EXACT POST-SELECTION INFERENCE WITH THE LASSO

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

Stanford University

We develop a framework for post-selection inference with the lasso. At the core of our framework is a result that characterizes the exact (non-asymptotic) distribution of linear combinations/contrasts of truncated normal random variables. This result allows us to (i) obtain honest confidence intervals for the selected coefficients that account for the selection procedure, and (ii) devise a test statistic that has an exact (non-asymptotic) Unif$(0,1)$ distribution when 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, paving the way for intervals and tests of the effect size. 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. The problem is especially acute if there are more variables than observations, when it is impossible to even fit linear regression.

A practitioner might wish to use the data to select the relevant variables and then make inference on the selected variables. As an example, one might fit a linear model, observe which coefficients are significant at level α, and report $(1-\alpha)$-confidence intervals for only the significant coefficients. However, these intervals fail to take into account the randomness in the selection procedure. In particular, the intervals do not have the stated coverage once one marginalizes over the selected model.

To see this formally, assume the usual linear model

$$y = \mu + \epsilon,\ \mu = X\beta^0,\ \epsilon \sim N(0, \sigma^2 I), \tag{1.1}$$

where $X \in \mathbb{R}^{n\times p}$ is the design matrix and $\beta^0 \in \mathbb{R}^p$. Let $\hat{E} \subset \{1, ..., p\}$ denote a (random) set of selected variables. Suppose the goal is inference about β^0_j. Then, we do not even form intervals for β^0_j when $j \notin \hat{E}$, so the first issue is

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

to define an interval when $j \notin \hat{E}$ in order to evaluate the coverage of this procedure. There is no obvious way to do this so that the marginal coverage is $1-\alpha$. Furthermore, as $\hat{E}$ varies, the target of the ordinary least-squares (OLS) estimator $\hat{\beta}^{OLS}_{\hat{E}}$ is not β^0, but rather

$$\beta^\star_{\hat{E}} := X^+_{\hat{E}}\mu,$$

where $X^+_{\hat{E}}$ denotes the Moore-Penrose pseudoinverse of $X_{\hat{E}}$. We see that $X_{\hat{E}}\beta^\star_{\hat{E}} = P_{\hat{E}}\mu$, the projection of μ onto the columns of $X_{\hat{E}}$, so $\beta^\star_{\hat{E}}$ represents the coefficients in the best linear model using only the variables in $\hat{E}$. In general, $\beta^\star_{\hat{E},j} \neq \beta^0_j$ unless $\hat{E}$ contains the support set of β^0, i.e., $\hat{E} \supset S := \{j : \beta^0_j \neq 0\}$. Since $\hat{\beta}^{OLS}_{\hat{E},j}$ may not be estimating β^0_j at all, there is no reason to expect a confidence interval based on it to cover β^0_j. Berk et al. (2013) provide an explicit example of the non-normality of $\hat{\beta}^{OLS}_{\hat{E},j}$ in the post-selection context. In short, inference in the linear model has traditionally been incompatible with model selection.

1.1. *Inference Conditional on the Model.* To resolve these issues, we propose inference conditional on the selected model $\hat{E}$. Conditioning on $\hat{E}$ avoids the thorny issue of how to compare coefficients across two different models $\hat{E}_1 \neq \hat{E}_2$. Our framework allows post-selection inference about $\eta^T\mu$, where $\eta = \eta_{\hat{E}}$ is allowed to depend on the model, since many interesting post-selection hypotheses are formulated after observing which variables have been selected. In particular, by choosing $\eta = X^+_{\hat{E}}e_j$, one obtains inference about $\eta^T\mu = \beta^\star_{\hat{E},j}$, although our framework applies even when the linear model is misspecified. Furthermore, our procedure is *exact*, meaning that the distributional results are non-asymptotic and exact in finite samples.

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. The problem was most recently considered by Berk et al. (2013), who cast post-selection inference in terms of simultaneous inference over all possible submodels.

In some sense, conditioning on the model is natural in post-selection inference. If one considers the evidently valid method of splitting the data into two halves, selecting the model using only the former, and making inferences

based on the latter, the inferences are still conditional on the (random) model selected on the first half (Wasserman, 2014). Even though in this case, the inference is independent of the selected model, the inferential questions that are raised are still contingent upon the random model that was selected in the first stage.

1.2. *The Lasso.* In this paper, we focus on a particular model selection procedure, the lasso (Tibshirani, 1996), which achieves model selection by setting coefficients to zero exactly. This is accomplished by adding an ℓ_1 penalty term to the usual least-squares objective:

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1, \tag{1.2}$$

where $\lambda \geq 0$ is a penalty parameter that controls the tradeoff between fit to the data and sparsity of the coefficients. 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.

We apply our framework for post-selection inference about $\eta_{\hat{E}}^T \mu$ to form confidence intervals for $\beta^\star_{\hat{E},j}$ and to test whether the the fitted model captures all relevant signal variables.

1.3. *Related Work.* Most of the theoretical work on fitting high-dimensional linear models focuses on *consistency*. The flavor of these results is that under certain assumptions on X, the lasso fit $\hat{\beta}$ is close to the unknown β^0 (Negahban et al., 2012) and selects the correct model (Wainwright, 2009; Zhao and Yu, 2006). A comprehensive survey of the literature can be found in Bühlmann and van de Geer (2011).

There is also some recent work on obtaining confidence intervals and significance testing for penalized M-estimators such as the lasso. One class of methods uses sample splitting or subsampling to obtain confidence intervals and p-values. Recently, Meinshausen and Bühlmann (2010) proposed *stability selection* as a general technique designed to improve the performance of a variable selection algorithm. The basic idea is, instead of performing variable selection on the whole data set, to perform variable selection on random subsamples of the data of size $\binom{n}{2}$ and include the variables that are selected most often on the subsamples.

A separate line of work establishes the asymptotic normality of a corrected estimator obtained by "inverting" the KKT conditions (Javanmard and Montanari, 2013; van de Geer et al., 2013; Zhang and Zhang, 2014).

The corrected estimator $\hat{b}$ usually has the form

$$\hat{b} = \hat{\beta} + \lambda \Theta \hat{z},$$

where $\hat{z}$ is a subgradient of the penalty at $\hat{\beta}$ and Θ is an approximate inverse to the Gram matrix $X^T X$. This approach is very general and easily handles M-estimators that minimize the sum of a smooth convex loss and a convex penalty. The two main drawbacks to this approach are:

1. the confidence intervals are valid only when the M-estimator is consistent
2. obtaining Θ is usually much more expensive than obtaining $\hat{\beta}$.

Most closely related to our work is the pathwise signficance testing framework laid out in Lockhart et al. (2014). They establish a test for whether a newly added coefficient is a relevant variable. This method only allows for testing at λ that are LARS knot values. This is a considerable restriction, since the lasso is often not solved with the LARS algorithm. Furthermore, the test is asymptotic, makes strong assumptions on X, and the weak convergence assumes that all relevant variables are already included in the model. They do not discuss forming confidence intervals for the selected variables. Section 5.2 establishes a nonasymptotic test for the same null hypothesis, while only assuming X is in general position.

In contrast, we provide a test that is exact, allows for arbitrary λ, and arbitrary design matrix X. By extension, we do not make any assumptions on n and p, and do not require the lasso to be a consistent estimator of β^0. Furthermore, the computational expense to conduct our test is negligible compared to the cost of obtaining the lasso solution.

Like all of the preceding works, our test assumes that the noise variance σ^2 is known or can be estimated. In the low-dimensional setting $p \ll n$, σ^2 can be estimated from the residual sum-of-squares of the saturated model. Strategies in high dimensions are discussed in Fan et al. (2012) and Reid et al. (2013). In Section 8, we also provide a strategy for estimating σ^2 based on the framework we develop.

1.4. *Outline of Paper.* We begin by defining several important quantities related to the lasso in Section 2; most notably, we define the selected model $\hat{E}$ in terms of the equicorrelation set of the lasso solution. Section 3 provides an alternative characterization of the selection procedure for the lasso in terms of affine constraints on y, i.e., $Ay \leq b$. Therefore, the distribution of y conditional on the selected model is the distribution of a Gaussian vector conditional on its being in a polytope. In Section 4, we generalize

and show that for $y \sim N(\mu, \Sigma)$, the distribution of $\eta^T y \mid Ay \leq b$ is roughly a truncated Gaussian random variable, and derive a pivot for $\eta^T \mu$. In Section 5, we specialize again to the lasso, deriving confidence intervals for $\beta^\star_{\hat{E},j}$ and hypothesis tests of the selected model as special cases of $\eta^T \mu$. Section 6 presents an example of these methods applied to a dataset.

In Section 7, we consider a refinement that produces narrower confidence intervals. Finally, Section 8 collects a number extensions of the framework. In particular, we demonstrate:

- the relationship between the approaches of (Javanmard and Montanari, 2013; van de Geer et al., 2013; Zhang and Zhang, 2014) and ours, deriving (conditional) intervals for the estimand that they consider. The bias of this estimand tends to zero under certain assumptions.
- modifications needed for the elastic net (Zou and Hastie, 2005).
- different norms as test statistics for the "goodness of fit" test discussed in Section 5.
- estimation of σ^2 based on fitting the lasso with a sufficiently small λ.
- composite null hypotheses.
- fitting the lasso for a sequence of λ values and its effect on our basic tests and intervals.

2. Preliminaries. Necessary and sufficient conditions for $(\hat{\beta}, \hat{z})$ to be solutions to the lasso problem (1.2) are the Karush-Kuhn-Tucker (KKT) conditions:

$$X^T(X\hat{\beta} - y) + \lambda \hat{z} = 0, \tag{2.1}$$

$$\hat{z}_i \in \begin{cases} \text{sign}(\hat{\beta}_i) & \text{if } \hat{\beta}_i \neq 0 \\ [-1, 1] & \text{if } \hat{\beta}_i = 0 \end{cases}. \tag{2.2}$$

where $\hat{z} := \partial || \cdot ||_1(\hat{\beta})$ denotes the subgradient of the ℓ_1 norm at $\hat{\beta}$. We consider the *equicorrelation set* (Tibshirani, 2013)

$$\hat{E} = \{i \in \{1, \ldots, p\} : |\hat{z}_i| = 1\}, \tag{2.3}$$

so-named because by examining only the rows corresponding to $\hat{E}$ in (2.1), we obtain the relation

$$X_{\hat{E}}^T(y - X\hat{\beta}) = -\lambda \hat{z}_{\hat{E}},$$

where $X_{\hat{E}}$ is the submatrix of X consisting of the columns in $\hat{E}$. Hence

$$|X_{\hat{E}}^T(y - X\hat{\beta})| = \lambda,$$

i.e. the variables in this set have equal (absolute) correlation with the residual $y - X\hat{\beta}$. Since $\hat{z}_i \in \{-1, 1\}$ for any $\hat{\beta}_i \neq 0$, all variables with non-zero coefficients are contained in the equicorrelation set.

Recall that we are interested in inference for $\eta^T \mu$ in the model (1.1) for some direction $\eta = \eta_{\hat{E}} \in \mathbb{R}^n$, which is allowed to depend on the selected variables $\hat{E}$. In most applications, we will assume $\mu = X\beta^0$, although our results hold even if the linear model is not correctly specified.

A natural estimate for $\eta^T \mu$ is $\eta^T y$. As mentioned previously, we allow $\eta = \eta_{\hat{E}}$ to depend on the random selection procedure, so our goal is post-selection inference based on

$$\eta^T y \mid \{\hat{E} = E\}.$$

For reasons that will become clear, a more tractable quantity is the distribution conditional on both the selected variables and their signs

$$\eta^T y \mid \{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}.$$

Note that confidence intervals and hypothesis tests that are valid conditional on the finer partition $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$ will also be valid for $\{\hat{E} = E\}$, by summing over the possible signs z_E:

$$\mathbb{P}(\,\cdot \,\big|\, \hat{E} = E) = \sum_{z_E} \mathbb{P}(\,\cdot \,\big|\, (\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E))\; \mathbb{P}(\hat{z}_{\hat{E}} = z_E \,\big|\, \hat{E} = E).$$

From this, it is clear that controlling $\mathbb{P}(\,\cdot \,\big|\, (\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E))$ to be, say, less than α (as in the case of hypothesis testing) will ensure $\mathbb{P}(\,\cdot \,\big|\, \hat{E} = E) \leq \alpha$.

It may not be obvious yet why we condition on $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$ instead of $\{\hat{E} = E\}$. In the next section, we show that the former can be restated in terms of affine constraints on y, i.e., $\{Ay \leq b\}$. We revisit the problem of conditioning only on $\{\hat{E} = E\}$ in Section 7.

3. Characterizing Selection for the Lasso. Recall from the previous section that our goal is inference conditional on $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$. In this section, we show that this "selection" event can be rewritten in terms of affine constraints on y, i.e.,

$$\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\} = \{A(E, z_E)y \leq b(E, z_E)\}$$

for a suitable matrix $A(E, z_E)$ and vector $b(E, z_E)$. Therefore, the conditional distribution $y \mid \{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$ is simply $y \,\big|\, \{A(E, z_E)y \leq b(E, z_E)\}$. This key theorem follows from two intermediate results.

LEMMA 3.1. *Without loss of generality, assume the columns of X are in general position. Let $E \subset \{1,\dots,p\}$ and $z_E \in \{-1,1\}^{|E|}$ be a candidate set of variables and signs, respectively. Define*

$$U = U(E, z_E) := (X_E^T X_E)^{-1}(X_E^T y - \lambda z_E) \tag{3.1}$$

$$W = W(E, z_E) := X_{-E}^T (X_E^T)^+ z_E + \frac{1}{\lambda} X_{-E}^T (I - P_E) y. \tag{3.2}$$

Then the selection procedure can be rewritten in terms of U and W as:

$$\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\} = \{\operatorname{sign}(U(E, z_E)) = z_E, \|W(E, z_E)\|_\infty < 1\} \tag{3.3}$$

PROOF. First, we rewrite the KKT conditions (2.1) and (2.2) by partitioning them according to the equicorrelation set $\hat{E}$:

$$\begin{aligned} X_{\hat{E}}^T (X_{\hat{E}} \hat{\beta}_{\hat{E}} - y) + \lambda \hat{z}_{\hat{E}} &= 0 \\ X_{-\hat{E}}^T (X_{\hat{E}} \hat{\beta}_{\hat{E}} - y) + \lambda \hat{z}_{-\hat{E}} &= 0 \\ \operatorname{sign}(\hat{\beta}_{\hat{E}}) = \hat{z}_{\hat{E}},\ \hat{z}_{-\hat{E}} &\in (-1, 1). \end{aligned}$$

Since the KKT conditions are necessary and sufficient for a solution, we obtain that $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$ if and only if there exist U and W satisfying:

$$X_E^T (X_E U - y) + \lambda z_E = 0 \tag{3.4}$$

$$X_{-E}^T (X_E U - y) + \lambda W = 0 \tag{3.5}$$

$$\operatorname{sign}(U) = z_E,\ W \in (-1, 1). \tag{3.6}$$

Solving (3.4) and (3.5) for U and W yields (3.1) and (3.2). That U and W are subject to (3.6) yields (3.3). □

Lemma 3.1 is remarkable because it says that the selection event $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$ is equivalent to affine constraints on y. To see this, note that both U and W are affine functions of y, so $\{\operatorname{sign}(U) = z_E, \|W\|_\infty < 1\}$ can be written as affine constraints $\{A(E, z_E) y \le b(E, z_E)\}$. The following proposition provides explicit formulas for A and b.

PROPOSITION 3.2. *Let U and W be defined as in (3.1) and (3.2). Then:*

$$\{\operatorname{sign}(U) = z_E, \|W\|_\infty < 1\} = \left\{ \begin{pmatrix} A_0(E, z_E) \\ A_1(E, z_E) \end{pmatrix} y < \begin{pmatrix} b_0(E, z_E) \\ b_1(E, z_E) \end{pmatrix} \right\} \tag{3.7}$$

where A_0, b_0 encode the "inactive" constraints $\{\|W\|_\infty < 1\}$, and A_1, b_1 encode the "active" constraints $\{\text{sign}(U) = z_E\}$. These matrices have the explicit forms:

$$A_0(E, z_E) = \frac{1}{\lambda}\begin{pmatrix} X_{-E}^T(I - P_E) \\ -X_{-E}^T(I - P_E) \end{pmatrix} \qquad b_0(E, z_E) = \begin{pmatrix} \mathbf{1} - X_{-E}^T(X_E^T)^+ z_E \\ \mathbf{1} + X_{-E}^T(X_E^T)^+ z_E \end{pmatrix}$$
$$A_1(E, z_E) = -\,\mathbf{diag}(z_E)(X_E^T X_E)^{-1} X_E^T \quad b_1(E, z_E) = -\lambda\,\mathbf{diag}(z_E)(X_E^T X_E)^{-1} z_E$$

PROOF. First, we write

$$\{\text{sign}(U) = z_E\} = \{\mathbf{diag}(z_E)U > 0\}.$$

From here, it is straightforward to derive the above expressions from the definitions of U and W given in (3.1) and (3.2). □

Combining Lemma 3.1 with Proposition 3.2, we obtain the following.

THEOREM 3.3. *The selection procedure can be rewritten in terms of affine constraints on y:*

$$\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\} = \{A(E, z_E)y \le b(E, z_E)\}.$$

To summarize, we have shown that in order to understand the distribution of $y \sim N(\mu, \Sigma)$ conditional on the selection procedure $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$, it suffices to study the distribution of y conditional on being in the polytope $\{Ay \le b\}$. The next section derives a pivot for $\eta^T\mu$ for such distributions, which will be useful for constructing confidence intervals and hypothesis tests in Section 5.

4. A Pivot for Gaussian Vectors Subject to Affine Constraints. The distribution of a Gaussian vector $y \sim N(\mu, \Sigma)$ conditional on affine constraints $\{Ay \le b\}$, while explicit, still involves the intractable normalizing constant $\mathbb{P}(Ay \le b)$. In this section, we derive a one-dimensional pivotal quantity for μ, in particular for $\eta^T\mu$. This pivot is $F^{[\mathcal{V}^-, \mathcal{V}^+]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y)$, which has a $\text{Unif}(0, 1)$ distribution conditional on $Ay \le b$.

The key step to deriving this pivot is the following lemma:

LEMMA 4.1. *The conditioning set can be rewritten in terms of $\eta^T y$ as follows:*

$$\{Ay \le b\} = \{\mathcal{V}^-(y) \le \eta^T y \le \mathcal{V}^+(y), \mathcal{V}^0(y) \ge 0\}$$

where

$$\alpha = \frac{A\Sigma\eta}{\eta^T\Sigma\eta} \tag{4.1}$$

$$\mathcal{V}^- = \mathcal{V}^-(y) = \max_{j:\ \alpha_j<0} \frac{b_j - (Ay)_j + \alpha_j\eta^T y}{\alpha_j} \tag{4.2}$$

$$\mathcal{V}^+ = \mathcal{V}^+(y) = \min_{j:\ \alpha_j>0} \frac{b_j - (Ay)_j + \alpha_j\eta^T y}{\alpha_j}. \tag{4.3}$$

$$\mathcal{V}^0 = \mathcal{V}^0(y) = \min_{j:\ \alpha_j=0} b_j - (Ay)_j \tag{4.4}$$

Moreover, $(\mathcal{V}^+, \mathcal{V}^-, \mathcal{V}^0)$ *are independent of* $\eta^T y$.

However, before stating the proof of this lemma, we show how it is used to obtain our main result.

THEOREM 4.2. *Let* $F^{[a,b]}_{\mu,\sigma^2}$ *denote the CDF of* $TN(\mu, \sigma, a, b)$, *i.e.:*

$$F^{[a,b]}_{\mu,\sigma^2}(x) = \frac{\Phi((x-\mu)/\sigma) - \Phi((a-\mu)/\sigma)}{\Phi((b-\mu)/\sigma) - \Phi((a-\mu)/\sigma)} \tag{4.5}$$

where Φ *is the CDF of a* $N(0,1)$ *random variable. Then* $F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y)$ *is a pivotal quantity, conditional on* $\{Ay \le b\}$*:*

$$F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y) \ \big|\ \{Ay \le b\} \sim \mathrm{Unif}(0,1) \tag{4.6}$$

where $\mathcal{V}^-$ *and* $\mathcal{V}^+$ *are defined in* (4.2) *and* (4.3).

PROOF. First, note that for *fixed* v^- and v^+,

$$F^{[v^-,v^+]}_{\eta^T\mu,\eta^T\Sigma\eta}(\eta^T y) \ \big|\ \{v^- \le \eta^T y \le v^+\} \sim \mathrm{Unif}(0,1). \tag{4.7}$$

This follows from the probability integral transform, since $\eta^T y \ \big|\ \{v^- \le \eta^T y \le v^+\} \sim TN(\eta^T\mu, \eta^T\Sigma\eta, v^-, v^+)$.

To allow for random $\mathcal{V}^-$ and $\mathcal{V}^+$, we first use Lemma 4.1 to rewrite:

$$\begin{aligned}\mathbb{P}\left(F^{[\mathcal{V}^-,\mathcal{V}^+]}(\eta^T y) \le t \ \big|\ Ay \le b\right) \\ = \mathbb{P}\left(F^{[\mathcal{V}^-,\mathcal{V}^+]}(\eta^T y) \le t \ \big|\ \mathcal{V}^- \le \eta^T y \le \mathcal{V}^+, V^0 \ge 0\right)\end{aligned}$$

where we have suppressed the mean and variance in the notation of F. Next, we condition on $\mathcal{V}^- = v^-$ and $\mathcal{V}^+ = v^+$ and integrate over their (conditional) distribution. Letting $\mathbb{Q}$ denote the law of $(\mathcal{V}^-, \mathcal{V}^+)$ conditional on the event $\{\mathcal{V}^- \le \eta^T y \le \mathcal{V}^+, \mathcal{V}^0 \ge 0\}$:

$$= \int \mathbb{P}\left(F^{[v^-,v^+]}(\eta^T y) \le t \;\middle|\; v^- \le \eta^T y \le v^+, \mathcal{V}^0 \ge 0, \mathcal{V}^- = v^-, \mathcal{V}^+ = v^+\right) d\mathbb{Q}(v^-, v^+)$$

Now using the independence of $(\mathcal{V}^-, \mathcal{V}^+, \mathcal{V}^0)$ and $\eta^T y$:

$$= \int \mathbb{P}\left(F^{[v^-,v^+]}(\eta^T y) \le t \;\middle|\; v^- \le \eta^T y \le v^+\right) d\mathbb{Q}(v^-, v^+)$$

By (4.7), $\mathbb{P}\left(F^{[v^-,v^+]}(\eta^T y) \le t \;\middle|\; v^- \le \eta^T y \le v^+\right) = t$, so:

$$= t \int d\mathbb{Q}(v^-, v^+) = t.$$

This establishes the conditional Unif$(0,1)$ distribution of F. □

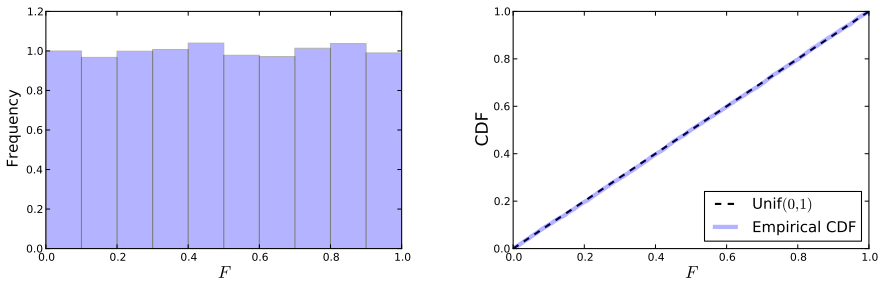

Fig 1: Histogram and empirical distribution of $F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y)$ obtained by sampling $y \sim N(\mu, \Sigma)$ constrained to $\{Ay \le b\}$. The distribution is very close to Unif$(0,1)$.

Proof of Lemma 4.1. First, we subtract the conditional expectation $\mathbb{E}(Ay|\eta^T y)$ from both sides of $Ay \le b$ to obtain a quantity independent of $\eta^T y$ on the left-hand side:

$$Ay - \mathbb{E}(Ay|\eta^T y) \le b - \mathbb{E}(Ay|\eta^T y)$$

The conditional expectation is $\mathbb{E}(Ay|\eta^T y) = A\mu + \alpha(\eta^T y - \eta^T \mu)$, where α is defined in (4.1). Substituting this into the above equation, we obtain:

$$\begin{aligned} Ay - A\mu - \alpha(\eta^T y - \eta^T \mu) &\leq b - A\mu - \alpha(\eta^T y - \eta^T \mu) \\ Ay - b - \alpha\eta^T y &\leq -\alpha\eta^T y \end{aligned}$$

Since $\alpha \in \mathbb{R}^q$, this is in fact q separate inequalities. For inequality j, we divide both sides by $-\alpha_j$, obtaining two cases depending on whether $\alpha_j \lessgtr 0$:

$$\begin{aligned} \eta^T y &\geq \frac{b_j - (Ay)_j + \alpha_j \eta^T y}{\alpha_j} && \alpha_j < 0 \\ \eta^T y &\leq \frac{b_j - (Ay)_j + \alpha_j \eta^T y}{\alpha_j} && \alpha_j > 0 \\ 0 &\geq b_j - (Ay)_j && \alpha_j = 0 \end{aligned}$$

To obtain the tightest possible bound, we take the maximum of the lower bounds and the minimum of the upper bounds:

$$\underbrace{\max_{j:\ \alpha_j<0} \frac{b_j - (Ay)_j + \alpha_j \eta^T y}{\alpha_j}}_{\mathcal{V}^-} \leq \eta^T y \leq \underbrace{\min_{j:\ \alpha_j>0} \frac{b_j - (Ay)_j + \alpha_j \eta^T y}{\alpha_j}}_{\mathcal{V}^+} \tag{4.8}$$

Moreover, $\mathcal{V}^+$ and $\mathcal{V}^-$ are the minimum and maximum of quantities independent of $\eta^T y$, so they are also independent of $\eta^T y$.

□

Although the proof of Lemma 4.1 is elementary, the geometric picture gives more intuition as to why $\mathcal{V}^+$ and $\mathcal{V}^-$ are independent of $\eta^T y$. Without loss of generality, we assume $||\eta||_2 = 1$ and $y \sim N(\mu, I)$ (since otherwise we could replace y by $\Sigma^{-\frac{1}{2}} y$). Now we can decompose y into two independent components, a 1-dimensional component $\eta^T y$ and an $(n-1)$-dimensional component orthogonal to η:

$$y = \eta^T y + P_{\eta^\perp} y.$$

The case of $n = 2$ is illustrated in Figure 2. $\mathcal{V}^-$ and $\mathcal{V}^+$ are independent of $\eta^T y$, since they are functions of $P_{\eta^\perp}$ only, which is independent of $\eta^T y$.

In Figure 3, we plot the density of the truncated Gaussian, noting that its shape depends on the location of μ relative to $[a, b]$ as well as the width relative to σ.

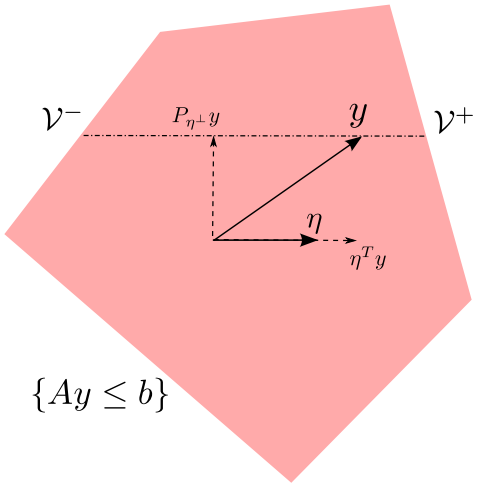

Fig 2: A picture demonstrating that the set $\{Ay \leq b\}$ can be characterized by $\{\mathcal{V}^- \leq \eta^T y \leq \mathcal{V}^+\}$. Assuming $\Sigma = I$ and $||\eta||_2 = 1$, $\mathcal{V}^-$ and $\mathcal{V}^+$ are functions of $P_{\eta^\perp} y$ only, which is independent of $\eta^T y$.

5. Application to Inference for the Lasso. In this section, we apply the theory developed in in Sections 3 and 4 to the lasso. In particular, we will construct confidence intervals for the active variables and test the chosen model based on the pivot developed in Section 4.

To summarize the developments so far, recall that our model says that $y \sim N(\mu, \sigma^2 I)$. The distribution of interest is $y \mid \{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$. By Theorem 3.1, this is equivalent to $y \mid \{A(E, z_E)y \leq b(E, z_E)\}$ defined in Proposition 3.2. Now we can apply Theorem 4.2 to obtain the (conditional) pivot

$$F^{[\mathcal{V}^-, \mathcal{V}^+]}_{\eta^T\mu,\ \sigma^2||\eta||_2^2}(\eta^T y) \mid \{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\} \sim \text{Unif}(0, 1) \tag{5.1}$$

for any η, where $\mathcal{V}^-$ and $\mathcal{V}^+$ are defined in (4.2) and (4.3). Note that $A(E, z_E)$ and $b(E, z_E)$ appear in this pivot through $\mathcal{V}^-$ and $\mathcal{V}^+$. This pivot will play a central role in all of the applications that follow.

5.1. *Confidence Intervals for the Active Variables.* In this section, we describe how to form confidence intervals for the components of $\beta^\star_{\hat{E}} = X^+_{\hat{E}}\mu$. If we choose

$$\eta_j = (X^T_{\hat{E}})^+ e_j, \tag{5.2}$$

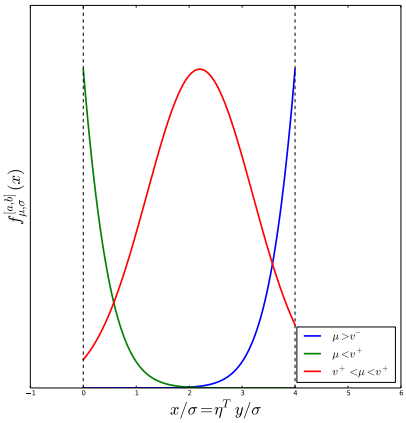

Fig 3: The density of the truncated Gaussian with distribution $F^{[v^-,v^+]}_{\mu,\sigma^2}$ depends on the width of $[v^-, v^+]$ relative to σ as well as the location of μ relative to $[v^-, v^+]$. When μ is firmly inside the interval, the distribution resembles a Gaussian. As μ varies drifts outside $[v^-, v^+]$, the density begins to converge to an exponential distribution with mean inversely proportional to the distance between μ and its projection onto $[v^-, v^+]$.

then $\eta_j^T \mu = \beta^\star_{\hat{E},j}$, so the above framework provides a method for inference about the j^{th} variable in the model $\hat{E}$. Note that this reduces to inference about the true β^0_j if $\hat{E} \supset S := \{j : \beta^0_j \neq 0\}$, as discussed in Section 1. The conditions under which this holds is well known in the literature, cf. Bühlmann and van de Geer (2011).

By applying Theorem 4.2, we obtain the following (conditional) pivot for $\beta^\star_{\hat{E},j}$:

$$F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\beta^\star_{\hat{E},j},\ \sigma^2||\eta_j||^2}(\eta_j^T y) \ \Big|\ \{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\} \sim \text{Unif}(0,1).$$

Note that j and η_j are both random—but only through $\hat{E}$, a quantity which is fixed after conditioning—so Theorem 4.2 holds even for this "random" choice of η. The obvious way to obtain an interval is to "invert" the pivot. In other words, since

$$\mathbb{P}\left(\frac{\alpha}{2} \leq F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\beta^\star_{\hat{E},j},\ \sigma^2||\eta_j||^2}(\eta_j^T y) \leq 1 - \frac{\alpha}{2} \ \Big|\ \{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}\right) = \alpha,$$

one can define a $(1-\alpha)$ (conditional) confidence interval for $\beta^\star_{\hat{E},j}$ as

$$\left\{\beta^\star_{\hat{E},j} : \frac{\alpha}{2} \leq F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\beta^\star_{\hat{E},j},\ \sigma^2||\eta_j||^2}(\eta_j^T y) \leq 1-\frac{\alpha}{2}\right\}.$$

In fact, F is monotone decreasing in $\beta^\star_{\hat{E},j}$, so to find its endpoints, one need only solve for the root of a smooth one-dimensional function. The monotonicity is a consequence of the fact that the truncated Gaussian distribution is a natural exponential family and hence has monotone likelihood ratio in μ. The details can be found in Appendix A.

We now formalize the above observations in the following result, an immediate consequence of Theorem 4.2.

COROLLARY 5.1. *Let η_j be defined as in (5.2), and let $L^j_\alpha = L^j_\alpha(\eta_j, \hat{E}, \hat{z}_{\hat{E}})$ and $U^j_\alpha = U^j_\alpha(\eta_j, \hat{E}, \hat{z}_{\hat{E}})$ be the (unique) values satisfying*

$$F^{[\mathcal{V}^-,\mathcal{V}^+]}_{L^j_\alpha,\ \sigma^2||\eta_j||^2}(\eta_j^T y) = 1-\alpha \qquad F^{[\mathcal{V}^-,\mathcal{V}^+]}_{U^j_\alpha,\ \sigma^2||\eta_j||^2}(\eta_j^T y) = \alpha$$

Then $[L^j_\alpha, U^j_\alpha]$ is a $(1-\alpha)$ confidence interval for $\eta_j^T\mu$, conditional on $(\hat{E}, \hat{z}_{\hat{E}})$:

$$\mathbb{P}\left(\eta_j^T\mu \in [L^j_\alpha, U^j_\alpha] \;\middle|\; \{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}\right) = 1-\alpha. \tag{5.3}$$

The above discussion has focused on constructing intervals for a single j. If we repeat the procedure for each $j \in \hat{E}$, our intervals in fact control the false coverage rate (FCR) of Benjamini and Yekutieli (2005).

COROLLARY 5.2. *For each $j \in \hat{E}$,*

$$\mathbb{P}\left(\eta_j^T\mu \in [L^j_\alpha, U^j_\alpha]\right) = 1-\alpha. \tag{5.4}$$

Furthermore, the FCR of the intervals $\left\{[L^j_\alpha, U^j_\alpha]\right\}_{j\in\hat{E}}$ is α.

If $\eta^T y$ are not near the boundaries $[\mathcal{V}^-, \mathcal{V}^+]$, then the intervals will be relatively short. This is shown in Figure 4. Figure 5 shows two simulations that demonstrate our intervals cover at the nominal rate. We leave an exhaustive study of such intervals for the lasso to future work, noting that the truncation framework described can be used to form intervals with exact coverage properties.

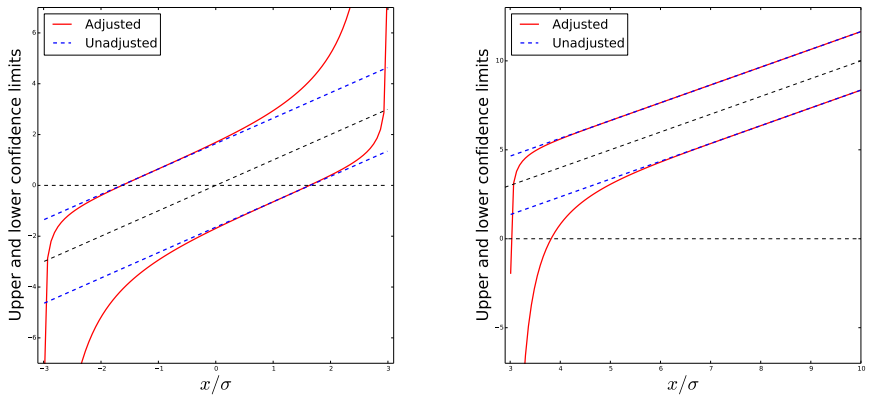

Fig 4: Upper and lower bounds of 90% confidence intervals based on $[a, b] = [-3\sigma, 3\sigma]$ as a function of the observation x/σ. We see that as long as the observation x/σ is roughly 0.5σ away from either boundary, the size of the intervals is comparable to an unadjusted confidence interval.

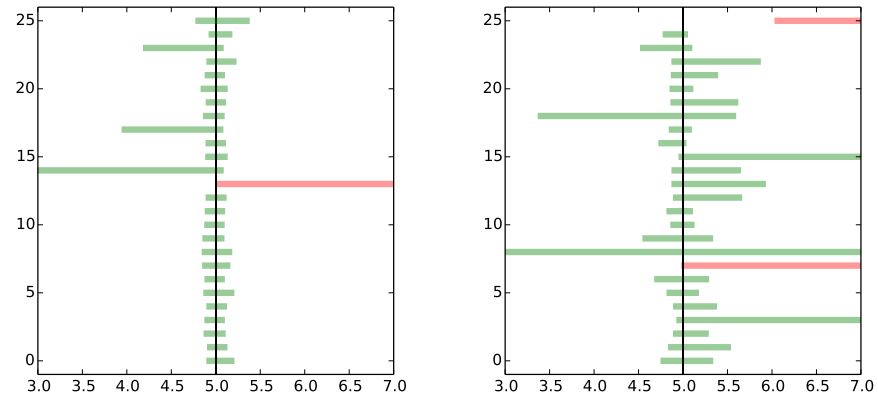

Fig 5: 90% confidence intervals for $\eta_1^T \mu$ for a small ($n = 100$, $p = 50$) and a large ($n = 100$, $p = 200$) uncorrelated Gaussian design, computed over 25 simulated data sets. The true model has five non-zero coefficients, all set to 5.0, and the noise variance is 0.25. A green bar means the confidence interval covers the true value while a red bar means otherwise.

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

$$H_0 : \beta^0_{-\hat{E}} = \mathbf{0}. \tag{5.5}$$

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

$$H_0 : X^T_{-\hat{E}}(I - P_{\hat{E}})\mu = \mathbf{0}. \tag{5.6}$$

This quantity is the partial correlation of the non-selected variables with μ, adjusting for the variables in $\hat{E}$. This is more general because if we assume $\mu = X\beta^0$ for some β^0 and X is full rank, then rejecting (5.6) implies that there is an $i \in \text{supp}(\beta^0)$ not in $\hat{E}$, so we would also reject (5.5).

The natural approach is to compare the observed partial correlations $X^T_{-E}(I - P_E)y$ to $\mathbf{0}$. However, the framework of Section 4 only allows tests of μ in a single direction η. To make use of that framework, we can choose η such that it selects the maximum magnitude of $X^T_{-E}(I - P_E)y$. 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 $j^\star := \text{argmax}_j \, |e_j^T X^T_{-E}(I - P_E)y|$ and $s_j := \text{sign}(e_j^T X^T_{-E}(I - P_E)y)$, we set

$$\eta_{j^\star} = s_{j^\star}(I - P_E)X_{-E}e_{j^\star}, \tag{5.7}$$

and test $H_0 : \eta^T_{j^\star}\mu = 0$. However, the results in Section 4 cannot be directly applied to this setting because $j^\star$ and $s_{j^\star}$ are random variables that are not measurable with respect to $(\hat{E}, \hat{z}_{\hat{E}})$.

To resolve this issue, we propose a test conditional not only on $(\hat{E}, \hat{z}_{\hat{E}})$, but also on the index and sign of the maximizer:

$$\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E), \, (j^\star, s_{j^\star}) = (j, s)\}. \tag{5.8}$$

A test that is level α conditional on (5.8) for all (E, z_E) and (j, s) is also level α conditional on $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$.

In order to use the results of Section 4, we must show that (5.8) can be written in the form $A(E, z_E, j, s)y \le b(E, z_E, j, s)$. This is indeed possible, and the following proposition provides an explicit construction.

PROPOSITION 5.3. *Let A_0, b_0, A_1, b_1 be defined as in Proposition 3.2. Then:*

$$\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E),\ (j^\star, s_{j^\star}) = (j, s)\} = \left\{ \begin{pmatrix} A_0(E, z_E) \\ A_1(E, z_E) \\ A_2(E, j, s) \end{pmatrix} y < \begin{pmatrix} b_0(E, z_E) \\ b_1(E, z_E) \\ \mathbf{0} \end{pmatrix} \right\}$$

where $A_2(E, j, s)$ is defined as

$$A_2(E, j, s) = -s \begin{pmatrix} D_j(E) \\ S_j(E) \end{pmatrix} X_{-E}^T (I - P_E)$$

and D_j and S_j are $(|E| - 1) \times |E|$ operators that compute the difference and sum, respectively, of the j^{th} element with the other elements, e.g.,

$$D_1 = \begin{pmatrix} 1 & -1 & & & \\ 1 & & -1 & & \\ & & & \ddots & \\ 1 & & & & -1 \end{pmatrix} \qquad S_1 = \begin{pmatrix} 1 & 1 & & & \\ 1 & & 1 & & \\ & & & \ddots & \\ 1 & & & & 1 \end{pmatrix}.$$

PROOF. The constraints $\{A_0 y < b_0\}$ and $\{A_1 y < b_1\}$ come from Proposition (3.2) and encode the constraints $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$. We show that the last two sets of constraints encode $\{(j^\star, s_{j^\star}) = (j, s)\}$.

Let $r := X_{-E}^T (I - P_E) y$ denote the vector of partial correlations. If $s = +1$, then $|r_j| > |r_i|$ for all $i \neq j$ if and only if $r_j - r_i > 0$ and $r_j + r_i > 0$ for all $i \neq j$. We can write this as $D_j r > 0$ and $S_j r > 0$. If $s = -1$, then the signs are flipped: $D_j r < 0$ and $S_j r < 0$. This establishes

$$\{(j^\star, s_{j^\star}) = (j, s)\} = \left\{ -s \begin{pmatrix} D_j \\ S_j \end{pmatrix} r < \mathbf{0} \right\} = \{A_2 y < \mathbf{0}\}.$$

□

Because of Proposition 5.3, we can now obtain the following result as a simple consequence of Theorem 4.2, which says that $F_{0, \sigma^2 ||\eta_{j^\star}||^2}^{[\mathcal{V}^-, \mathcal{V}^+]}(\eta_{j^\star}^T y) \sim \text{Unif}(0, 1)$, conditional on the set (5.8) and H_0. We reject when $F_{0, \sigma^2 ||\eta_j^*||^2}^{[\mathcal{V}^-, \mathcal{V}^+]}(\eta_{j^\star}^T y)$ is large because $F_{0,\ \sigma^2 ||\eta_j^*||^2}^{[\mathcal{V}^-, \mathcal{V}^+]}(\cdot)$ is monotone increasing in the argument and $\eta_{j^*}^T \mu$ is likely to be positive under the alternative.

COROLLARY 5.4. *Let H_0 and $\eta_{j^\star}$ be defined as in (5.7). Then, the test which rejects when*

$$\left\{ F_{0,\ \sigma^2 ||\eta_j^*||^2}^{[\mathcal{V}^-, \mathcal{V}^+]}(\eta_{j^\star}^T y) > 1 - \alpha \right\}$$

is level α*, conditional on* $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E), (j^\star, s_{j^\star}) = (j, s)\}$*. That is,*

$$\mathbb{P}\left(F^{[\mathcal{V}^-,\mathcal{V}^+]}_{0,\ \sigma^2||\eta_{j^\star}||^2}(\eta_{j^\star}^T y) > 1-\alpha \ \middle|\ \{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E),\ (j^\star, s_{j^\star}) = (j, s)\} \cap H_0\right) = \alpha.$$

In particular, since this holds for every (E, z_E, j, s)*, this test also controls Type I error conditional only on* $(\hat{E}, \hat{z}_{\hat{E}})$*, and unconditionally:*

$$\mathbb{P}\left(F^{[\mathcal{V}^-,\mathcal{V}^+]}_{0,\ \sigma^2||\eta_{j^\star}||^2}(\eta_{j^\star}^T y) > 1-\alpha \ \middle|\ \{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\} \cap H_0\right) = \alpha$$
$$\mathbb{P}\left(F^{[\mathcal{V}^-,\mathcal{V}^+]}_{0,\ \sigma^2||\eta_{j^\star}||^2}(\eta_{j^\star}^T y) > 1-\alpha \ \middle|\ H_0\right) = \alpha.$$

Figures 6 and 7 show the results of four simulation studies that demonstrate that the p-values are uniformly distributed when $H_{0,\lambda}$ is true and stochastically smaller than $\text{Unif}(0,1)$ when it is false.

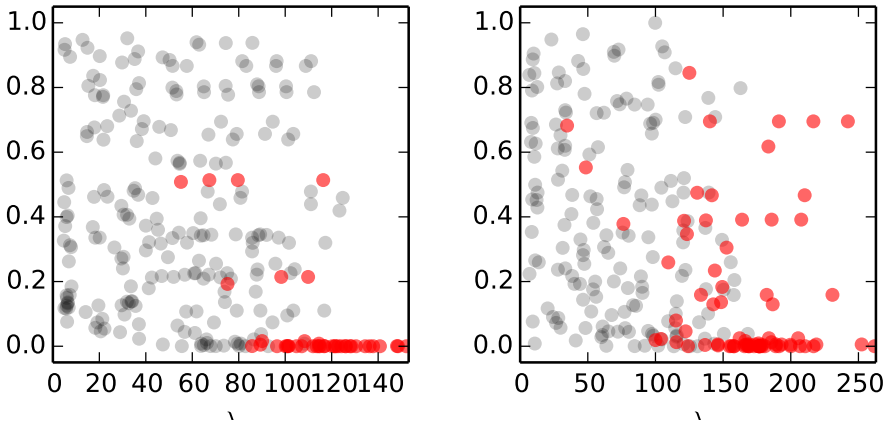

Fig 6: P-values for $H_{0,\lambda}$ at various λ values for a small ($n = 100$, $p = 50$) and a large ($n = 100$, $p = 200$) uncorrelated Gaussian design, computed over 50 simulated data sets. The true model has three non-zero coefficients, all set to 1.0, and the noise variance is 2.0. We see the p-values are $\text{Unif}(0,1)$ when the selected model includes the truly relevant predictors (black dots) and are stochastically smaller than $\text{Unif}(0,1)$ when the selected model omits a relevant predictor (red dots).

6. Data Example. We illustrate the application of inference for the lasso to the diabetes data set from Efron et al. (2004). First, all variables were standardized. Then, we chose λ according to the strategy in Negahban

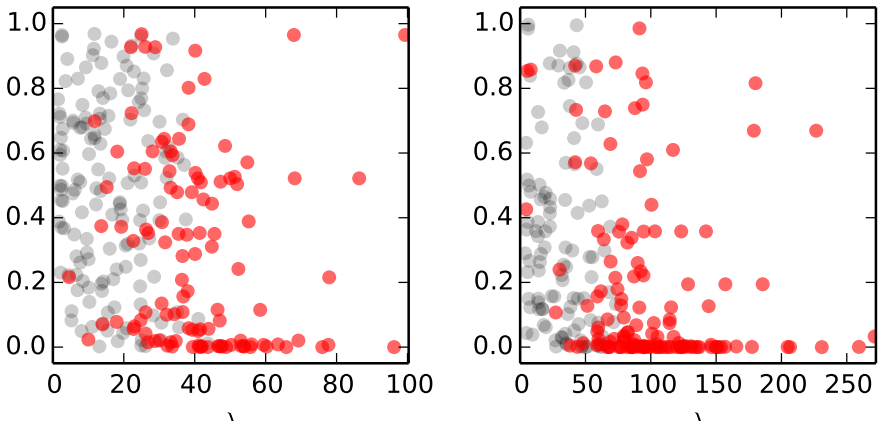

Fig 7: P-values for $H_{0,\lambda}$ at various λ values for a small ($n = 100$, $p = 50$) and a large ($n = 100$, $p = 200$) *correlated* ($\rho = 0.7$) Gaussian design, computed over 50 simulated data sets. The true model has three non-zero coefficients, all set to 1.0, and the noise variance is 2.0. Since the predictors are correlated, the relevant predictors are not always selected first. However, the p-values remain uniformly distributed when $H_{0,\lambda}$ is true and stochastically smaller than $\text{Unif}(0, 1)$ otherwise.

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 intervals are shown in Figure 8, alongside the unadjusted confidence intervals produced by fitting OLS to the four selected variables, ignoring the selection. The latter is not a valid confidence interval conditional on the model. 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 model selection and the other for inference. This is a competitor method that also produces valid confidence intervals conditional on the model. In this case, data splitting selected the same four variables, and the confidence intervals were formed based on OLS on the half of the data set not used for model selection.

We can make two main observations from Figure 8.

1. The adjusted intervals provided by our method essentially reproduces the OLS intervals for the strong effects, whereas data splitting results in a loss of power by roughly a factor of $\sqrt{2}$ (since only $n/2$ observations are used in the inference).
2. One variable, `S3`, which would have been deemed significant using the

OLS intervals, is no longer significant after adjustment. This demonstrates that taking model selection into account can have substantive impacts on the conclusions that are made.

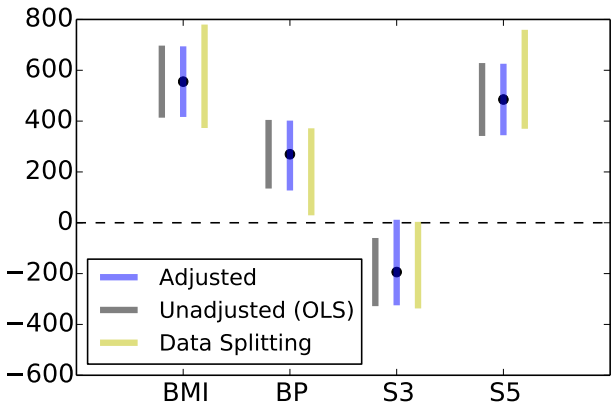

Fig 8: Inference for the four variables selected by the lasso ($\lambda = 190$) on the diabetes data set. The point estimate and adjusted confidence intervals using the approach in Section 5 are shown in blue. The gray show the OLS intervals, which ignore selection. The yellow lines show the intervals produced by splitting the data into two halves, forming the interval based on only half of the data.

7. Minimal Post-Selection Inference. We have described how to perform post-selection inference for the lasso conditional on both the equicorrelation set and signs $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$. However, recall from Section 1 that the goal was inference conditional solely on the model, i.e., $\{\hat{E} = E\}$. In this section, we extend our framework to this setting, which we call minimal post-selection inference because we condition on the minimal set necessary for the random η to be measurable. This results in more precise confidence intervals at the expense of greater computational cost.

To this end, we note that $\{\hat{E} = E\}$ is simply

$$\bigcup_{z_E \in \{-1,1\}^{|E|}} \{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\},$$

where the union is taken over all choices of signs. Therefore, the distribution of y conditioned on only the equicorrelation set $\{\hat{E} = E\}$ is a Gaussian

vector constrained to a union of polytopes

$$y \mid \bigcup_{z_E \in \{-1,1\}^{|E|}} \{A(E, z_E)y \leq b(E, z_E)\},$$

where $A(E, z_E)$ and $b(E, z_E)$ are given by (3.2).

To obtain inference about $\eta^T \mu$, we follow the arguments in Section 4 to obtain that this conditional distribution is equivalent to

$$\eta^T y \mid \bigcup_{z_E \in \{-1,1\}^{|E|}} \{\mathcal{V}^-_{z_E}(y) \leq \eta^T y \leq \mathcal{V}^+_{z_E}(y), \mathcal{V}^0_{z_E}(y) \geq 0\}, \tag{7.1}$$

where $\mathcal{V}^-_{z_E}$, $\mathcal{V}^+_{z_E}$, $\mathcal{V}^0_{z_E}$ are defined according to (4.2), (4.3), (4.4) with $A = A(E, z_E)$ and $b = b(E, z_E)$. Moreover, all of these quantities are still independent of $\eta^T y$, so instead of having a Gaussian truncated to a single interval $[\mathcal{V}^-, \mathcal{V}^+]$ as in Section 4, we now have a Gaussian truncated to the union of intervals $\bigcup_{z_E}[\mathcal{V}^-_{z_E}, \mathcal{V}^+_{z_E}]$. The geometric intuition is illustrated in Figure 9.

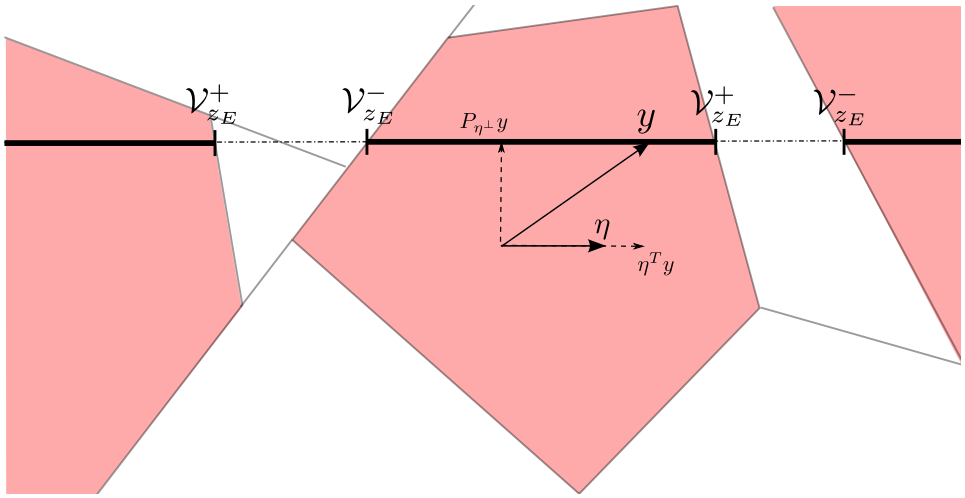

Fig 9: A picture demonstrating the effect of taking a union over signs. The polytope in the middle corresponds to the $(\hat{E}, \hat{z}_{\hat{E}})$ that was observed and is the same polytope as in Figure 2. The difference is that we now consider potential (E, z_E) in addition to the one that was observed. The polytopes for the other (E, z_E) which have the same equicorrelation set $\hat{E}$ are red. The conditioning set is the union of these polytopes. We see that for y to be in this union, $\eta^T y$ must be in $\bigcup_{z_E}[\mathcal{V}^-_{z_E}, \mathcal{V}^+_{z_E}]$. The key point is that all of the $\mathcal{V}^-_{z_E}$ and $\mathcal{V}^+_{z_E}$ are still functions of only $P_{\eta^\perp} y$ and so are independent of $\eta^T y$.

Finally, the probability integral transform once again yields a pivot:

$$F^{\bigcup_{z_E}[\mathcal{V}^-_{z_E}(y),\mathcal{V}^+_{z_E}(y)]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y) \mid \{\hat{E} = E\} \sim \text{Unif}(0,1).$$

It is now more useful to think of the notation of F as indicating the truncation set $C \subset \mathbb{R}$:

$$F^{C}_{\mu,\sigma^2}(x) := \frac{\Phi((-\infty, x] \cap C)}{\Phi(C)}, \tag{7.2}$$

where Φ is the law of a $N(0,1)$ random variable. We summarize these results in the following theorem.

THEOREM 7.1. *Let $F^{\bigcup_i [a_i,b_i]}_{\mu,\sigma^2}$ be the CDF of a normal truncated to the union of intervals $\bigcup_i [a_i, b_i]$, i.e., given by (7.2). Then:*

$$F^{\bigcup_{z_E}[\mathcal{V}^-_{z_E}(y),\mathcal{V}^+_{z_E}(y)]}_{\eta^T\mu,\ \eta^T\Sigma\eta}(\eta^T y) \mid \{\hat{E} = E\} \sim \text{Unif}(0,1), \tag{7.3}$$

where $\mathcal{V}^-_{z_E}(y)$ and $\mathcal{V}^+_{z_E}(y)$ are defined in (4.2) and (4.3) with $A = A(E, z_E)$ and $b = b(E, z_E)$.

The derivations of the confidence intervals and hypothesis tests in Section 5 remain valid using (7.3) as the pivot instead of (5.1). Figure 10 illustrates the effect of minimal post-selection inference in a simulation study, as compared with the "simple" inference described previously. The intervals are similar in most cases, but one can obtain great gains in precision using the minimal intervals when the simple intervals are very wide.

However, the tradeoff for this increased precision is greater computational cost. We computed $\mathcal{V}^-_{z_E}$ and $\mathcal{V}^+_{z_E}$ for all $z_E \in \{-1, 1\}^{|E|}$, which is only feasible when $|E|$ is fairly small. In what follows, we revert to the simple intervals described in Section 5, but extensions to the minimal inference setting are straightforward.

8. Extensions.

8.1. *Intervals for coefficients in full model.* Assuming the linear model $\mu = X\beta^0$ is correct, one might be interested in confidence intervals for β^0_j instead of $\beta^\star_{\hat{E},j}$. In this case, the target is not selection-dependent, so when $p < n$, one can simply fit least squares and the standard OLS intervals will be valid (provided that one has a reasonable estimate of σ^2 when $p \approx n$). However, this approach is not possible when $p > n$.

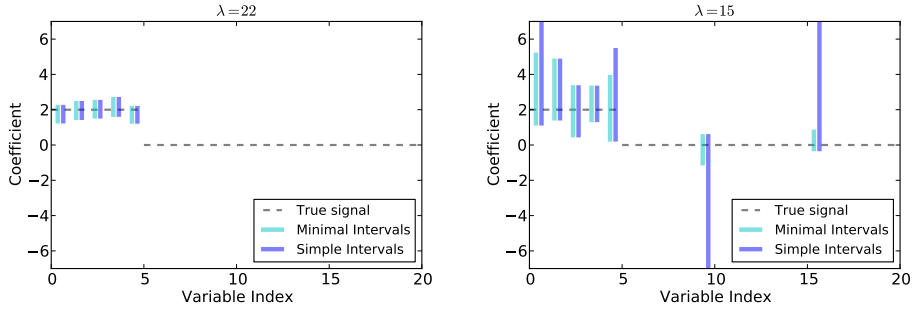

Fig 10: Comparison of the minimal and simple intervals as applied to the same simulated data set for two values of λ. The simulated data featured $n = 25$, $p = 50$, and 5 true non-zero coefficients; only the first 20 coefficients are shown. (We have included variables with no intervals to emphasize that inference is only on the selected variables.) We see that the simple intervals are virtually as good as the minimal intervals most of the time; the advantage of the minimal intervals is realized when the estimate is unstable and the simple intervals are very long, as in the right plot.

One proposal that has been suggested by numerous authors (Javanmard and Montanari, 2013; van de Geer et al., 2013; Zhang and Zhang, 2014) is to use a debiased version of the lasso estimator:

$$\hat{\beta}^d = \hat{\beta} + \Theta X^T (y - X\hat{\beta})$$

where $\hat{\beta}$ denotes the lasso estimate and Θ is an approximate inverse of $X^T X$. From the representation

$$\hat{\beta}^d = \beta^0 + \Theta X^T \epsilon + \underbrace{(I - \Theta X^T X)(\hat{\beta} - \beta_0)}_{\hat{\Delta}}, \ \epsilon \sim N(0, \Sigma), \tag{8.1}$$

the authors provide sufficient conditions for $\hat{\Delta} \to 0$, thereby making $\hat{\beta}^d$ asymptotically unbiased for β_0.

We now show that in our framework, $\hat{\beta}^d$ is estimating a selection dependent target $\beta^d = \beta^d(\hat{E}, \hat{z}_{\hat{E}})$. By now familiar manipulations of the optimality conditions (2.1) and (2.2),

$$\hat{\beta}_{\hat{E}} = X_{\hat{E}}^+ y - \lambda (X_{\hat{E}}^T X_{\hat{E}})^{-1} \hat{z}_{\hat{E}}.$$

Assuming that $X = [X_{\hat{E}} : X_{-\hat{E}}]$, this allows us to write

$$\hat{\beta}^d = (I - \Theta X^T X) \begin{pmatrix} X_{\hat{E}}^+ y - \lambda (X_{\hat{E}}^T X_{\hat{E}})^{-1} \hat{z}_{\hat{E}} \\ \mathbf{0} \end{pmatrix} + \Theta X^T y \tag{8.2}$$

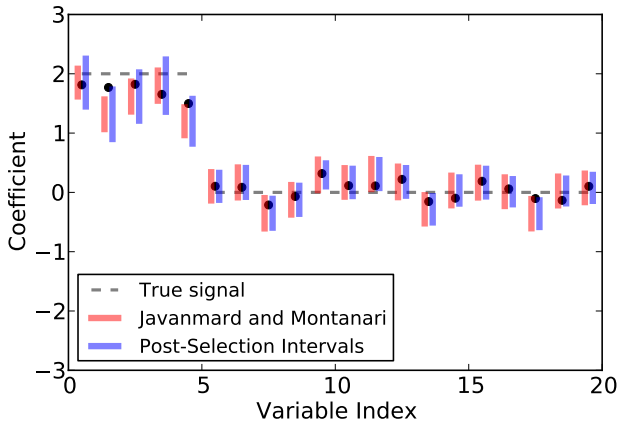

Fig 11: Confidence intervals for the coefficients in a design with $n = 25$, $p = 50$, and 5 non-zero coefficients. Only the first 20 coefficients are shown. The dotted line represents the true signal, and the points represent the (biased) post-selection target. The colored bars denote the intervals.

On the set $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$, $\hat{\beta}^d$ is simply an affine function of y. Thus, we can write $\hat{\beta}^d_j$ in the form $\eta_j^T y + h_j$, which can be thought of in the post-selection framework as estimating $\beta^d_j := \eta_j^T \mu + h_j$. If $\mu = X\beta^0$, then one obtains

$$\beta^d = \beta^0 + \underbrace{(I - \Theta X^T X)\left(\begin{pmatrix} X^+_{\hat{E}}\mu - \lambda(X^T_{\hat{E}}X_{\hat{E}})^{-1}\hat{z}_{\hat{E}} \\ \mathbf{0} \end{pmatrix} - \beta^0\right)}_{\Delta}.$$

Δ can be thought of as the population analog (conditional on $\hat{E}, \hat{z}_{\hat{E}}$) of $\hat{\Delta}$ in (8.1). The post-selection framework of Section 4 allows us to construct exact conditional intervals for $\beta^d_j = \beta^0_j + \Delta_j$. Under the assumptions about the choice of Θ and high-dimensional consistency in Javanmard and Montanari (2013), $\Delta_j \to 0$, so β^d_j is equivalent to β^0_j.

Figure 11 shows the results of a simulation study. It makes clear that the intervals of Javanmard and Montanari (2013) do not necessarily cover the truth, but rather a biased target. Our post-selection intervals also cover the same biased target conditional on $(\hat{E}, \hat{z}_{\hat{E}})$.

8.2. *Elastic net.* One problem with the lasso is that it tends to select only one variable out of a set of correlated variables, resulting in estimates which

are unstable. The elastic net (Zou and Hastie, 2005) adds an ℓ_2 penalty to the lasso objective in order to stabilize the estimates:

$$\hat{\beta}^e = \underset{\beta}{\operatorname{argmin}} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 + \frac{\gamma}{2} \|\beta\|_2^2 . \tag{8.3}$$

Using a nearly identical argument to the one in Section 3, we see that necessary and sufficient conditions for $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$ are the existence of $U(E, z_E)$ and $W(E, z_E)$ satisfying

$$\begin{aligned} (X_E^T X_E + \gamma I)U - X_E^T y + \lambda z_E &= 0 \\ X_{-E}^T X_E U - X_{-E}^T y + \lambda W &= 0 \\ \operatorname{sign}(U) = z_E, \ W &\in (-1, 1). \end{aligned}$$

Solving for U and W, we see that the selection event can be written

$$\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\} = \left\{ \begin{pmatrix} A_0(E, z_E) \\ A_1(E, z_E) \end{pmatrix} y < \begin{pmatrix} b_0(E, z_E) \\ b_1(E, z_E) \end{pmatrix} \right\} \tag{8.4}$$

where A_0, A_1, b_0, and b_1 are the same as in Proposition 3.2, except replacing $(X_E^T X_E)^{-1}$, which appears in the expressions through P_E and $(X_E^T)^+$, by the "damped" version $(X_E^T X_E + \gamma I)^{-1}$.

Having rewritten the selection event in the form (8.4), we can once again apply the framework of Section 4 to obtain a test for the elastic net conditional on this event.

8.3. *Alternative norms as test statistics.* In Section 5.2 we used the test statistic

$$T_\infty = \|X_{-\hat{E}}^T (I - P_{\hat{E}}) y\|_\infty$$

and its conditional distribution on $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$ to test whether we had missed any large partial correlations in using $\hat{E}$ as the estimated active set. If we have indeed missed some variables in E there is no reason to suppose that the mean of $X_{-E}^T (I - P_E) y$ is sparse; hence the ℓ_∞ norm may not be the best norm to use as a test statistic.

In principle, we could have used virtually any norm, as long as we can say something about the distribution of this norm conditional on $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$. Problems of this form are considered in Taylor et al. (2013). For example, if we consider the quadratic

$$T_2 = \|X_{-E}^T (I - P_E) y\|_2$$

the general approach in Taylor et al. (2013) derives the conditional distribution of T_2 conditioned on

$$\eta_2^* = \arg\max_{\|\eta\|_2 \leq 1} \eta^T (X_{-E}^T (I - P_E) y).$$

In general, this distribution will be a χ^2 subject to random truncation as in Section 4 (see the group lasso examples in Taylor et al. (2013)). Adding the constraints encoded by $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$ affects only the random truncation $[\mathcal{V}^-, \mathcal{V}^+]$.

8.4. *Estimation of* σ^2. As noted above, all of our results rely on a reliable estimate of σ^2. While there are several approaches to estimating σ^2 in the literature, the truncated Gaussian theory described in this work itself provides a natural estimate.

Suppose the linear model is correct ($\mu = X\beta^0$). Then, on the event $\{\hat{E} = E, \hat{E} \supset S\}$, which we assume, the residual

$$(I - P_E) y$$

is a (multivariate) truncated Gaussian with mean $\mathbf{0}$, with law

$$\mathbb{P}_{C,\sigma^2}(B) = \mathbb{P}(Z \in B | Z \in C), \qquad Z \sim N(\mathbf{0}, \sigma^2 I).$$

As σ^2 varies, one obtains a one-parameter exponential family with density

$$\frac{d\mathbb{P}_{C,\sigma^2}}{dz} = e^{-\alpha \|z\|_2^2 - \Lambda_C(\alpha)} 1_C(z)$$

and natural parameter $\alpha = \sigma^2/2$. On the event $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$, we set

$$C = \{y : A(E, z_E) y \leq b(E, z_E)\},$$

and then choose α (or equivalently, σ^2) to satisfy the score equation

$$\mathbb{E}_{C,\hat{\sigma}^2}(\|Z\|_2^2) = \|(I - P_E) y\|_2^2. \tag{8.5}$$

This amounts to a maximum likelihood estimate of σ^2. The expectation on the left is generally impossible to do analytically, but there exist fast algorithms for sampling from $\mathbb{P}_{C,\sigma^2}$, c.f. Geweke (1991); Rodriguez-Yam et al. (2004). A rough outline of a naive version of such algorithms is to pick a direction such as e_i one of the coordinate axes. Based on the current state of Z, draw a new entry for the Z_i from the appropriate univariate truncated normal determined from the cutoffs described in Section 4. We repeat this procedure to evaluate the expectation on the left, and use gradient descent to find $\hat{\sigma}^2$.

8.5. *Composite Null Hypotheses.* In Section 5, we considered hypotheses of the form $H_0 : \eta_{j\star}^T \mu = 0$, which said that the partial correlation of the variables in $-E$ with y, adjusting for the variables in E, was exactly 0. This may be unrealistic, and in practice, we may want to allow some tolerance for the partial correlation.

We consider testing instead the *composite* hypothesis

$$H_0 : |\eta_{j\star}^T \mu| \leq \delta_0. \tag{8.6}$$

The following result characterizes a test for H_0.

PROPOSITION 8.1. *The test which rejects when* $F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\delta_0,\ \sigma^2||\eta_{j\star}||^2}(\eta^T y) > 1-\alpha$ *is exact level* α.

PROOF. Let $\delta := \eta_{j\star}^T \mu$. Define $T_{\delta_0} := \inf_{|\delta|\leq\delta_0} F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\delta,\ \sigma^2||\eta_{j\star}||^2}(\eta^T y)$. Then:

$$\begin{aligned}
\text{Type I error} &:= \sup_{|\delta|\leq\delta_0} \mathbb{P}_\delta(T_{\delta_0} > 1-\alpha) \\
&\leq \sup_{|\delta|\leq\delta_0} \mathbb{P}_\delta\left(F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\delta,\ \sigma^2||\eta_{j\star}||^2}(\eta^T y) > 1-\alpha\right) \\
&= \alpha
\end{aligned}$$

Next, we have that $T_{\delta_0} = F^{[\mathcal{V}^-,\mathcal{V}^+]}_{\delta_0,\ \sigma^2||\eta_{j\star}||^2}(\eta^T y)$, i.e., the infimum is achieved at $\delta = \delta_0$, so calculating T_{δ_0} is a simple matter of evaluating F_{δ_0}. This follows from the fact that F_δ is monotone decreasing in δ (c.f. Appendix A).

Finally, the Type I error is exactly α because the reverse inequality also holds:

$$\text{Type I error} \geq \mathbb{P}_{\delta_0}(T_{\delta_0} > 1-\alpha) = \alpha.$$

□

Although the test is exact level α, the significance level of a test for a composite null is a "worst-case" Type I error; for most values of μ such that $|\eta^T\mu| \leq \delta_0$, the Type I error will be less than α, so the test will be conservative. Of course, what we lose in power, we gain in robustness to the assumption that $\eta^T\mu = 0$ exactly.

8.6. *How long a lasso should you use?.* Procedures for fitting the lasso, such as `glmnet` (Friedman et al., 2010), solve (1.2) for a decreasing sequence of λ values starting from $\lambda_1 = \|X^T y\|_\infty$. The framework developed so far provides a means to decide when to stop along the regularization path, i.e.,

when the lasso has done enough "fitting." In this section, we describe a path-wise testing procedure for the lasso,

The path-wise procedure is simple. At each value of λ:

1. Solve the lasso and obtain an equicorrelation set $\hat{E}_\lambda$ and signs $\hat{z}_{\hat{E}_\lambda}$.
2. Test $H_{0,\lambda} : X^T_{\hat{E}_\lambda}(I - P_{\hat{E}_\lambda})(\mu) = 0$ at level α. Rather than being conditional on only $(\hat{E}_\lambda, \hat{z}_{\hat{E}_\lambda})$, this test is conditional on the entire sequence of equicorrelation sets and signs $\{(\hat{E}^m, \hat{z}^m) = (E^m, z^m)\}$, as we describe below.

As λ decreases, we expect to reject the null hypotheses as the fit improves and stop once the first null hypothesis has been accepted.

To understand the properties of this procedure, we formalize it as a multiple testing problem. For each value $\lambda_1, ..., \lambda_m$, we test H_{0,λ_i}. We test these hypotheses sequentially and stop after the first hypothesis has been accepted. Implicitly, this means that we accept all the remaining hypotheses.

Our next result shows that this procedure controls the family-wise error rate (FWER) at level α. Let V denote that number of false rejections. Then FWER is defined as $\mathbb{P}(V \geq 1)$. The practical implication of this result is the model selected by this procedure will be larger than the true model with probability α.

PROPOSITION 8.2. *The path-wise testing procedure controls FWER at level α.*

PROOF. Let $\hat{E}^m$ and $\hat{z}^m$ denote the complete sequence of equicorrelation sets and signs at $\lambda_1, \ldots, \lambda_m$, i.e.,

$$\hat{E}^m = \{\hat{E}_{\lambda_1}, \ldots, \hat{E}_{\lambda_m}\}$$
$$\hat{z}^m = \{\hat{z}_{\hat{E}_{\lambda_1}}, \ldots, \hat{z}_{\hat{E}_{\lambda_m}}\}.$$

We seek to control the family-wise error rate (FWER) when testing the hypotheses $H_{0,\lambda_1}, \ldots, H_{0,\lambda_m}$, i.e., $\mathbb{P}(V \geq 1)$. We partition the space over all possible sequences $\hat{E}^m$ and $\hat{z}^m$:

$$\mathbb{P}(V \geq 1) = \sum_{(E^m, z^m)} \mathbb{P}\left(V \geq 1 \;\middle|\; (\hat{E}^m, \hat{z}^m) = (E^m, z^m)\right) \mathbb{P}\left((\hat{E}^m, \hat{z}^m) = (E^m, z^m)\right).$$

Since $\sum_{(E^m, z^m)} \mathbb{P}\left((\hat{E}^m, \hat{z}^m) = (E^m, z^m)\right) = 1$, we can ensure FWER $\leq \alpha$ by ensuring

$$\mathbb{P}\left(V \geq 1 \;\middle|\; (\hat{E}^m, \hat{z}^m) = (E^m, z^m)\right) \leq \alpha \text{ for any } (E^m, z^m).$$

Let λ_k denote the first λ_i for which H_{0,λ_i} is true. Then the event $V \geq 1$ is equivalent to the event that we reject H_{0,λ_k} because the preceding hypotheses $H_{0,\lambda_1}, \ldots, H_{0,\lambda_{k-1}}$ are all false so we cannot make a false discovery before the k^{th} hypothesis. Thus

$$\mathbb{P}\left(V \geq 1 \mid (\hat{E}^m, \hat{z}^m) = (E^m, z^m)\right) = \mathbb{P}\left(\text{reject } H_{0,\lambda_k} \mid (\hat{E}^m, \hat{z}^m) = (E^m, z^m)\right).$$

Therefore, we can control FWER at level α by ensuring

$$\mathbb{P}\left(\text{reject } H_{0,\lambda} \mid (\hat{E}^m, \hat{z}^m) = (E^m, z^m)\right) \leq \alpha$$

for each $\lambda \in \{\lambda_1, \ldots, \lambda_k\}$. □

To perform a test of $H_{0,\lambda}$ conditioned on $\{(\hat{E}^m, \hat{z}^m) = (E^m, z^m)\}$, we apply the framework of Section 4. Let

$$\{A(E_i, s_i)y < b(E_i, s_i)\}$$

be the affine constraints that characterize the event $\{(\hat{E}_{\lambda_i}, \hat{z}_{\lambda_i}) = (E_i, z_i)\}$ from Proposition 3.2. The event $\{(\hat{E}^m, \hat{z}^m) = (E^m, z^m)\}$ is equivalent to the intersection of all of these constraints:

$$\underbrace{\begin{bmatrix} A(E_1, z_1) \\ \vdots \\ A(E_m, z_m) \end{bmatrix}}_{A(E^m, z^m)} y < \underbrace{\begin{bmatrix} b(E_1, z_1) \\ \vdots \\ b(E_m, z_m) \end{bmatrix}}_{b(E^m, z^m)}.$$

Now Theorem 4.2 applies, and we can obtain the usual pivot as a test statistic.

9. Conclusion. We have described a method for making inference about $\eta^T \mu$ in the linear model based on the lasso estimator, where η is chosen adaptively after model selection. The confidence intervals and tests that we propose are conditional on $\{(\hat{E}, \hat{z}_{\hat{E}}) = (E, z_E)\}$. In contrast to existing procedures on inference for the lasso, we provide a pivot whose conditional distribution can be characterized exactly (non-asymptotically). This pivot can be used to derive confidence intervals and hypothesis tests based on lasso estimates anywhere along the solution path, not necessarily just at the knots of the LARS path as in Lockhart et al. (2014). Finally, our test is computationally simple: the quantities required to form the test statistic are readily available from the solution of the lasso.

Acknowledgements. 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 (4.5). 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

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

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.

Bühlmann, P. L. and van de Geer, S. A. (2011). *Statistics for High-dimensional Data*. Springer.

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

Fan, J., Guo, S. and Hao, N. (2012). Variance estimation using refitted cross-validation in ultrahigh dimensional regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, **74** 37–65.

Fisher, R. (1956). On a test of significance in pearson's biometrika tables (no. 11). *Journal of the Royal Statistical Society: Series B (Methodological)*, **18** pp. 56–60.

Friedman, J., Hastie, T. and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, **33** 1.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints. In *Computer Sciences and Statistics Proceedings of the 23d Symposium on the Interface*. Defense Technical Information Center, 571–578.

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

Knight, K. and Fu, W. (2000). Asymptotics for lasso-type estimators. *Annals of Statistics*.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, **21** pp. 21–59.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Annals of Statistics*, **34** 2554–2591.

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

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Methodological)*, **72** 417–473.

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.

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

Reid, S., Tibshirani, R. and Friedman, J. (2013). A study of error variance estimation in lasso regression. *Preprint*.

Robinson, G. K. (1979). Conditional properties of statistical procedures. *Annals of Statistics*.

Rodriguez-Yam, G., Davis, R. A. and Scharf, L. L. (2004). Efficient gibbs sampling of truncated multivariate normal with application to constrained linear regression. Tech. rep., Department of Statistics, Colorado State University. URL http://www.stat.columbia.edu/~rdavis/papers/CLR.pdf.

Taylor, J., Loftus, J. and Tibshirani, R. J. (2013). Tests in adaptive regression via the kac-rice formula. ArXiv:1308.3020. Submitted.

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*.

Wainwright, M. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, **55** 2183–2202.

Wasserman, L. (2014). Assumption-free high-dimensional inference. *Annals of Statistics*.

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.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, **7** 2541–2563.

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