*The Annals of Statistics*
2016, Vol. 44, No. 3, 907–927
DOI: 10.1214/15-AOS1371


# EXACT POST-SELECTION INFERENCE, WITH APPLICATION TO THE LASSO

BY JASON D. LEE[*,1], DENNIS L. SUN[+,2], YUEKAI SUN[*,3] AND JONATHAN E. TAYLOR[‡,4]

*University of California, Berkeley*[*], *California Polytechnic State University*[+] *and Stanford University*[‡]

We develop a general approach to valid inference after model selection. 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 inferences to be made about the effects. However, in many applications, 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 problematic when there are more variables than observations, since then the model is unidentifiable (at least in the setting where the predictors are assumed fixed).

In such settings, it is tempting to let the data decide which variables to include in the model. For example, one common approach when the number of variables is not too large is to fit a linear model with all variables included, observe which ones are significant at level $\alpha$, and then refit the linear model with only those variables included. The problem with this 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 a particular realization of the data.

To formalize the problem, consider the standard linear regression setup, where the response $\mathbf{y} \in \mathbb{R}^n$ is generated from a multivariate normal distribution:

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

Received January 2015; revised September 2015.

[1]Supported by a National Defense Science and Engineering Graduate Fellowship and a Stanford Graduate Fellowship.

[2]Supported by a Ric Weiland Graduate Fellowship and the Stanford Genome Training Program (SGTP; NIH/NHGRI).

[3]Supported in part by the NIH Grant U01GM102098.

[4]Supported by the NSF Grant DMS-12-08857, and by AFOSR Grant 113039.

*MSC2010 subject classifications.* Primary 62F03, 62J07; secondary 62E15.

*Key words and phrases.* Lasso, confidence interval, hypothesis test, model selection.

where $\boldsymbol{\mu}$ is modeled as a linear function of predictors $\mathbf{x}_1, \ldots, \mathbf{x}_p \in \mathbb{R}^n$, and $\sigma^2$ is assumed known. (We consider the more realistic case where $\sigma^2$ is unknown in Section 8.1.) We choose a subset $M \subset \{1, \ldots, p\}$ and ask for the linear combination of the predictors in $M$ that minimizes the expected error, that is,

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

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 $\beta_j^M$ 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 model*. In general, the coefficient of a predictor cannot be compared across different models.

Thus, "inference after selection" is ambiguous in linear regression because the target of inference changes 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 $\boldsymbol{\beta}^M$ changes with the model is deeply troubling, since it seems to imply that the parameters are random. However, the randomness is actually in the *choice* of which parameters to consider, not in the parameters themselves. Imagine that there are a priori $p2^{p-1}$ well-defined population parameters, one for each coefficient in all $2^p$ possible models:

$$\{\beta_j^M : M \subset \{1, \ldots, p\}, j \in M\}.$$

We only ever form inferences for the parameters $\beta_j^{\hat{M}}$ in the model $\hat{M}$ we select. This adaptive choice of which parameters to consider can lead to inferences with undesirable frequency properties, as noted by Benjamini and Yekutieli (2005) and Benjamini, Heller and Yekutieli (2009).

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

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

but the event inside the probability is not well-defined because $\beta_j^M$ is undefined when $j \notin M$. Two ways around this issue are suggested by Berk et al. (2013):

1. *Conditional coverage*: Since we form an interval for $\beta_j^M$ if and only if model $M$ is selected, that is, $\hat{M} = M$, it makes sense to condition on this event. Hence,

we might require that our confidence interval $C_j^M$ satisfy

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

The benefit of this approach is that we avoid ever having to compare coefficients across two different models $M \neq M'$.

Another way to understand conditioning on the model is to consider *data splitting* [Cox (1975)], an approach to post-selection inference that most statisticians would agree is valid. In data splitting, the data is divided into two halves, with one half used to select the model and the other used to conduct inference. Fithian, Sun and Taylor (2014) argues that inferences obtained by data splitting are only valid conditional on the model that was selected on the first half of the data. Therefore, conditional coverage is a reasonable frequency property to require of a post-selection confidence interval.

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

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

but this is very stringent when many predictors are involved.

Instead of controlling the probability of making *any* error, we can control the expected proportion of errors—although "proportion of errors" is ambiguous in the event that we select zero variables. Benjamini and Yekutieli (2005) simply declare the error to be zero when $|\hat{M}| = 0$:

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

while Storey (2003) suggests conditioning on $|\hat{M}| > 0$:

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

The two criteria are closely related. Since $\text{FCR} = \text{pFCR} \cdot \mathbb{P}(|\hat{M}| > 0)$, pFCR control implies FCR control.

The two ways above are related: conditional coverage (2.1) implies pFCR (2.4) (and hence, FCR) control.

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

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

*Then* $\text{FCR} \leq \text{pFCR} \leq \alpha$.

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

$$\begin{aligned}
\text{pFCR} &= \mathbb{E}\biggl[\mathbb{E}\biggl[\frac{|\{j \in \hat{M} : \beta_j^{\hat{M}} \notin C_j^{\hat{M}}\}|}{|\hat{M}|}\Big|\hat{M}\biggr]\Big||\hat{M}| > 0\biggr] \\
&= \mathbb{E}\biggl[\frac{\sum_{j\in\hat{M}} \mathbb{P}(\beta_j^{\hat{M}} \notin C_j^{\hat{M}}|\hat{M})}{|\hat{M}|}\Big||\hat{M}| > 0\biggr] \\
&\le \mathbb{E}\biggl[\frac{\alpha|\hat{M}|}{|\hat{M}|}\Big||\hat{M}| > 0\biggr] \\
&= \alpha.
\end{aligned}$$

□

Theorem 2 in Weinstein, Fithian and Benjamini (2013) proves a special case of Lemma 2.1 for a particular selection procedure, and Proposition 11 in Fithian, Sun and Taylor (2014) provides a more general result, but this result is sufficient for our purposes: to establish that conditional coverage is a sensible criterion to consider in post-selection inference.

Although the criterion is easy to state, how do we construct an interval with conditional coverage? This requires that we understand the conditional distribution

$$\mathbf{y}|\{\hat{M}(\mathbf{y}) = M\}, \qquad \mathbf{y} \sim N(\mu, \sigma^2 I).$$

One of the main contributions of this paper is to show that this distribution is indeed possible to characterize, making valid post-selection inference feasible in the context of linear regression.

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

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

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

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

which will allow, more generally, conditional inference for parameters of the form $\boldsymbol{\eta}_M^T\boldsymbol{\mu}$. In particular, the regression coefficients $\beta_j^M = \mathbf{e}_j^T X_M^+\boldsymbol{\mu}$ 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 polyhedra. More precisely, the event

$\{\hat{M} = M, \hat{\mathbf{s}}_M = \mathbf{s}_M\}$, that specifies the model *and* the signs of the selected variables, is a polyhedron of the form

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

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

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

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 $F^{\mathbf{z}}(\boldsymbol{\eta}^T \mathbf{y})$ whose distribution given $\{A\mathbf{y} \leq \mathbf{b}\}$ is 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. While their significance tests are specifically intended for the path context, our framework allows more general questions about the model the lasso selects: we can test the model at any value of $\lambda$ 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 nonzero 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 $\boldsymbol{\beta}^0$. Pötscher and Schneider (2010) consider confidence intervals for the hard-thresholding and soft-thresholding estimators in the case of orthogonal design. Our approach instead regards the selected model as a linear approximation to the truth, a view shared by Berk et al. (2013) and Miller (2002).

The idea of post-selection inference conditional on the selected model appears in Pötscher (1991), although the notion of inference conditional on certain *relevant subsets* dates back to Fisher (1956); see also Robinson (1979). Leeb and Pötscher (2005, 2006) obtained a number of negative results about estimating the distribution of a post-selection estimator, although they note their results do not necessarily preclude the possibility of post-selection inference. Benjamini and Yekutieli (2005) also consider conditioning on the selection event, although they argue that this is too conservative. To the contrary, we show that conditioning on the selected model can produce reasonable confidence intervals in a wide variety of situations.

Inference conditional on selection has also appeared in literature on the *winner's curse*: Sampson and Sill (2005), Sill and Sampson (2009), Zhong and Prentice

(2008), Zollner and Pritchard (2007). These works are not really associated with model selection in linear regression, though they employ a similar approach to inference.

**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 $\ell_1$ penalty on the coefficients:

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

The $\ell_1$ 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 $\lambda \geq 0$. However, the distribution of the lasso estimator $\hat{\beta}$ is known only in the less interesting $n \gg p$ case Knight and Fu (2000), and even then, only asymptotically. Inference based on the lasso estimator is still an open question.

Because the lasso produces sparse solutions, we can define model "selected" by the lasso to be simply the set of predictors with nonzero 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 $\hat{\mathbf{s}}$ to be solutions to the lasso problem (4.1), it is necessary and sufficient that they satisfy the Karush–Kuhn–Tucker (KKT) conditions:

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

$$\begin{aligned} \hat{s}_i &= \operatorname{sign}(\hat{\beta}_j) \qquad \text{if } \hat{\beta}_j \neq 0, \\ \hat{s}_i &\in [-1, 1] \qquad \text{if } \hat{\beta}_j = 0. \end{aligned} \tag{4.3}$$

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 identified the model $\hat{M}$ with the equicorrelation set. Since $|\hat{s}_i| = 1$ for any $\hat{\beta}_i \neq 0$, the equicorrelation set does in fact contain all predictors with nonzero coefficients, although it may also include some predictors with zero coefficients. However, for almost every $\lambda$, the equicorrelation set is precisely the set of predictors with nonzero coefficients.

It turns out that it is easier to first characterize $\{(\hat{M}, \hat{\mathbf{s}}) = (M, \mathbf{s})\}$ 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 $\mathbf{s} \in \{-1, 1\}^{|M|}$ be a candidate set of variables and their signs, respectively. Define the random variables*

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

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

*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* $\mathbf{w}$ *and* $\mathbf{u}$ *as*

$$\big\{(\hat{M}, \hat{\mathbf{s}}) = (M, \mathbf{s})\big\} = \big\{\operatorname{sign}\big(\mathbf{w}(M, \mathbf{s})\big) = \mathbf{s}, \big\|\mathbf{u}(M, \mathbf{s})\big\|_\infty < 1\big\}. \tag{4.7}$$

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 $\hat{M}$":

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

□

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

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

$$\{\hat{M} = M, \hat{\mathbf{s}} = \mathbf{s}\} = \{A(M, \mathbf{s}) \mathbf{y} \le \mathbf{b}(M, \mathbf{s})\}.$$

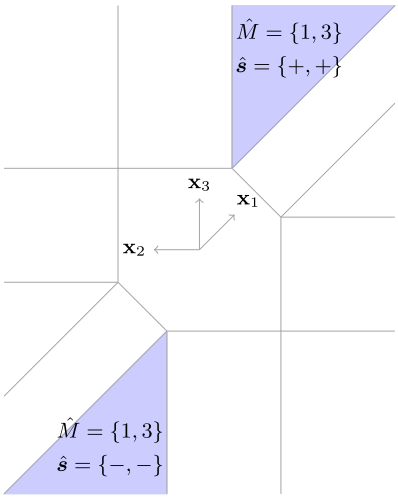


FIG. 1. *A geometric picture illustrating Theorem 4.3 for $n = 2$ and $p = 3$. The lasso partitions $\mathbb{R}^n$ into polyhedra according to the selected model and signs.*

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

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

Figure 1 illustrates Theorem 4.3 and Corollary 4.4. The lasso partitions of $\mathbb{R}^n$ into polyhedra 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 polyhedra. Notice that the sign patterns $\{+,-\}$ and $\{-,+\}$ are not possible for the model $\{1,3\}$.

**5. Polyhedral conditioning sets.** In order to obtain inference conditional on the model, we need to understand the distribution of

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

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

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

since the conditioning event is a single polyhedron $\{A(M,\mathbf{s})\mathbf{y} \le \mathbf{b}(M,\mathbf{s})\}$. 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 $\beta_j^M$ has $(1-\alpha)$ coverage conditional on the model and signs

$$\mathbb{P}\big(\beta_j^M \in C_j^M | \hat{M} = M, \hat{\mathbf{s}} = \mathbf{s}\big) \ge 1 - \alpha,$$

it will also have $(1-\alpha)$ coverage conditional only on the model by the Law of Total Probability:

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

This section is divided into two subsections. First, we study how to condition on a single polyhedron; this will allow us to condition on $\{\hat{M} = M, \hat{\mathbf{s}} = \mathbf{s}\}$. Then we extend the framework to condition on a union of polyhedra, 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 polyhedron.* Suppose we observe $\mathbf{y} \sim N(\boldsymbol{\mu}, \Sigma)$, and $\boldsymbol{\eta} \in \mathbb{R}^n$ is some direction of interest. To understand the distribution of

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

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

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

where

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

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

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

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

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

*where*

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

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

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

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

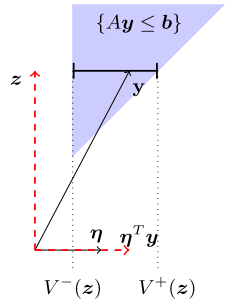


FIG. 2. *A geometric interpretation of why the event* $\{A\mathbf{y} \le \mathbf{b}\}$ *can be characterized as* $\{\mathcal{V}^-(\mathbf{z}) \le \boldsymbol{\eta}^T\mathbf{y} \le \mathcal{V}^+(\mathbf{z})\}$. *Assuming* $\Sigma = I$ *and* $\|\boldsymbol{\eta}\|_2 = 1$, $\mathcal{V}^-(\mathbf{z})$ *and*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

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

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

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

Lemma 5.1 tells us that

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

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

$$\text{"}\boldsymbol{\eta}^T\mathbf{y} | \{A\mathbf{y} \le \mathbf{b}\} \sim \mathrm{TN}(\boldsymbol{\eta}^T\boldsymbol{\mu}, \sigma^2\boldsymbol{\eta}^T\Sigma\boldsymbol{\eta}, \mathcal{V}^-(\mathbf{z}), \mathcal{V}^+(\mathbf{z}))\text{,"}$$

but this is technically incorrect, since the distribution on the right-hand side changes with $\mathbf{z}$. By conditioning on the value of $\mathbf{z}$, $\boldsymbol{\eta}^T\mathbf{y} | \{A\mathbf{y} \le \mathbf{b}, \mathbf{z} = \mathbf{z}_0\}$ is a truncated normal. We can then use the probability integral transform to obtain

a statistic $F^{\mathbf{z}}(\boldsymbol{\eta}^T\mathbf{y})$ that has a Unif(0, 1) distribution for any value of $\mathbf{z}$. Hence, $F^{\mathbf{z}}(\boldsymbol{\eta}^T\mathbf{y})$ will also have a Unif(0, 1) distribution marginally over $\mathbf{z}$. 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 $[a, b]$, that is,*

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

*where $\Phi$ is the CDF of a $N(0, 1)$ random variable. Then*

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

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

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

PROOF. First, apply Lemma 5.1:

$$\begin{aligned}[\boldsymbol{\eta}^T\mathbf{y}|A\mathbf{y}\le\mathbf{b}, \mathbf{z}=\mathbf{z}_0] &\overset{d}{=} [\boldsymbol{\eta}^T\mathbf{y}|\mathcal{V}^-(\mathbf{z}) \le \boldsymbol{\eta}^T\mathbf{y} \le \mathcal{V}^+(\mathbf{z}), \mathcal{V}^0(\mathbf{z}) \ge 0, \mathbf{z}=\mathbf{z}_0]\\ &\overset{d}{=} [\boldsymbol{\eta}^T\mathbf{y}|\mathcal{V}^-(\mathbf{z}_0) \le \boldsymbol{\eta}^T\mathbf{y} \le \mathcal{V}^+(\mathbf{z}_0), \mathcal{V}^0(\mathbf{z}_0) \ge 0, \mathbf{z}=\mathbf{z}_0].\end{aligned}$$

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

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

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

$$\begin{aligned}[F^{\mathbf{z}}(\boldsymbol{\eta}^T\mathbf{y})|A\mathbf{y}\le\mathbf{b}, \mathbf{z}=\mathbf{z}_0] &\overset{d}{=} [F^{\mathbf{z}_0}(\boldsymbol{\eta}^T\mathbf{y})|A\mathbf{y}\le\mathbf{b}, \mathbf{z}=\mathbf{z}_0]\\ &\sim \mathrm{Unif}(0,1).\end{aligned}$$

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

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

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

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

□

5.2. *Conditioning on a union of polyhedra.* We have just characterized the distribution of $\boldsymbol{\eta}^T\mathbf{y}$, conditional on $\mathbf{y}$ falling into a single polyhedron $\{A\mathbf{y} \le \mathbf{b}\}$. We obtain such a polyhedron if we condition on both the model and the signs $\{\hat{M} = M, \hat{\mathbf{s}} = \mathbf{s}\}$. If we want to only condition on the model $\{\hat{M} = M\}$, then we will have to understand the distribution of $\boldsymbol{\eta}^T\mathbf{y}$, conditional on $\mathbf{y}$ falling into a union of such polyhedra, that is,

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

As Figure 3 makes clear, the argument proceeds exactly as before, except that $\boldsymbol{\eta}^T\mathbf{y}$ 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 $\mathbf{s}$, 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 $F^S_{\mu,\sigma^2}$ denote the CDF of a $N(\mu, \sigma^2)$ random variable truncated to the set $S$. Then*

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

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

**6. Post-selection intervals for regression coefficients.** In this section, we combine the characterization of the lasso selection event in Section 4 with the

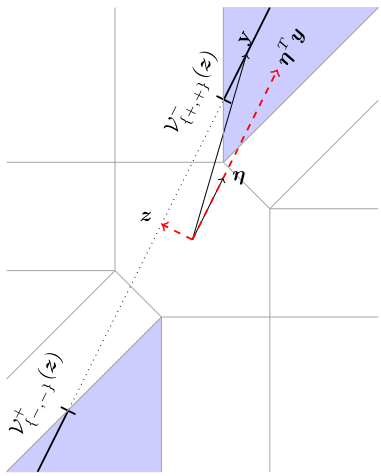


FIG. 3. *When we take the union over signs, the conditional distribution of $\boldsymbol{\eta}^T\mathbf{y}$ is truncated to a union of disjoint intervals. In this case, the Gaussian is truncated to the set $(-\infty, \mathcal{V}^+_{\{-,-\}}(\mathbf{z})] \cup [\mathcal{V}^-_{\{+,+\}}(\mathbf{z}), \infty)$.*

results about the distribution of a Gaussian truncated to a polyhedron (or union of polyhedra) 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 polyhedra:

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

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

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

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

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

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

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

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

$$F_{L,\sigma^2\|\boldsymbol{\eta}\|^2}^{\bigcup_{\mathbf{s}}[\mathcal{V}_{\mathbf{s}}^-(\mathbf{z}),\mathcal{V}_{\mathbf{s}}^+(\mathbf{z})]}(\boldsymbol{\eta}^T\mathbf{y}) = 1-\frac{\alpha}{2}, \qquad F_{U,\sigma^2\|\boldsymbol{\eta}\|^2}^{\bigcup_{\mathbf{s}}[\mathcal{V}_{\mathbf{s}}^-(\mathbf{z}),\mathcal{V}_{\mathbf{s}}^+(\mathbf{z})]}(\boldsymbol{\eta}^T\mathbf{y}) = \frac{\alpha}{2}.$$

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

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

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 $F_{L,\sigma^2\|\boldsymbol{\eta}\|^2}^{\bigcup_{\mathbf{s}}[\mathcal{V}_{\mathbf{s}}^-(\mathbf{z}),\mathcal{V}_{\mathbf{s}}^+(\mathbf{z})]}(\boldsymbol{\eta}^T\mathbf{y})$ is monotone decreasing in $\beta_j^M$ 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 the Appendix for details. □

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

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

Inverting this statistic will produce intervals that have $(1-\alpha)$ coverage conditional on $\{\hat{M}=M,\hat{\mathbf{s}}=\mathbf{s}\}$, 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 $[\mathcal{V}_{\mathbf{s}}^-(\mathbf{z}),\mathcal{V}_{\mathbf{s}}^+(\mathbf{z})]$ for each $\mathbf{s}\in\{-1,1\}^{|M|}$. The number of such sign patterns is $2^{|M|}$. While this might be feasible when $|M|$ is small, it is not feasible when we select hundreds of variables. Conditioning on the signs means that we only have to compute the interval $[\mathcal{V}_{\mathbf{s}}^-(\mathbf{z}),\mathcal{V}_{\mathbf{s}}^+(\mathbf{z})]$ for the sign pattern $\mathbf{s}$ 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 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 $\mu$ that would be consistent with that observation—hence, the wide intervals. Figure 5 shows confidence intervals for $\mu$ as a function of $Z$. When $Z$ is far from the endpoints of the truncation

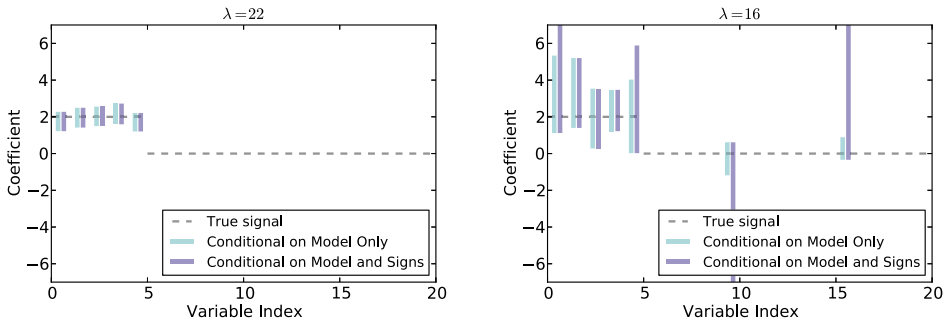


FIG. 4. *Comparison of the confidence intervals by conditioning on the model only* (*statistically more efficient*, *but computationally more expensive*) *and conditioning on both the model and signs* (*statistically less efficient*, *but computationally more feasible*). *Data were simulated for* $n=25$, $p=50$, *and* 5 *true nonzero 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*).

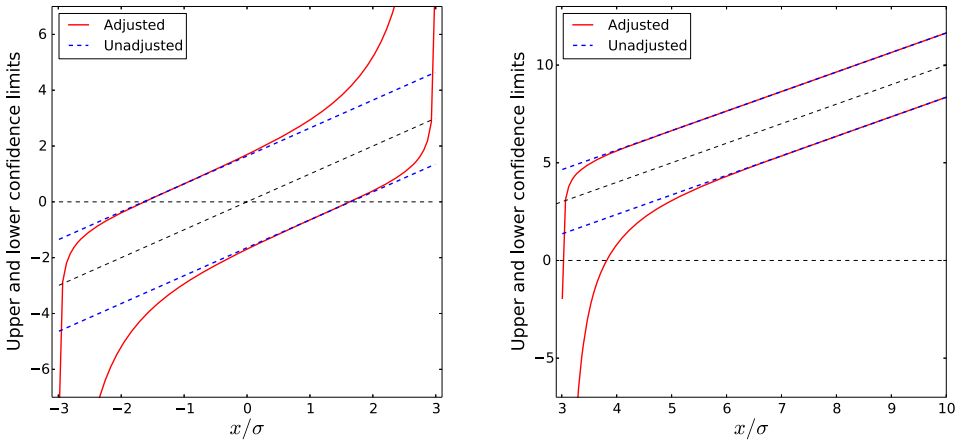


FIG. 5. *Upper and lower bounds of* 90% *confidence intervals for* $\mu$ *based on a single observation* $x/\sigma \sim \mathrm{TN}(0, 1, -3, 3)$. *We see that as long as the observation* $x$ *is roughly* $0.5\sigma$ *away from either boundary, the size of the intervals is comparable to the unadjusted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

interval, we basically recover the nominal OLS intervals (i.e., not adjusted for selection).

The implications are clear. When the signal is strong, $\boldsymbol{\eta}^T \mathbf{y}$ 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 $\boldsymbol{\eta}^T \mathbf{y}$ 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 at least $1 - \alpha$. The fact that we have found such an interval is not remarkable, since many such intervals have this property. However, given two intervals with the same coverage, we generally prefer the shorter one. This problem is considered in Fithian, Sun and Taylor (2014) where it is shown that $C_j^M$ is, with one small tweak, the shortest interval among all *unbiased* intervals with $1 - \alpha$ coverage.

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

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

Unbiasedness is a common restriction to ensure the existence of an optimal interval [Lehmann and Romano (2005)]. The shortest unbiased interval for $\beta_j^M$, among all intervals with conditional $1 - \alpha$ coverage, resembles to the interval $[L, U]$ in Theorem 6.1. There, the critical values $L$ and $U$ were chosen symmetrically so that the pivot has $\alpha/2$ area in either tail. However, it may be possible to obtain a shorter interval on average by allocating the a probability unequally between the two tails. Theorem 5 of Fithian, Sun and Taylor (2014) provides a general formula for obtaining shortest unbiased intervals in exponential families.

**7. Data example.** We apply our post-selection intervals to the diabetes data set from Efron et al. (2004). Since $p < n$ for this data set, we can estimate $\sigma^2$ using the residual sum of squares from the full regression model with all $p$ predictors. After standardizing all variables, we chose $\lambda$ according to the strategy in Negahban et al. (2012), $\lambda = 2\mathbf{E}(\|X^T\varepsilon\|_\infty)$. This expectation was computed by simulation, where $\varepsilon \sim N(0, \hat{\sigma}^2)$, 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, as discussed in Section 2. 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.

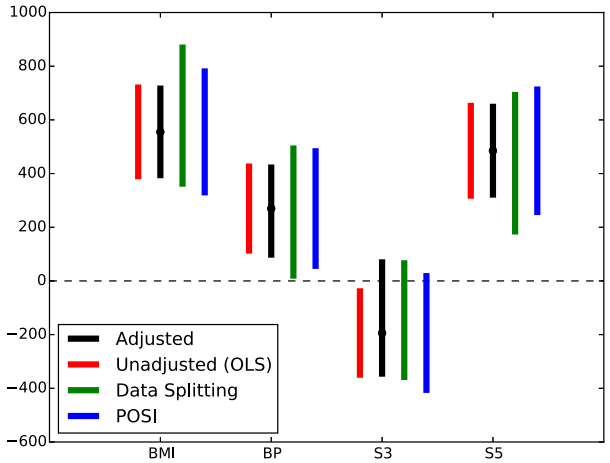


FIG. 6. *Inference for the four variables selected by the lasso* ($\lambda = 190$) *on the diabetes data set. The point estimate and adjusted confidence intervals using the approach in Section 6 are shown in black. The OLS intervals, which ignore selection, are shown in red. The green lines show the 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).

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. Extensions.

8.1. *Estimation of* $\sigma^2$. The above results rely on knowing $\sigma^2$ or at least having a good estimate of it. If $n > p$, then the variance $\hat{\sigma}^2$ of the residuals from fitting the full model is a consistent estimator and in general can be substituted for $\sigma^2$ to yield asymptotically valid confidence intervals. Formally, the condition is that the pivot is smooth with respect to $\sigma$. Geometrically speaking, the upper and lower truncation limits $\mathcal{V}^+$ and $\mathcal{V}^-$ must be well-separated (with high probability). We refer the interested reader to Section 2.3 in Tian and Taylor (2015) for details.

In the setting where $p > n$, obtaining an estimate of $\sigma^2$ is more challenging, but if the pivot satisfies a monotonicity property, plugging in an overestimate of the variance gives conservative confidence intervals. We refer the reader to Theorem 11 in Tibshirani et al. (2015) for details.

8.2. *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 $\ell_2$ penalty to the lasso objective, resulting in the elastic net [Zou and Hastie (2005)]:

$$\tilde{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - X\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 + \frac{\gamma}{2}\|\boldsymbol{\beta}\|_2^2. \tag{8.1}$$

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

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

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

$$\{\hat{M} = M, \hat{\mathbf{s}} = \mathbf{s}\} = \left\{ \begin{pmatrix} \tilde{A}_0(M,\mathbf{s}) \\ \tilde{A}_1(M,\mathbf{s}) \end{pmatrix} \mathbf{y} < \begin{pmatrix} \tilde{\mathbf{b}}_0(M,\mathbf{s}) \\ \tilde{\mathbf{b}}_1(M,\mathbf{s}) \end{pmatrix} \right\}, \tag{8.2}$$

where $\tilde{A}_k$ and $\tilde{\mathbf{b}}_k$ are analogous to $A_k$ and $\mathbf{b}_k$ in Proposition 4.2, except replacing $(X_M^T X_M)^{-1}$ by $(X_M^T X_M + \gamma I)^{-1}$ everywhere it appears. Notice that $(X_M^T X_M)^{-1}$ appears explicitly in $A_1$ and $\mathbf{b}_1$, and also implicitly in $A_0$ and $\mathbf{b}_0$, since $P_M$ and $(X_M^T)^+$ both depend on $(X_M^T X_M)^{-1}$.

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

**9. 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*, that is, 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.

## APPENDIX: 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 $\mu$.*

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

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

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 $(x, \infty)$ to obtain

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

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

**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.

## REFERENCES

BENJAMINI, Y., HELLER, R. and YEKUTIELI, D. (2009). Selective inference in complex research. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **367** 4255–4271. MR2546387

BENJAMINI, Y. and YEKUTIELI, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.* **100** 71–93. MR2156820

BERK, R., BROWN, L., BUJA, A., ZHANG, K. and ZHAO, L. (2013). Valid post-selection inference. *Ann. Statist.* **41** 802–837. MR3099122

COX, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika* **62** 441–444. MR0378189

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

FISHER, R. (1956). On a test of significance in Pearson's Biometrika Tables (No. 11). *J. Roy. Statist. Soc. Ser. B.* **18** 56–60. MR0082773

FITHIAN, W., SUN, D. and TAYLOR, J. (2014). Optimal inference after model selection. Preprint. Available at arXiv:1410.2597.

JAVANMARD, A. and MONTANARI, A. (2013). Confidence intervals and hypothesis testing for high-dimensional regression. Preprint. Available at arXiv:1306.3171.

KNIGHT, K. and FU, W. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28** 1356–1378. MR1805787

LEEB, H. and PÖTSCHER, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* **21** 21–59. MR2153856

LEEB, H. and PÖTSCHER, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.* **34** 2554–2591. MR2291510

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3rd ed. Springer, New York. MR2135927

LOCKHART, R., TAYLOR, J., TIBSHIRANI, R. and TIBSHIRANI, R. (2014). A significance test for the lasso (with discussion). *Ann. Statist.* **42** 413–468. MR3210970

MILLER, A. (2002). *Subset Selection in Regression*, 2nd ed. Chapman & Hall/CRC, Boca Raton, FL. MR2001193

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. *Statist. Sci.* **27** 538–557. MR3025133

PÖTSCHER, B. M. (1991). Effects of model selection on inference. *Econometric Theory* **7** 163–185. MR1128410

PÖTSCHER, B. M. and SCHNEIDER, U. (2010). Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electron. J. Stat.* **4** 334–360. MR2645488

ROBINSON, G. K. (1979). Conditional properties of statistical procedures. *Ann. Statist.* **7** 742–755. MR0532239

SAMPSON, A. R. and SILL, M. W. (2005). Drop-the-losers design: Normal case. *Biom. J.* **47** 257–268. MR2145117

SILL, M. W. and SAMPSON, A. R. (2009). Drop-the-losers design: Binomial case. *Comput. Statist. Data Anal.* **53** 586–595. MR2654594

STOREY, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the $q$-value. *Ann. Statist.* **31** 2013–2035. MR2036398

TAYLOR, J., LOCKHART, R., TIBSHIRANI, R. J. and TIBSHIRANI, R. (2014). Post-selection adaptive inference for least angle regression and the lasso. Preprint. Available at arXiv:1401.3889.

TIAN, X. and TAYLOR, J. (2015). Asymptotics of selective inference. Preprint. Available at arXiv:1501.03588.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B* **58** 267–288. MR1379242

TIBSHIRANI, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat.* **7** 1456–1490. MR3066375

TIBSHIRANI, R. J., RINALDO, A., TIBSHIRANI, R. and WASSERMAN, L. (2015). Uniform asymptotic inference and the bootstrap after model selection. Preprint. Available at arXiv:1506.06266.

VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. and DEZEURE, R. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. Preprint. Available at arXiv:1303.0518.

WEINSTEIN, A., FITHIAN, W. and BENJAMINI, Y. (2013). Selection adjusted confidence intervals with more power to determine the sign. *J. Amer. Statist. Assoc.* **108** 165–176. MR3174610

ZHANG, C.-H. and ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **76** 217–242. MR3153940

ZHONG, H. and PRENTICE, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics* **9** 621–634.

ZOLLNER, S. and PRITCHARD, J. K. (2007). Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.* **80** 605–615.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 301–320. MR2137327

J. D. LEE
UNIVERSITY OF CALIFORNIA, BERKELEY
465 SODA HALL
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: jasondlee@berkeley.edu

D. L. SUN
CALIFORNIA POLYTECHNIC STATE UNIVERSITY
BUILDING 25, ROOM 107D
SAN LUIS OBISPO, CALIFORNIA
USA
E-MAIL: dennisliusun@gmail.com

Y. SUN
UNIVERSITY OF CALIFORNIA, BERKELEY
367 EVANS HALL
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: yuekai@berkeley.edu

J. E. TAYLOR
STANFORD UNIVERSITY
390 SERRA MALL
STANFORD, CALIFORNIA
USA
E-MAIL: jonathan.taylor@stanford.edu