditional p-values are as in (29), (30), (31), and conditional confidence intervals as in (32), (33), (34). We do not rehash the details beyond this. Instead we focus on discussing the spacing test in detail, which is a special version of our selection-adjusted test, and depends on our most concise representation of the LAR selection event.

## 5.1 A brute force characterization of the polyhedral set

As we did with FS, we can write down a polyhedral characterization of the LAR selection set by "brute force". One caveat is that our characterization does not actually apply to $\mathcal{P}$ in (39), which describes the sequence of active variables and signs across the first $k$ steps of the LAR algorithm, but rather to the smaller set

$$\mathcal{P}' = \Big\{y : \hat{A}_k(y) = A_k,\, \hat{s}_{A_k}(y) = s_{A_k},\, \hat{S}_\ell(y) = S_\ell,\, \ell = 1, \ldots k\Big\}. \tag{40}$$

Explained in words, $S_\ell \subseteq \{1, \ldots p\} \times \{-1, 1\}$ contains the variable-sign pairs that were "in competition" to become the active variable-sign pair step $\ell$. A subtlety of LAR: it is not always the case that $S_\ell = \{1, \ldots p\} \setminus A_{\ell-1} \times \{-1, 1\}$, since some variable-sign pairs are automatically excluded from consideration, as they would have produced a knot value that is too large (larger than the previous knot $\lambda_{\ell-1}$). This is reflected by the indicator function in (37). We note that the smaller characterization in (40) is *perfectly viable for inference*, because any conditional statement over $\mathcal{P}'$ translates into a valid conditional statement over $\mathcal{P}$ (by marginalizing over $S_\ell$, $\ell = 1, \ldots k$). Recall the general discussion of marginalization in Section 2.3.

We now show how to represent $\mathcal{P}'$ in (40) in the form $\{y : \Gamma y \geq 0\}$. Starting with $k = 1$, we can express the optimality of $j_1, s_1$ in (35) as

$$c(j_1, s_1)^T y \geq c(j, s)^T y, \quad \text{for all } j \neq j_1,\, s \in \{-1, 1\},$$

where $c(j, s) = sX_j$. Therefore $\Gamma$ has $2(p-1)$ rows, of the form $c(j_1, s_1) - c(j, s)$. (In the first step, $S_1 = \{1, \ldots p\} \times \{-1, 1\}$, and we do not require extra rows of $\Gamma$ to explicitly represent it.) Further, suppose that the selection set can be represented in the desired manner, after $k-1$ steps. Then the optimality of $j_k, s_k$ in (37) can be expressed as

$$\begin{aligned}
&c(j_k, s_k, A_{k-1}, s_{A_{k-1}})^T y \geq c(j, s, A_{k-1}, s_{A_{k-1}})^T y \quad \text{for all } (j, s) \in S_k \setminus \{(j_k, s_k)\} \\
&c(j_k, s_k, A_{k-1}, s_{A_{k-1}})^T y \geq 0,
\end{aligned}$$

where $c(j, s, A_{k-1}, s_{A_{k-1}}) = (P^{\perp}_{A_{k-1}} X_j)/(s - X_j^T (X^+_{A_{k-1}})^T s_{A_{k-1}})$. The set $S_k$ is characterized by

$$c(j, s, A_{k-1}, s_{A_{k-1}})^T y \le c(j_{k-1}, s_{k-1}, A_{k-2}, s_{A_{k-2}})^T y \quad \text{for } (j, s) \in S_k$$
$$c(j, s, A_{k-1}, s_{A_{k-1}})^T y \ge c(j_{k-1}, s_{k-1}, A_{k-2}, s_{A_{k-2}})^T y \quad \text{for } (j, s) \notin S_k.$$

Therefore, the new $\Gamma$ is created by appending the following $|S_k| + 2(p - k + 1)$ rows to the previous matrix: $c(j_k, s_k, A_{k-1}, s_{A_{k-1}}) - c(j, s, A_{k-1}, s_{A_{k-1}})$, for $(j, s) \in S_k \setminus \{(j_k, s_k)\}$; $c(j_k, s_k, A_{k-1}, s_{A_{k-1}})$; $c(j_{k-1}, s_{k-1}, A_{k-2}, s_{A_{k-2}}) - c(j, s, A_{k-1}, s_{A_{k-1}})$, for $(j, s) \in S_k$; and finally $c(j, s, A_{k-1}, s_{A_{k-1}}) - c(j_{k-1}, s_{k-1}, A_{k-2}, s_{A_{k-2}})$, for $(j, s) \notin S_k$. In total, the number of rows of $\Gamma$ at step $k$ is bounded by $\sum_{\ell=1}^{k} (|S_\ell| + 2(p - \ell + 1)) \le 4pk - 2k(k-1)$.

A downside to mention is that the matrix $\Gamma$ grows very large in the proposed representation: it has roughly $4pk$ rows after $k$ steps (the FS representation is itself quite large too, with roughly $2pk$ rows after $k$ steps). This makes it somewhat cumbersome to form, and also to compute $\mathcal{V}^{\text{lo}}, \mathcal{V}^{\text{up}}$ as needed for the test and intervals. Next we suggest an alternative polyhedral representation for the LAR selection event, in an attempt to address this issue.

## 5.2 A refined characterization of the polyhedral set

An alternative characterization for the LAR selection event, after $k$ steps, is described below. The proof draws heavily on results from Lockhart et al. (2014), and is given in Appendix A.4.

**Lemma 5.** *Suppose that the LAR algorithm produces the list of active variables $A_k$ and signs $s_{A_k}$ after $k$ steps. Define $c(j, s, A_{k-1}, s_{A_{k-1}}) = (P^{\perp}_{A_{k-1}} X_j)/(s - X_j^T (X^+_{A_{k-1}})^T s_{A_{k-1}})$, with the convention $A_0 = s_{A_0} = \emptyset$, so that $c(j, s, A_0, s_{A_0}) = c(j, s) = sX_j$. Consider the following conditions:*

$$c(j_1, s_1, A_0, s_{A_0})^T y \ge c(j_2, s_2, A_1, s_{A_1})^T y \ge \ldots \ge c(j_k, s_k, A_{k-1}, s_{A_{k-1}})^T y \ge 0 \tag{41}$$
$$c(j_k, s_k, A_{k-1}, s_{A_{k-1}})^T y \ge M_k^+ \Big( j_k, s_k, c(j_{k-1}, s_{k-1}, A_{k-2}, s_{A_{k-2}})^T y \Big) \tag{42}$$
$$c(j_\ell, s_\ell, A_{\ell-1}, s_{A_{\ell-1}})^T y \le M_\ell^- \Big( j_\ell, s_\ell, c(j_{\ell-1}, s_{\ell-1}, A_{\ell-2}, s_{A_{\ell-2}})^T y \Big), \quad \text{for } \ell = 1, \ldots k \tag{43}$$
$$0 \ge M_\ell^0 \Big( j_\ell, s_\ell, c(j_{\ell-1}, s_{\ell-1}, A_{\ell-2}, s_{A_{\ell-2}})^T y \Big), \quad \text{for } \ell = 1, \ldots k \tag{44}$$
$$0 \le M_\ell^S \, y, \quad \text{for } \ell = 1, \ldots k. \tag{45}$$

*(Note that for $\ell = 1$ in (43), (44), we are meant to interpret $c(j_1, s_1, A_0, s_{A_0})^T y = \infty$.) The set of all $y$ satisfying the above conditions is the same as the set $\mathcal{P}'$ in (40).*

*Moreover, the quantity $M_k^+$ in (42) can be written as a maximum of linear functions of $y$, each $M_\ell^-$ in (43) can be written as a minimum of linear functions of $y$, each $M_\ell^0$ in (44) can be written as a maximum of linear functions of $y$, and each $M_\ell^S$ in (45) is a matrix. Hence (41)–(45) can be expressed as $\Gamma y \ge 0$ for a matrix $\Gamma$. The number of rows of $\Gamma$ is upper bounded by $4pk - 2k^2 - k$.*

At first glance, Lemma 5 seems to have done little for us over the polyhedral characterization from the last subsection: again, after $k$ steps, we are faced with a $\Gamma$ matrix that has on the order of $4pk$ rows. Meanwhile, at the risk of stating the obvious, the characterization in Lemma 5 is far more succinct (i.e., the $\Gamma$ matrix is much smaller) without the conditions in (43)–(45). Indeed, in certain special cases (e.g., orthogonal predictors) these conditions are vacuous, and so they do not contribute to the formation of $\Gamma$. But even outside of such cases, we have found that dropping the conditions (43)–(45) yields a highly accurate (and computationally efficient) approximation of the LAR selection set in practice. This is discussed next.

## 5.3 A simple approximation of the polyhedral set

It is not hard to see from their definitions in Appendix A.4 that when $X$ is orthogonal (i.e., when $X^T X = I$), we have $M_\ell^- = \infty$ and $M_\ell^0 = -\infty$, and furthermore, the matrix $M_\ell^S$ has zero rows,

for each $\ell$. This means that the conditions (43)–(45) are vacuous. The polyhedral characterization in Lemma 5, therefore, reduces to $\{y : \Gamma y \geq U\}$, where $\Gamma$ has only $k+1$ rows, defined by the $k+1$ constraints (41), (42), and $U$ is a random vector with components $U_1 = \ldots = U_k = 0$, and $U_{k+1} = M_k^+(j_k, s_k, c(j_{k-1}, s_{k-1}, A_{k-2}, s_{A_{k-2}})^T y)$.

For a general (nonorthogonal) $X$, we might still consider ignoring the conditions (43)–(45) and using the compact representation $\{y : \Gamma y \geq U\}$ induced by (41), (42). This is an approximation to the exact polyhedral characterization in Lemma 5, but it is a computationally favorable one, since $\Gamma$ has only $k+1$ rows (compared to about $4pk$ rows per the construction of the lemma). Roughly speaking, the constraints in (43)–(45) are often inactive (loose) among the full collection (41)–(45), so dropping them does not change the geometry of the set. We do not pursue formal arguments to this end (beyond the orthogonal case), but we do show convincing empirical arguments later which suggest that this approximation is justified.

Hence let us suppose for the moment that we are interested in the polyhedral set $\{y : \Gamma y \geq U\}$ with $\Gamma, U$ as defined above, serving as either an exact or approximate reduced representation of the full description from Lemma 5. Our focus now is the application of our polyhedral inference tools from Section 3 to $\{y : \Gamma y \geq U\}$. Recall that the established polyhedral theory considers sets of the form $\{y : \Gamma y \geq u\}$ for $u$ fixed. As the equivalence in (10) is a deterministic rather than a distributional result, it holds whether $U$ is random or fixed. However, the independence of the constructed $\mathcal{V}^{\mathrm{lo}}, \mathcal{V}^{\mathrm{up}}, \mathcal{V}^0$ and $v^T y$ is not as immediate. The quantities $\mathcal{V}^{\mathrm{lo}}, \mathcal{V}^{\mathrm{up}}, \mathcal{V}^0$ are now functions of $y$ and $U$, both of which are random. A important special case occurs when

$$v^T y, \;\; (I - \Sigma v / v^T \Sigma v) v^T y, \;\; U \text{ are independent.}$$

In this case $\mathcal{V}^{\mathrm{lo}}, \mathcal{V}^{\mathrm{up}}, \mathcal{V}^0$—which only depend on the latter two terms above—are clearly independent of $v^T y$. To be explicit, we state this result as a corollary.

**Corollary 1** (**Polyhedral selection as truncation, random** $U$)**.** *For any fixed $y, \Gamma, U, v$ such that $v^T \Sigma v \neq 0$,*

$$\Gamma y \geq U \iff \mathcal{V}^{\mathrm{lo}}(y, U) \leq v^T y \leq \mathcal{V}^{\mathrm{up}}(y, U),\ \mathcal{V}^0(y, U) \leq 0,$$

*where*

$$\begin{aligned}
\rho &= \frac{\Gamma \Sigma v}{v^T \Sigma v} \\
\mathcal{V}^{\mathrm{lo}}(y, U) &= \max_{j : \rho_j > 0} \frac{U_j - (\Gamma y)_j + \rho_j v^T y}{\rho_j} \\
\mathcal{V}^{\mathrm{up}}(y, U) &= \min_{j : \rho_j < 0} \frac{U_j - (\Gamma y)_j + \rho_j v^T y}{\rho_j} \\
\mathcal{V}^0(y, U) &= \max_{j : \rho_j = 0} U_j - (\Gamma y)_j.
\end{aligned}$$

*Moreover, assuming that $y$ and $U$ are random, and that*

$$v^T y, \;\; (I - \Sigma v / v^T \Sigma v) v^T y, \;\; U \;\; \textit{are independent}, \tag{46}$$

*the triplet $(\mathcal{V}^{\mathrm{lo}}, \mathcal{V}^{\mathrm{up}}, \mathcal{V}^0)(y, U)$ is independent of $v^T y$.*

Under the conditions (46) on $v^T y$ and $U$, the rest of the inferential treatment proceeds just as before, since Corollary 1 ensures that we have the required alternate truncated Gaussian representation of $\Gamma y \geq U$, with the random truncation limits $\mathcal{V}^{\mathrm{lo}}, \mathcal{V}^{\mathrm{up}}$ being independent of the univariate Gaussian $v^T y$.

In our LAR problem setup, $U$ is a given random variate (as described in the first paragraph of this subsection). The relevant question is of course: when do the conditions (46) hold? Fortunately, these hold with only very minor assumptions on $v$: this vector must lie in the column space of the LAR active variables at the current step.

**Lemma 6.** *Suppose that we have run $k$ steps of LAR, and represent the conditions* (41), (42) *in Lemma 5 as $\Gamma y \geq U$. Under our running regression model $y \sim N(\theta, \sigma^2 I)$, if $v$ is in the column space of the active variables $A_k$, written $v \in \mathrm{col}(X_{A_k})$, then the independence condition in* (46) *holds, so inference for $v^T\theta$ can be carried out with the same set of tools as developed in Section 3, conditional on $\Gamma y \geq U$.*

See Appendix A.5 for a proof of this lemma. For example, if we choose the contrast vector to be $v_k = (X^+_{A_k})^T e_k$, a case we have revisited throughout the paper, then this satisfies the conditions of Lemma 6. Therefore, for testing the significance of the projected regression coefficient of the latest selected LAR variable, conditional on $\Gamma y \geq U$, we may use the p-values and intervals derived in Section 3. We walk through this usage in the next subsection.

## 5.4 The spacing test

As remarked earlier, the polyhedral characterizations of the form $\{y : \Gamma y \geq 0\}$, derived in Sections 5.1 or 5.2 (where the $\Gamma$ matrix is big, having on the order of $4pk$ rows), can be followed up with the same inferential treatment as given for FS Section 4.2. The p-values and confidence intervals laid out in this section apply directly to LAR, but they employ a different $\Gamma$ matrix.

The (approximate) representation of the form $\{y : \Gamma y \geq U\}$ derived in Section 5.3 (where $\Gamma$ is small, having $k+1$ rows), is different. As we explained in the last subsection, such a representation can only be used to conduct inference over $v^T\theta$ for certain vectors $v$, namely, those lying in the span of current active LAR variables. The particular choice of contrast vector

$$v_k = c(j_k, s_k, A_{k-1}, s_{A_{k-1}}) = \frac{P^{\perp}_{A_{k-1}} X_{j_k}}{s_k - X_{j_k}^T (X^+_{A_{k-1}})^T s_{A_{k-1}}}, \tag{47}$$

paired with the compact representation $\{y : \Gamma y \geq U\}$, leads to a very special test that we name the *spacing test*. From the definition (47), and the well-known formula for partial regression coefficients, we see that the null hypothesis being considered is

$$H_0 : \ v_k^T \theta = 0 \iff H_0 : \ e_k^T X^+_{A_k} \theta = 0,$$

i.e., the spacing test is a test for the $k$th coefficient in the multiple regression of $\theta$ on $X_{A_k}$, just as we have investigated all along under the equivalent choice of contrast vector $v_k = (X^+_{A_k})^T e_k$. The main appeal of the spacing test lies in its simplicity. Letting

$$\omega_k = \big\| (X^+_{A_k})^T s_{A_k} - (X^+_{A_{k-1}})^T s_{A_{k-1}} \big\|_2. \tag{48}$$

the one-sided spacing test statistic is defined by

$$R_k = \frac{\Phi(\lambda_{k-1} \frac{\omega_k}{\sigma}) - \Phi(\lambda_k \frac{\omega_k}{\sigma})}{\Phi(\lambda_{k-1} \frac{\omega_k}{\sigma}) - \Phi(M_k^+ \frac{\omega_k}{\sigma})}, \tag{49}$$

and the two-sided spacing test statistic is defined by

$$T_k = 2 \cdot \min\{R_k, 1 - R_k\}. \tag{50}$$

Above, $\lambda_{k-1}$ and $\lambda_k$ are the knots at steps $k-1$ and $k$ in the LAR path, and $M_k^+$ is the random variable from Lemma 5. The statistics (49) and (50) come from our polyhedral testing framework, adapted to the case of a random vector $U$ (Corollary 1 and Lemma 6). They are both valid p-values for testing $H_0 : v_k^T \theta = 0$, and have exact conditional size. We emphasize this point by stating it in a theorem.

**Theorem 1** (**Spacing test**). *Suppose that we have run $k$ steps of LAR. Represent the conditions (41), (42) in Lemma 5 as $\Gamma y \geq U$. Specifically, we define $\Gamma$ to have the following $k+1$ rows:*

$$\begin{aligned}
&\Gamma_1 = c(j_1, s_1, A_0, s_{A_0}) - c(j_2, s_2, A_1, s_{A_1}) \\
&\Gamma_2 = c(j_2, s_2, A_1, s_{A_1}) - c(j_3, s_3, A_2, s_{A_2}) \\
&\ldots \\
&\Gamma_{k-1} = c(j_{k-1}, s_{k-1}, A_k, s_{A_k}) - c(j_k, s_k, A_k, s_{A_k}) \\
&\Gamma_k = \Gamma_{k+1} = c(j_k, s_k, A_k, s_{A_k}),
\end{aligned}$$

*and $U$ to have the following $k+1$ components:*

$$\begin{aligned}
&U_1 = U_2 = \ldots = U_k = 0 \\
&U_{k+1} = M_k^+ \Big( j_k, s_k, c(j_{k-1}, s_{k-1}, A_{k-2}, s_{A_{k-2}})^T y \Big).
\end{aligned}$$

*For testing the null hypothesis $H_0 : e_k^T X_{A_k}^+ \theta = 0$, the one-sided spacing statistic $R_k$ in (48), (49), and the two-sided statistic $T_k$ in (50), both serve as exact p-values conditional on $\Gamma y \geq U$:*

$$\begin{aligned}
\mathbb{P}_{e_k^T X_{A_k}^+ \theta = 0} \Big( R_k \leq \alpha \,\Big|\, \Gamma y \geq U \Big) = \alpha \\
\mathbb{P}_{e_k^T X_{A_k}^+ \theta = 0} \Big( T_k \leq \alpha \,\Big|\, \Gamma y \geq U \Big) = \alpha,
\end{aligned}$$

*for any $0 \leq \alpha \leq 1$.*

*Remark 1.* The p-values from our polyhedral testing theory depend the truncation limits $\mathcal{V}^{\mathrm{lo}}, \mathcal{V}^{\mathrm{up}}$ (e.g., see (29) for the case of FS), which in turn depend on the polyhedral representation (again, see (27), (28) for FS). For the special polyhedron $\{y : \Gamma y \geq U\}$ considered in the theorem, it turns out that $\mathcal{V}^{\mathrm{lo}} = M_k^+$ and $\mathcal{V}^{\mathrm{up}} = \lambda_{k-1}$, which is fortuituous, as it means that no extra computation is needed to form $\mathcal{V}^{\mathrm{lo}}, \mathcal{V}^{\mathrm{up}}$ (beyond that already needed for the path and $M_k^+$). Furthermore, for the contrast vector $v_k$ in (47), it turns out that $\|v_k\|_2 = 1/\omega_k$. These two facts completely explain the spacing test statistics (49), (50), and the proof of Theorem 1 reduces to checking these facts, which is done in Appendix A.6.

*Remark 2.* The event $\Gamma y \geq U$ is not exactly equivalent to a LAR selection event at the $k$th step. Recall that, as defined, this only encapsulates the first part (41), (42) of a longer set of conditions (41)–(45) that provides the exact characterization, as explained in Lemma 5. However, in practice, we have found that (41), (42) often provide a very strong (essentially perfect) approximation to the LAR selection event. This is demonstrated empirically in the next section.

*Remark 3.* The one-sided spacing statistic in (49) tests the null hypothesis $H_0 : v_k^T \theta = 0$ versus the alternative $H_1 : v_k^T \theta > 0$, where $v_k$ is the vector in (47). As we pointed out above, $v_k^T \theta = 0$ is the same as $e_k^T X_{A_k}^+ \theta = 0$, so the null $H_0$ is a test for the last coefficient in a projected regression model of $\theta$ on the active variables. How do we interpret the one-sided alternative $H_1 : v_k^T \theta > 0$? Because $v_k^T y = \lambda_k \geq 0$, the denominator of $v_k$ in (47) must have the same sign as $X_{j_k}^T P_{A_{k-1}}^{\perp} y$, i.e., the same sign as $e_k^T X_{A_k}^+ y$. Hence

$$H_1 : \;\; v_k^T \theta > 0 \iff H_1 : \;\; \mathrm{sign}(e_k^T X_{A_k}^+ y) \cdot e_k^T X_{A_k}^+ \theta > 0,$$

i.e., the alternative hypothesis $H_1$ states that the projected population regression coefficient of the last selected variable is nonzero, and shares the sign of the sample regression coefficient of this last selected variable.

The spacing test statistics (49), (50) are very simple and concrete, but they do still depend on the random variable $M_k^+$. The quantity $M_k^+$ is computable in $O(p)$ operations (see Appendix A.4

for its definition), but it is not an output of standard software for computing the LAR path (e.g., the R package `lars`). To further simplify matters, therefore, we might consider replacing $M_k^+$ by the next knot in the LAR path, $\lambda_{k+1}$. The motivation: sometimes, but not always, $M_k^+$ and $\lambda_{k+1}$ will be equal. Altogether, as the next theorem shows, replacing $M_k^+$ by $\lambda_{k+1}$ yields an asymptotically valid test. The proof is given in Appendix A.7.

**Theorem 2** (**Asymptotically valid spacing test**). *After $k$ steps along the LAR path, define the modified one-sided spacing test statistic*

$$\tilde{R}_k = \frac{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(\lambda_k \frac{\omega_k}{\sigma})}{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(\lambda_{k+1}\frac{\omega_k}{\sigma})}, \tag{51}$$

*and the corresponding modified two-sided test statistic*

$$\tilde{T}_k = 2 \cdot \min\{\tilde{R}_k, 1 - \tilde{R}_k\}. \tag{52}$$

*Here, $\omega_k$ is as defined in (48), and $\lambda_{k-1}, \lambda_k, \lambda_{k+1}$ are the LAR knots at steps $k-1, k, k+1$ of the path, respectively. Denote by $\Gamma y \geq U$ the compact polyhedral representation at step $k+1$ of the LAR path, as described in Theorem 1 at step $k$, where $U$ has last component $M_{k+1}^+$.*

*Now consider an asymptotic regime in which $k$ is fixed as $n, p \to \infty$, and*

$$\begin{aligned} \omega_{k+1}\lambda_k &\leq \omega_k \lambda_{k-1} \\ \omega_k \lambda_k &\leq \omega_{k+1}\lambda_{k+1}, \end{aligned}$$

*with probability tending to 1, conditional on $\Gamma y \geq U$. Assume also that*

$$\frac{\Phi(M_{k+1}^+\frac{\omega_k}{\sigma}) - \Phi(M_{k+1}^+\frac{\omega_{k+1}}{\sigma})}{\left[\Phi(\lambda_{k+1}\frac{\omega_{k+1}}{\sigma}) - \Phi(M_{k+1}^+\frac{\omega_{k+1}}{\sigma})\right]\left[\Phi(\lambda_{k+1}\frac{\omega_{k+1}}{\sigma}) - \Phi(M_{k+1}^+\frac{\omega_k}{\sigma})\right]} \to 0,$$

*denoting convergence in probability, conditional on $\Gamma y \geq U$. Then $\tilde{R}_k, \tilde{T}_k$ are asymptotically valid conditional p-values for testing the null hypothesis $H_0 : e_k^T X_{A_k}^+ \theta = 0$ and $e_{k+1}^T X_{A_{k+1}}^T \theta = 0$:*

$$\begin{aligned} \mathbb{P}_0(\tilde{R}_k \leq \alpha) &\leq \alpha + o(1) \\ \mathbb{P}_0(\tilde{T}_k \leq \alpha) &\leq \alpha + o(1), \end{aligned}$$

*for any $0 \leq \alpha \leq 1$, where $o(1)$ denotes terms converging to zero, and we abbreviate*

$$\mathbb{P}_0(\cdot) = \mathbb{P}_{e_k^T X_{A_k}^+ \theta = 0,\, e_{k+1}^T X_{A_{k+1}}^T \theta = 0}\Big( \cdot \;\Big|\; \Gamma y \geq U \Big),$$

*for the null conditional probability measure.*

*Remark.* The null hypothesis in the theorem test whether both the projected regression coefficient of the variable entered at the current step $k$, and of the variable entered at the next step $k+1$, are zero. This is a stronger null than that considered by the exact finite-sample spacing test, but it is needed in the proof to control the left tail of the modified spacing statistic $\tilde{R}_k$ at step $k$; we do this by tying it to the exact spacing statistic $R_{k+1}$ at step $k+1$. An interesting relationship shown later is that the modified one-sided spacing statistic is asymptotically equivalent to the covariance statistic. The null hypothesis of the covariance test also concerns more than just the variable entered at the current step.

It is not hard to verify that the one-sided statistic in (51) is a monotone decreasing function of $\lambda_k - \lambda_{k+1}$, the spacing between LAR knots at steps $k$ and $k+1$, hence the name "spacing" test. Similarly, the exact one-sided statistic in (49) measures the magnitude of the spacing $\lambda_k - M_k^+$.

# 6 Examples

## 6.1 Conditional size of FS and LAR tests

We first examine the conditional type I error properties of the selection-adjusted FS and spacing test. We generated standard Gaussian predictors $X$ with $n = 50$, $p = 10$. We fixed true regression coefficients $\beta^* = (3.5, -2.5, 0, \ldots 0)$, and we set $\sigma^2 = 1$. We then drew observations according to $y \sim N(X\beta^*, \sigma^2 I)$, and ran FS and LAR. Both procedures chose $[1, -2]$—which we write to denote predictor 1 with a positive sign, and then predictor 2 with a negative sign—in the first two steps. Over 5000 repetitions (i.e., 5000 draws of $y$ from the regression setup), the model $[1, -2]$ was chosen about 25% and 20% of the time by the two procedures, respectively.

Figure 4 shows the conditional p-values for the third step of each method, in which we test the partial regression coefficient of the third variable to enter. The colors indicate whether the first two steps have found the correct model $[1, -2]$, or not. We see that the p-values are uniform when the first two steps have found the correct model, and smaller than uniform when the first two steps have not. This is the desired behavior, as we would want to reject at the third step, in the latter case.

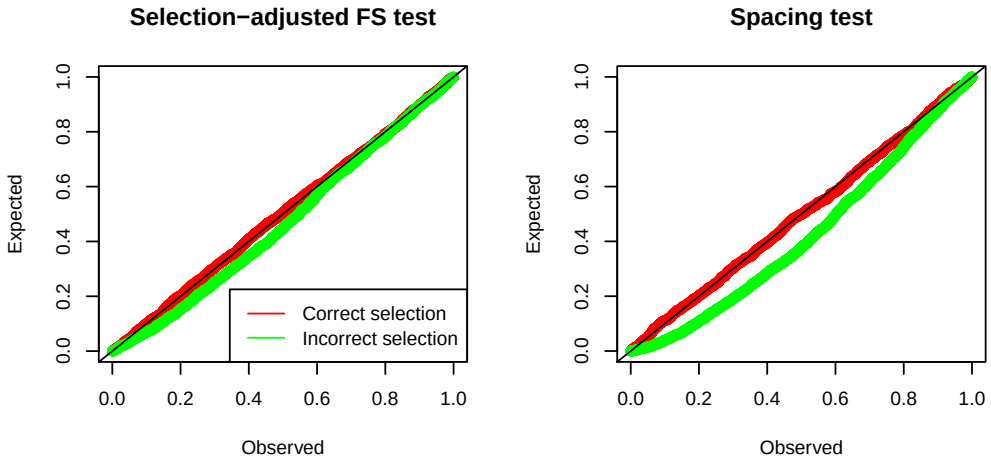


Figure 4: *Simulated data with two true active variables. Shown are p-values for the third selection step, with the colors indicating whether the first two steps have found the correct model or not.*

## 6.2 Coverage of LAR selection intervals

Now we examine the spacing test in more detail. We generated standard Gaussian predictors with $n = 100$, $p = 10$, and set $\beta^* = (6, 3, 0, \ldots 0)$ and $\sigma^2 = 0.25$. We applied LAR to $y \sim N(X\beta^*, \sigma^2 I)$, and repeated this 1000 times (i.e., over 1000 draws of $y$). The nonzero components of $\beta^*$ were large enough that the variables 1 and 2 were always chosen in this order. Table 2 shows the proportion of p-values from the spacing test greater than the nominal size $\alpha$, with $\alpha$ ranging from 0.01 to 0.20. The type I error for steps 3 to 9 is close to the nominal level in each case, as to be expected.

Figure 5 shows 90% selection intervals for the first two predictors, for the model fit at step two of LAR. The plot shows 100 repetitions, with the intervals in red indicating cases in which the true underlying coefficient value (dotted line) was not covered. The empirical coverage is close to the nominal 90% level.

For Figure 6, the setup is the same except that $\beta^* = (1, -1, 0, 0, \ldots)$. The true signal is weaker now: at the first step, predictors 1 and 2 were chosen 3% and 39% of the time, and at the second

| Nominal size | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| 0.01 | 1.000 | 0.987 | 0.012 | 0.008 | 0.009 | 0.006 | 0.004 | 0.010 | 0.007 |
| 0.025 | 1.000 | 0.993 | 0.026 | 0.020 | 0.027 | 0.017 | 0.022 | 0.023 | 0.025 |
| 0.05 | 1.000 | 0.996 | 0.052 | 0.044 | 0.052 | 0.046 | 0.042 | 0.052 | 0.046 |
| 0.10 | 1.000 | 0.997 | 0.094 | 0.090 | 0.107 | 0.097 | 0.098 | 0.098 | 0.101 |
| 0.20 | 1.000 | 0.998 | 0.212 | 0.205 | 0.235 | 0.214 | 0.205 | 0.228 | 0.226 |

Table 2: *Demonstration of the spacing test. We report the proportion of p-values less than the nominal size, ranging from 0.01 to 0.20, over 1000 repetitions. There are two truly active predictors, and we see that the size is about right from the third predictor on.*

step, it was 23% for each. The remaining selections were distributed across the other 8 predictors. The figure shows the resulting selection intervals, with the value of the appropriate underlying true coefficient indicated by a dotted line. Intervals in green cover the true value, while the red ones do not. The coverage is still about 90%, as expected.

### 6.3 Comparison to the max-$|t|$-test

This last example reveals some unique properties of the selection-adjusted FS and LAR p-values. For testing the significance of variables entered by FS, Buja & Brown (2014) suggested a test they call the max-$|t|$-test. Here is a description. At the $k$th step of FS, where $A_{k-1}$ is the current active list (with $k-1$ active variables), let

$$t_{\max}(y) = \max_{j \notin A_{k-1}} \frac{\|X_j^T P_{A_{k-1}}^{\perp} y\|_2}{\|P_{A_{k-1}} X_j\|_2 \sigma}. \tag{53}$$

As the distribution of $t_{\max}(y)$ is generally intractable, we simulate $\epsilon \sim N(0, \sigma^2 I)$, and use this to estimate null probability that $t_{\max}(\epsilon) > t_{\max}(y)$.

To compare methods, we generated Gaussian predictors $X$ with $n = 50$ and $p = 10$, from a population model with pairwise correlation $0.5^{|j-j'|}$ between dimensions $j$ and $j'$. We set $\beta^* = 0$ and $\sigma^2 = 1$, and as usual drew $y \sim N(X\beta^*, \sigma^2 I)$.

Figure 7 displays the p-values from three tests, over 1000 repetitions (1000 draws of $y$): the max-$|t|$-test in the top row, the selection-adjusted FS test in the middle, and the spacing test in the bottom. The tests are displayed at each of the first four steps of the procedure (FS in the top two rows, LAR in the bottom row). All three tests look good at the first step, but the max-$|t|$-test becomes more and more conservative for later steps. The reason is that selection-adjusted FS and spacing test p-values properly account for all selection events up to and including step $k$. To be more explicit, the max-$|t|$-test ignores the fact that at the second step, the observed $t_{\max}(y)$ is the *second largest* value of the statistic in the data, and erroneously compares it to a reference distribution of *largest* $t_{\max}(\epsilon)$ values over a smaller set. This creates a conservative bias in the p-value. Remarkably, the selection-adjusted FS and spacing tests are able carry out full conditioning on past selection events exactly.

## 7 Relationship to the covariance test

There is an interesting connection between the LAR spacing test and the covariance test of Lockhart et al. (2014). We first review the covariance test and then discuss this connection.

After $k$ steps of LAR, let $A_k$ denote the list of active variables and $s_{A_k}$ denote the sign list, the same notation as we have been using thus far. The covariance test provides a significance test for the $k$th step of LAR. More precisely, it assumes an underlying linear model $\theta = X\beta^*$, and tests the

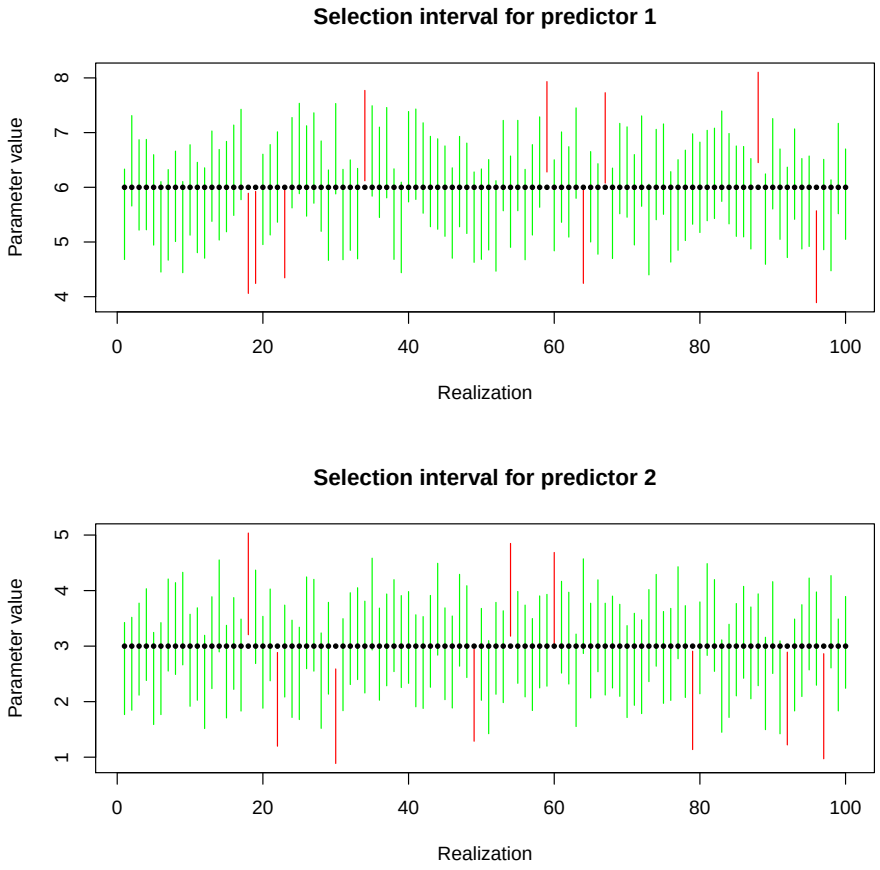


Figure 5: *Strong signal case: selection intervals for first and second predictors entered, computed at the second LAR step, across 100 realizations. In each plot, the values of the corresponding true underlying coefficients are indicated by the black dots: these are equal to 6 and 3, respectively. Green intervals cover the true value, while the red ones do not.*

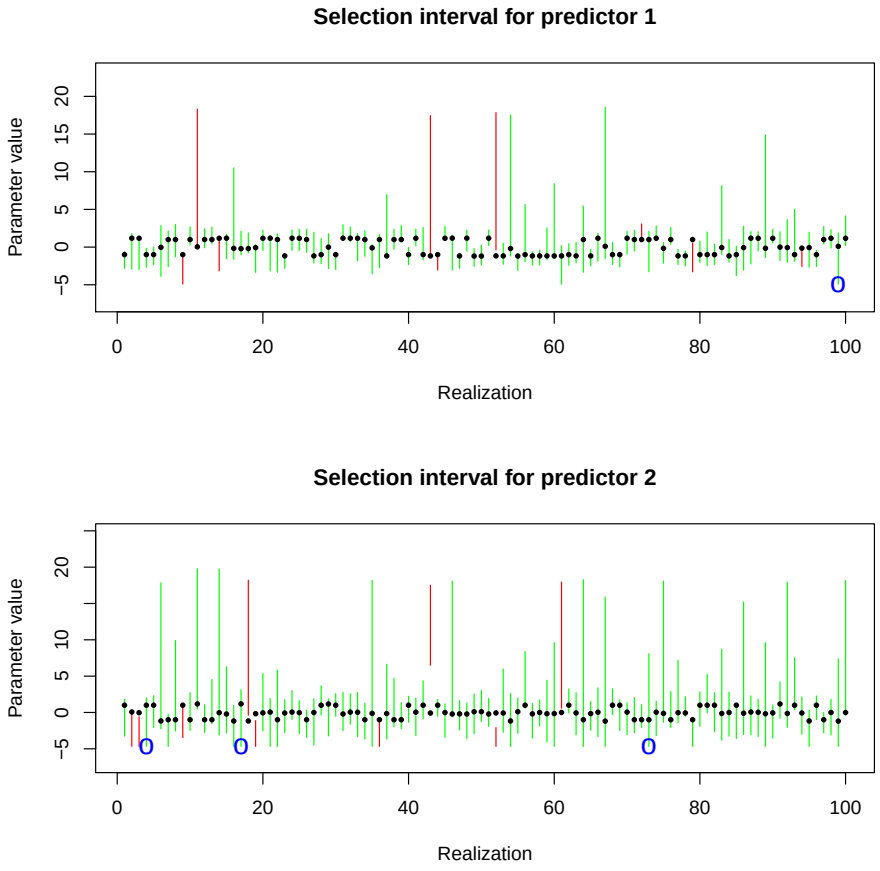


Figure 6: *Weak signal case: selection intervals for first and second predictors entered, at the second LAR step, across 100 realizations, as in Figure 5. Some endpoints of the intervals end up at* $-\infty$*, indicated by the blue circles.*

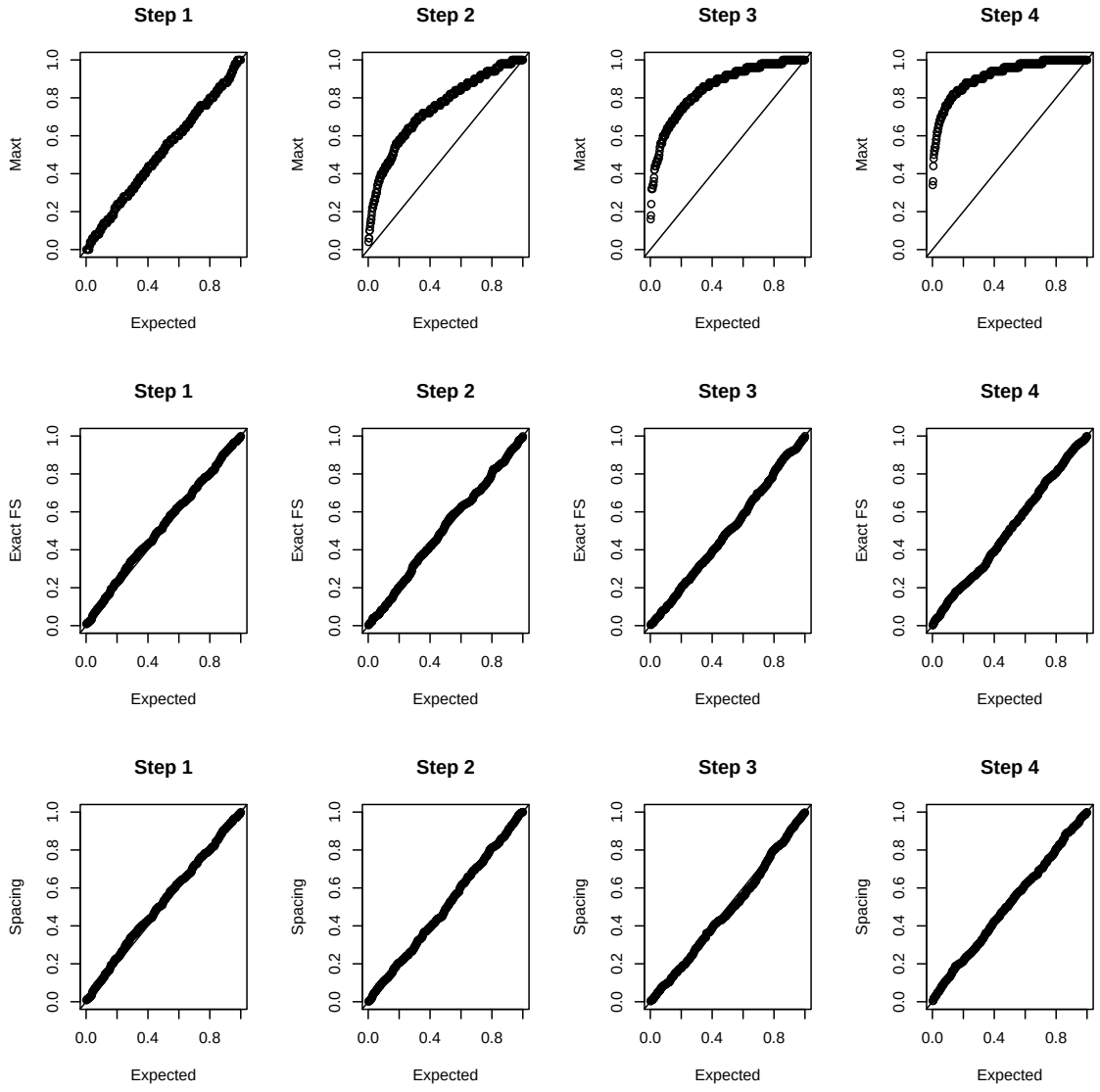


Figure 7: *Simulation of p-values for the first four steps of the max-|t|-test (top row), exact selection-adjusted FS test (middle row), and spacing test (bottom row).*

null hypothesis

$$H_0: \ A_{k-1} \supseteq \text{supp}(\beta^*),$$

where $\text{supp}(\beta^*)$ denotes the support of set of $\beta^*$ (the true active set). In words, this tests simultaneously the significance of *any variable entered at step $k$ and later*.

Though its original definition is motivated from a difference in (empirical) covariance between LAR fitted values, the covariance statistic can be written in an equivalent form that is suggestive of a connection to the spacing test. This form, at step $k$ of the LAR path, is

$$C_k = \omega_k^2 \cdot \lambda_k(\lambda_k - \lambda_{k+1})/\sigma^2, \tag{54}$$

where $\lambda_k, \lambda_{k+1}$ are the LAR knots at steps $k$ and $k+1$ of the path, and $\omega_k$ is the weight in (48). (All proofs characterizing the null distribution of the covariance statistic in Lockhart et al. (2014) use this equivalent definition.) The main result (Theorem 3) in Lockhart et al. (2014) is that under correlation restrictions on the predictors $X$ and other conditions, the covariance statistic (54) has a conversative Exp(1) limiting distribution under the null hypothesis. Roughly, they show that

$$\lim_{n,p\to\infty} \mathbb{P}_{A_{k-1}\supseteq \text{supp}(\beta^*)}\Big(C_k > t \,\Big|\, \hat{A}_k(y) = A_k,\, \hat{s}_{A_k}(y) = s_{A_k}\Big) \leq e^{-t},$$

for all $t \geq 0$.

A susprising result, perhaps, is that the covariance test in (54) and the spacing test in (51) are asymptotically equivalent. The proof for this equivalence uses relatively straightforward calculations with Mills' inequalities, and is deferred until Appendix A.8.

**Theorem 3** (**Asymptotic equivalence between spacing and covariance tests**). *After a fixed number $k$ steps of LAR, the one-sided spacing p-value in* (51) *and the covariance statistic in* (54) *are asymptotically equivalent, in the following sense. Assume an asymptotic region in which*

$$\lambda_k/\lambda_{k-1} \to 0$$
$$\lambda_k/\lambda_{k+1} \to 1,$$

*denoting convergence in probability. The spacing statistic, transformed by the inverse* Exp(1) *survival function, satisfies*

$$-\log\left(\frac{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(\lambda_k\frac{\omega_k}{\sigma})}{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(\lambda_{k+1}\frac{\omega_k}{\sigma})}\right) = \frac{\omega_k^2}{\sigma^2}\lambda_k(\lambda_k - \lambda_{k+1}) + o(1).$$

*Said differently, the asymptotic p-value of the covariance statistic, under the* Exp(1) *limit, satisfies*

$$\exp\left(-\frac{\omega_k^2}{\sigma^2}\lambda_k(\lambda_k - \lambda_{k+1})\right) = \left(\frac{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(\lambda_k\frac{\omega_k}{\sigma})}{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(\lambda_{k+1}\frac{\omega_k}{\sigma})}\right)(1 + o(1)).$$

*Above, we use $o(1)$ to denote terms converging to zero in probability.*

*Remark.* The asymptotic equivalence described in this theorem raises an interesting and unforeseen point about the one-sided nature of the covariance test. That is, the covariance statistic is seen to be asymptotically tied to the one-sided spacing p-value in (51), which, recalling Remark 3 after Theorem 1, we can interpret as testing the null hypothesis $H_0 : e_k^T X_{A_k}^+ \theta = 0$ against the one-sided alternative $H_1 : \text{sign}(e_k^T X_{A_k}^+ y) \cdot e_k^T X_{A_k}^+ \theta > 0$. Hence, the covariance test in (54) is implicitly aligned to have power when the selected variable at the $k$th step has a sign matching that of the projected population effect of this variable.

# 8 Discussion

In a regression model with Gaussian errors, we developed a method for exact inference, conditional on a polyhedral constraint on the observations $y$. Since both the FS and LAR algorithms admit polyhedral representations for their model selections events, our framework produces exact p-values and confidence intervals post model selection for either of these two adaptive regression procedures. A particularly special and simple case arises when we use our framework to test significance of the projected regression coefficient, in the population, of the latest selected variable at a given step of LAR. This leads to the spacing test, which is asymptotically equivalent to the covariance test of Lockhart et al. (2014), but (like the rest of our framework) produces exact p-values in finite samples. An R language package, implementing the proposals in this paper, will be made publicly available soon.

While this work focused mainly on FS and LAR, our results in principle extend to any method whose selection events can be characterized by polyhedral constraints on $y$. This includes the model selection events from the lasso path, i.e., the LAR path but including variable deletions, and the generalized lasso path (Tibshirani & Taylor 2011) as well. Details of this construction will appear in future work. Finally, an extension to generalized regression models is also possible. Although a rigorous treatment of this extension is beyond the scope of this paper, we broadly describe the idea in the final subsection below.

## 8.1 Extension to general likelihood-based models

The selection-adjusted forward stepwise inference can be extended to other models including generalized linear models and the proportional hazards model. The extension is simple if we focus on forward entry of predictors according to the score test.

Consider a likelihood-based regression model with natural parameter $\eta = X\beta$ and log-likelihood $\ell(\beta)$. Let $\hat{U}(\beta)$ and $\hat{\mathcal{I}}(\beta)$ be the sample score vector and information matrix, and $U(\beta)$ and $\mathcal{I}(\beta)$ be the population versions. Assuming the existence of some true coefficient vector $\beta^*$, under standard asymptotics with $p$ fixed and $n \to \infty$, we have

$$\hat{U}(\beta^*) \xrightarrow{d} N\big(0, \mathcal{I}(\beta^*)\big).$$

In finite samples, we can use a normal approximation to the score to build a sequential p-value procedure. At the first step, the usual score test enters the variable $j_1$ with sign $s_1$ that maximizes $s \cdot \hat{U}_j(0)/\hat{\mathcal{I}}_j(0)^{1/2}$ over all variables $j$ and signs $s$. Suppose that wish to test the hypothesis $\beta^*_{j_1} = 0$, by testing that $U_{j_1}(0) = 0$. To do this, we use the approximation

$$\hat{U}(0) \sim N\big(0, \hat{\mathcal{I}}(0)\big), \tag{55}$$

and apply the polyhedral selection and inference lemmas, with constraints of the form

$$s_1 c_{j_1}^T \hat{U}(0) \geq \pm c_j^T \hat{U}(0) \quad \text{for all } j \neq j_1.$$

Above, $c_j$ contains $\hat{\mathcal{I}}_j(0)^{-1/2}$ in the $j$th component and is 0 in all other components. These constraints simply express the fact that variable $j_1$ yields the largest value of the score test.

At a future step $k > 1$, write $j_k$ for the predictor entered by the score test, and write $\hat{\beta}^{(k)}$ for the coefficient estimate with active variables fixed at their MLEs, and all others (including $j_k$) set to zero. Suppose that we wish to test $\beta^*_{j_k} = 0$, i.e., to test $U_{j_k}(\hat{\beta}^{(k)}) = 0$. We use an approximation

$$U_{j_k}(\hat{\beta}^{(k)}) \sim N\Big(0, \hat{\mathcal{I}}_{j_k,-j_k} \hat{\mathcal{I}}^{-1}_{-j_k,-j_k} \hat{\mathcal{I}}_{-j_k,j_k}\Big). \tag{56}$$

All sample information matrices above are to be evaluated at 0, which we omit for notational simplicity. Note that, unlike the forward stepwise construction in the ordinary regression case, the

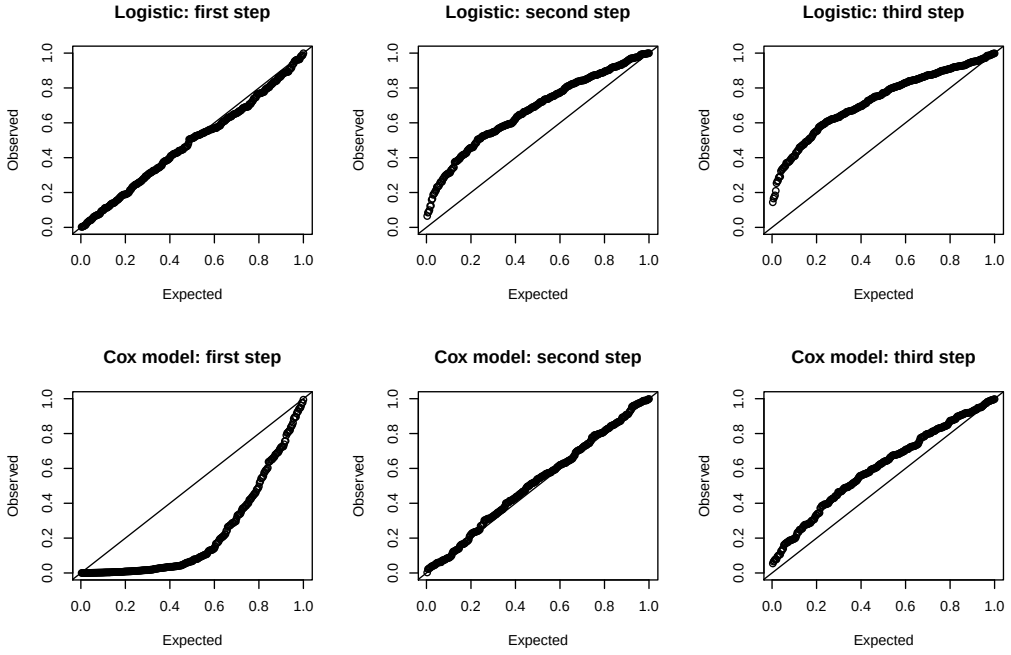


Figure 8: *Simulation of p-values for logistic regression in global null setting (top panels), and for Cox's proportional hazards model in a setting with one strong predictor (bottom panels).*

normal approximation here is only really valid for the score evaluated at $\hat{\beta}^{(k)}$ at a general step $k$. Effectively, this means that we can only condition on the selection at the current step, not those in all previous steps. As a result, the inferences drawn tend to become conservative as we go further into the null regime, as the next example shows.

Figure 8 considers simulated data with $n = 50$ and $p = 4$. In the top panels we have applied logistic regression to a setting where all true coefficients are zero; in the bottom panels, we applied Cox's proportional hazards model in a setting with one large coefficient. Over 500 repetitions, we see that the p-values from the selection-adjusted test shows good agreement with $\text{Unif}(0, 1)$ in the first null step, in each case (step 1 for the logistic model, step 2 for the Cox model). In subsequent steps it becomes conservative. Since we are relying on approximate normality in (55) and (56), the type I error will not be exact in finite samples, but rather only correct asymptotically. Some theory is needed to work out the details.

## Acknowledgements

We would like to thank Andreas Buja, Max Grazier G'Sell, Alessandro Rinaldo, and Larry Wasserman for helpful comments and discussion. We would also like to thank the editors and referees whose comments led to a complete overhaul of this paper! 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 Proofs

## A.1 Proof of Lemma 1

Rewrite the inequality $\Gamma y \geq u$ as

$$\Gamma\bigg(\frac{\Sigma v v^T}{v^T\Sigma v}y + \Big(I - \frac{\Sigma v v^T}{v^T\Sigma v}\Big)y\bigg) \geq u.$$

Note that when $\Sigma = I$ this is just decomposing up $y$ into the projection onto $v$ and its orthogonal subspace. For general $\Sigma$, it is a kind of Mahalanobis projection. Rearranging the above, and defining $\rho = \Gamma\Sigma v / v^T\Gamma v$ as in the lemma, we get

$$\rho v^T y \geq u - \Gamma y + \rho v^T y.$$

This is actually a set of $n$ inequalities, one for each component, which we can express as

$$\begin{aligned}
v^T y &\geq \frac{u_j - (\Gamma y)_j + \rho_j v^T y}{\rho_j} \quad \text{for } \rho_j > 0 \\
v^T y &\leq \frac{u_j - (\Gamma y)_j + \rho_j v^T y}{\rho_j} \quad \text{for } \rho_j < 0 \\
0 &\geq u_j - (\Gamma y)_j \quad \text{for } \rho_j < 0,
\end{aligned}$$

or more concisely

$$\begin{aligned}
v^T y &\geq \mathcal{V}^{\text{lo}} = \max_{j:\rho_j>0} \frac{u_j - (\Gamma y)_j + \rho_j v^T y}{\rho_j} \\
v^T y &\leq \mathcal{V}^{\text{up}} = \min_{j:\rho_j<0} \frac{u_j - (\Gamma y)_j + \rho_j v^T y}{\rho_j} \\
0 &\geq \mathcal{V}^0 = \max_{j:\rho_j=0} u_j - (\Gamma y)_j.
\end{aligned}$$

Finally, note that $\mathcal{V}^{\text{lo}}, \mathcal{V}^{\text{up}}, \mathcal{V}^0$ are all functions of $y - \Sigma v v^T y / v^T\Sigma v$, which is independent of $v^T y$ assuming that $y \sim N(\theta, \Sigma)$.

## A.2 Proof of Lemma 2

By a classic result, we know that if $Z \sim N(\mu, \sigma^2)$ truncated to lie in between $[a, b]$, where $a, b$ are fixed bounds, and $F^{[a,b]}_{\mu,\sigma^2}$ is its CDF, then $F^{[a,b]}_{\mu,\sigma^2}(Z) \sim \text{Unif}(0, 1)$. From this it follows that

$$\mathbb{P}\Big(F^{[a,b]}_{v^T\theta, v^T\Sigma v}(v^T y) \leq \alpha \;\Big|\; v^L \leq v \leq v^U\Big) = \alpha,$$

for any $0 \leq \alpha \leq 1$, and fixed bounds $v^L, v^U$. Therefore, integrating over $\mathcal{V}^{\text{lo}} = v^L$, $\mathcal{V}^{\text{up}} = v^U$, and $\mathcal{V}^0 \leq 0$, and using the independence of $\mathcal{V}^{\text{lo}}, \mathcal{V}^{\text{up}}, \mathcal{V}^0$ and $v^T y$ from Lemma 1, gives the result in the lemma.

## A.3 Proofs of Lemmas 3 and 4

First we discuss Lemma 3. The fact that the statistic in (17) is conditionally uniform under the null, in (18), is a direct result of the pivot statement (16) in Lemma 2 (and the fact that $U \sim \text{Unif}(0, 1)$ implies $1 - U \sim \text{Unif}(0, 1)$). The coverage result in (20) relies on the monotonicity of the truncated

normal survival function $1 - F^{[a,b]}_{\mu,\sigma^2}(x)$ as a function of $\mu$. (See, e.g., the appendix of Lee et al. (2013) for a proof.) From the definition of $\delta_\alpha$ in (19),

$$\begin{aligned} v^T\theta \geq \delta_\alpha &\iff 1 - F^{[\mathcal{V}^{\mathrm{lo}},\mathcal{V}^{\mathrm{up}}]}_{v^T\theta, v^T\Sigma v}(v^T y) \geq 1 - F^{[\mathcal{V}^{\mathrm{lo}},\mathcal{V}^{\mathrm{up}}]}_{\delta_\alpha, v^T\Sigma v}(v^T y) \\ &\iff 1 - F^{[\mathcal{V}^{\mathrm{lo}},\mathcal{V}^{\mathrm{up}}]}_{v^T\theta, v^T\Sigma v}(v^T y) \geq \alpha, \end{aligned}$$

the latter event having conditional probability $1 - \alpha$.

As for the Lemma 4, that the two-sided statistic in (21) is conditionally uniform under the null in (22) follows again from the pivot (16), as well as the easily verifiable fact that $U \sim \mathrm{Unif}(0,1)$ implies $2 \cdot \min\{U, 1-U\} \sim \mathrm{Unif}(0,1)$. The confidence interval associated with (21) is the set of all contrast values $v^T\theta'$ for which the p-value for testing $H_0 : v^T\theta = v^T\theta'$ exceeds the nominal level $\alpha$. As $2 \cdot \min\{U, 1-U\} \geq \alpha \iff \alpha/2 \leq 1 - U \leq 1 - \alpha/2$, and the truncated normal survival function is monotone in its mean argument, we arrive at the interval construction in (23), (24), (25).

## A.4 Proof of Lemma 5

We first consider an alternative characterization for the optimality of a pair $(j_k, s_k)$ at an iteration $k$ of LAR. This characterization is already proved in Lemma 7 of Lockhart et al. (2014), and we present it here for reference.

**Lemma 7** (Lemma 7 of Lockhart et al. 2014)**.** *Consider an iteration $\ell$ of LAR, and define*

$$\Sigma_{j,j'} = c(j, s, A_{\ell-1}, s_{A_{\ell-1}})^T c(j', s', A_{\ell-1}, s_{A_{\ell-1}}),$$

*for variables $j, j' \notin A_{\ell-1}$ and signs $s, s'$. (Note that $\Sigma_{j,j'}$ actually depends on $s, s'$, but we suppress this notationally for brevity.) Also define*

$$\begin{aligned} S^+_\ell(j,s,\rho) &= \Big\{(j',s') : j' \notin A \cup \{j\},\ 1 - \Sigma_{j,j'}/\Sigma_{jj} > 0,\ c(j', s', A_{\ell-1}, s_{A_{\ell-1}})^T y \leq \rho\Big\} \\ S^-_\ell(j,s,\rho) &= \Big\{(j',s') : j' \notin A \cup \{j\},\ 1 - \Sigma_{j,j'}/\Sigma_{jj} < 0,\ c(j', s', A_{\ell-1}, s_{A_{\ell-1}})^T y \leq \rho\Big\} \\ S^0_\ell(j,s,\rho) &= \Big\{(j',s') : j' \notin A \cup \{j\},\ 1 - \Sigma_{j,j'}/\Sigma_{jj} = 0,\ c(j', s', A_{\ell-1}, s_{A_{\ell-1}})^T y \leq \rho\Big\}, \end{aligned}$$

*and*

$$\begin{aligned} M^+_\ell(j,s,\rho) &= \max_{(j',s') \in S^+_\ell(j,s,\rho)} \frac{c(j', s', A_{\ell-1}, s_{A_{\ell-1}})^T y - (\Sigma_{j,j'}/\Sigma_{j,j}) \cdot c(j, s, A_{\ell-1}, s_{A_{\ell-1}})^T y}{1 - (\Sigma_{j,j'}/\Sigma_{j,j})} \\ M^-_\ell(j,s,\rho) &= \min_{(j',s') \in S^-_\ell(j,s,\rho)} \frac{c(j', s', A_{\ell-1}, s_{A_{\ell-1}})^T y - (\Sigma_{j,j'}/\Sigma_{j,j}) \cdot c(j, s, A_{\ell-1}, s_{A_{\ell-1}})^T y}{1 - (\Sigma_{j,j'}/\Sigma_{j,j})} \\ M^0_\ell(j,s,\rho) &= \max_{(j',s') \in S^0_\ell(j,s,\rho)} c(j', s', A_{\ell-1}, s_{A_{\ell-1}})^T y - (\Sigma_{j,j'}/\Sigma_{j,j}) \cdot c(j, s, A_{\ell-1}, s_{A_{\ell-1}})^T y. \end{aligned}$$

*Then LAR selects $j_\ell$ and $s_\ell$ at iteration $\ell$ if and only if*

$$c(j_\ell, s_\ell, A_{\ell-1}, s_{A_{\ell-1}})^T y \leq \lambda_{\ell-1}, \tag{57}$$

$$c(j_\ell, s_\ell, A_{\ell-1}, s_{A_{\ell-1}})^T y \geq 0, \tag{58}$$

$$c(j_\ell, s_\ell, A_{\ell-1}, s_{A_{\ell-1}})^T y \geq M^+_\ell(j_\ell, s_\ell, \lambda_{\ell-1}), \tag{59}$$

$$c(j_\ell, s_\ell, A_{\ell-1}, s_{A_{\ell-1}})^T y \leq M^-_\ell(j_\ell, s_\ell, \lambda_{\ell-1}), \tag{60}$$

$$0 \geq M^0_\ell(j_\ell, s_\ell, \lambda_{\ell-1}). \tag{61}$$

*Further, the triplet $(M^+_\ell(j_\ell, s_\ell), M^-_\ell(j_\ell, s_\ell), M^0_\ell(j_\ell, s_\ell))$ is independent of $c(j_\ell, s_\ell, A_{\ell-1}, s_{A_{\ell-1}})^T y$, for fixed $j_\ell, s_\ell$.*

We now recall another result of Lockhart et al. (2014) that is important in our context.

**Lemma 8** (Lemma 9 of Lockhart et al. 2014)**.** *At any iteration $\ell$ of LAR, we have*

$$M_\ell^+(j_\ell, s_\ell, \lambda_{\ell-1}) \leq c(j_{\ell+1}, s_{\ell+1}, A_\ell, s_{A_\ell})^T y.$$

Finally, then, to build a list of constraints that are equivalent to the LAR algorithm selecting variables $A_k$ and signs $s_{A_k}$ through step $k$, we intersect conditions (57)–(61) from Lemma 7 over $\ell = 1, \ldots k$. Notice that for each $\ell < k$, the inequality in (59) can be dropped, because it is implied by (57) and the result of Lemma 8,

$$c(j_\ell, s_\ell, A_{\ell-1}, s_{A_{\ell-1}})^T y \geq c(j_{\ell+1}, s_{\ell+1}, A_\ell, s_{A_\ell})^T y \geq M^+\Big(j_\ell, s_\ell, c(j_\ell, s_\ell, A_{\ell-1}, s_{A_{\ell-1}})^T y\Big).$$

This gives rise to conditions (41)–(44) in Lemma 5. The last condition (45) is needed to specify the sets $S_\ell$, $\ell = 1, \ldots k$, and its construction follows that given in Section 4.1.

## A.5 Proof of Lemma 6

Recall that we assume $y \sim N(\theta, \sigma^2 I)$. One can check from its definition in Lemma 7 that $M_k^+$ is a maximum over linear functions of $y$ that are orthogonal to $\mathrm{col}(X_{A_k})$, the span of active variables. If $v \in \mathrm{col}(A_k)$, then these latter linear functions, $v^T y$, and $(I - v/\|v\|_2^2)v^T y$ are orthogonal Gaussian random variables, implying independence.

## A.6 Proof of Theorem 1

As pointed out in Remark 1 after the theorem, we must only prove that $\mathcal{V}^{\mathrm{lo}} = M_k^+$, $\mathcal{V}^{\mathrm{up}} = \lambda_{k-1}$, and $\|v_k\|_2 = 1/\omega_k$, and then the result follows from Corollary 1 and the polyhedral inference lemmas from Section 3. Well, $\mathcal{V}^{\mathrm{lo}}$ and $\mathcal{V}^{\mathrm{up}}$ are the maximum and minimum of the quantities

$$\frac{U_j - (\Gamma P_{v_k^\perp} y)_j}{(\Gamma v_k)_j} \cdot \|v_k\|_2^2, \tag{62}$$

over all $j$ such that $(\Gamma v_k)_j > 0$ and $(\Gamma v_k)_j < 0$, respectively. As defined in (47), $v_k$ is proportional to $P_{A_{k-1}}^\perp$. The first $k-1$ rows of $\Gamma$ are contained in $\mathrm{col}(X_{A_{k-1}})$, so $(\Gamma v_k)_j = 0$ for $j = 1, \ldots k-1$. For the $k$th row, we have $(\Gamma v_k)_k = -\|v_k\|_2^2$, and for the $(k+1)$st row, we have $(\Gamma v_k)_{k+1} = \|v_k\|_2^2$. The quantities $\mathcal{V}^{\mathrm{up}}$ and $\mathcal{V}^{\mathrm{lo}}$, therefore, are singularly defined in terms of the $k$th and $(k+1)$st rows. As $U_k = 0$ and $(\Gamma P_{v_k^\perp} y)_k = \lambda_{k-1}$, we see from (62) that $\mathcal{V}^{\mathrm{up}} = \lambda_{k-1}$. Similarly, as $U_{k+1} = M_k^+$ and $(\Gamma P_{v_k^\perp} y)_{k+1} = 0$, we see from (62) that $\mathcal{V}^{\mathrm{lo}} = M_k^+$.

Lastly, the result $\|v_k\|_2 = 1/\omega_k$ can be seen by direct calculation, and is given in Lemma 10 of Lockhart et al. (2014).

## A.7 Proof of Theorem 2

By Lemma 5 (Lemma 7 of Appendix A.4), we know that $\lambda_{k+1} \geq M_{k+1}^+$. Hence, since the truncated Gaussian survival function is monotone increasing in its lower truncation limit, we have

$$\tilde{R}_k = \frac{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(\lambda_k \frac{\omega_k}{\sigma})}{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(\lambda_{k+1}\frac{\omega_k}{\sigma})} \geq \frac{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(\lambda_k \frac{\omega_k}{\sigma})}{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(M_{k+1}^+\frac{\omega_k}{\sigma})}.$$

Since $\lambda_{k-1}\omega_k \geq \lambda_k \omega_{k+1}$ with probability tending to 1 by assumption, and the truncated Gaussian survival function is monotone increasing in its upper truncation limit, we have

$$\frac{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(\lambda_k \frac{\omega_k}{\sigma})}{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(M_{k+1}^+\frac{\omega_k}{\sigma})} \geq \frac{\Phi(\lambda_k\frac{\omega_{k+1}}{\sigma}) - \Phi(\lambda_k \frac{\omega_k}{\sigma})}{\Phi(\lambda_k\frac{\omega_{k+1}}{\sigma}) - \Phi(M_{k+1}^+\frac{\omega_k}{\sigma})},$$

with probability tending to 1. Also, $<$ because $\lambda_k\omega_k \leq \lambda_{k+1}\omega_{k+1}$ with probability tending to 1 by assumption, it follows that

$$\frac{\Phi(\lambda_k\frac{\omega_{k+1}}{\sigma}) - \Phi(\lambda_k\frac{\omega_k}{\sigma})}{\Phi(\lambda_k\frac{\omega_{k+1}}{\sigma}) - \Phi(M^+_{k+1}\frac{\omega_k}{\sigma})} \geq \frac{\Phi(\lambda_k\frac{\omega_{k+1}}{\sigma}) - \Phi(\lambda_{k+1}\frac{\omega_{k+1}}{\sigma})}{\Phi(\lambda_k\frac{\omega_{k+1}}{\sigma}) - \Phi(M^+_{k+1}\frac{\omega_k}{\sigma})},$$

with probability tending to 1. Therefore we may write

$$\tilde{R}_k \geq R_{k+1} + W,$$

with probability tending to 1, where $R_{k+1}$ is the usual (unmodified) spacing statistic at step $k+1$, and $W$ is the remainder term

$$\frac{\Phi(M^+_{k+1}\frac{\omega_k}{\sigma}) - \Phi(M^+_{k+1}\frac{\omega_{k+1}}{\sigma})}{\big[\Phi(\lambda_{k+1}\frac{\omega_{k+1}}{\sigma}) - \Phi(M^+_{k+1}\frac{\omega_{k+1}}{\sigma})\big]\big[\Phi(\lambda_{k+1}\frac{\omega_{k+1}}{\sigma}) - \Phi(M^+_{k+1}\frac{\omega_k}{\sigma})\big]}.$$

The key point is that the statistic $R_{k+1}$ is $\mathrm{Unif}(0,1)$ under the null hypothesis. This, and the fact that $W$ converges in probability to zero by assumption, implies

$$\mathbb{P}_0(\tilde{R}_k \leq \alpha) \leq \mathbb{P}_0(R_{k+1} + W \leq \alpha) + o(1) \leq \alpha + o(1),$$

for any $0 \leq \alpha \leq 1$. This establishes the one-sided asymptotic result.

As for the two-sided result, notice that

$$\mathbb{P}_0(\tilde{T}_k \leq \alpha) = \mathbb{P}_0(\tilde{R}_k \leq \alpha/2) + \mathbb{P}_0(1 - \tilde{R}_k \leq \alpha/2).$$

The first term is bounded by $\alpha/2 + o(1)$, and the second term is bounded by $\mathbb{P}_0(1 - R_k \leq \alpha/2) = \alpha/2$, because $M^+_k \leq \lambda_{k+1}$ by Lemma 8 in Appendix A.4, and the truncated Gaussian CDF is monotone decreasing in its lower truncation limit, so that $1 - \tilde{R}_k \geq 1 - R_k$.

## A.8 Proof of Theorem 3

We begin with a helpful lemma based on Mills' inequalities.

**Lemma 9.** *Suppose that $x, y \to \infty$ but $x/y \to 0$. Then, for $\Phi$ the standard normal CDF and $\phi$ the standard normal density,*

$$\big(\Phi(y) - \Phi(x)\big) \cdot \frac{x}{\phi(x)} \to 1.$$

*Proof.* Write

$$\Phi(y) - \Phi(x) = 1 - \Phi(x) - \big(1 - \Phi(y)\big),$$

and apply Mills' inequality to each of the survival terms individually, yielding

$$\frac{\phi(x)}{x}\frac{1}{1+1/x^2} - \frac{\phi(y)}{y} \leq \Phi(y) - \Phi(x) \leq \frac{\phi(x)}{x} - \frac{\phi(y)}{y}\frac{1}{1+1/y^2}.$$

Multiplying through by $x/\phi(x)$, and using the assumption that $x/y \to 0$, gives the result. □

Now as

$$\tilde{R}_k = \frac{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(\lambda_k\frac{\omega_k}{\sigma})}{\Phi(\lambda_{k-1}\frac{\omega_k}{\sigma}) - \Phi(\lambda_{k+1}\frac{\omega_k}{\sigma})},$$

and we assume that $\lambda_k/\lambda_{k-1} \to 0$ in probability, we can utilize Lemma 9 to write

$$\tilde{R}_k = (1 + o(1)) \cdot \frac{\phi(\lambda_k\frac{\omega_k}{\sigma})}{\phi(\lambda_{k+1}\frac{\omega_k}{\sigma})} \cdot \frac{\lambda_{k+1}}{\lambda_k},$$

where $o(1)$ denotes a term converging to zero in probability. Thus

$$\begin{aligned} -\log(\tilde{R}_k) &= \frac{\omega_k^2}{\sigma^2}\frac{\lambda_k^2 - \lambda_{k+1}^2}{2} - \log\Big(\frac{\lambda_{k+1}}{\lambda_k}\Big) + o(1) \\ &= \frac{\omega_k^2}{\sigma^2}\frac{\lambda_k(\lambda_k - \lambda_{k+1})}{2} + \frac{\omega_k^2}{\sigma^2}\frac{\lambda_{k+1}(\lambda_k - \lambda_{k+1})}{2} - \log\Big(\frac{\lambda_{k+1}}{\lambda_k}\Big) + o(1). \end{aligned}$$

Using the assumption that $\lambda_k/\lambda_{k+1} \to 1$ in probability, this becomes

$$-\log(\tilde{R}_k) = \frac{\omega_k^2}{\sigma^2}\lambda_k(\lambda_k - \lambda_{k+1}) + o(1),$$

as desired.

# References


Berk, R., Brown, L., Buja, A., Zhang, K. & Zhao, L. (2013), 'Valid post-selection inference', *Annals of Statistics* **41**(2), 802–837.

Buhlmann, P. (2012), Statistical significance in high-dimensional linear models. arXiv: 1202.1377.

Buja, A. & Brown, L. (2014), 'Discussion: A significance test for the lasso', *The Annals of Statistics* **42**(2), 509–517.

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

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.

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.

Lee, J., Sun, D., Sun, Y. & Taylor, J. (2013), Exact post-selection inference with the lasso. arXiv: 1311.6238.

Lockhart, R., Taylor, J., Tibshirani, R. J. & Tibshirani, R. (2014), 'A significance test for the lasso', **42**(2), 413–468.

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

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

Tibshirani, R. J. (2013), 'The lasso problem and uniqueness', *Electronic Journal of Statistics* **7**, 1456–1490.

Tibshirani, R. J. & Taylor, J. (2011), 'The solution path of the generalized lasso', *Annals of Statistics* **39**(3), 1335–1371.

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

Wasserman, L. & Roeder, K. (2009), 'High-dimensional variable selection', *Annals of Statistics* **37**(5), 2178–2201.

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