# Summary and discussion of: "Exact Post-selection Inference for Forward Stepwise and Least Angle Regression"

Statistics Journal Club 36-825

Jisu Kim and Veeranjaneyulu Sadhanala

## 1 Introduction

In this report we summarize the recent paper [Taylor et al., 2014] which proposes new inference tools for methods that perform variable selection and estimation in an adaptive regression. Although this paper mainly studies forward stepwise regression (FS) and least angle regression (LAR), the approach in this paper is not limited to these cases. This paper describes how to carry out exact inference after any polyhedron selection mechanism, and the approach is applicable to both FS and LAR.

We introduce the problem at a high-level now. Suppose observations $y \in \mathbb{R}^n$ are drawn from a Gaussian model,

$$y = \theta + \epsilon, \epsilon \sim N(0, \sigma^2 I).$$

where $\theta$ is a function(not necessarily linear) of a fixed matrix $X \in \mathbb{R}^{n \times p}$ of predictor variables. Suppose that an adaptive variable selection and estimation procedure such as FS is performed up to step $k$ and we are interested in testing whether the $k^{th}$ variable entering the model is significant or not, i.e., whether $\theta$ depends on $k^{th}$ variable or not.

Classical $t$-test has been widely used for such testing, but it does not work for our adaptive regression setting. To see this, consider the variable selected after the first step of FS, assuming that $\theta(X) = X\beta$ for $\beta$. Let $j_1$ be the index of first entered variable. In the $t$-test we are in fact testing

$$H_0 : \beta_{j_1} = 0 \text{ vs } H_1 : \beta_{j_1} \neq 0$$

since $t$-test assumes that the tested variable is fixed a priori, regardless of observation $y$. Then, $t$-test is equivalent to $\chi^2_1$ test, with the test statistic

$$W = (RSS_\emptyset - RSS_{j_1})/\sigma^2,$$

where $RSS_\emptyset$ is the residual sum of squares($RSS$) for empty model, and $RSS_{j_1}$ is the $RSS$ for the model using predictor variable $X_{j_1}$. However, this comparison makes sense only when $y$ is such that the first entered variable is $X_{j_1}$, that is, when $\hat{j}_1(y) = j_1$, where $\hat{j}_1(y)$ is the index of the first entered predictor variable, as a function of $y$. When $\hat{j}_1(y) = i \neq j_1$, i.e. when

the first entered varaible is different from $j_1$, then the test statistic is $(RSS - RSS_i)/\sigma^2$. In fact, forward stepwise procedure always chooses $X_i$ that maximizes $(RSS - RSS_i)/\sigma^2$, so

$$W = \max_i (RSS - RSS_i)/\sigma^2.$$

Hence

$$W \geq (RSS - RSS_{j_i})/\sigma^2,$$

and usually $\gg$ holds, so using the distribution of $(RSS - RSS_{j_i})/\sigma^2$ to test $W$ will always give lower p-value than its actual p-value. Hence although it is quite widely accepted to use classical $t$-test or $\chi^2_1$ test for model selection, those classical tests will always underestimate p-value and will tend to choose a more complex model beyond optimal level in adaptive regression.

A possible remedy for the above problem is to not consider all $y$, but only those that give $j_1$ as the first selected variable, that is, $\hat{j}_1(y) = j_1$. Hence we would like to test

$$H_0 : \beta_{j_1} = 0 \text{ vs } H_1 : \beta_{j_1} \neq 0$$

conditioned on $y \in \mathcal{P} := \{y : \hat{j}_1(y) = j_1\}$. In the case of FS, the set $\mathcal{P}$ is $\{y : RSS_{j_1} \leq RSS_i, \forall i\}$ which is polyhedral. It turns out that such a set is polyhedral in the case of LAR algorithm too. This paper develops methods for hypothesis testing and finding confidence intervals with $y$ conditioned on a polyhedral set. We summarize the main results of the paper in the next section.

# 2 Summary of Results

We start with the results for the general polyhedral set selection and then summarize how they are applied to FS and LAR. In particular, we briefly describe the spacing test which seems to be practically useful as it is applicable to many state-of-the-art algorithms and is computationally efficient.

## 2.1 An equivalent representation of a polyhedral set

Consider the 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 inequality is interpreted elementwise. The following lemma from the paper provides an alternative representation for $\mathcal{P}$ which is key to developing the test statistic.

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

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

*where*

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

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

*Proof.* For simplicity, suppose $\Sigma = I$ and see Figure 1 sourced from the paper. Decompose $y = P_v y + P_{v^\perp} y$, where $P_v y = vv^T y/\|v\|_2^2$ is the projection of $y$ along $v$, and $P_{v^\perp} y = y - P_v y$ is the projection onto orthocomplement of $v$. Accordingly, we view $y$ as a deviation from $P_{v^\perp} y$, of amount $v^T y$, along the line determined by $v$. Then starting from $P_{v^\perp} y$, condition $v^T y \leq \mathcal{V}^{\text{up}}(y)$ describes how far $P_v y$ can deviate along the direction of $v$. before $y$ leaves the polygon. Similarily, $v^T y \geq \mathcal{V}^{\text{lo}}(y)$ corresponds to how far $P_v y$ can deviate along the direction of $-v$. And lastly, $\mathcal{V}^{\text{up}}(y) \leq 0$ comes from faces that are parallel to $v$, which is just to check that $y$ lies on the correct side of these faces.

Furthermore, $\mathcal{V}^{\text{lo}}(y)$, $\mathcal{V}^{\text{up}}(y)$, $\mathcal{V}^0(y)$ depends only on $P_{v^\perp}(y)$, and $P_{v^\perp}(y)$ is independent of $P_v(y)$, so $(\mathcal{V}^{\text{lo}}, \mathcal{V}^{\text{up}}, \mathcal{V}^0)(y)$ is independent of $v^T y$, which will be a key ingredient for Lemma 2 in the paper. □

## 2.2 Statistic for testing and pivoting

We would like to test whether the variable entering into the model at the $k$th step is significant, given the variables selected and their signs in the previous steps. An intuitively reasonable test for that would be to test if $(X_{A_k})^+ y = 0$. The authors design a general hypothesis test with $H_0 : v^T y = 0$ conditioned on $y \in \mathcal{P}$. From Lemma 1, this is equivalent to testing:

$$v^T y = 0 | \mathcal{V}^{\text{lo}}(y) \leq v^T y \leq \mathcal{V}^{\text{up}}(y), \mathcal{V}^0(y) \leq 0$$

Thus distribution of $v^T y | y \in \mathcal{P}$ is a truncated Gaussian with random truncation limits. Note that for any random variable $X$ with cdf $F$, $F(X)$ follows uniform distribution. The cdf of $X$ which follows $N(\mu, \sigma^2)$ truncated to the interval $[a, b]$ is given by

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

Therefore $F_{\mu,\sigma^2}^{[a,b]}(X)$ follows uniform distribution. The mean and variance for $v^T y$ are $v^T\theta$ and $v^T\Sigma v$ respectively. So, if $\mathcal{V}^{\text{up}}(y), \mathcal{V}^{\text{lo}}(y), \mathcal{V}^0(y)$ were not random, then

$$F_{v^T\theta, v^T\Sigma v}^{[\mathcal{V}^{\text{lo}}(y), \mathcal{V}^{\text{up}}(y)]}(v^T y) \tag{1}$$

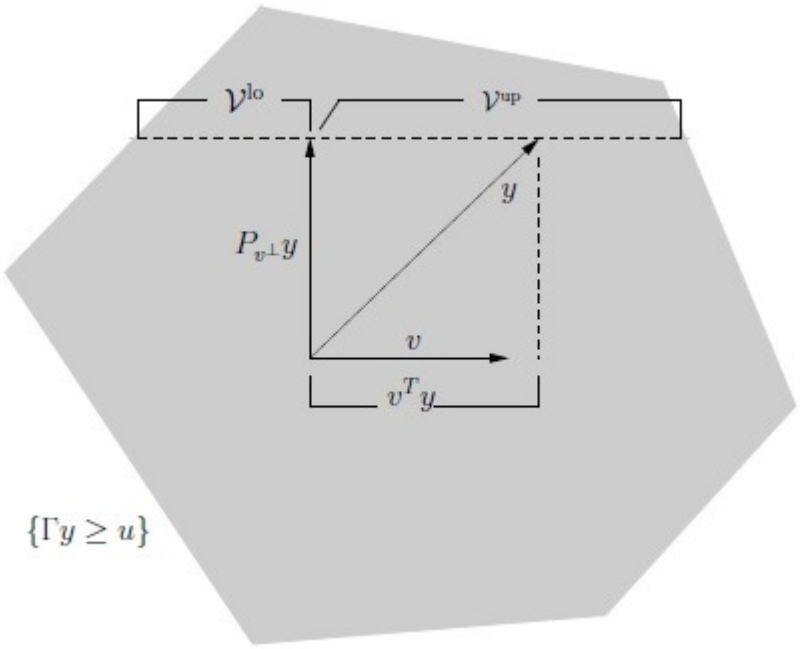


Figure 1: ([Taylor et al., 2014]) Geometry of polyhedral selection as truncation. For simplicity $\Sigma = I$. The shaded gray area is the polyhedral set $\{y : \Gamma y \geq u\}$. By breaking up $y$ into $P_v y$ (its projection onto $v$) and $P_{v^\perp} y$ (its projection onto the orthogonal complement of $v$), we see that $\Gamma y \geq u$ holds if and only if $v^T y$ does not deviate ftoo far from $P_{v^\perp} y$, hence trapping it in between bounds $\mathcal{V}^{\text{lo}}, \mathcal{V}^{\text{up}}$.

would have been close to being uniform. Utilizing the independence of $\mathcal{V}^{\text{lo}}(y), \mathcal{V}^{\text{up}}(y), \mathcal{V}^0(y)$ and $v^T y$, the authors show that this statistic indeed follows uniform distribution exactly. This is Lemma 2 in the paper. Relying on this result, they build the one-sided and two-sided confidence intervals for $v^T\theta$(Lemmas 3,4 in [Taylor et al., 2014]).

They apply these general tests for the cases of FS and LAR. In both the cases, the authors first show that the set $\mathcal{P}$ is a polyhedron, then derive $\mathcal{V}^{\text{lo}}, \mathcal{V}^{\text{up}}$ in a tractable form and compute the one-sided and two-sided confidence intervals. In the case of LARS, $\mathcal{V}^{\text{lo}}, \mathcal{V}^{\text{up}}$ turn out to be very easy to compute and the authors design a specialized test called spacing test.

## 2.3 Polyhedral set for Forward stepwise regression

In Forward stepwise procedure, in each iteration, we select the variable that reduces the residual sum of squares(RSS) the most. It is not difficult to work out that after $k$ steps, the set $\mathcal{P} = \{y|\hat{A}_k(y) = A_k, \hat{s}_{A_k}(y) = s_{A_k}\}$ is a polyhedral set. Assume without loss of generality that all columns in $X$ have unit length. For $k = 1$, the variable selection criterion boils down to

$$s_1\langle X_{j_1}, y\rangle \geq \pm\langle X_j, y\rangle \text{ for } j \neq j_1$$

where the sign $s_1 = \text{sign}(\langle X_{j_1}, y\rangle)$ . In other words, the variable with the highest correlation with $y$ is selected first. This condition can be easily written in the form $\Gamma y \geq u$ with $u = 0$. For $k > 1$, the selection criterion boils down to a similar condition in terms of the residuals left after removing projections on $X_{A_{k-1}}$:

$$s_k\langle \tilde{X}_{j_k}, \tilde{y}\rangle \geq \pm\langle \tilde{X}_j, \tilde{y}\rangle \text{ for } j \neq j_k$$

The condition can be interpreted as selecting the variable with the highest correlation with the residual $\tilde{y}$ after removing its projection onto $X_{A_k}$. Here $\tilde{x} = P^{\perp}_{A_{k-1}}x/\|P^{\perp}_{A_{k-1}}x\|$ with $P^{\perp}_{A_{k-1}}$ denoting the matrix which projects a vector into the space orthogonal to the columns $A_{k-1}$ of $X$. Plugging in this relation, the selection criterion can again easily be rewritten in the form $\Gamma y \geq u$ with $u = 0$. This shows that $\mathcal{P}$ is a polyhedral set. Conditioned on $y \in \mathcal{P}$, we can thus apply Lemma 2 of the paper to find the p-value of the $k$th variable selected and Lemmas 3,4 of the paper to find one-sided and two-sided selection intervals for $e_k^T X^+_{A_k}\theta$.

## 2.4 Polyhedral selection set and Spacing test for LAR

In this section, we first describe the Least angle regression(LAR) algorithm, and then summarize how the paper shows that $\mathcal{P}$ is (almost) a polyhedral set and how the spacing test is developed.

LAR can be viewed as a "democratic" version of forward stepwise regression. To explain the geometrical intuition of LAR, we will assume that the predictor variables $X$ is normalized,i.e., $\forall j,\ \|X_j\|_2 = 1$. Algorithm 1 is an adapted description of LAR from [Hastie et al., 2001].

**Algorithm 1** Least Angle Regression([Hastie et al., 2001])

1: Standardize the predictors to have unit norm. Initialize the residual $r = y, \beta_1, \cdots, \beta_p = 0$
2: Find the predictor $X_j$ most correlated with $r$.
3: Move $\beta_j$ from 0 towards its least-squares coefficient $\langle X_j, r \rangle$, until some other competitor $X_k$ has as much correlation with the current residual as does $X_j$.
4: Move $\beta_j$ and $\beta_k$ in the direction defined by their joint least squares coefficient of the current residual on $(X_j, X_k)$, until some other competitor $X_l$ has as much correlation with the current residual.
5: Continue in this way until all $p$ predictors have been entered.

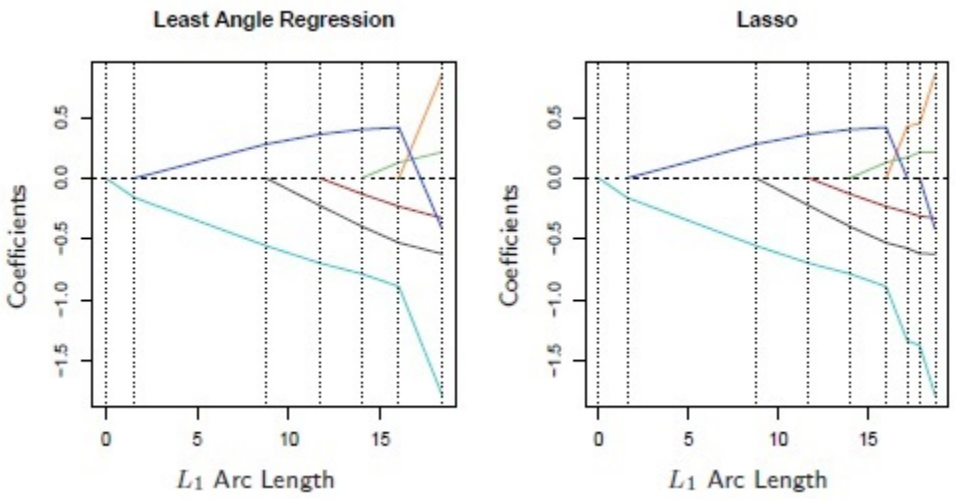


Figure 2: ([Hastie et al., 2001]) Left panel shows the LAR coefficient profiles on the simulated data, as a function of the $L_1$ arc length. The right panel shows the Lasso profile. They are identical until the dark-blue coefficient crosses zero at an arc length of about 18.

Let $\hat{y}_k$ be the fitted value at step $k$ with predictors in $A_k$. Then LAR profile evolves such that incremental fitted value vector $\hat{y}_k - \hat{y}_{k-1}$ makes the smallest (and equal) angle with each of the predictors in $A_k$. The name "least angle" arises from this geometrical interpretation.

LAR is intimately tied to 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. Hence, LAR coefficient profile is identical to the lasso profile until one coefficient crosses zero, as illustrated in Figure 2([Hastie et al., 2001])

For LAR, the set of observation vectors $y$ that result in a given sequence of active variable and sign lists,

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

is not directly considered, but rather a smaller set

$$\mathcal{P}' = \{y : \ \hat{A}_k(y) = A_k, \ \hat{s}_{A_k}(y) = s_{A_k}, \ \hat{S}_l(y) = S_l, \ l = 1, \cdots, k\} \tag{2}$$

is considered, where $S_l \subset \{1, \cdots, p\} \times \{-1, 1\}$ contains the variable-sign pairs that were "in competition" to become the active variable-sign pair step $l$. This is perfectly viable

for inference, because any conditional statement over $\mathcal{P}'$ translates into a valid conditional statement over $\mathcal{P}$ by marginalizing over $S_l$, $l = 1, \cdots, k$.

Then (2) can be exactly represented as

$$\{y : \ \Gamma' y \geq U'\} \tag{3}$$

where $U'$ is some random variable. Number of rows of $\Gamma'$ in (3) is upper bounded by $4pk - 2k^2 - k$(Lemma 5 in [Taylor et al., 2014]).

When $X$ is orthogonal, some rows of $\Gamma'$ in (3) are vacuous and only $k+1$ rows are meaningful. For nonorthogonal $X$, those rows can be still ignored, which leads to the compact representation

$$\{y : \ \Gamma y \geq U\}, \tag{4}$$

with $\Gamma$ in (4) having $k+1$ rows. This approximation is empirically justified. Also, Lemma 1 continues to hold for sets of $\{y : \ \Gamma y \geq U\}$ for random $U$, when we additionally have that $v^T y, (I - \Sigma v / v^T \Sigma v) v^T y, U$ are independent. In LAR, this additional condition is satisfied when $v$ lies in the column space of the LAR active variables at current step.

The particular choice of contrast vector

$$v_k = \frac{P^{\perp}_{A_{k-1}} X_{j_k}}{s_k - X^T_{j_k} (X^+_{A_{k-1}})^T s_{A_{k-1}}} \tag{5}$$

satisfies above condition, and $v_k$ paired with the compact representation $\{y : \ \Gamma y \geq U\}$ leads to a special test that called *spacing test*. $\mathcal{V}^{\text{lo}}, \mathcal{V}^{\text{up}}$ can be computed cheaply in this setting. From the definition (5) of $v_k$, the null hypothesis being considered is

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

hence the spacing test is a test for the $k$th entered coefficient in the multiple regresison of $\theta$ on $X_{A_k}$. Letting

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

the one-sided spacing test statistic is defined by

$$R_k = \frac{\Phi(\lambda_{k-1} \frac{w_k}{\sigma}) - \Phi(\lambda_k \frac{w_k}{\sigma})}{\Phi(\lambda_{k-1} \frac{w_k}{\sigma}) - \Phi(M_k^+ \frac{w_k}{\sigma})} \tag{6}$$

and the two-sided spacing test statistic is defined by

$$T_k = 2 \min\{R_k, 1 - R_k\} \tag{7}$$

In the above, $\lambda_{k-1}$ and $\lambda_k$ are knots at steps $k-1$ and $k$ in LAR path, and $M_k^+$ is the random variable from Lemma 5 in [Taylor et al., 2014]. $R_k$'s and $T_k$'s are both valid $p$-values for testing $H_0 : v_k^T \theta = 0$.

The spacing test statistics (6), (7) still depend on the random variable $M_k^+$. It is computable in $O(p)$ operations, but it is not an output of standard software for computing the LAR path. To simplify further, $M_k^+$ can be replaced by the next knot in the LAR path $\lambda_{k+1}$. This yields an asymptotically valid test.

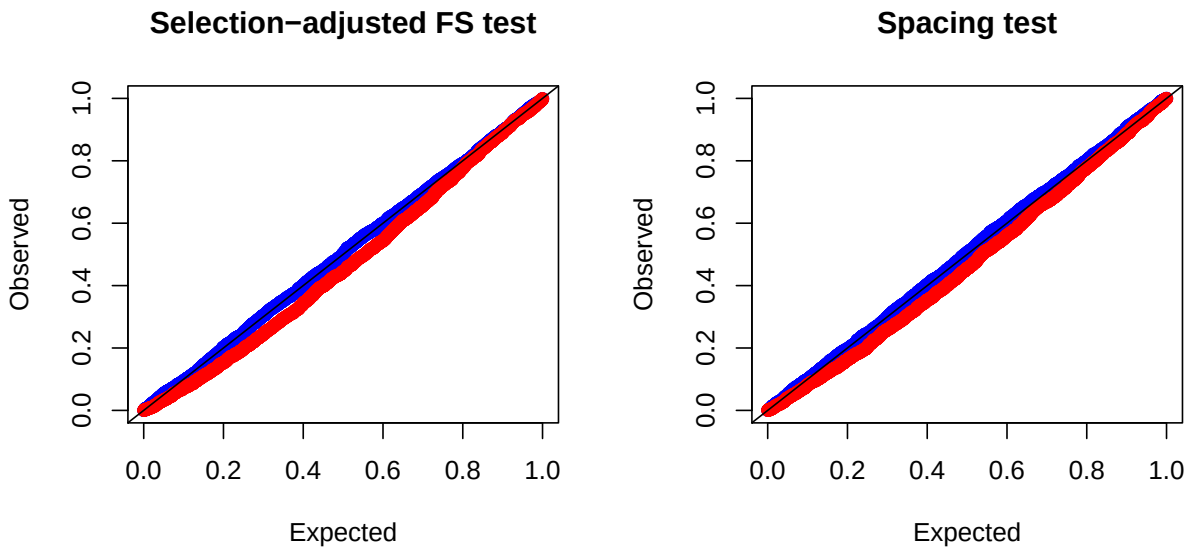


Figure 3: p-values of the third variable when the true model contains signal only in the first two variables. The blue line indicates the cases where the correct model is selected in the first two steps and the red line indicates the cases where the correc model is not selected.

# 3 Simulations

We simulated and verified some of the experiments in the paper.

For the first experiment, we drew 5000 $y$'s from $N(X\beta^*, \sigma^2 = 1)$ where the elements of the predictor matrix $X \in \mathbb{R}^{n\times p}$ with $n = 50, p = 10$ are drawn from standard normal distribution and $\beta^* = (3.5, -2.5, 0, \cdots, 0)$. For both FS and LAR, we computed the conditional p-values for the partial regression coefficient of the third variable entering into the active set. The observed p-values are plotted against the expected ones in Figure 3 For the simlulations where the first two variables are correctly selected with the correct order and sign, the p-values are uniform as desired. This is indicated by he blue line. For other simulations, the p-values are smaller than uniform which is also desired because we would like to reject the third variable with some chance. If we naively use a $t$-test without conditioning, we expect the $p$-values to be less than uniform even when the correct model is selected in the first two steps.

In the second experiment, we checked if the selection intervals in LAR capture the true model value. We drew 100 $y$'s from $N(X\beta^*, \sigma^2 I)$ where $n = 100, p = 10$ and $\sigma^2 = 0.25$ this time. In the true model $\beta^*$, we set the first two components to non-zero and the rest to zero. First, to ensure that there is strong signal we set $\beta^* = (6, 3, 0, \cdots, 0)$ and get the 90% confidence intervals after the second step of LAR, for the two variables selected. The confidence intervals are plotted in Figure 4 for the 100 repetitions. The black lines indicate the cases where the confidence intervals do not cover the true model value. For both the

variables, the empirical coverage is close to the expected 90% coverage. We continued our second experiment with a weak signal in the true model, by setting $\beta^* = (1, -1, 0, \cdots, 0)$. We observe a similar empirical coverage in this setting too. In our third experiment, we compared the p-values given by FS, LAR and another test from [Buja and Brown, 2014] called max-$|t|$-test. We generated a random $X$ with $n = 50, p = 10$ and orthogonal columns and sampled $y$ 1000 times from $N(X\beta^*, I)$ with the true model $\beta^* = 0$. See Figure 6 for the plots for the first four variables. As the true model does not have any signal, we expect the p-values to be uniformly distributed. All the three tests give uniformly distributed p-values for the first variable. But max-$t$-test gives larger and larger p-values as we go from second to fourth variable. So max-$t$-test has an undesirable conservative bias whereas the polyhedral tests do not have that problem because of the conditioning.

# 4 Conclusion

Concluding, the paper develops an elegant and computationally efficient test statistic to measure the significance of the $k$th variable entering the model when the response variables are normally distributed for methods which result in polyhedral selection sets. The methodology has promising practical utility. It might be possible to extend the methodology to general convex sets as it seems $\mathcal{V}^{\text{lo}}(y), \mathcal{V}^{\text{up}}(y)$ and $\mathcal{V}^0(y)$ depend only on $P_v^\perp(y)$ in that case too.

# Acknowledgement

We thank Rob Tibshirani and Ryan Tibshirani for providing most of the simulation code.

# References

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

[Hastie et al., 2001] Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer Series in Statistics. Springer New York Inc., New York, NY, USA.

[Taylor et al., 2014] Taylor, J., Lockhart, R., Tibshirani, R. J., and Tibshirani, R. (2014). Exact post-selection inference for forward stepwise and least angle regression.

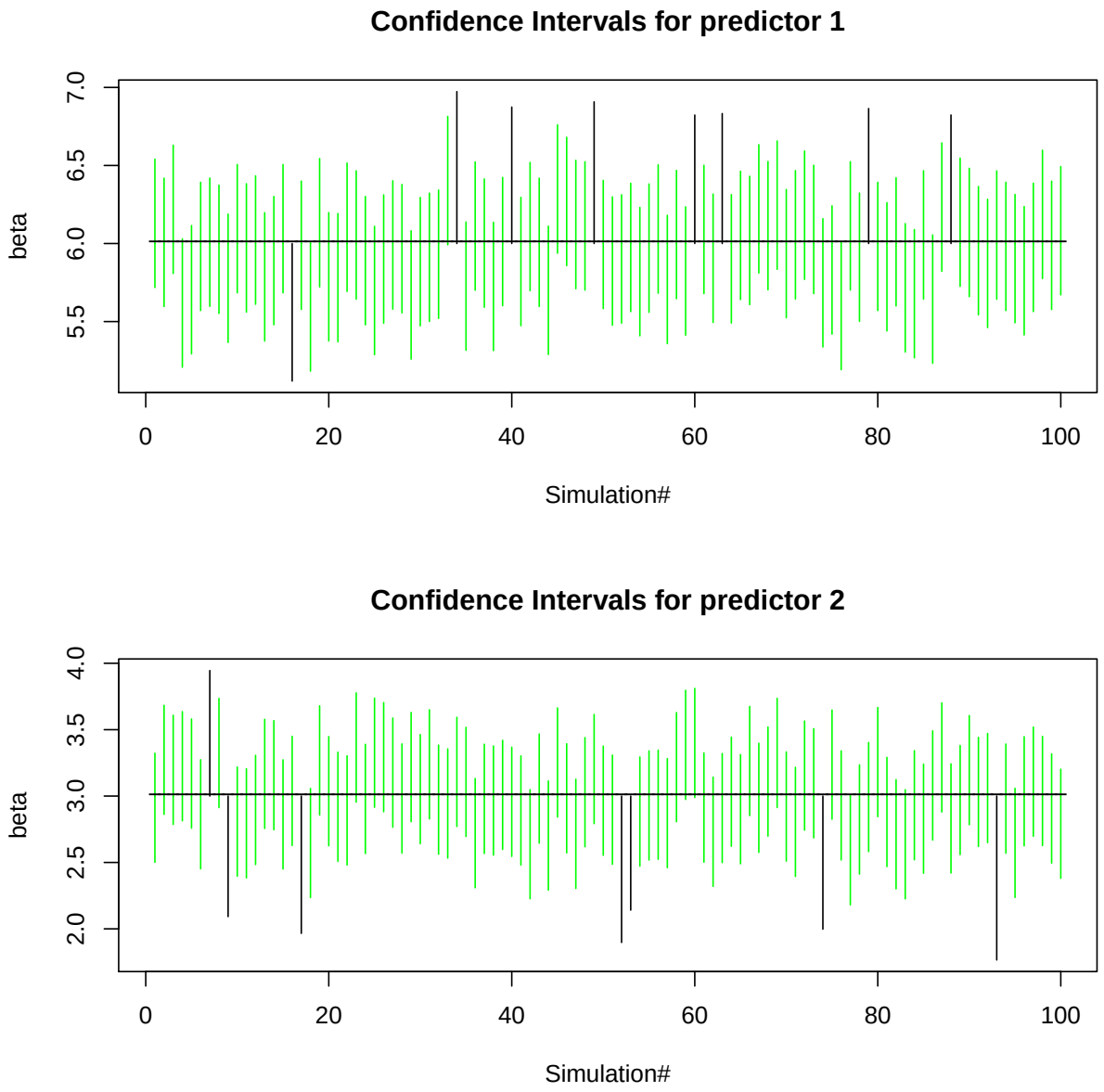


Figure 4: Selection intervals for first and second variables in the case with strong signal over 100 simulations

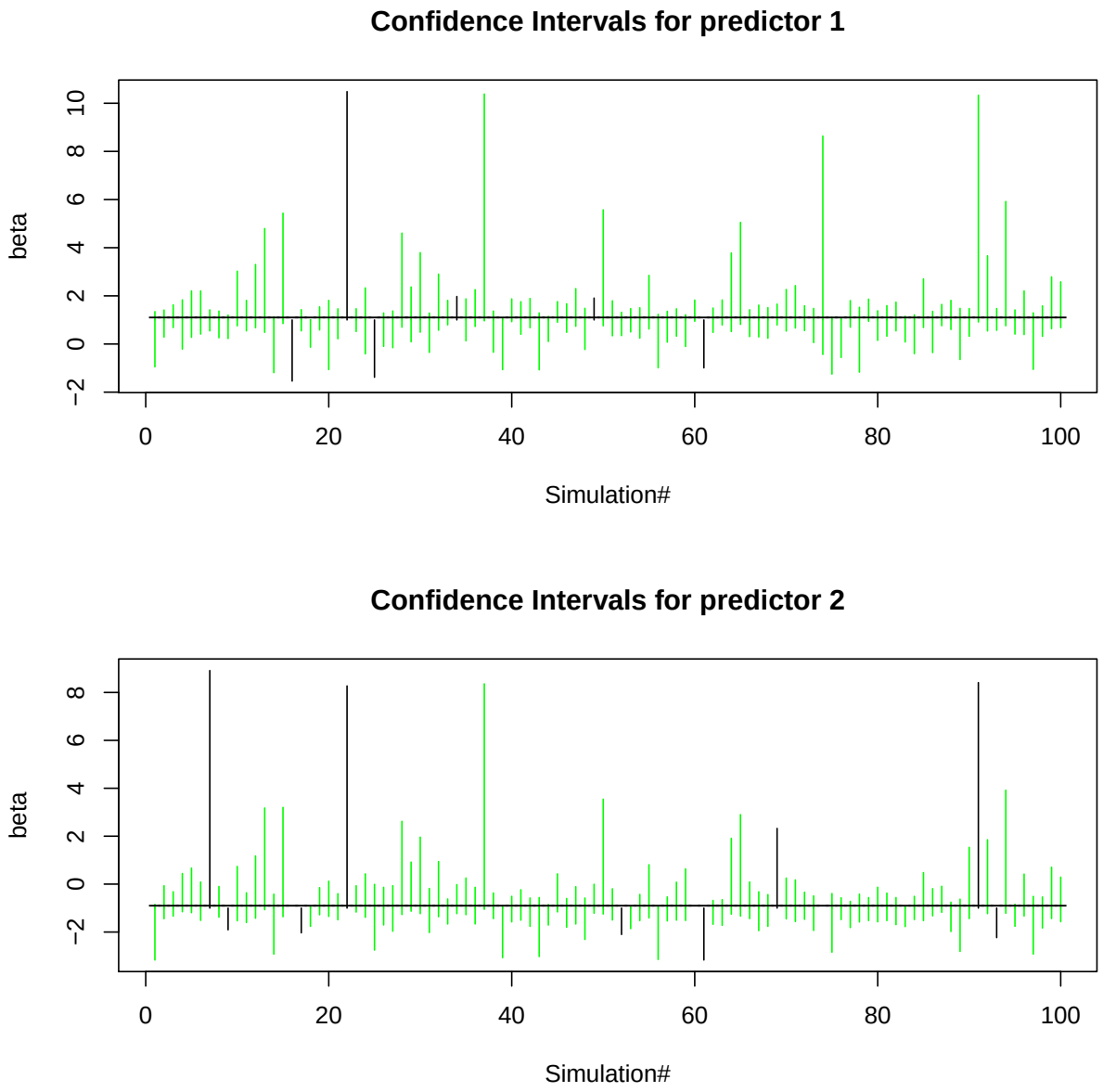


Figure 5: Selection intervals for first and second variables in the case with weak signal over 100 simulations

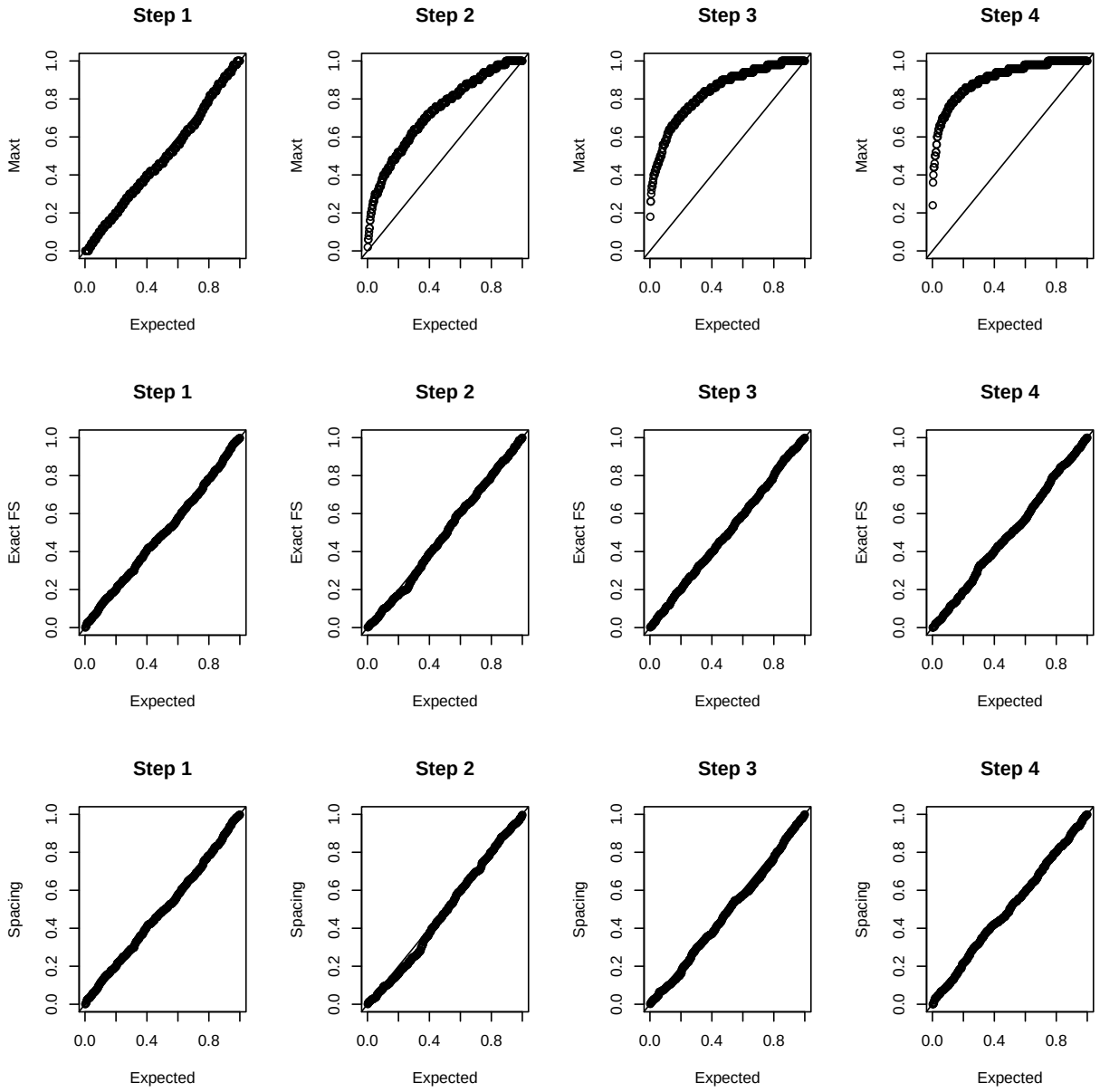


Figure 6: Comparison of the p-values given by max-$|t|$-test and the polyhedral test for FS and LAR, for the first four variables to enter the model when $\beta^* = 0$