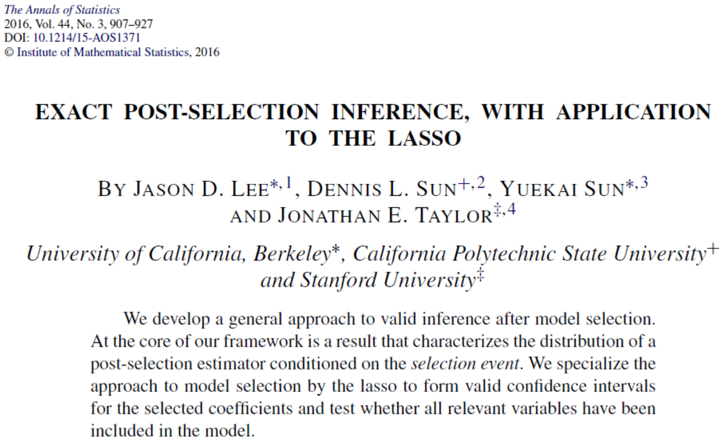
Lee, Sun, Sun and Taylor (2016, Annals)

Student Presentation in Master's Thesis Workshop 1

Yasuyuki Matsumura (Kyoto University)

April 30th, 2025

https://yasu0704xx.github.io



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.

Contents

Introduction

Post-Selection

- High-Dimensional Data: In empirical studies, researchers start with a large pool of candidate variables, and they do not know a priori which are relevant. This is especially problematic when there are more variables than observations, since then the model cannot be identified.
- Post-Selection: In such settings, researchers often let the data decide which variables to include in the model.
  - Fit a linear model with all variables included, observe which ones are significant at level α, and then refit the linear model with only those variables included.
  - Information Criterion (AIC, BIC), Regularization (LASSO) etc.

Inference after Selection (1)

- However, a problem is arising with such post-selection: 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.

Inference after Selection (2)

- Let us formalize the problem: Consider the standard linear regression setup, where the response $\boldsymbol{y} \in \mathbb{R}^n$ is generated from a multivariate normal distribution:

$$\boldsymbol{y} \sim \text{Normal}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n), \tag{1.1}$$

where $\boldsymbol{\mu}$ is modeled as a linear function of predictors $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_p \in \mathbb{R}^n$, and σ^2 is assumed known.[1]

[1] We consider the more realistic case where σ^2 is unknown in Section 8.1.

Inference after Selection (3)

- We choose a subset $M \subset 1, \cdots, p$ and ask for the linear combination of the predictors in M that minimizes the expected error.
- That is,

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

  where $\boldsymbol{X}_M^+ \equiv \left(\boldsymbol{X}_M^T \boldsymbol{X}_M\right)^{-1} \boldsymbol{X}_M^T$ is the pseudo-inverse of $\boldsymbol{X}_M$.
- Note that (1.2) implies that the targets β_j^M and $\beta_j^{M'}$ in different models ($M \neq M'$) are in general different.
- This is simply a restatement of the well-known fact: 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.

Inference after Selection (4)

- Thus, “inference after selection” is ambiguous in linear regression because the target of inference changes with the selected model. [2]

[2] Berk, Brown, Buja, Zhang, and Zhao (2013) [4]

Post-Selection Inference in Linear Regression

Post-Selection Inference in Linear Regression

- The target of inference changes with the selected model. Note that the rondomness is actually in the choice of which parameters to consider (not in the parameters themselves).
- Consider that there are *a priori* $p^{2^{p-1}}$ well-defined population parameters, one for each coefficient in all 2^p possible models:

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

  We only ever form inferences for the selected parameters $\beta_j^{\hat{M}}$ in the selected model $\hat{M}$.

- This adaptive choice can lead undesirable frequency properties. [3]

[3] Benjamini and Yekutieli (2005) [3], Benjamini, Heller and Yekutieli (2009) [2]

Confidence Interval

- We want a confidence interval $C_j^{\hat{M}}$ for a parameter $\beta_j^{\hat{M}}$. We hope $C_j^{\hat{M}}$ has the following property:

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

- However, the event $\beta_j^{\hat{M}} \in C_j^{\hat{M}}$ is not well-defined because β_j^M is undefined when $j \notin M$.
- Berk et al. (2013) [4] suggest two ways around this issue: conditional coverage and simultaneous coverage.

Conditional Coverage

- We form an interval for β_j^M <u>if and only if</u> model M is selected. That is, it makes sense to condition on this event.
- Hence, we might require that the confidence interval C_j^M satisfy

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

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

Conditional Coverage: Data Spliting

- Another way to understand conditioning on the model is to consider data spliting, [4] an approach to post-selection inference that most statisticians would agree is valid.
- In data spliting, the data is divided into two halves, with one half used to select the model and the other used to conduct inference.
- Inferences obtained by data spliting are only valid conditional on the model that was selected on the first half of the data. [5]
- Therefore, conditional coverage is a reasonable frequency property to require of a post-selection confidence interval.

[4] Cox (1975) [5]

[5] Fithian, Sun and Taylor (2014) [8]

Simultaneous Coverage

- It also makes sense to talk about events that are defined simultaneously over all $j \in \hat{M}$.
- Berk et al. (2013) [4] propose controlling the familywise error rate

$$\text{FWER} = \mathbb{P}\left(\beta_j^{\hat{M}} \notin C_j^{\hat{M}} \text{ for any } j \in \hat{M}\right), \qquad (2.2)$$

but this is very stringent when many predictors are involved.

Simultaneous Coverage: FCR & pFCR

- It is restrictive to control FWER (the probability of making *any* error). Instead, 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) [3] simply declare the error to be zero when $|M| = 0$:

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

- Storey (2003) [22] suggests conditioning on $|\hat{M}| > 0$:

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

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

Revisiting Conditional Coverage

- The two ways above (conditional coverage and simultaneous coverage) are related: Conditional coverage (2.1) implies pFCR (2.4) control (and hence, FCR (2.3) control).

Lemma 2.1

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

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

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

Proof of Lemma 2.1

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

$$\begin{aligned}
\text{pFCR} &= \mathbb{E}\left[\mathbb{E}\left[\frac{|j \in \hat{M} \ : \ \beta_j^{\hat{M}} \notin C_j^{\hat{M}}|}{|\hat{M}|}\middle| |\hat{M}| > 0\right]\right] \\
&= \mathbb{E}\left[\frac{\sum_{j \in \hat{M}} \mathbb{P}\left(\beta_j^{\hat{M}} \notin C_j^{\hat{M}} \middle| \hat{M}\right)}{|\hat{M}|}\middle| |\hat{M}| > 0\right] \\
&\cdots\cdots \text{by } (2.1) \cdots\cdots \\
&\leq \mathbb{E}\left[\frac{\alpha|\hat{M}|}{|\hat{M}|}\middle| |\hat{M}| > 0\right] \\
&= \alpha.
\end{aligned}$$

□

Around Lemma 2.1

- Theorem 2 in Weinstein, Fithian and Benjamini (2013) [29] proves a special case of Lemma 2.1 for a particular selection procedure.
- Proposition 11 in Fithian, Sun and Taylor (2014) [8] provides a more general resuly.
- However, the result of Lemma 2.1 is sufficient to establish that conditional coverage is a sensible criterion to consider in post-selection inference.

How to Construct CI?

- To construct an interval with an interval with conditional coverage, we need to understand the conditional distribution

$$\boldsymbol{y} \mid \left\{\hat{M}(\boldsymbol{y}) = M\right\}, \quad \boldsymbol{y} \sim \mathrm{Normal}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}).$$

- One of the main contributions of Lee et al. (2016) [1] is to show that this distribution is indeed possible to characterize, making valid post-selection inference feasible in the context of linear regression.

Outline

Post-Selection Intervals with Conditional Coverage

- We have argued that post-selection intervals for regression coefficients should have $1 - \alpha$ coverage conditional on the selected model:

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

- To obtain an interval C_j^M with this property, we study the conditional distribution

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

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

Outline (1)

- Lee et al. (2016) [1] focus on the specific case where 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 $\{\hat{M} = M\}$, that specifies the model and the signs of the selected variables, is a polyhedron of the form

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

Outline (2)

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

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

  We do this in Section 5.

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

Related Work (1)

- The resulting post-selection test has a similar structure to the pathwise significance tests of Lockhart et al. (2014) [14] and Taylor et al. (2014) [23], which also are conditional tests.
- 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 λ or form confidence intervals for an individual coeeficient in the model.

Related Work (2)

- Javanmard and Montanari (2013) [9], van de Geer et al. (2013) [28], Zhang and Zhang (2014) [30] are also parallel literatures on confidence intervals for coefficients in high dimensional linear models based on the lasso estimator.
    - 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 coefficients $\boldsymbol{\beta}^0$ and $\boldsymbol{\beta}^{\hat{M}}$ will not be the same unless $\hat{M}$ happens to contain all nonzero coefficients of $\boldsymbol{\beta}^0$.
    - Inference for $\boldsymbol{\beta}^{\hat{M}}$ requires assumptions about correctness of the linear model and sparsity of $\boldsymbol{\beta}^0$.
- Potscher and Schneider (2010) [18] consider confidence intervals for the hard-thresholding and soft-thresholding estimators in the case of orthogonal design.
- Instead, our approach regards the selected model as a linear approximation to the truth, a view shared by Berk et al. (2013) [4] and Miller (2002) [15].

Related Work (3)

- The idea of post-selection inference conditional on the selected model appears in Potscher (1991) [17], although the notion of inference conditional on certain relevant subsets dates back to Fisher (1956) [7]. See also Robinson (1979) [19].
- Leeb and Potscher (2005, 2006) [11], [12] 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.
- To the contrary, we show that conditioning on the selected model can produce reasonable confidence intervals in a wide variety of situations.

Related Work (4)

- Inference conditional on selection has also apppeared in literature on the *winner's curse*: Sampson and Sill (2005) [20], Sill and Sampson (2009) [21], Zhong and Prentice (2008) [31], Zollner and Pritchard (2007). These works are not really associated with model selection in linear regression, though they employ a similar approach to inference.

The Lasso and its Selection Event

Lasso

- We apply our post-selection inference procedure to the model selected by the lasso (Tibshirani, 1996 [25]).
- The lasso estimate is the solution to the usual least squares problem with an additional ℓ_1 penalty on the coefficients:

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

- The ℓ_1 penalty shrinks many of the coefficients to exactly zero. 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 $\hat{\boldsymbol{\beta}}$ is known only when $n >> p$ (Knight and Fu, 2000 [10]), and even then, only asymptotically.

Model-Selection by Lasso

- Since the lasso produces sparse solutions, we can define model selected by the lasso to be the set of predictors with nonzero coefficients:

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

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

KKT Conditions

- Let us focus on characterizing the event $\{\hat{M} = M\}$.
- Let $\hat{\boldsymbol{s}}$ be a vector of signs of $\hat{\boldsymbol{\beta}}$.
- Since $\hat{\boldsymbol{\beta}}$ is the solution to the lasso problem (4.1), it is necessary and sufficient that they satisfy the KKT conditions:

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

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

Equicorrelation Set

- Following Tibshirani (2013) [26], we consider the equicorrelation set

$$\hat{M} = \{i \in \{1, \cdots, p\} \, : \, |\hat{s}_j| = 1\}. \tag{4.4}$$

- 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 contains all predictors with nonzero coefficients, although it may also include some predictors with zero coefficients.
- However, for almost every λ, the equicorrelation set is precisely the set of predictors with nonzero coefficients.
- Then, it is easier to first characterize $\{(\hat{M}, \hat{\boldsymbol{s}}) = (M, \boldsymbol{s})\}$ and obtain $\{\hat{M} = M\}$ as a corollary by taking a union over the possible signs.

Lemma 4.1

Lemma 4.1

- Assume the columns of $\boldsymbol{X}$ are in general posotion (Tibshirani, 2013 [26]).
- Let $M \subset \{1, \cdots, p\}$ and $\boldsymbol{s} \in \{-1, 1\}^{|M|}$ be a candidate set of variables and their signs, respectively.
- Define the random variables

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

$$\boldsymbol{u}(M, \boldsymbol{s}) := \boldsymbol{X}^T_{-M}(\boldsymbol{X}_M)^{+}\boldsymbol{s} + \frac{1}{\lambda}\boldsymbol{X}^T_{-M}(1 - \boldsymbol{P}_M)\boldsymbol{y}, \tag{4.6}$$

where $\boldsymbol{P}_M \equiv \boldsymbol{X}_M(\boldsymbol{X}_M \boldsymbol{X}_M)^{-1}\boldsymbol{X}_M$ is projection onto the column span of $\boldsymbol{X}_M$.

Lemma 4.1 (cont'd)

- Then, the selection procedure can be rewritten in terms of $\boldsymbol{w}$ and $\boldsymbol{u}$ as

$$\begin{aligned}&\{(\hat{M}, \hat{\boldsymbol{s}}) = (M, \boldsymbol{s})\}\\ &\qquad = \{\text{sign}\left(\boldsymbol{w}(M, \boldsymbol{s})\right) = \boldsymbol{s},\ ||\boldsymbol{u}(M, \boldsymbol{s})||_\infty < 1\}. \quad (4.7)\end{aligned}$$

Proof of Lemma 4.1

- 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 $\hat{M}$":

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

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

$$\begin{aligned} \boldsymbol{X}_M^T(\boldsymbol{X}_M\boldsymbol{w} - \boldsymbol{y}) + \lambda\boldsymbol{s} &= 0, \\ \boldsymbol{X}_{-M}^T(\boldsymbol{X}_M\boldsymbol{w} - \boldsymbol{y}) + \lambda\boldsymbol{u} &= 0, \\ \text{sign}(\boldsymbol{w}) = \boldsymbol{s}, \quad ||\boldsymbol{u}||_\infty &< 1. \end{aligned}$$

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

$$\begin{aligned} \boldsymbol{w} &= (\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1}(\boldsymbol{X}_M^T\boldsymbol{y} - \lambda\boldsymbol{s}), \\ \boldsymbol{u} &= \boldsymbol{X}_{-M}^T(\boldsymbol{X}_M^T)^+\boldsymbol{s} + \frac{1}{\lambda}\boldsymbol{X}_{-M}^T(1 - \boldsymbol{P}_M)\boldsymbol{y}, \\ \text{sign}(\boldsymbol{w}) = \boldsymbol{s}, &\quad ||\boldsymbol{u}||_{-\infty} < 1, \end{aligned}$$

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

Proposition 4.2

- Lemma 4.1 says that the event $\{(\hat{M}, \hat{\boldsymbol{s}}) = (M, \boldsymbol{s})\}$ can be rewritten as affine constraints on $\boldsymbol{y}$.
- This is because $\boldsymbol{w}$ and $\boldsymbol{u}$ are already affine functions of $\boldsymbol{y}$, and the constraints sign$(\cdot) = \boldsymbol{s}$ and $||\cdot||_\infty < 1$ can also be rewritten in terms of affine constraints.
- Proposition 4.2 makes this explicit.

Proposition 4.2

- Let $\boldsymbol{w}$ and $\boldsymbol{u}$ be defined as in (4.5) and (4.6).
- Then,

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

where $A_0, \boldsymbol{b}_0$ encode the inactive constraints $\{||\boldsymbol{u}||_\infty < 1\}$, and $A_1, \boldsymbol{b}_1$ encode the active constraints $\{\text{sign}(\boldsymbol{w}) = \boldsymbol{s}\}$.

Proposition 4.2 (cont'd)

- These matrices have the explicit forms

$$A_0(M, \boldsymbol{s}) = \frac{1}{\lambda} \begin{pmatrix} \boldsymbol{X}_{-M}^T (1 - \boldsymbol{P}_M) \\ -\boldsymbol{X}_{-M}^T (1 - \boldsymbol{P}_M) \end{pmatrix},$$

$$\boldsymbol{b}_0(M, \boldsymbol{s}) = \begin{pmatrix} \boldsymbol{1} - \boldsymbol{X}_{-M}^T (\boldsymbol{X}_M^T)^+ \boldsymbol{s} \\ \boldsymbol{1} + \boldsymbol{X}_{-M}^T (\boldsymbol{X}_M^T)^+ \boldsymbol{s} \end{pmatrix},$$

$$A_1(M, \boldsymbol{s}) = -\mathsf{diag}(\boldsymbol{s})(\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^T,$$

$$\boldsymbol{b}_1(M, \boldsymbol{s}) = -\lambda \mathsf{diag}(\boldsymbol{s})(\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{s}.$$

Proof of Proposition 4.2

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

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

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

$$\begin{aligned}\{||\boldsymbol{u}||_\infty < 1\} &= \{-\mathbf{1} < \boldsymbol{X}^T_{-M}(\boldsymbol{X}_M)^+\boldsymbol{s} + \frac{1}{\lambda}\boldsymbol{X}^T_{-M}(1 - \boldsymbol{P}_M)\boldsymbol{y} < \mathbf{1}\} \\ &= \{A_0(M,\boldsymbol{s})\boldsymbol{y} < \boldsymbol{b}_0(M,\boldsymbol{s})\}.\end{aligned}$$

□

Characterization of $\{\hat{M} = M, \hat{\boldsymbol{s}} = s\}$

- Combining Lemma 4.1 and Proposition 4.2, we obtain the following.

Proposition 4.3

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

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

Characterization of $\{\hat{M} = M\}$

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

Corollary 4.4

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

Illustration of Theorem 4.3 & Corollary 4.4

- Fig. 1 illustrates the lasso partitions of $\mathbb{R}^n$ into polyhedra according to the selected model and signs of the coefficients (Suppose $p = 2, n = 3$).
- The shaded area corresponds to the event $\{\hat{M} = \{1, 3\}\}$, which is a union of two polyherda.
- Note that the signs patterns $\{+, -\}$ and $\{-, +\}$ are not possible for the model $\{1, 3\}$ (Such polyherda do not exist).

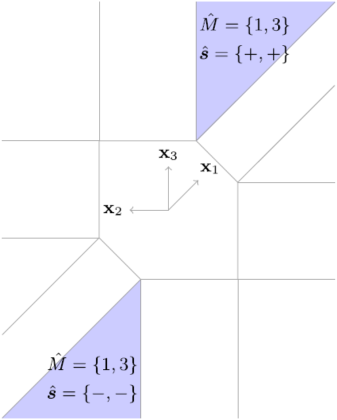


FIG. 1. *A geometric picture illustrating Theorem 4.3 for $n = 2$ and $p = 3$. The lasso partitions $\mathbb{R}^n$ into polyhedra according to the selected model and signs.*

Polyhedral Conditioning Sets

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

- In order to obtain inference conditional on the model, we need to understand the distribution of

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

- As we saw in the previous section, $\{\hat{M} = M\}$ is a union of polyhedra, which makes it easier to condition on both the model and the signs,

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

  since the conditioning event is a single polyhedron (Theorem 4.3):

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

- Note that inferences that are valid conditional on this finer event will also be valid conditional on $\{\hat{M} = M\}$.
- Indeed, if a confidence interval C_j^M for β_j^M has $(1-\alpha)$ coverage conditional on the model and signs:

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

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

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

- This section is devided into two subsections:
- First, we study how to condition on a single polyhedron. This will allow us to condition on $\{\hat{M} = M, \hat{\boldsymbol{s}} = \boldsymbol{s}\}$.
- 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 in general be more efficient (i.e., narrower intervals, more powerful tests), at the price of more computation.

Conditioning on a Single Polyhedron

- Suppose that we observe $\boldsymbol{y} \sim \text{Normal}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\eta} \in \mathbb{R}^n$ is some direction of interest.
- To understand the distribution of

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

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

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

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

- It is easy to verify that $\boldsymbol{z}$ is uncorrelated with, and hence independent of, $\boldsymbol{\eta}^T \boldsymbol{y}$. Note that, in the case where $\Sigma = \sigma^2 \boldsymbol{I}_n$, $\boldsymbol{z}$ is simply the residual $(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{\eta}})\boldsymbol{y}$ from projecting $\boldsymbol{y}$ onto $\boldsymbol{\eta}$.
- Now, we can rewrite $\{A\boldsymbol{y} \leq \boldsymbol{b}\}$ in terms of $\boldsymbol{\eta}^T \boldsymbol{y}$ and $\boldsymbol{z}$.

Lemma 5.1

Lemma 5.1

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

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

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

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

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

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

Proof of Lemma 5.1

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

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

where in the last step we have devided the components into three categories depending on whether $(A\boldsymbol{c})_j > 0$, < 0, or $= 0$, since this affects the direction of the inequality (or whether we can divide at all).

- Since $\boldsymbol{\eta}^T \boldsymbol{y}$ is the same quantity for all j, it must be at least the maximum of the lower bounds $\mathcal{V}^-(\boldsymbol{z})$, and no more than the minimum of the upper bounds $\mathcal{V}^+(\boldsymbol{z})$.

$\square$

- Lemma 5.1 tells us that

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

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

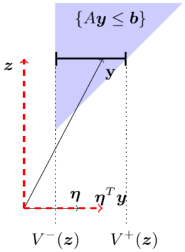


FIG. 2. *A geometric interpretation of why the event* $\{A\mathbf{y} \le \mathbf{b}\}$ *can be characterized as* $\{\mathcal{V}^-(\mathbf{z}) \le \boldsymbol{\eta}^T\mathbf{y} \le \mathcal{V}^+(\mathbf{z})\}$. *Assuming* $\Sigma = I$ *and* $\|\boldsymbol{\eta}\|_2 = 1$, $\mathcal{V}^-(\mathbf{z})$ *and*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

- We would like to be able to say

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

  but this is technically incorrect, since the distribution on the right-hand side changes with $\boldsymbol{z}$.
- By conditioning on the value of $\boldsymbol{z}$, $\boldsymbol{\eta}^T\boldsymbol{y}|\{A\boldsymbol{y} \le \boldsymbol{b}, \boldsymbol{z} = \boldsymbol{z}_0\}$ is a truncated normal.
- We can then use the probability integral transform to obtain a statistic $\boldsymbol{F}^{\boldsymbol{z}}(\boldsymbol{\eta}^T\boldsymbol{y})$ that has a Unif$(0, 1)$ distribution for any value of $\boldsymbol{z}$.
- Hence, $\boldsymbol{F}^{\boldsymbol{z}}(\boldsymbol{\eta}^T\boldsymbol{y})$ will also have a Unif$(0, 1)$ distribution marginally over $\boldsymbol{z}$.
- We make this precise in Theorem 5.2.

Theorem 5.2

Theorem 5.2

- Let $\boldsymbol{F}_{\mu,\sigma^2}^{[a,b]}$ denote the CDF of a Normal(μ, σ^2) random variable truncated to the interval $[a, b]$, that is,

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

  where Ω is the CDF of a Normal$(0, 1)$ random variable.

- Then,

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

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

Theorem 5.2 (cont'd)

- Furthermore,

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

Proof of Theorem 5.2

- First, apply Lemma 5.1:

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

- The only random quantities left are $\boldsymbol{\eta}^T \boldsymbol{y}$ and $\boldsymbol{z}$.

- Now we can eliminate $\boldsymbol{z} = \boldsymbol{z}_0$ from the condition using independence:

$$\begin{aligned}
&[\boldsymbol{\eta}^T \boldsymbol{y} | A\boldsymbol{y} \leq \boldsymbol{b},\, \boldsymbol{z} = \boldsymbol{z}_0] \\
&\quad \overset{d}{=} [\boldsymbol{\eta}^T \boldsymbol{y} | \mathcal{V}^-(\boldsymbol{z}_0) \leq \boldsymbol{\eta}^T \boldsymbol{y} \leq \mathcal{V}^+(\boldsymbol{z}_0)] \\
&\quad \sim \text{Truncated Normal}(\boldsymbol{\eta}^T \boldsymbol{y},\, \sigma^2 ||\boldsymbol{\eta}||^2,\, \mathcal{V}^-(\boldsymbol{z}_0),\, \mathcal{V}^+(\boldsymbol{z}_0)).
\end{aligned}$$

- Let

$$\boldsymbol{F}^{\boldsymbol{z}}(\boldsymbol{\eta}^T \boldsymbol{y}) \equiv \boldsymbol{F}^{[\mathcal{V}^-(\boldsymbol{z}), \mathcal{V}^+(\boldsymbol{z})]}_{\boldsymbol{\eta}^T \boldsymbol{\mu},\, \boldsymbol{\eta}^T \Sigma \boldsymbol{\eta}}(\boldsymbol{\eta}^T \boldsymbol{y}).$$

- We can apply the probability integral transform to the above result to obtain

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

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

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

  for any $\boldsymbol{z}_0$.

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

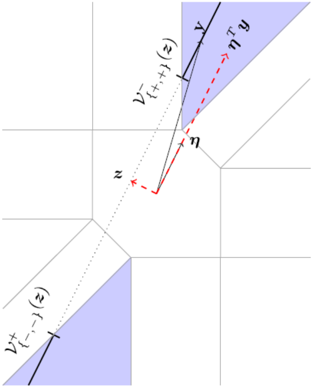
$$\begin{aligned} p_{\boldsymbol{F}^{\boldsymbol{z}}(\boldsymbol{\eta}^T\boldsymbol{y})|\boldsymbol{z}}(t) &= \int p_{\boldsymbol{F}^{\boldsymbol{z}}(\boldsymbol{\eta}^T\boldsymbol{y})|\boldsymbol{z}}(t|\boldsymbol{z}_0) p_{\boldsymbol{z}}(\boldsymbol{z}_0) d\boldsymbol{z}_0 \\ &= \int 1_{[0,1]}(t) p_{\boldsymbol{z}}(\boldsymbol{z}_0)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

□

Conditioning on a Union of Polyhedra

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

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



$\mathbf{y}$
$\boldsymbol{\eta}^T\mathbf{y}$
$\mathcal{V}^-_{\{+,+\}}(\mathbf{z})$
$\boldsymbol{\eta}$
$\mathbf{z}$
$\mathcal{V}^+_{\{-,-\}}(\mathbf{z})$

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* $(-\infty, \mathcal{V}^+_{\{-,-\}}(\mathbf{z})] \cup [\mathcal{V}^-_{\{+,+\}}(\mathbf{z}), \infty)$.

- As Figure 3 makes clear, the argumment proceeds exactly as before, except that $\boldsymbol{\eta}^T\boldsymbol{y}$ is now truncated to a union of intervals, instead of a single interval.
- There is a $\mathcal{V}^-$ and a $\mathcal{V}^+$ for each possible sign patterns $\boldsymbol{s}$, so we index the intervals by the signs.
- This leads immediately to Theorem 5.3.

Theorem 5.3

Theorem 5.3

- Let F^{S}_{μ,σ^2} denote the CDF of a Normal(μ, σ^2) random variable truncated to the set S.
- Then,

$$\boldsymbol{F}^{\cup_s [\mathcal{V}_s^-(\boldsymbol{z}), \mathcal{V}_s^+(\boldsymbol{z})]}_{\boldsymbol{\eta}^T \boldsymbol{\mu}, \boldsymbol{\eta}^T \Sigma \boldsymbol{\eta}}(\boldsymbol{\eta}^T \boldsymbol{y}) | \{A_s \boldsymbol{y} \le \boldsymbol{b}_s\} \sim \text{Uniform}(0,1), \tag{5.11}$$

where $\mathcal{V}_s^-(\boldsymbol{z})$ and $\mathcal{V}_s^+(\boldsymbol{z})$ are defined in (5.4) and (5.5) and $A = A_s$ and $b = b_s$.

Proof of Theorem 5.3

- The proof is essentially the same as that of Theorem 5.2.

Post-Selection Intervals for Regression Coefficients

Post-Selction Intervals for Lasso-Selected Coefficients (1)

- Here we combine the characterization of the lasso selection event (obtained in Section 4) with results about the distribution of a Gaussian truncated to a polyhedron or union of polyhedra (obtained 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:

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

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

Post-Selction Intervals for Lasso-Selected Coefficients (2)

- Recall that our goal is to form confidence intervals for $\beta_j^M = \boldsymbol{e}_j^T \boldsymbol{X}_M^+ \boldsymbol{\mu}$, with $(1-\alpha)$ coverage conditional on $\{\hat{M} = M\}$.
- Taking $\boldsymbol{\eta} = (\boldsymbol{X}_M^T \boldsymbol{e}_j)$, we can use Theorem 5.3 to obtain
$$\boldsymbol{F}_{\beta_j^M, \sigma^2 ||\boldsymbol{\eta}||^2}^{\cup_s [\mathcal{V}_s^-(\boldsymbol{z}), \mathcal{V}_s^+(\boldsymbol{z})]}(\boldsymbol{\eta}^T \boldsymbol{y}) | \{\hat{M} = M\} \sim \text{Uniform}(0,1).$$
- This gives us a test statistic for testing any hypothesized value of β_j^M.
- We can invert this test to obtain a confidence set
$$C_j^M \equiv \left\{ \beta_j^M \ : \ \frac{\alpha}{2} \leq \boldsymbol{F}_{\beta_j^M, \sigma^2 ||\boldsymbol{\eta}||^2}^{\cup_s [\mathcal{V}_s^-(\boldsymbol{z}), \mathcal{V}_s^+(\boldsymbol{z})]}(\boldsymbol{\eta}^T \boldsymbol{y}) \leq 1 - \frac{\alpha}{2} \right\}. \tag{6.1}$$
- In fact, the set C_j^M is an interval, as formalized in Theorem 6.1.

Theorem 6.1

Theorem 6.1

- Let $\boldsymbol{\eta} = (\boldsymbol{X}_M^+)^T \boldsymbol{e}_j$.
- Let L and U be the (unique) values satisfying

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

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

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

Proof of Theorem 6.1

- By construction,

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

  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 $\boldsymbol{F}_{L,\sigma^2||\boldsymbol{\eta}||^2}^{\cup_s[\mathcal{V}_s^-(\boldsymbol{z}),\mathcal{V}_s^+(\boldsymbol{z})]}(\boldsymbol{\eta}^T\boldsymbol{y})$ is monotone decreasing in β_j^M so that it crosses $1 - \alpha$ and α at unique values.
- This follows from the fact that the truncated Gaussian distribution has monotone likelihood ratio in the mean parameter.
- See Appendix A for details.

□

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 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

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

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

Conditioning on a Single Polyhedron will be Less Efficient

- 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

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

- Inverting this statistic will produce intervals that have $(1-\alpha)$ coverage conditional on $\{\hat{M} = M, \boldsymbol{s} \mathrel{\hat{=}} \boldsymbol{s}\}$, and hence $(1-\alpha)$ coverage conditional on $\{\hat{M} = M\}$.
- These intervals will be less efficient; They will in general be wider.
- One may be willing to sacrifice statistical efficiency for computational efficiency.

Computational Cost

- The main cost in computing intervals according to Theorem 6.1 is determining the intervals $[\mathcal{V}_{\boldsymbol{s}}^{-}(\boldsymbol{z}), \mathcal{V}_{\boldsymbol{s}}^{+}(\boldsymbol{z})]$ for each $\boldsymbol{s} \in \{-1, 1\}^{|M|}$.
- The number of such sign patterns is $2^{|M|}$.
- While this might be feasible when $|M|$ is small, it is not feasible when we select hundreds of variables.
- Conditioning on the signs means that we only have to compute intervals $[\mathcal{V}_{\boldsymbol{s}}^{-}(\boldsymbol{z}), \mathcal{V}_{\boldsymbol{s}}^{+}(\boldsymbol{z})]$ for the sign pattern $\boldsymbol{s}$ that was acutually observed.

Trade-Off in Statistical Efficiency

- Figure 4 shows the tradeoff in statistical efficiency.
- When the signal is strong, as in th eleft-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 acutual noise variables, as expected.

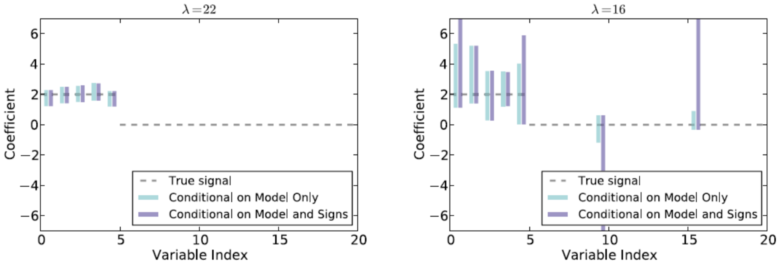


FIG. 4. *Comparison of the confidence intervals by conditioning on the model only (statistically more efficient, but computationally more expensive) and conditioning on both the model and signs (statistically less efficient, but computationally more feasible). Data were simulated for* $n = 25$, $p =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

Why are Post-Selection Intervals sometimes Wide?

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

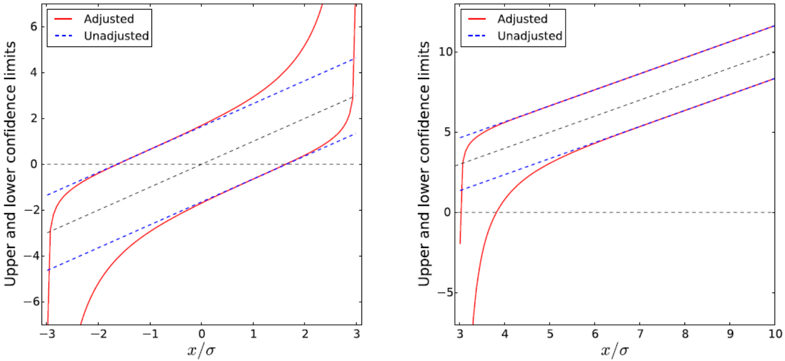


FIG. 5. *Upper and lower bounds of 90% confidence intervals for μ based on a single observation $x/\sigma \sim \mathrm{TN}(0, 1, -3, 3)$. We see that as long as the observation x is roughly 0.5σ away from either boundary, the size of the intervals is comparable to the unadjusted OLS confidence interval.*

Implications on Efficiency

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

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 tha same coverage, we generally prefer the shorter one.
- This problem is considered in Fithian, Sun and Taylor (2014) [8] where it is shown that C_j^M is, with one small tweak, the shortest interval among all unbiased intervals with $(1 - \alpha)$ coverage.

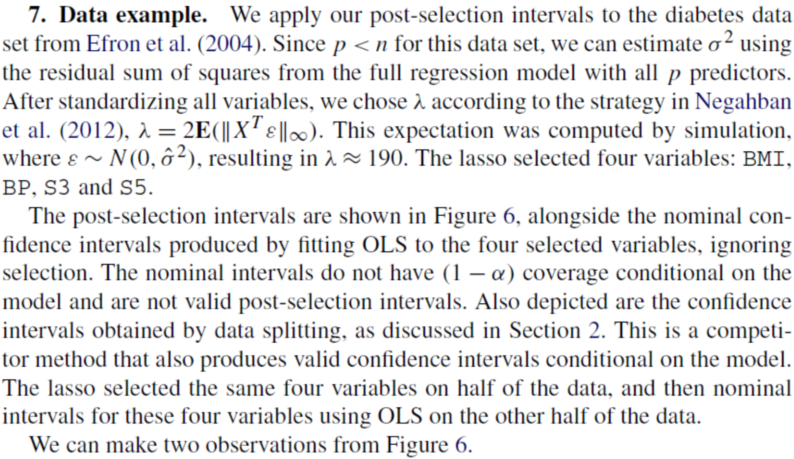
- Unbiasedness is a common restriction to ensure the existence of an optimal interval (Lehmann and Romano, 2005 [13]). An unbiased interval C for a parameter θ is one which cobers no other parameter θ' with probability more than $1-\alpha$:

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

- The shortest unbiased interval for β_j^M, among all intervals with conditional $1-\alpha$ coverage, resembles to the interval $[L, U]$ in Theorem 6.1. The critical values L and U were chosen symmetrically so that the pivot has $\frac{\alpha}{2}$ area in either tail.
- However, it may be possible to obtain a shorter intervals on average by allocating the probability unequally between the two tails. Theorem 5 of Fithian. Sun and Taylor (2014) [8] provides a general formula for obtaining shortest unbiased intervals in exponential families.

Data Example

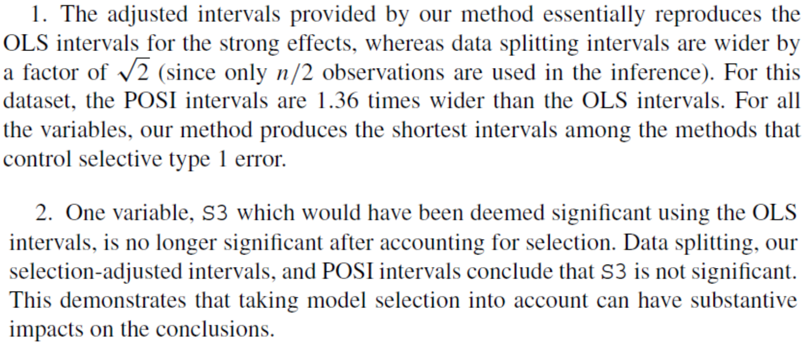
Simulation



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 σ^2 using the residual sum of squares from the full regression model with all p predictors. After standardizing all variables, we chose λ 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.

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.

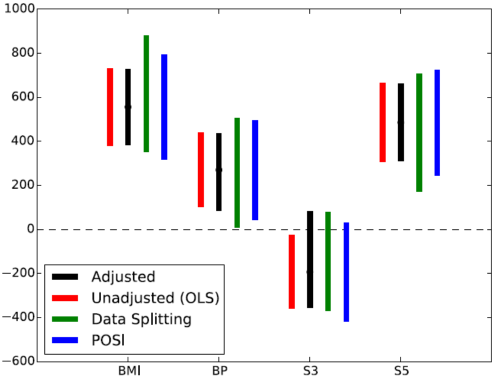
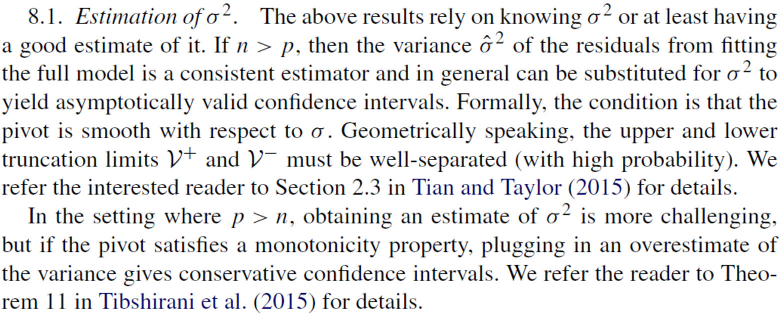


FIG. 6. *Inference for the four variables selected by the lasso* ($\lambda = 190$) *on the diabetes data set. The point estimate and adjusted confidence intervals using the approach in Section* 6 *are shown in black. The OLS intervals, which ignore selection, are shown in red. The green lines show the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

Extensions

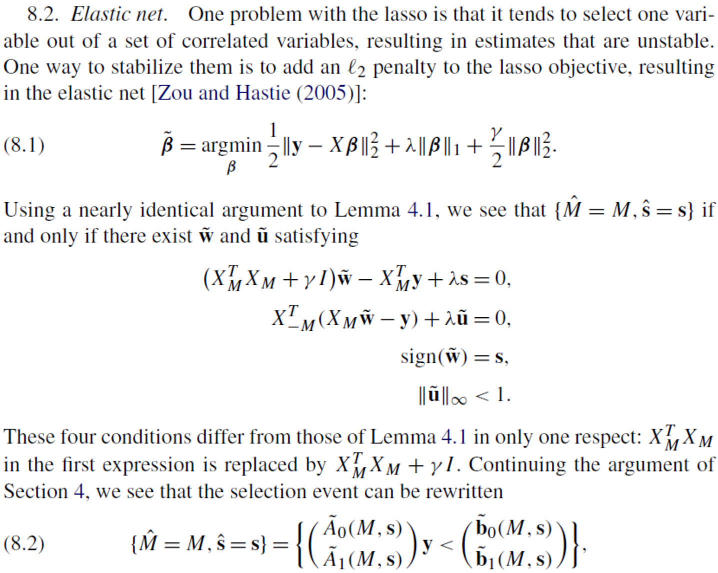
Estimation of σ^2



8.1. *Estimation of* σ^2. The above results rely on knowing σ^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 σ^2 to yield asymptotically valid confidence intervals. Formally, the condition is that the pivot is smooth with respect to σ. 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 σ^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.

Elastic Net



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 ℓ_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

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

Conclusion

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.

References

References i

- Lee, J. D., Sun, D. L., Sun Y. and Taylor J. E. (2016). Exact post-selection inference, with application to the lasso. *Annals of Statistics 44 (3)*, 907-927.
- Benjamini, Y., Heller, R. and Yekutieli, D. (2009). Selective inference in complex research. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **367**(1906), 4255-4271.
- Benjamini, Y. and Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, **100**(469), 71-93.

References ii

- Berk, R., Brown, L., Buja, A., Zhang, K. and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, **41**(2), 802-837.
- Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, **62**(2), 441-444.
- Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, **32**(2), 407-499.
- Fisher, R. A. (1956). On a test of significance in Pearson ’ s Biometrika Tables (No. 11). *Journal of the Royal Statistical Society. Series B (Methodological)*, **18**(1), 56-60.

References iii

- Fithian, W., Sun, D. and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint*, arXiv:1410.2597.
- 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. *The Annals of Statistics*, **28**(5), 1356-1378.
- Leeb, H. and Potscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory*, **21**(1), 21-59.

References iv

- Leeb, H. and Potscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, **34**(5), 2554-2591.
- Lehmann, E. L. and Romano, J. P. (2005). *Testing Statistical Hypotheses* (3rd ed.). Springer.
- Lockhart, R., Taylor, J., Tibshirani, R. J. and Tibshirani, R. (2014). A significance test for the lasso. *The Annals of Statistics*, **42**(2), 413-468.
- Miller, A. (2002). *Subset Selection in Regression* (2nd ed.). Chapman & Hall/CRC.

References v

- 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**(4), 538-557.
- Potscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, **7**(2), 163-185.
- Potscher, B. M. and Schneider, U. (2010). Confidence sets based on penalized maximum likelihood estimators in Gaussian regression. *Electronic Journal of Statistics*, **4**, 334-360.
- Robinson, G. K. (1979). Conditional properties of statistical procedures. *The Annals of Statistics*, **7**(4), 742-755.

References vi

- Sampson, A. R. and Sill, M. W. (2005). Drop-the-losers design: Normal case. *Biometrical Journal*, **47**(2), 257-268.
- Sill, M. W. and Sampson, A. R. (2009). Drop-the-losers design: Binomial case. *Computational Statistics & Data Analysis*, **53**(3), 586-595.
- Storey, J. D. (2003). The positive false discovery rate: A Bayesian interpretation and the q-value. *The Annals of Statistics*, **31**(6), 2013-2035.
- Taylor, J., Lockhart, R., Tibshirani, R. J. and Tibshirani, R. (2014). Post-selection adaptive inference for least angle regression and the lasso. *arXiv preprint*, arXiv:1401.3889.

References vii

- Tian, X. and Taylor, J. (2015). Asymptotics of selective inference. *arXiv preprint*, arXiv:1501.03588.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**(1), 267-288.
- Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electronic Journal of Statistics*, **7**, 1456-1490.
- Tibshirani, R. J., Rinaldo, A., Tibshirani, R. and Wasserman, L. (2015). Uniform asymptotic inference and the bootstrap after model selection. *arXiv preprint*, arXiv:1506.06266.

References viii

- van de Geer, S., Buhlmann, P., Ritov, Y. and Dezeure, R. (2013). On asymptotically optimal confidence regions and tests for high-dimensional models. *arXiv preprint*, arXiv:1303.0518.
- Weinstein, A., Fithian, W. and Benjamini, Y. (2013). Selection adjusted confidence intervals with more power to determine the sign. *Journal of the American Statistical Association*, **108**(501), 165-176.
- Zhang, C.-H. and Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(1), 217-242.

References ix

- Zhong, H. and Prentice, R. L. (2008). Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics*, **9**(4), 621-634.

- Zollner, S. and Pritchard, J. K. (2007). Overcoming the winner ’ s curse: Estimating penetrance parameters from case-control data. *American Journal of Human Genetics*, **80**(4), 605-615.

- Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**(2), 301-320.