EXACT POST-SELECTION INFERENCE, WITH APPLICATION TO THE LASSO

By Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor

Stanford University

We develop a general approach to valid inference after model selection. In a nutshell, our approach produces post-selection inferences with the same frequency guarantees as those given by data splitting but are more powerful. At the core of our framework is a result that characterizes the distribution of a post-selection estimator conditioned on the *selection event*. We specialize the approach to model selection by the lasso to form valid confidence intervals for the selected coefficients and test whether all relevant variables have been included in the model.

1. Introduction. As a statistical technique, linear regression is both simple and powerful. Not only does it provide estimates of the "effect" of each variable, but it also quantifies the uncertainty in those estimates, allowing for inference about the effects. However, in many exploratory investigations, a practitioner starts with a large pool of candidate variables, such as genes or demographic features, and does not know *a priori* which are relevant. This is especially a problem when there are more variables than observations, since then the (full) linear model is unidentified.

We might wish to use the data to select a subset of variables. One approach is to fit a linear model with all variables included (assuming this is possible), observe which ones are significant at level α, and then refit the linear model with only those variables included. The problem with this approach is that the p-values can no longer be trusted, since the variables that are selected will tend to be those that are significant. Intuitively, we are "overfitting" to this realization of the data.

To formalize the problem, consider the standard linear regression setup, where the response $\boldsymbol{y} \in \mathbb{R}^n$ is generated

$$\boldsymbol{y} \sim N(\boldsymbol{\mu}, \sigma^2 I_n) \tag{1.1}$$

and $\boldsymbol{\mu}$ is modeled as a linear function of predictors $\boldsymbol{x}_1, ..., \boldsymbol{x}_p \in \mathbb{R}^n$. We choose a subset of the predictors $M \subset [p]$ and ask for inferences about the

AMS 2000 subject classifications: Primary 62F03, 62J07; secondary 62E15
Keywords and phrases: lasso, confidence interval, hypothesis test, model selection

submodel coefficients, i.e. the linear combination of predictors in M that minimizes the expected error:

$$\boldsymbol{\beta}^M \equiv \arg\min_{\boldsymbol{\beta}} \ \mathbb{E}||\boldsymbol{y} - X_M\boldsymbol{\beta}||^2 = X_M^+\boldsymbol{\mu}, \tag{1.2}$$

where $X_M^+ \equiv (X_M^T X_M)^{-1} X_M^T$ is the pseudo-inverse of X_M. Notice that (1.2) implies that the targets β_j^M and $\beta_j^{M'}$ in different models $M \neq M'$ are in general different. This is simply a restatement of the well-known fact that a regression coefficient describes the effect of a predictor, *adjusting for the other predictors in the (sub)model*.

Alternatively, we may ask for inferences about a selected subset of the *full model coefficients:*

$$\boldsymbol{\beta}^{[p]} \equiv \arg\min_{\boldsymbol{\beta}} \ \mathbb{E}||\boldsymbol{y} - X\boldsymbol{\beta}||^2 = X^+\boldsymbol{\mu};$$

i.e. we ask for valid selective inferences about the $\boldsymbol{\beta}^{[p]}$. For example, we may ask for inferences about the k most "significant" coefficients, e.g. the k coefficients with the largest absolute Z-score. In this setting, the targets $\beta_j^{[p]}$ remain consistent in different models.

Thus, "inference after selection" is ambiguous in linear regression because the target of inference may change with the selected model (Berk et al., 2013). In the next section, we discuss several ways to resolve this ambiguity.

2. Post-Selection Inference in Linear Regression. At first blush, the fact that the target changes with the model is deeply disturbing, since it seems to imply that the parameters are random. However, it is not so much the target that is random as our selection of targets. This is evidently the case when asking for selective inference about the full model coefficients. When asking for inference about the submodel coefficients, note that there are $p2^{p-1}$ possible submodel coefficient, one for each coefficient in all 2^p possible models:

$$\{\beta_j^M : M \subset [p], j \in M\}.$$

However, we only form inferences for a subset of these parameters—specifically, the parameters $\beta_j^{\hat{M}}$ in the model $\hat{M}$ we select. To avoid repetition, we continue the discussion focusing on inference for the submodel targets.

To be concrete, suppose we want a confidence interval $C_j^{\hat{M}}$ for a submodel coefficient $\beta_j^{\hat{M}}$. What frequency properties should $C_j^{\hat{M}}$ have? By analogy to the classical setting, we might want

$$\mathbb{P}(\beta_j^{\hat{M}} \in C_j^{\hat{M}}) \geq 1 - \alpha,$$

but this is ill-posed because β_j^M does not exist when $j \notin M$. Berk et al. (2013) suggest the following alternatives:

1. Conditional Coverage: Since we form an interval for β_j^M if and only if we select model M, i.e., $\hat{M} = M$, it makes sense to condition on this event. Hence, we might require that our confidence interval C_j^M satisfy

$$\mathbb{P}(\beta_j^M \in C_j^M \mid \hat{M} = M) \geq 1 - \alpha. \tag{2.1}$$

   The benefit of this approach is that we avoid ever having to compare coefficients across two different models $M \neq M'$.
2. Simultaneous Coverage: It also makes sense to talk about events that are defined simultaneously over all $j \in \hat{M}$. Berk et al. (2013) propose controlling the familywise error rate

$$FWER \equiv \mathbb{P}(\beta_j^{\hat{M}} \notin C_j^{\hat{M}} \text{ for any } j \in \hat{M}), \tag{2.2}$$

   but it may be too stringent when there are many predictors involved. Instead of controlling the probability of making any error, we can control the expected proportion of errors—although the "proportion of errors" is ambiguous when we select no variables. We can simply declare the error to be zero when $|\hat{M}| = 0$ (Benjamini and Yekutieli, 2005):

$$FCR \equiv \mathbb{E}\left[\frac{\left|\{j \in \hat{M} : \beta_j^{\hat{M}} \notin C_j^{\hat{M}}\}\right|}{|\hat{M}|}; |\hat{M}| > 0\right], \tag{2.3}$$

   or condition on $|\hat{M}| > 0$ (Storey, 2003):

$$pFCR \equiv \mathbb{E}\left[\frac{\left|\{j \in \hat{M} : \beta_j^{\hat{M}} \notin C_j^{\hat{M}}\}\right|}{|\hat{M}|}\middle| |\hat{M}| > 0\right]. \tag{2.4}$$

   Since $FCR = pFCR \cdot \mathbb{P}(|\hat{M}| > 0)$, $pFCR$ control implies FCR control.

Remarkably, conditional coverage (2.1) implies pFCR (2.4) (and hence, FCR) control.

THEOREM 2.1. *Consider a family of intervals $\{C_j^{\hat{M}}\}_{j \in \hat{M}}$ that each have conditional $(1-\alpha)$ coverage:*

$$\mathbb{P}(\beta_j^{\hat{M}} \notin C_j^{\hat{M}} | \hat{M} = M) \leq \alpha, \text{ for all } M \text{ and } j \in M.$$

Then, $FCR \leq pFCR \leq \alpha$.

PROOF. Condition on $\hat{M}$ and iterate expectations.

$$\begin{aligned} pFCR &= \mathbb{E}\left[\mathbb{E}\left[\frac{\left|\{j \in \hat{M} : \beta_j^{\hat{M}} \notin C_j^{\hat{M}}\}\right|}{|\hat{M}|}\middle| \hat{M}\right] \;\middle|\; |\hat{M}| > 0\right] \\ &= \mathbb{E}\left[\frac{\sum_{j\in\hat{M}} \mathbb{P}\left(\beta_j^{\hat{M}} \notin C_j^{\hat{M}} | \hat{M}\right)}{|\hat{M}|} \;\middle|\; |\hat{M}| > 0\right] \\ &\leq \mathbb{E}\left[\frac{\alpha|\hat{M}|}{|\hat{M}|} \;\middle|\; |\hat{M}| > 0\right] \\ &= \alpha. \end{aligned}$$

□

3. Outline of Our Approach. We have argued that post-selection intervals for regression coefficients should have $1-\alpha$ coverage conditional on the model:

$$\mathbb{P}(\beta_j^M \in C_j^M \mid \hat{M} = M) \geq 1 - \alpha,$$

both because this criterion is interesting in its own right and because it implies FCR control. To obtain an interval with this property, we study the conditional distribution

$$\boldsymbol{\eta}_M^T \boldsymbol{y} \mid \{\hat{M} = M\}, \tag{3.1}$$

which will allow, more generally, conditional inference for parameters of the form $\boldsymbol{\eta}_M^T \boldsymbol{\mu}$. In particular, $\beta_j^M = \boldsymbol{e}_j^T X_M^+ \boldsymbol{\mu}$ can be written in this form, as can many other linear contrasts.

Our paper focuses on the specific case where the lasso is used to select the model $\hat{M}$. We begin in Section 4 by characterizing the event $\{\hat{M} = M\}$ for the lasso. As it turns out, this event is a union of polytopes. More precisely, the event $\{\hat{M} = M, \hat{\boldsymbol{s}}_M = \boldsymbol{s}_M\}$, that specifies the model *and* the signs of the selected variables, is a polytope of the form

$$\{\boldsymbol{y} \in \mathbb{R}^n : A(M, \boldsymbol{s}_M)\boldsymbol{y} \leq \boldsymbol{b}(M, \boldsymbol{s}_M)\}.$$

Therefore, if we condition on both the model and the signs, then we only need to study

$$\boldsymbol{\eta}^T \boldsymbol{y} \mid \{A\boldsymbol{y} \leq \boldsymbol{b}\}. \tag{3.2}$$

We do this in Section 5. It turns out that this conditional distribution is essentially a (univariate) truncated Gaussian. We use this to derive a statistic $F^{\boldsymbol{z}}(\boldsymbol{\eta}^T \boldsymbol{y})$ whose distribution given $\{A\boldsymbol{y} \leq \boldsymbol{b}\}$ is $\text{Unif}(0, 1)$.

3.1. *Related Work.* The resulting post-selection test has a similar structure to the pathwise significance tests of Lockhart et al. (2014) and Taylor et al. (2014), which also are conditional tests. However, the intended application of our test is different. Whereas they test the specific hypothesis of whether a newly added coefficient along the LARS path is non-zero, our framework allows more general questions about the model the lasso selects: we can test the model at any value of λ or form confidence intervals for an individual coefficient in the model.

There is also a parallel literature on confidence intervals for coefficients in high-dimensional linear models based on the lasso estimator (Javanmard and Montanari, 2013; van de Geer et al., 2013; Zhang and Zhang, 2014). The difference between their work and ours is that they do not address post-selection inference; their target is $\boldsymbol{\beta}^0$, the coefficients in the true model, rather than $\boldsymbol{\beta}^{\hat{M}}$, the coefficients in the selected model. The two will not be the same unless $\hat{M}$ happens to contain all non-zero coefficients of $\boldsymbol{\beta}^0$. Although inference for $\boldsymbol{\beta}^0$ is appealing, it requires assumptions about correctness of the linear model and sparsity of $\boldsymbol{\beta}^0$. Our approach instead regards the selected model as a linear approximation to the truth, a view shared by Berk et al. (2013).

4. The Lasso and Its Selection Event. In this paper, we apply our post-selection inference procedure to the model selected by the lasso (Tibshirani, 1996). The lasso estimate is the solution to the usual least squares problem with an additional ℓ_1 penalty on the coefficients:

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1. \tag{4.1}$$

The ℓ_1 penalty shrinks many of the coefficients to exactly zero, and the tradeoff between sparsity and fit to the data is controlled by the penalty parameter $\lambda \geq 0$.

Because the lasso produces sparse solutions, we can define model "selected" by the lasso to be simply the set of predictors with non-zero coefficients:

$$\hat{M} = \{j : \hat{\beta}_j \neq 0\}.$$

Then, post-selection inference seeks to make inferences about $\boldsymbol{\beta}^M$, given $\{\hat{M} = M\}$, as defined in (1.2).

The rest of this section focuses on characterizing this event $\{\hat{M} = M\}$. We begin by noting that in order for a vector of coefficients $\hat{\boldsymbol{\beta}}$ and a vector of signs $\hat{\boldsymbol{s}}$ to be solutions to the lasso problem (4.1), it is necessary and

sufficient that they satisfy the Karush-Kuhn-Tucker (KKT) conditions:

$$X^T(X\hat{\boldsymbol{\beta}} - \boldsymbol{y}) + \lambda \hat{\boldsymbol{s}} = 0, \tag{4.2}$$

$$\hat{s}_i = \text{sign}(\hat{\beta}_j) \quad \text{if } \hat{\beta}_j \neq 0$$

$$\hat{s}_i \in [-1, 1] \qquad \text{if } \hat{\beta}_j = 0 \tag{4.3}$$

Following Tibshirani (2013), we consider the *equicorrelation set*

$$\hat{M} \equiv \{i \in \{1, \ldots, p\} : |\hat{s}_i| = 1\}. \tag{4.4}$$

Notice that we have implicitly defined the model $\hat{M}$ to be equicorrelation set. Since $|\hat{s}_i| = 1$ for any $\hat{\beta}_i \neq 0$, the equicorrelation set does in fact contain all predictors with non-zero coefficients, although it may also include some predictors with zero coefficients. However, for almost every λ, the equicorrelation set is precisely the set of predictors with non-zero coefficients.

It turns out that it is easier to first characterize $\{(\hat{M}, \hat{\boldsymbol{s}}) = (M, \boldsymbol{s})\}$ and obtain $\{\hat{M} = M\}$ as a corollary by taking a union over the possible signs. The next result is an important first step.

LEMMA 4.1. *Assume the columns of X are in general position (Tibshirani, 2013). Let $M \subset \{1, \ldots, p\}$ and $\boldsymbol{s} \in \{-1, 1\}^{|M|}$ be a candidate set of variables and their signs, respectively. Define the random variables*

$$\boldsymbol{w}(M, \boldsymbol{s}) := (X_M^T X_M)^{-1}(X_M^T \boldsymbol{y} - \lambda \boldsymbol{s}) \tag{4.5}$$

$$\boldsymbol{u}(M, \boldsymbol{s}) := X_{-M}^T (X_M^T)^+ \boldsymbol{s} + \frac{1}{\lambda} X_{-M}^T (I - P_M)\boldsymbol{y}. \tag{4.6}$$

where $P_M \equiv X_M(X_M^T X_M)^{-1} X_M$ is projection onto the column span of X_M. Then the selection procedure can be rewritten in terms of $\boldsymbol{w}$ and $\boldsymbol{u}$ as:

$$\left\{(\hat{M}, \hat{\boldsymbol{s}}) = (M, \boldsymbol{s})\right\} = \left\{\text{sign}(\boldsymbol{w}(M, \boldsymbol{s})) = \boldsymbol{s},\ ||\boldsymbol{u}(M, \boldsymbol{s})||_\infty < 1\right\} \tag{4.7}$$

PROOF. First, we rewrite the KKT conditions (4.2) by partitioning them according to the equicorrelation set $\hat{M}$, adopting the convention that $-\hat{M}$ means "variables not in $\hat{M}$."

$$\begin{aligned} X_{\hat{M}}^T (X_{\hat{M}} \hat{\boldsymbol{\beta}}_{\hat{M}} - \boldsymbol{y}) + \lambda \hat{\boldsymbol{s}}_{\hat{M}} &= 0 \\ X_{-\hat{M}}^T (X_{\hat{M}} \hat{\boldsymbol{\beta}}_{\hat{M}} - \boldsymbol{y}) + \lambda \hat{\boldsymbol{s}}_{-\hat{M}} &= 0 \\ \text{sign}(\hat{\boldsymbol{\beta}}_{\hat{M}}) &= \hat{\boldsymbol{s}} \\ ||\hat{\boldsymbol{s}}_{-\hat{M}}||_\infty &< 1. \end{aligned}$$

Since the KKT conditions are necessary and sufficient for a solution, we obtain that $\{(\hat{M}, \hat{\boldsymbol{s}}) = (M, \boldsymbol{s})\}$ if and only if there exist $\boldsymbol{w}$ and $\boldsymbol{u}$ satisfying:

$$\begin{aligned} X_M^T(X_M\boldsymbol{w} - \boldsymbol{y}) + \lambda \boldsymbol{s} &= 0 \\ X_{-M}^T(X_M\boldsymbol{w} - \boldsymbol{y}) + \lambda \boldsymbol{u} &= 0 \\ \operatorname{sign}(\boldsymbol{w}) &= \boldsymbol{s} \\ ||\boldsymbol{u}||_\infty &< 1. \end{aligned}$$

We can solve the first two equations for $\boldsymbol{w}$ and $\boldsymbol{u}$ to obtain the equivalent set of conditions

$$\begin{aligned} \boldsymbol{w} &= (X_M^T X_M)^{-1}(X_M^T \boldsymbol{y} - \lambda \boldsymbol{s}) \\ \boldsymbol{u} &= X_{-M}^T (X_M^T)^+ \boldsymbol{s} + \frac{1}{\lambda} X_{-M}^T (I - P_M)\boldsymbol{y} \\ \operatorname{sign}(\boldsymbol{w}) &= \boldsymbol{s} \\ ||\boldsymbol{u}||_\infty &< 1, \end{aligned}$$

where the first two are the definitions of $\boldsymbol{w}$ and $\boldsymbol{u}$ given in (4.5) and (4.6), and the last two are the conditions on $\boldsymbol{w}$ and $\boldsymbol{u}$ given in (4.7). □

Lemma 4.1 is remarkable because it says that the event $\{(\hat{M}, \hat{\boldsymbol{s}}) = (M, \boldsymbol{s})\}$ can be rewritten as affine constraints on $\boldsymbol{y}$. This is because $\boldsymbol{w}$ and $\boldsymbol{u}$ are already affine functions of $\boldsymbol{y}$, and the constraints $\operatorname{sign}(\cdot) = \boldsymbol{s}$ and $||\cdot||_\infty < 1$ can also be rewritten in terms of affine constraints. The following proposition makes this explicit.

PROPOSITION 4.2. *Let $\boldsymbol{w}$ and $\boldsymbol{u}$ be defined as in (4.5) and (4.6). Then:*

$$\{\operatorname{sign}(\boldsymbol{w}) = \boldsymbol{s}, \|\boldsymbol{u}\|_\infty < 1\} = \left\{ \begin{pmatrix} A_0(M, \boldsymbol{s}) \\ A_1(M, \boldsymbol{s}) \end{pmatrix} \boldsymbol{y} < \begin{pmatrix} \boldsymbol{b}_0(M, \boldsymbol{s}) \\ \boldsymbol{b}_1(M, \boldsymbol{s}) \end{pmatrix} \right\} \tag{4.8}$$

where $A_0, \boldsymbol{b}_0$ encode the "inactive" constraints $\{\|\boldsymbol{u}\|_\infty < 1\}$, and $A_1, \boldsymbol{b}_1$ encode the "active" constraints $\{\operatorname{sign}(\boldsymbol{w}) = \boldsymbol{s}\}$. These matrices have the explicit forms:

$$\begin{aligned} A_0(M, \boldsymbol{s}) &= \frac{1}{\lambda} \begin{pmatrix} X_{-M}^T(I - P_M) \\ -X_{-M}^T(I - P_M) \end{pmatrix} & \boldsymbol{b}_0(M, \boldsymbol{s}) &= \begin{pmatrix} \mathbf{1} - X_{-M}^T (X_M^T)^+ \boldsymbol{s} \\ \mathbf{1} + X_{-M}^T (X_M^T)^+ \boldsymbol{s} \end{pmatrix} \\ A_1(M, \boldsymbol{s}) &= -\operatorname{diag}(\boldsymbol{s})(X_M^T X_M)^{-1} X_M^T & \boldsymbol{b}_1(M, s) &= -\lambda \operatorname{diag}(\boldsymbol{s})(X_M^T X_M)^{-1} \boldsymbol{s} \end{aligned}$$

PROOF. First, substituting expression (4.5) for $\boldsymbol{w}$, we rewrite the "active" constraints as

$$\begin{aligned}\{\text{sign}(\boldsymbol{w}) = \boldsymbol{s}\} &= \{\text{diag}(\boldsymbol{s})\boldsymbol{w} > 0\} \\ &= \{\text{diag}(\boldsymbol{s})(X_M^T X_M)^{-1}(X_M^T \boldsymbol{y} - \lambda \boldsymbol{s}) > 0\} \\ &= \{A_1(M, \boldsymbol{s})\boldsymbol{y} < \boldsymbol{b}_1(M, \boldsymbol{s})\}.\end{aligned}$$

Next, substituting expression (4.6) for $\boldsymbol{u}$, we rewrite the "inactive" constraints as

$$\begin{aligned}\{||\boldsymbol{u}||_\infty < 1\} &= \left\{-\mathbf{1} < X_{-M}^T (X_M^T)^+ \boldsymbol{s} + \frac{1}{\lambda} X_{-M}^T (I - P_M)\boldsymbol{y} < \mathbf{1}\right\} \\ &= \{A_0(M, \boldsymbol{s})\boldsymbol{y} < \boldsymbol{b}_0(M, \boldsymbol{s})\}\end{aligned}$$

□

Combining Lemma 4.1 with Proposition 4.2, we obtain the following.

THEOREM 4.3. *Let* $A(M, \boldsymbol{s}) = \begin{pmatrix} A_0(M, \boldsymbol{s}) \\ A_1(M, \boldsymbol{s}) \end{pmatrix}$ *and* $b(M, \boldsymbol{s}) = \begin{pmatrix} \boldsymbol{b}_0(M, \boldsymbol{s}) \\ \boldsymbol{b}_1(M, \boldsymbol{s}) \end{pmatrix}$, *where* A_i *and* b_i *are defined in Proposition 4.2. Then:*

$$\{\hat{M} = M, \hat{\boldsymbol{s}} = \boldsymbol{s}\} = \{A(M, \boldsymbol{s})\boldsymbol{y} \leq \boldsymbol{b}(M, \boldsymbol{s})\}.$$

As a corollary, $\{\hat{M} = M\}$ is simply the union of the above events over all possible sign patterns.

COROLLARY 4.4. $$\{\hat{M} = M\} = \bigcup_{\boldsymbol{s} \in \{-1,1\}^{|M|}} \{A(M, \boldsymbol{s})\boldsymbol{y} \leq \boldsymbol{b}(M, \boldsymbol{s})\}.$$

Figure 1 illustrates Theorem 4.3 and Corollary 4.4. The lasso partitions of $\mathbb{R}^n$ into polytopes according to the model it selects and the signs of the coefficients. The shaded area corresponds to the event $\{\hat{M} = \{1, 3\}\}$, which is a union of two polytopes. Notice that the sign patterns $\{+, -\}$ and $\{-, +\}$ are not possible for the model $\{1, 3\}$.

5. Conditioning on Polytopes. In order to obtain inference conditional on the model, we need to understand the distribution of

$$\boldsymbol{\eta}_M^T \boldsymbol{y} \mid \{\hat{M} = M\}.$$

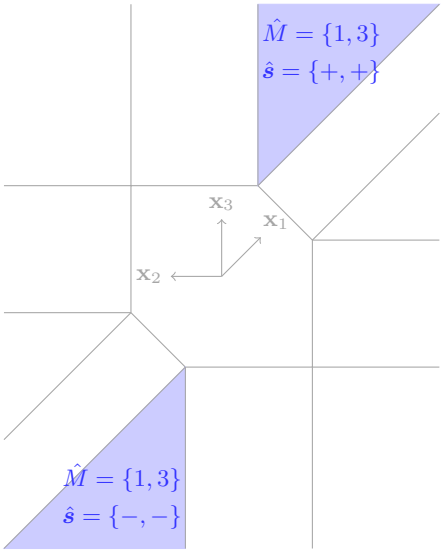

Fig 1: A geometric picture illustrating Theorem 4.3 for $n = 2$ and $p = 3$. The lasso partitions $\mathbb{R}^n$ into polytopes according to the selected model and signs.

However, as we saw in the previous section, $\{\hat{M} = M\}$ is a union of polytopes, so it is easier to condition on both the model *and the signs*,

$$\boldsymbol{\eta}_M^T \boldsymbol{y} \mid \{\hat{M} = M, \hat{\boldsymbol{s}} = \boldsymbol{s}\},$$

where the conditioning event is a single polytope $\{A(M, \boldsymbol{s})\boldsymbol{y} \leq \boldsymbol{b}(M, \boldsymbol{s})\}$. Notice that inferences that are valid conditional on this finer event will also be valid conditional on $\{\hat{M} = M\}$. For example, if a confidence interval C_j^M for β_j^M has $(1 - \alpha)$ coverage conditional on the model and signs

$$\mathbb{P}(\beta_j^M \in C_j^M \mid \hat{M} = M, \hat{\boldsymbol{s}} = \boldsymbol{s}) \geq 1 - \alpha,$$

then it will also have $(1 - \alpha)$ coverage conditional only on the model:

$$\begin{aligned}
\mathbb{P}(\beta_j^M \in C_j^M \mid \hat{M} = M) &= \sum_{\boldsymbol{s}} \mathbb{P}(\beta_j^M \in C_j^M \mid \hat{M} = M, \hat{\boldsymbol{s}} = \boldsymbol{s}) \mathbb{P}(\hat{\boldsymbol{s}} = \boldsymbol{s} \mid \hat{M} = M) \\
&\geq \sum_{\boldsymbol{s}} (1 - \alpha) \mathbb{P}(\hat{\boldsymbol{s}} = \boldsymbol{s} \mid \hat{M} = M) \\
&= 1 - \alpha.
\end{aligned}$$

This section is divided into two subsections. First, we study how to condition on a single polytope; this will allow us to condition on $\{\hat{M} = M, \hat{\boldsymbol{s}} = \boldsymbol{s}\}$. Then, we look at how to extend the framework to condition on a union of polytopes, which will allow us to condition only on the model $\{\hat{M} = M\}$. The inferences obtained by conditioning on the model will in general be more efficient (i.e., narrower intervals, more powerful tests), at the price of more computation.

5.1. *Conditioning on a Single Polytope.* Suppose we observe $\boldsymbol{y} \sim N(\boldsymbol{\mu}, \Sigma)$, and $\boldsymbol{\eta} \in \mathbb{R}^n$ is some direction of interest. To understand the distribution of

$$\boldsymbol{\eta}^T \boldsymbol{y} \mid \{A\boldsymbol{y} \leq \boldsymbol{b}\}, \tag{5.1}$$

we rewrite $\{A\boldsymbol{y} \leq \boldsymbol{b}\}$ in terms of $\boldsymbol{\eta}^T\boldsymbol{y}$ and a component $\boldsymbol{z}$ which is independent of $\boldsymbol{\eta}^T\boldsymbol{y}$. That component is

$$\boldsymbol{z} \equiv (I_n - \boldsymbol{c}\boldsymbol{\eta}^T)\boldsymbol{y}, \tag{5.2}$$

where

$$\boldsymbol{c} \equiv \Sigma\boldsymbol{\eta}(\boldsymbol{\eta}^T\Sigma\boldsymbol{\eta})^{-1}. \tag{5.3}$$

It is easy to verify that $\boldsymbol{z}$ is uncorrelated with, and hence independent of, $\boldsymbol{\eta}^T\boldsymbol{y}$. Although definition (5.2) may seem unmotivated, in the case where $\Sigma = \sigma^2 I_n$, $\boldsymbol{z}$ is simply the residual $(I_n - P_{\boldsymbol{\eta}})\boldsymbol{y}$ from projecting $\boldsymbol{y}$ onto $\boldsymbol{\eta}$.

We can now rewrite $\{A\boldsymbol{y} \leq \boldsymbol{b}\}$ in terms of $\boldsymbol{\eta}^T\boldsymbol{y}$ and $\boldsymbol{z}$.

LEMMA 5.1. *Let $\boldsymbol{z}$ be defined as in (5.2) and $\boldsymbol{c}$ as in (5.3). Then, the conditioning set can be rewritten as follows:*

$$\{A\boldsymbol{y} \leq \boldsymbol{b}\} = \{\mathcal{V}^-(\boldsymbol{z}) \leq \boldsymbol{\eta}^T\boldsymbol{y} \leq \mathcal{V}^+(\boldsymbol{z}), \mathcal{V}^0(\boldsymbol{z}) \geq 0\}$$

where

$$\mathcal{V}^-(\boldsymbol{z}) \equiv \max_{j:(A\boldsymbol{c})_j < 0} \frac{b_j - (A\boldsymbol{z})_j}{(A\boldsymbol{c})_j} \tag{5.4}$$

$$\mathcal{V}^+(\boldsymbol{z}) \equiv \min_{j:(A\boldsymbol{c})_j > 0} \frac{b_j - (A\boldsymbol{z})_j}{(A\boldsymbol{c})_j} \tag{5.5}$$

$$\mathcal{V}^0(\boldsymbol{z}) \equiv \min_{j:(A\boldsymbol{c})_j = 0} b_j - (A\boldsymbol{z})_j. \tag{5.6}$$

Note that $\mathcal{V}^-$, $\mathcal{V}^+$, and $\mathcal{V}^0$ refer to functions. Since they are functions of $\boldsymbol{z}$ only, (5.4)–(5.6) are independent of $\boldsymbol{\eta}^T\boldsymbol{y}$.

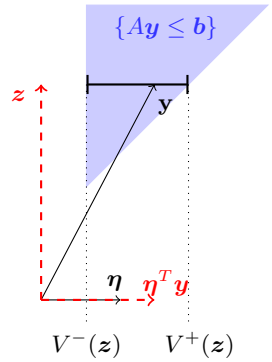

Fig 2: A geometric interpretation of why the event $\{A\boldsymbol{y} \le \boldsymbol{b}\}$ can be characterized as $\{\mathcal{V}^-(\boldsymbol{z}) \le \eta^T y \le \mathcal{V}^+(\boldsymbol{z})\}$. Assuming $\Sigma = I$ and $||\eta||_2 = 1$, $\mathcal{V}^-(\boldsymbol{z})$ and $\mathcal{V}^+(\boldsymbol{z})$ are functions of $\boldsymbol{z}$ only, which is independent of $\boldsymbol{\eta}^T\boldsymbol{y}$.

PROOF. We can decompose $\boldsymbol{y} = \boldsymbol{c}(\boldsymbol{\eta}^T\boldsymbol{y}) + \boldsymbol{z}$ and rewrite the polytope as

$$\begin{aligned}
\{A\boldsymbol{y} \le \boldsymbol{b}\} &= \{A(\boldsymbol{c}(\boldsymbol{\eta}^T\boldsymbol{y}) + \boldsymbol{z}) \le \boldsymbol{b}\} \\
&= \{A\boldsymbol{c}(\boldsymbol{\eta}^T\boldsymbol{y}) \le \boldsymbol{b} - A\boldsymbol{z}\} \\
&= \left\{(A\boldsymbol{c})_j(\boldsymbol{\eta}^T\boldsymbol{y}) \le b_j - (A\boldsymbol{z})_j \text{ for all } j\right\} \\
&= \left\{\begin{array}{ll} \boldsymbol{\eta}^T\boldsymbol{y} \le \frac{b_j - (A\boldsymbol{z})_j}{(A\boldsymbol{c})_j} & \text{for } j : (A\boldsymbol{c})_j > 0 \\ \boldsymbol{\eta}^T\boldsymbol{y} \ge \frac{b_j - (A\boldsymbol{z})_j}{(A\boldsymbol{c})_j} & \text{for } j : (A\boldsymbol{c})_j < 0 \\ 0 \le b_j - (A\boldsymbol{z})_j & \text{for } j : (A\boldsymbol{c})_j = 0 \end{array}\right\},
\end{aligned}$$

where in the last step, we have divided the components into three categories depending on whether $(A\boldsymbol{c})_j \gtreqless 0$, since this affects the direction of the inequality (or whether we can divide at all). Since $\boldsymbol{\eta}^T\boldsymbol{y}$ is the same quantity for all j, it must be at least the maximum of the lower bounds and no more than the minimum of the upper bounds, which is precisely the definition of $\mathcal{V}^-(\boldsymbol{z})$ and $\mathcal{V}^+(\boldsymbol{z})$. Finally, $b_j - (A\boldsymbol{z})_j \ge 0$ for all $j : (A\boldsymbol{c})_j = 0$ is encoded by $\mathcal{V}^0(\boldsymbol{z}) \ge 0$. □

Lemma 5.1 tells us that

$$\left[\boldsymbol{\eta}^T\boldsymbol{y} \mid \{A\boldsymbol{y} \le \boldsymbol{b}\}\right] \overset{d}{=} \left[\boldsymbol{\eta}^T\boldsymbol{y} \mid \{\mathcal{V}^-(\boldsymbol{z}) \le \boldsymbol{\eta}^T\boldsymbol{y} \le \mathcal{V}^+(\boldsymbol{z}), \mathcal{V}^0(\boldsymbol{z}) \ge 0\}\right] \tag{5.7}$$

Since $\mathcal{V}^+(\boldsymbol{z}), \mathcal{V}^-(\boldsymbol{z}), \mathcal{V}^0(\boldsymbol{z})$ are independent of $\boldsymbol{\eta}^T\boldsymbol{y}$, they behave as "fixed" quantities. Thus, $\boldsymbol{\eta}^T\boldsymbol{y}$ is conditionally like a normal random variable, truncated to be between $\mathcal{V}^-(\boldsymbol{z})$ and $\mathcal{V}^+(\boldsymbol{z})$. We would like to be able to say

$$\text{“ } \boldsymbol{\eta}^T\boldsymbol{y} \mid \{A\boldsymbol{y} \le \boldsymbol{b}\} \sim TN(\boldsymbol{\eta}^T\boldsymbol{\mu}, \sigma^2\boldsymbol{\eta}^T\Sigma\boldsymbol{\eta}, \mathcal{V}^-(\boldsymbol{z}), \mathcal{V}^+(\boldsymbol{z})),\text{”}$$

but this is technically incorrect, since the distribution on the right-hand side changes with $\boldsymbol{z}$. By conditioning on the value of $\boldsymbol{z}$, $\boldsymbol{\eta}^T\boldsymbol{y} \mid \{A\boldsymbol{y} \leq \boldsymbol{b}, \boldsymbol{z} = \boldsymbol{z}_0\}$ is a truncated normal. We then use the probability integral transform to obtain a statistic $F^{\boldsymbol{z}}(\boldsymbol{\eta}^T\boldsymbol{y})$ that has a $\text{Unif}(0,1)$ distribution for any value of $\boldsymbol{z}$. Hence, $F^{\boldsymbol{z}}(\boldsymbol{\eta}^T\boldsymbol{y})$ will also have a $\text{Unif}(0,1)$ distribution marginally over $\boldsymbol{z}$. We make this precise in the next theorem.

THEOREM 5.2. *Let $F^{[a,b]}_{\mu,\sigma^2}$ denote the CDF of a $N(\mu,\sigma^2)$ random variable truncated to the interval $[a,b]$, i.e.:*

$$F^{[a,b]}_{\mu,\sigma^2}(x) = \frac{\Phi((x-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)} \tag{5.8}$$

where Φ is the CDF of a $N(0,1)$ random variable. Then:

$$F^{[\mathcal{V}^-(\boldsymbol{z}),\mathcal{V}^+(\boldsymbol{z})]}_{\boldsymbol{\eta}^T\boldsymbol{\mu},\ \boldsymbol{\eta}^T\Sigma\boldsymbol{\eta}}(\boldsymbol{\eta}^T\boldsymbol{y}) \ \big|\ \{A\boldsymbol{y} \leq \boldsymbol{b}\} \sim \text{Unif}(0,1) \tag{5.9}$$

where $\mathcal{V}^-$ and $\mathcal{V}^+$ are defined in (5.4) and (5.5). Furthermore,

$$\left[\boldsymbol{\eta}^T\boldsymbol{y} \middle| A\boldsymbol{y} \leq \boldsymbol{b}, \boldsymbol{z} = \boldsymbol{z}_0\right] \sim TN(\boldsymbol{\eta}^T\boldsymbol{\mu}, \sigma^2||\boldsymbol{\eta}||^2, \mathcal{V}^-(\boldsymbol{z}_0), \mathcal{V}^+(\boldsymbol{z}_0)).$$

PROOF. First, apply Lemma 5.1:

$$\begin{aligned}\left[\boldsymbol{\eta}^T\boldsymbol{y} \middle| A\boldsymbol{y} \leq \boldsymbol{b}, \boldsymbol{z} = \boldsymbol{z}_0\right] &\overset{d}{=} \left[\boldsymbol{\eta}^T\boldsymbol{y} \middle| \mathcal{V}^-(\boldsymbol{z}) \leq \boldsymbol{\eta}^T\boldsymbol{y} \leq \mathcal{V}^+(\boldsymbol{z}), \mathcal{V}^0(\boldsymbol{z}) \geq 0, \boldsymbol{z} = \boldsymbol{z}_0\right] \\ &\overset{d}{=} \left[\boldsymbol{\eta}^T\boldsymbol{y} \middle| \mathcal{V}^-(\boldsymbol{z}_0) \leq \boldsymbol{\eta}^T\boldsymbol{y} \leq \mathcal{V}^+(\boldsymbol{z}_0), \mathcal{V}^0(\boldsymbol{z}_0) \geq 0, \boldsymbol{z} = \boldsymbol{z}_0\right]\end{aligned}$$

The only random quantities left are $\boldsymbol{\eta}^T\boldsymbol{y}$ and $\boldsymbol{z}$. Now we can eliminate $\boldsymbol{z} = \boldsymbol{z}_0$ from the condition using independence:

$$\begin{aligned}\left[\boldsymbol{\eta}^T\boldsymbol{y} \middle| A\boldsymbol{y} \leq \boldsymbol{b}, \boldsymbol{z} = \boldsymbol{z}_0\right] &\overset{d}{=} \left[\boldsymbol{\eta}^T\boldsymbol{y} \middle| \mathcal{V}^-(\boldsymbol{z}_0) \leq \boldsymbol{\eta}^T\boldsymbol{y} \leq \mathcal{V}^+(\boldsymbol{z}_0)\right] \\ &\sim TN(\boldsymbol{\eta}^T\boldsymbol{\mu}, \sigma^2||\boldsymbol{\eta}||^2, \mathcal{V}^-(\boldsymbol{z}_0), \mathcal{V}^+(\boldsymbol{z}_0))\end{aligned}$$

Letting $F^{\boldsymbol{z}}(\boldsymbol{\eta}^T\boldsymbol{y}) \equiv F^{[\mathcal{V}^-(\boldsymbol{z}),\mathcal{V}^+(\boldsymbol{z})]}_{\boldsymbol{\eta}^T\boldsymbol{\mu},\ \boldsymbol{\eta}^T\Sigma\boldsymbol{\eta}}(\boldsymbol{\eta}^T\boldsymbol{y})$, we can apply the probability integral transform to the above result to obtain

$$\begin{aligned}\left[F^{\boldsymbol{z}}(\boldsymbol{\eta}^T\boldsymbol{y}) \middle| A\boldsymbol{y} \leq \boldsymbol{b}, \boldsymbol{z} = \boldsymbol{z}_0\right] &\overset{d}{=} \left[F^{\boldsymbol{z}_0}(\boldsymbol{\eta}^T\boldsymbol{y}) \middle| A\boldsymbol{y} \leq \boldsymbol{b}, \boldsymbol{z} = \boldsymbol{z}_0\right] \\ &\sim \text{Unif}(0,1)\end{aligned}$$

If we let p_X denote the density of a random variable X given $\{A\boldsymbol{y} \leq \boldsymbol{b}\}$, what we have just shown is that

$$p_{F^{\boldsymbol{z}}(\boldsymbol{\eta}^T\boldsymbol{y})|\boldsymbol{z}}(t|\boldsymbol{z}_0) \equiv \frac{p_{F^{\boldsymbol{z}}(\boldsymbol{\eta}^T\boldsymbol{y}),\boldsymbol{z}}(t,\boldsymbol{z}_0)}{p_{\boldsymbol{z}}(\boldsymbol{z}_0)} = 1_{[0,1]}(f)$$

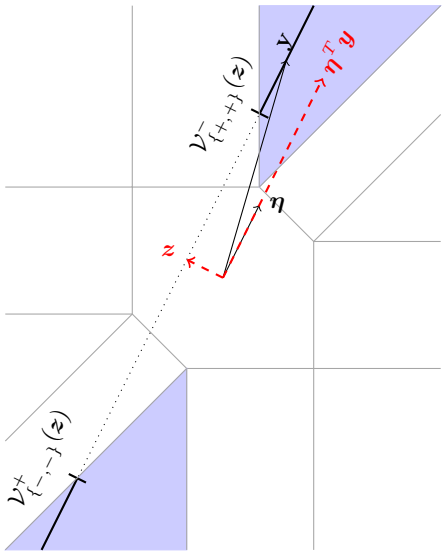

Fig 3: When we take the union over signs, the conditional distribution of $\boldsymbol{\eta}^T\boldsymbol{y}$ is truncated to a union of disjoint intervals. In this case, the Gaussian is truncated to the set $(-\infty, \mathcal{V}^+_{\{-,-\}}(\boldsymbol{z})] \cup [\mathcal{V}^-_{\{+,+\}}(\boldsymbol{z}), \infty)$.

for any $\boldsymbol{z}_0$. The desired result now follows by integrating over $\boldsymbol{z}_0$:

$$\begin{aligned} p_{F^{\boldsymbol{z}}(\boldsymbol{\eta}^T\boldsymbol{y})}(t) &= \int p_{F^{\boldsymbol{z}}(\boldsymbol{\eta}^T\boldsymbol{y})|\boldsymbol{z}}(t|\boldsymbol{z}_0)\, p_{\boldsymbol{z}}(\boldsymbol{z}_0)\, d\boldsymbol{z}_0 \\ &= \int 1_{[0,1]}(t)\, p_{\boldsymbol{z}}(\boldsymbol{z}_0)\, d\boldsymbol{z}_0 \\ &= 1_{[0,1]}(t). \end{aligned}$$

□

5.2. *Conditioning on a Union of Polytopes.* We have just characterized the distribution of $\boldsymbol{\eta}^T\boldsymbol{y}$, conditional on $\boldsymbol{y}$ falling into a single polytope $\{A\boldsymbol{y} \leq \boldsymbol{b}\}$. We obtain such a polytope if we condition on both the model and the signs $\{\hat{M} = M, \hat{\boldsymbol{s}} = \boldsymbol{s}\}$. If we want to only condition on the model $\{\hat{M} = M\}$, then we will have to understand the distribution of $\boldsymbol{\eta}^T\boldsymbol{y}$, conditional on $\boldsymbol{y}$ falling into a union of such polytopes, i.e.,

$$\boldsymbol{\eta}^T\boldsymbol{y} \,\Big|\, \bigcup_{\boldsymbol{s}} \{A_{\boldsymbol{s}}\boldsymbol{y} \leq \boldsymbol{b}_{\boldsymbol{s}}\}. \tag{5.10}$$

As Figure 3 makes clear, the argument proceeds exactly as before, except that $\boldsymbol{\eta}^T\boldsymbol{y}$ is now truncated to a union of intervals, instead of a single interval. There is a $\mathcal{V}^-$ and a $\mathcal{V}^+$ for each possible sign pattern $\boldsymbol{s}$, so we index the intervals by the signs. This leads immediately to the next theorem, whose proof is essentially the same as that of Theorem 5.2.

THEOREM 5.3. *Let F^S_{μ,σ^2} denote the CDF of a $N(\mu,\sigma^2)$ random variable truncated to the set S. Then:*

$$F^{\bigcup_{\boldsymbol{s}}[\mathcal{V}^-_{\boldsymbol{s}}(\boldsymbol{z}),\mathcal{V}^+_{\boldsymbol{s}}(\boldsymbol{z})]}_{\boldsymbol{\eta}^T\boldsymbol{\mu},\ \boldsymbol{\eta}^T\Sigma\boldsymbol{\eta}}(\boldsymbol{\eta}^T\boldsymbol{y}) \;\Big|\; \bigcup_{\boldsymbol{s}} \{A_{\boldsymbol{s}}\boldsymbol{y} \le \boldsymbol{b}_{\boldsymbol{s}}\}. \sim \text{Unif}(0,1), \tag{5.11}$$

where $\mathcal{V}^-_{\boldsymbol{s}}(\boldsymbol{z})$ and $\mathcal{V}^+_{\boldsymbol{s}}(\boldsymbol{z})$ are defined in (5.4) and (5.5) and $A = A_{\boldsymbol{s}}$ and $b = b_{\boldsymbol{s}}$.

6. Post-Selection Intervals for Regression Coefficients. In this section, we combine the characterization of the lasso selection event in Section 4 with the results about the distribution of a Gaussian truncated to a (union of) polytope(s) in Section 5 to form post-selection intervals for lasso-selected regression coefficients. The key link is that the lasso selection event can be expressed as a union of polytopes:

$$\begin{aligned}\{\hat{M} = M\} &= \bigcup_{\boldsymbol{s}\in\{-1,1\}^{|M|}} \{\hat{M} = M, \hat{\boldsymbol{s}} = \boldsymbol{s}\} \\ &= \bigcup_{\boldsymbol{s}\in\{-1,1\}^{|M|}} \{A(M,\boldsymbol{s})\boldsymbol{y} \le \boldsymbol{b}(M,\boldsymbol{s})\},\end{aligned}$$

where $A(M,\boldsymbol{s})$ and $\boldsymbol{b}(M,\boldsymbol{s})$ are defined in Theorem 4.3. Therefore, conditioning on selection is the same as conditioning on a union of polytopes, so the framework of Section 5 applies.

Recall that our goal is to form confidence intervals for $\beta_j^M = \boldsymbol{e}_j^T X_M^+ \boldsymbol{\mu}$, with $(1-\alpha)$-coverage conditional on $\{\hat{M} = M\}$. Taking $\boldsymbol{\eta} = (X_M^+)^T\boldsymbol{e}_j$, we can use Theorem 5.3 to obtain

$$F^{\bigcup_{\boldsymbol{s}}[\mathcal{V}^-_{\boldsymbol{s}}(\boldsymbol{z}),\mathcal{V}^+_{\boldsymbol{s}}(\boldsymbol{z})]}_{\beta_j^M,\sigma^2||\boldsymbol{\eta}||^2}(\boldsymbol{\eta}^T\boldsymbol{y}) \;\big|\; \{\hat{M} = M\} \sim \text{Unif}(0,1).$$

This gives us a test statistic for testing any hypothesized value of $\boldsymbol{\beta}_j^M$. We can invert this test to obtain a confidence set

$$C_j^M \equiv \left\{\beta_j^M : \frac{\alpha}{2} \le F^{\bigcup_{\boldsymbol{s}}[\mathcal{V}^-_{\boldsymbol{s}}(\boldsymbol{z}),\mathcal{V}^+_{\boldsymbol{s}}(\boldsymbol{z})]}_{\beta_j^M,\sigma^2||\boldsymbol{\eta}||^2}(\boldsymbol{\eta}^T\boldsymbol{y}) \le 1-\frac{\alpha}{2}\right\}. \tag{6.1}$$

In fact, the set C_j^M is an *interval*, as formalized in the next result.

THEOREM 6.1. *Let* $\boldsymbol{\eta} = (X_M^+)^T \boldsymbol{e}_j$. *Let* L *and* U *be the (unique) values satisfying*

$$F^{\bigcup_s [\mathcal{V}_s^-(\boldsymbol{z}), \mathcal{V}_s^+(\boldsymbol{z})]}_{L,\ \sigma^2||\boldsymbol{\eta}||^2}(\boldsymbol{\eta}^T \boldsymbol{y}) = 1 - \frac{\alpha}{2} \qquad F^{\bigcup_s [\mathcal{V}_s^-(\boldsymbol{z}), \mathcal{V}_s^+(\boldsymbol{z})]}_{U,\ \sigma^2||\boldsymbol{\eta}||^2}(\boldsymbol{\eta}^T \boldsymbol{y}) = \frac{\alpha}{2}$$

Then $[L, U]$ *is a* $(1-\alpha)$ *confidence interval for* β_j^M, *conditional on* $\{\hat{M} = M\}$, *i.e.,*

$$\mathbb{P}\left(\beta_j^M \in [L, U] \;\middle|\; \hat{M} = M\right) = 1 - \alpha. \tag{6.2}$$

PROOF. By construction, $\mathbb{P}_{\beta_j^M}(\beta_j^M \in C_j^M | \hat{M} = M) = 1 - \alpha$, where C_j^M is defined in (6.1). The claim is that the set C_j^M is in fact the interval $[L, U]$. To see this, we need to show that the test statistic $F^{\bigcup_s [\mathcal{V}_s^-(\boldsymbol{z}), \mathcal{V}_s^+(\boldsymbol{z})]}_{L,\ \sigma^2||\boldsymbol{\eta}||^2}(\boldsymbol{\eta}^T \boldsymbol{y})$ is monotone decreasing in β_j^M so that it crosses $1 - \frac{\alpha}{2}$ and $\frac{\alpha}{2}$ at unique values. This follows from the fact that the truncated Gaussian distribution has monotone likelihood ratio in the mean parameter. See Appendix A for details. □

Alternatively, we could have conditioned on the signs, in addition to the model, so that we would only have to worry about conditioning on a single polytope. We also showed in Section 5 that

$$F^{[\mathcal{V}_s^-(\boldsymbol{z}), \mathcal{V}_s^+(\boldsymbol{z})]}_{\boldsymbol{\beta}_j^M, \sigma^2||\boldsymbol{\eta}||^2}(\boldsymbol{\eta}^T \boldsymbol{y}) \mid \{\hat{M} = M, \hat{\boldsymbol{s}} = \boldsymbol{s}\} \sim \text{Unif}(0, 1).$$

Inverting this statistic will produce intervals that have $(1-\alpha)$ coverage conditional on $\{\hat{M} = M, \hat{\boldsymbol{s}} = \boldsymbol{s}\}$, and hence, $(1-\alpha)$ coverage conditional on $\{\hat{M} = M\}$. However, these intervals will be less efficient; they will in general be wider. However, one may be willing to sacrifice statistical efficiency for computational efficiency. Notice that the main cost in computing intervals according to Theorem 6.1 is determining the intervals $[\mathcal{V}_s^-(\boldsymbol{z}), \mathcal{V}_s^+(\boldsymbol{z})]$ for each $\boldsymbol{s} \in \{-1, 1\}^{|M|}$. The number of such sign patterns is $2^{|M|}$. While this might be feasible when $|M|$ is, say, less than 15, it is not feasible when we select hundreds of variables. Conditioning on the signs means that we only have to compute the interval $[\mathcal{V}_s^-(\boldsymbol{z}), \mathcal{V}_s^+(\boldsymbol{z})]$ for the sign pattern $\boldsymbol{s}$ that was actually observed.

Figure 4 shows the tradeoff in statistical efficiency. When the signal is strong, as in the left-hand plot, there is virtually no difference between the intervals obtained by conditioning on just the model, or the model and signs. On the other hand, in the right-hand plot, we see that we can obtain very

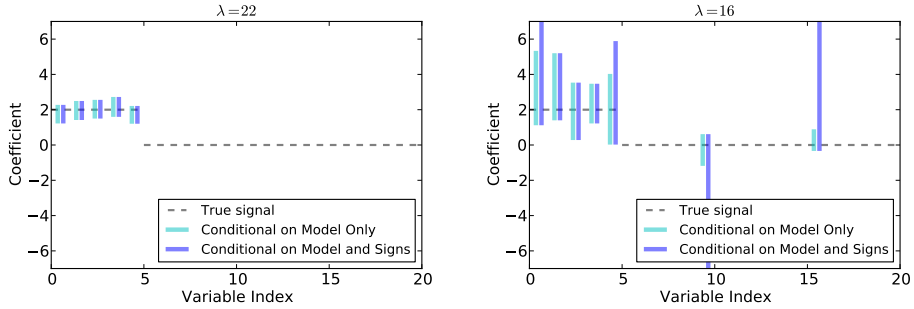

Fig 4: Comparison of the confidence intervals by conditioning on the model only (statistically more efficient, but computationally more expensive) and conditioning on both the model and signs (statistically less efficient, but computationally more feasible). Data were simulated for $n = 25$, $p = 50$, and 5 true non-zero coefficients; only the first 20 coefficients are shown. (Variables with no intervals are included to emphasize that inference is only on the selected variables.) Conditioning on the signs in addition to the model results in no loss of statistical efficiency when the signal is strong (left) but is problematic when the signal is weak (right).

wide intervals when the signal is weak. The widest intervals are for actual noise variables, as expected.

To understand why post-selection intervals are sometimes very wide, notice that when a truncated Gaussian random variable Z is close to the endpoints of the truncation interval $[a, b]$, there are many means μ that would be consistent with that observation—hence, the wide intervals. Figure 5 shows confidence intervals for μ as a function of Z. When Z is far from the endpoints of the truncation interval, we basically recover the nominal OLS intervals (i.e., not adjusted for selection).

The implications are clear. When the signal is strong, $\boldsymbol{\eta}^T \boldsymbol{y}$ will be far from the endpoints of the truncation region, so we obtain the nominal OLS intervals. On the other hand, when a variable just barely entered the model, then $\boldsymbol{\eta}^T \boldsymbol{y}$ will be close to the edge of the truncation region, and the interval will be wide.

6.1. *Optimality.* We have derived a confidence interval C_j^M whose conditional coverage, given $\{\hat{M} = M\}$, is $1-\alpha$. The fact that we have found such an interval is not remarkable, since many such intervals have this property, including the trivial interval $(-\infty, \infty)$. However, given two intervals with the same coverage, we generally prefer the shorter one. We now show that C_j^M

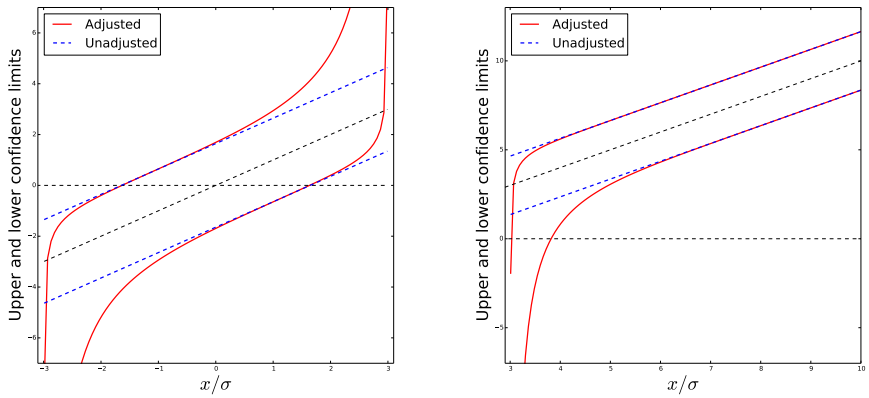

Fig 5: Upper and lower bounds of 90% confidence intervals for μ based on a single observation $x/\sigma \sim TN(0, 1, -3, 3)$. We see that as long as the observation x is roughly 0.5σ away from either boundary, the size of the intervals is comparable to the unadjusted OLS confidence interval.

is, with one small tweak, the interval with expected shortest length among all *unbiased* intervals with $1 - \alpha$ coverage.

An *unbiased* interval C for a parameter θ is one which covers no other parameter θ' with probability more than $1 - \alpha$, i.e.,

$$\mathbb{P}_\theta(\theta' \in C) \leq 1 - \alpha, \text{ for all } \theta,\, \theta' \neq \theta. \tag{6.3}$$

Unbiasedness is a common restriction to ensure that there is an optimal interval or test at all (Lehmann and Romano, 2005). The interval with expected shortest length for β_j^M, among all intervals with conditional $1 - \alpha$ coverage, is similar to the interval $[L, U]$ in Theorem 6.1. The only difference is that the critical values L and U were chosen symmetrically so that the pivot has $\alpha/2$ area in either tail. However, allocating the α probability equally to either tail may not be optimal in general. The next theorem provides the general recipe for constructing optimal intervals:

THEOREM 6.2. *Let* $\boldsymbol{\eta} = (X_M^+)^T \boldsymbol{e}_j$ *and define the set*

$$C_{j,opt}^M \equiv \left\{ \beta_j^M : \boldsymbol{\eta}^T \boldsymbol{y} \in R(\boldsymbol{z}) \right\}, \tag{6.4}$$

where the rejection region $R(\boldsymbol{z})$ *is defined in* (6.6) *below and is a function of* $\boldsymbol{z}$ (5.2)*. Then* $C_{j,opt}^M$ *is the interval with expected shortest length for* β_j^M

with $1-\alpha$ *coverage, conditional on* $\{\hat{M}=M\}$, *i.e.,*

$$\mathbb{P}\left(\beta_j^M \in C_{j,opt}^M \;\middle|\; \hat{M}=M\right)=1-\alpha. \tag{6.5}$$

PROOF. Without loss of generality, assume $||\boldsymbol{\eta}||_2=1$ and $\sigma=1$. Note that the distribution of $\boldsymbol{y}$, conditional on $\{\hat{M}=M\}$, is an exponential family. That is, with respect to some base measure ν on $\{\hat{M}=M\}$ and for some log-normalizing function ψ_M, its density is

$$\begin{aligned}\boldsymbol{y} &\sim \exp\left\{\boldsymbol{\mu}^T\boldsymbol{y}-\psi_M(\boldsymbol{\mu})\right\}\,d\nu(\boldsymbol{y})\\ &=\exp\left\{\boldsymbol{\mu}^T\left[P_{\boldsymbol{\eta}}\boldsymbol{y}+(I-P_{\boldsymbol{\eta}})\boldsymbol{y}\right]-\psi_M(\boldsymbol{\mu})\right\}\,d\nu(\boldsymbol{y})\\ &=\exp\left\{(\boldsymbol{\eta}^T\boldsymbol{\mu})(\boldsymbol{\eta}^T\boldsymbol{y})+((I-P_{\boldsymbol{\eta}})\boldsymbol{\mu})^T\boldsymbol{z}-\psi_M(\boldsymbol{\mu})\right\}\,d\nu(\boldsymbol{y}).\end{aligned}$$

This is an exponential family with one parameter of interest $(\boldsymbol{\eta}^T\boldsymbol{\mu})$ and $(n-1)$ nuisance parameters, represented by $(I-P_{\boldsymbol{\eta}})\boldsymbol{\mu}$.[1] Classical theory (cf. Theorem 4.4.1 in Lehmann and Romano (2005)) says that the uniformly most powerful unbiased (UMPU) test of $H_0:\boldsymbol{\eta}^T\boldsymbol{\mu}=\beta_j^M$ versus $H_1:\boldsymbol{\eta}^T\boldsymbol{\mu}\neq\beta_j^M$ is obtained by conditioning on $\boldsymbol{z}$ and rejecting for values of $\boldsymbol{\eta}^T\boldsymbol{y}$ that are too large or too small. In other words, we reject for $\boldsymbol{\eta}^T\boldsymbol{y}\in R(\boldsymbol{z})$, where

$$R(\boldsymbol{z})\equiv(-\infty,C_1(\boldsymbol{z})]\cup[C_2(\boldsymbol{z}),\infty), \tag{6.6}$$

where $C_1(\boldsymbol{z})$ and $C_2(\boldsymbol{z})$ are chosen to ensure:

1. The test is level α. Letting $\varphi(x)=\frac{e^{-x^2}}{\sqrt{2\pi}}$ and $S(\boldsymbol{z})\equiv\bigcup_{\boldsymbol{s}}[\mathcal{V}_{\boldsymbol{s}}^-(\boldsymbol{z}),\mathcal{V}_{\boldsymbol{s}}^+(\boldsymbol{z})]$,
$$\frac{\int_{R(\boldsymbol{z})\cap S(\boldsymbol{z})}\varphi(x-\beta_j^M)\,dx}{\int_{S(\boldsymbol{z})}\varphi(x-\beta_j^M)\,dx}=\alpha$$
for almost every $\boldsymbol{z}$.
2. Since the test is unbiased, its power function must be minimized under H_0, so its derivative at β_j^M must be 0, yielding the condition
$$\frac{\int_{R(\boldsymbol{z})\cap S(\boldsymbol{z})}x\varphi(x-\beta_j^M)\,dx}{\int_{S(\boldsymbol{z})}\varphi(x-\beta_j^M)\,dx}=\alpha\frac{\int_{S(\boldsymbol{z})}x\varphi(x-\beta_j^M)\,dx}{\int_{S(\boldsymbol{z})}\varphi(x-\beta_j^M)\,dx}.$$

This gives us the UMPU test of $H_0:\boldsymbol{\eta}^T\boldsymbol{y}=\beta_j^M$ for every β_j^M. Finally, the UMAU interval can be obtained by inverting the UMPU test, i.e.,

$$C_{j,opt}^M=\left\{\beta_j^M:\boldsymbol{\eta}^T\boldsymbol{y}\in R(\boldsymbol{z})\right\}.$$

The construction above is standard, and the details can be found in Chapter 4 of Lehmann and Romano (2005). By the Ghosh-Pratt theorem, this is the unbiased interval with expected shortest length. □

[1] Although this is technically a vector in $\mathbb{R}^n$, it is only $(n-1)$-dimensional.

7. Data Example. We apply our post-selection intervals to the diabetes data set from Efron et al. (2004). After standardizing all variables, we chose λ according to the strategy in Negahban et al. (2012), $\lambda = 2\,\mathbf{E}(\|X^T\epsilon\|_\infty)$, using an estimate of σ from the full model, resulting in $\lambda \approx 190$. The lasso selected four variables: `BMI`, `BP`, `S3`, and `S5`.

The post-selection intervals are shown in Figure 6, alongside the nominal confidence intervals produced by fitting OLS to the four selected variables, ignoring selection. The nominal intervals do not have $(1-\alpha)$ coverage conditional on the model and are not valid post-selection intervals. Also depicted are the confidence intervals obtained by *data splitting*; that is, if one splits the n observations into two halves, then uses one half for selection and the other for inference. This is a competitor method that also produces valid confidence intervals conditional on the model. The lasso selected the same four variables on half of the data, and then nominal intervals for these four variables using OLS on the other half of the data.

We can make two observations from Figure 6.

1. The adjusted intervals provided by our method essentially reproduces the OLS intervals for the strong effects, whereas data splitting intervals are wider by a factor of $\sqrt{2}$ (since only $n/2$ observations are used in the inference). For this dataset, the POSI intervals are 1.36 times wider than the OLS intervals. For all the variables, our method produces the shortest intervals among the methods that control selective type 1 error.
2. One variable, `S3`, which would have been deemed significant using the OLS intervals, is no longer significant after accounting for selection. Data splitting, our selection-adjusted intervals, and POSI intervals conclude that `S3` is not significant. This demonstrates that taking model selection into account can have substantive impacts on the conclusions.

8. Testing the Lasso-Selected Model. Having observed that the lasso selected the variables $\hat{M}$, another relevant question is whether it has captured all of the signal in the model, i.e.,

$$H_0 : \boldsymbol{\beta}^0_{-\hat{M}} = \mathbf{0}. \tag{8.1}$$

We consider a slightly more general question, which does not assume the correctness of the linear model $\boldsymbol{\mu} = X\boldsymbol{\beta}^0$ and also takes into account whether the non-selected variables can improve the fit:

$$H_0 : X^T_{-\hat{M}}(I - P_{-\hat{M}})\boldsymbol{\mu} = \mathbf{0}. \tag{8.2}$$

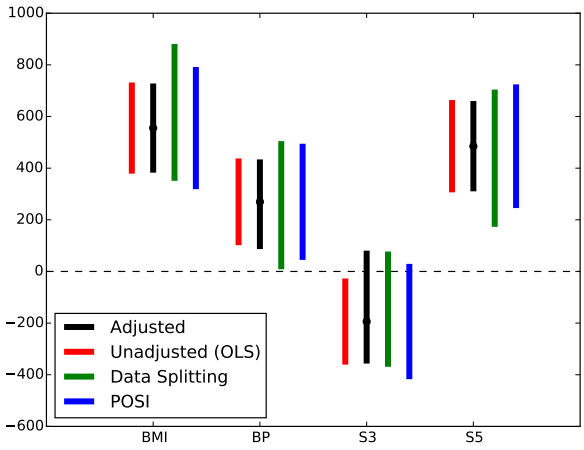

Fig 6: Inference for the four variables selected by the lasso ($\lambda = 190$) on the diabetes data set. The point estimate and adjusted confidence intervals using the approach in Section 6 are shown in black. The OLS intervals, which ignore selection, are shown in red. The green lines show the intervals produced by splitting the data into two halves, forming the interval based on only half of the data. The blue line corresponds to the POSI method of Berk et al. (2013)

.

This quantity is the partial correlation of the non-selected variables with μ, adjusting for the variables in $\hat{M}$. This is more general because if we assume $\boldsymbol{\mu} = X\boldsymbol{\beta}^0$ for some $\boldsymbol{\beta}^0$ and X is full rank, then rejecting (8.2) implies that there exists $i \in \text{supp}(\boldsymbol{\beta}^0)$ not in $\hat{M}$, so we would also reject (8.1).

The natural approach is to compare the observed partial correlations $X_{-M}^T(I - P_M)\boldsymbol{y}$ to $\mathbf{0}$. However, the framework of Section 5 only allows tests of $\boldsymbol{\mu}$ in a single direction $\boldsymbol{\eta}$. To make use of that framework, we can choose $\boldsymbol{\eta}$ such that it selects the maximum magnitude of $X_{-M}^T(I - P_M)\boldsymbol{y}$. In particular, this direction provides the most evidence against the null hypothesis of zero partial correlation, so if the null hypothesis cannot be rejected in this direction, it would not be rejected in any direction.

Letting $\hat{\boldsymbol{v}} := \text{argmax}_{\boldsymbol{v} \in B_\infty} \boldsymbol{v}^T X_{-M}^T (I - P_M)\boldsymbol{y}$, ($B_\infty$ is the ℓ_∞ ball) we set

$$\boldsymbol{\eta} = (I - P_M)X_{-M}\hat{\boldsymbol{v}}, \tag{8.3}$$

and test $H_0 : \boldsymbol{\eta}^T\boldsymbol{\mu} = 0$. To ensure $\boldsymbol{\eta}$ is measurable with respect to the

selection event, we condition on not only on $(\hat{M}, \hat{\boldsymbol{s}})$, but also on $\hat{\boldsymbol{v}}$:

$$\{(\hat{M}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{v}}) = (M, \boldsymbol{s}, \boldsymbol{v})\}. \tag{8.4}$$

A test that is level α conditional on (8.4) for all $(\hat{M}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{v}})$ is also level α conditional on $(\hat{M}, \hat{\boldsymbol{s}})$.

In order to use the results of Section 5, we must show that (8.4) can be written in the form $A(M, \boldsymbol{s}, \boldsymbol{v})y \leq b(M, \boldsymbol{s}, \boldsymbol{v})$. This is indeed possible, and the following proposition provides an explicit construction.

PROPOSITION 8.1. *Let $A_0, \boldsymbol{b}_0, A_1, \boldsymbol{b}_1$ be defined as in Proposition 4.2. Then:*

$$\{(\hat{M}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{v}}) = (M, \boldsymbol{s}, \boldsymbol{v})\} = \left\{ \begin{pmatrix} A_0(M, \boldsymbol{s}) \\ A_1(M, \boldsymbol{s}) \\ A_2(M, \boldsymbol{v}) \end{pmatrix} \boldsymbol{y} < \begin{pmatrix} \boldsymbol{b}_0(M, \boldsymbol{s}) \\ \boldsymbol{b}_1(M, \boldsymbol{s}) \\ \boldsymbol{0} \end{pmatrix} \right\}$$

where $A_2(M, \boldsymbol{v})$ is defined as

$$A_2(M, \boldsymbol{v}) = D(\boldsymbol{v}) X_{-M}^T (I - P_M),$$

where the rows of $D(\boldsymbol{v}) \in \mathbf{R}^{(2|M|-1)\times|M|}$ are given by $(\boldsymbol{w} - \boldsymbol{v})^T$ for all $\boldsymbol{w} \in \text{ext}(\mathcal{B}_\infty)$.

PROOF. The constraints $\{A_0 y < \boldsymbol{b}_0\}$ and $\{A_1 y < \boldsymbol{b}_1\}$ come from Proposition (4.2) and encode the constraints $\{(\hat{M}, \hat{\boldsymbol{s}}) = (M, \boldsymbol{s})\}$. We show that the last two sets of constraints encode $\{\hat{\boldsymbol{v}} = \boldsymbol{v}\}$.

Let $\boldsymbol{r} := X_{-M}^T (I - P_M)\boldsymbol{y}$ denote the vector of partial correlations. Since B_∞ is a polytope, the maximum of $\text{argmax}_{\boldsymbol{v} \in B_\infty} \boldsymbol{v}^T X_{-M}^T (I - P_M)\boldsymbol{y}$ is attained at an extreme point. Thus $\{\hat{\boldsymbol{v}} = \boldsymbol{v}\} = \{D(\boldsymbol{v})\boldsymbol{r} < 0\}$. □

Because of Proposition 8.1, we can now obtain the following result as a simple consequence of Theorem 5.2, which says that $F_{0,\sigma^2||\boldsymbol{\eta}||^2}^{[\mathcal{V}^-,\mathcal{V}^+]}(\boldsymbol{\eta}^T\boldsymbol{y}) \sim \text{Unif}(0,1)$, conditional on the set (8.4) and H_0. We reject when $F_{0,\sigma^2||\boldsymbol{\eta}||^2}^{[\mathcal{V}^-,\mathcal{V}^+]}(\boldsymbol{\eta}^T\boldsymbol{y})$ is large because $F_{0,\ \sigma^2||\boldsymbol{\eta}||^2}^{[\mathcal{V}^-,\mathcal{V}^+]}(\cdot)$ is monotone increasing in the argument and $\boldsymbol{\eta}^T\boldsymbol{\mu}$ is likely to be positive under the alternative.

COROLLARY 8.2. *Let H_0 and $\boldsymbol{\eta}$ be defined as in (8.3). Then, the test which rejects when*

$$\left\{ F_{0,\ \sigma^2||\boldsymbol{\eta}||^2}^{[\mathcal{V}^-,\mathcal{V}^+]}(\boldsymbol{\eta}^T\boldsymbol{y}) > 1 - \alpha \right\}$$

is level α, conditional on $\{(\hat{M}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{v}}) = (M, \boldsymbol{s}, \boldsymbol{v})\}$. That is,

$$\mathbb{P}_0\left(F^{[\mathcal{V}^-,\mathcal{V}^+]}_{0,\ \sigma^2||\boldsymbol{\eta}||^2}(\boldsymbol{\eta}^T\boldsymbol{y}) > 1-\alpha \;\middle|\; \{(\hat{M}, \hat{\boldsymbol{s}}, \hat{\boldsymbol{v}}) = (M, \boldsymbol{s}, \boldsymbol{v})\}\right) = \alpha.$$

In particular, since this holds for every $(M, \boldsymbol{s}, \boldsymbol{v})$, this test also controls Type I error conditional only on $(\hat{M}, \hat{\boldsymbol{s}})$, and unconditionally:

$$\mathbb{P}_0\left(F^{[\mathcal{V}^-,\mathcal{V}^+]}_{0,\ \sigma^2||\boldsymbol{\eta}||^2}(\boldsymbol{\eta}^T\boldsymbol{y}) > 1-\alpha \;\middle|\; \{(\hat{M}, \hat{\boldsymbol{s}}) = (M, \boldsymbol{s})\}\right) = \alpha$$
$$\mathbb{P}_0\left(F^{[\mathcal{V}^-,\mathcal{V}^+]}_{0,\ \sigma^2||\boldsymbol{\eta}||^2}(\boldsymbol{\eta}^T\boldsymbol{y}) > 1-\alpha\right) = \alpha.$$

9. Extension to the Elastic Net. One problem with the lasso is that it tends to select one variable out of a set of correlated variables, resulting in estimates that are unstable. One way to stabilize them is to add an ℓ_2 penalty to the lasso objective, resulting in the elastic net (Zou and Hastie, 2005):

$$\tilde{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \frac{1}{2}\|\boldsymbol{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 + \frac{\gamma}{2}\|\boldsymbol{\beta}\|_2^2. \tag{9.1}$$

Using a nearly identical argument to Lemma 4.1, we see that $\{\hat{M} = M, \hat{\boldsymbol{s}} = \boldsymbol{s}\}$ if and only if there exist $\tilde{\boldsymbol{w}}$ and $\tilde{\boldsymbol{u}}$ satisfying

$$\begin{aligned}(X_M^T X_M + \gamma I)\tilde{\boldsymbol{w}} - X_M^T\boldsymbol{y} + \lambda\boldsymbol{s} &= 0\\ X_{-M}^T(X_M\tilde{\boldsymbol{w}} - \boldsymbol{y}) + \lambda\tilde{\boldsymbol{u}} &= 0\\ \operatorname{sign}(\tilde{\boldsymbol{w}}) &= \boldsymbol{s}\\ ||\tilde{\boldsymbol{u}}||_\infty &< 1.\end{aligned}$$

These four conditions differ from those of Lemma 4.1 in only one respect: $X_M^T X_M$ in the first expression is replaced by $X_M^T X_M + \gamma I$. Continuing the argument of Section 4, we see that the selection event can be rewritten

$$\{\hat{M} = M, \hat{\boldsymbol{s}} = \boldsymbol{s}\} = \left\{\begin{pmatrix}\tilde{A}_0(M,\boldsymbol{s})\\ \tilde{A}_1(M,\boldsymbol{s})\end{pmatrix}\boldsymbol{y} < \begin{pmatrix}\tilde{\boldsymbol{b}}_0(M,\boldsymbol{s})\\ \tilde{\boldsymbol{b}}_1(M,\boldsymbol{s})\end{pmatrix}\right\} \tag{9.2}$$

where $\tilde{A}_k$ and $\tilde{\boldsymbol{b}}_k$ are analogous to A_k and $\boldsymbol{b}_k$ in Proposition 4.2, except with $(X_M^T X_M)^{-1}$ replaced by $(X_M^T X_M + \gamma I)^{-1}$. Notice that this quantity not only appears explicitly in A_1 and $\boldsymbol{b}_1$, but also appears implicitly in A_0 and $\boldsymbol{b}_0$ through P_M and $(X_M^T)^+$.

Now that we have rewritten the selection event in the form $\{A\boldsymbol{y} \le \boldsymbol{b}\}$, we can once again apply the framework in Section 5 to obtain a test for the elastic net conditional on this event.

10. Conclusion. Model selection and inference have long been regarded as conflicting goals in linear regression. Following the lead of Berk et al. (2013), we have proposed a framework for post-selection inference that *conditions on which model was selected*, i.e., the event $\{\hat{M} = M\}$. We characterize this event for the lasso and derive optimal and exact confidence intervals for linear contrasts $\boldsymbol{\eta}^T\boldsymbol{\mu}$, conditional on $\{\hat{M} = M\}$. With this general framework, we can form post-selection intervals for regression coefficients, equipping practitioners with a way to obtain "valid" intervals even after model selection.

Acknowledgements. We thank Will Fithian, Sam Gross, and Josh Loftus for helpful comments and discussions. In particular, Will Fithian provided insights that led to the geometric intuition of our procedure shown in Figure 2. J. Lee was supported by a National Defense Science and Engineering Graduate Fellowship and a Stanford Graduate Fellowship. D. L. Sun was supported by a Ric Weiland Graduate Fellowship and the Stanford Genome Training Program (SGTP; NIH/NHGRI). Y. Sun was partially supported by the NIH, grant U01GM102098. J.E. Taylor was supported by the NSF, grant DMS 1208857, and by the AFOSR, grant 113039.

APPENDIX A: MONOTONICITY OF F

LEMMA A.1. *Let $F_\mu(x) := F^{[a,b]}_{\mu,\sigma^2}(x)$ denote the cumulative distribution function of a truncated Gaussian random variable, as defined as in (5.8). Then $F_\mu(x)$ is monotone decreasing in μ.*

PROOF. First, the truncated Gaussian distribution with CDF $F_\mu := F^{[a,b]}_{\mu,\sigma^2}$ is a natural exponential family in μ, since it is just a Gaussian with a different base measure. Therefore, it has monotone likelihood ratio in μ. That is, for all $\mu_1 > \mu_0$ and $x_1 > x_0$:

$$\frac{f_{\mu_1}(x_1)}{f_{\mu_0}(x_1)} > \frac{f_{\mu_1}(x_0)}{f_{\mu_0}(x_0)}$$

where $f_{\mu_i} := dF_{\mu_i}$ denotes the density. (Instead of appealing to properties of exponential families, this property can also be directly verified.)

This implies

$$f_{\mu_1}(x_1)f_{\mu_0}(x_0) > f_{\mu_1}(x_0)f_{\mu_0}(x_1) \qquad\qquad x_1 > x_0.$$

Therefore, the inequality is preserved if we integrate both sides with respect

to x_0 on $(-\infty, x)$ for $x < x_1$. This yields:

$$\int_{-\infty}^{x} f_{\mu_1}(x_1) f_{\mu_0}(x_0)\, dx_0 > \int_{-\infty}^{x} f_{\mu_1}(x_0) f_{\mu_0}(x_1)\, dx_0 \qquad x < x_1$$

$$f_{\mu_1}(x_1) F_{\mu_0}(x) > f_{\mu_0}(x_1) F_{\mu_1}(x) \qquad x < x_1$$

Now we integrate both sides with respect to x_1 on (x, ∞) to obtain:

$$(1 - F_{\mu_1}(x)) F_{\mu_0}(x) > (1 - F_{\mu_0}(x)) F_{\mu_1}(x)$$

which establishes $F_{\mu_0}(x) > F_{\mu_1}(x)$ for all $\mu_1 > \mu_0$. □

References.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, **100** 71–81.

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

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, **32** 407–499.

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

Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. 3rd ed. Springer.

Lockhart, R., Taylor, J., Tibshirani, R. and Tibshirani, R. (2014). A significance test for the lasso (with discussion). *Annals of Statistics*.

Negahban, S. N., Ravikumar, P., Wainwright, M. J. and Yu, B. (2012). A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical Science*, **27** 538–557.

Storey, J. D. (2003). The positive false discovery rate: A bayesian interpretation and the q-value. *Annals of statistics* 2013–2035.

Taylor, J., Lockhart, R., Tibshirani, R. J. and Tibshirani, R. (2014). Post-selection adaptive inference for least angle regression and the lasso. *arXiv preprint arXiv:1401.3889*.

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

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

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv preprint arXiv:1303.0518*.

Zhang, C.-H. and Zhang, S. (2014). Confidence intervals for low-dimensional parameters with high-dimensional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, **76** 217–242.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Methodological)*, **67** 301–320.

Institute for Computational and Mathematical Engineering
Stanford University
Stanford, California
E-mail: jdl17@stanford.edu,
yuekai@stanford.edu

Department of Statistics
Stanford University
Stanford, California
E-mail: dlsun@stanford.edu,
jonathan.taylor@stanford.edu