Selective inference

Patrick Breheny

April 18

Introduction

- In our final lecture on inferential approaches for penalized regression, we will discuss two rather recent approaches:
    - The *covariance test* for testing the significance of additional terms along a covariate path
    - *Selective inference*, or *post-selection inference*, in which we carry out tests/construct confidence intervals conditional on the selected model

Motivation

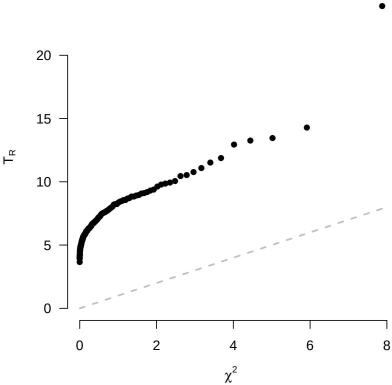
- The classical test for the significance of adding a variable to a model is to calculate the test statistic

$$T_R = \frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\sigma^2}$$

  and compare it to a χ_1^2 distribution (for simplicity, let's assume σ^2 is known)
- This is valid, of course, when the variable is prespecified
- In the model selection context, however, it is far too liberal, as we have seen

Failure of χ^2 result

$n = p = 100$, $\boldsymbol{\beta} = \mathbf{0}$, significance of first predictor added:



$\mathbb{P}(T > 3.84) > 0.99$

Covariance test: Motivation

- The goal of the covariance test is to develop a test statistic for an added variable whose distribution can be characterized despite the fact that we searched over a large number of candidate predictors to find it
- Let $\lambda_1 > \lambda_2 \cdots$ denote the values of the regularization parameter at which a new variable enters the model, and let $\mathcal{A}_{k-1}$ denote the active set, not including the variable added at step k
- Now, let us consider two solutions at the same value of λ:
  - $\widehat{\boldsymbol{\beta}}(\lambda_{k+1})$
  - $\widehat{\boldsymbol{\beta}}_{\mathcal{A}_{k-1}}(\lambda_{k+1})$

Covariance test

- Consider, then, the following quantity:

$$T_C = \frac{\mathbf{y}^T\mathbf{X}\widehat{\boldsymbol{\beta}}(\lambda_{k+1}) - \mathbf{y}^T\mathbf{X}\widehat{\boldsymbol{\beta}}_{\mathcal{A}_{k-1}}(\lambda_{k+1})}{\sigma^2};$$

  i.e., how much of the covariance between the fitted model and outcome can be attributed to the new predictor, as opposed to the increase in covariance we would get from simply lowering λ without adding any new variables

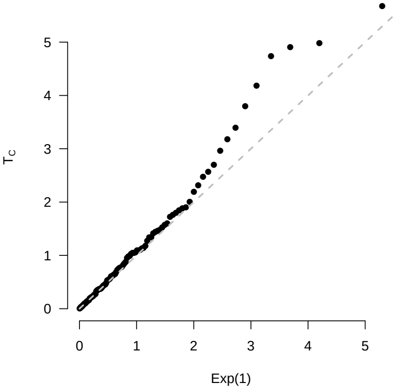
- Under the null hypothesis that all variables with $\beta_j \neq 0$ are included in $\mathcal{A}_{k-1}$, it can be shown that

$$T_C \xrightarrow{\text{d}} \text{Exp}(1)$$

  as n and $p \to \infty$

Accuracy of the approximating distribution

As before, $n = p = 100$, $\boldsymbol{\beta} = 0$, first predictor added:



$\mathbb{P}(T > 3.00) = 0.08$

Results: Example data

Applying the covariance test to the example data from our previous lecture, we have

	T_c	p
A1	13.0701	< 0.0001
A2	60.0133	< 0.0001
A3	0.4991	0.6108
A6	5.0203	0.0113
C29	0.0351	0.9655
A4	0.3442	0.7109
B3	0.0438	0.9572
A5	5.5515	0.0075
C4	0.1197	0.8875
B1	0.0477	0.9535

Remarks

- When σ^2 is not known, but can be estimated with rdf degrees of freedom, the estimate can be substituted for σ^2 and the covariance test statistic will follow a $F_{2,\mathrm{rdf}}$ distribution
- The distribution of the test statistic is thus somewhat inflated by searching over a number of candidate predictors, but far less so for the lasso than for forward selection due to the shrinkage imposed by the lasso
- It is possible to translate the sequential tests into an FDR rule, giving a false discovery rate for the included variables at a given point; for our example, however, this only allows the first two variables to enter

Selective inference

- Our final, and most recent, approach to inference is *selective inference*, or *post-selection inference*
- The idea here is to explicitly condition on the model we have selected in carrying out inference
- Similar to the previous approach, selective inference adjusts for the fact that we have "cherry-picked" the top k variables and eliminated $p-k$ other variables to arrive at this model
- Selective inference, however, is more comprehensive, offering post-selection confidence intervals and p-values for all of the terms in the selected model

The lasso selection event

- To proceed, we will need to explicitly characterize the condition that the lasso selects a given model $\mathcal{A}$ and eliminated the variables in set $\mathcal{B} = \mathcal{A}^C$
- **Theorem:** For a fixed value of λ, the event that the lasso sets $\widehat{\beta}_j = 0$ for all $j \in \mathcal{B}$ can be written

$$\left| \frac{1}{n} \mathbf{X}_{\mathcal{B}}^T (\mathbf{I} - \mathbf{P}_{\mathcal{A}}) \mathbf{y} + \lambda \mathbf{X}_{\mathcal{B}}^T \mathbf{X}_{\mathcal{A}} (\mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}})^{-1} \mathbf{s} \right| \leq \lambda,$$

where $\mathbf{P}_{\mathcal{A}} = \mathbf{X}_{\mathcal{A}} (\mathbf{X}_{\mathcal{A}}^T \mathbf{X}_{\mathcal{A}})^{-1} \mathbf{X}_{\mathcal{A}}^T$, $\mathbf{s} = \operatorname{sign}(\widehat{\boldsymbol{\beta}}_{\mathcal{A}})$, and the above inequality applies elementwise

Remarks

- Thus, the event that the lasso selects a certain model (and assigns its nonzero coefficients certain signs) can be written as a set of linear constraints $\{\mathbf{A}\mathbf{y} \leq \mathbf{b}\}$
- In other words, the set $\{\mathbf{y}|\mathbf{A}\mathbf{y} \leq \mathbf{b}\}$ is the set of random response vectors $\mathbf{y}$ that would yield the same active set and coefficient signs as the model we've selected

Selective inference: Big picture

- Now, suppose that $\mathbf{y} \sim \mathbf{N}(\boldsymbol{\mu}, \sigma^2\mathbf{I})$
- The main idea of selective inference is to make inference on $\boldsymbol{\mu}$, or more generally a linear combination $\theta = \boldsymbol{\eta}^T\boldsymbol{\mu}$, conditional on the event $\{\mathbf{A}\mathbf{y} \leq \mathbf{b}\}$
- For example, we would likely be interested in $(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}^T\boldsymbol{\mu}$; to address this question we would set $\boldsymbol{\eta}$ equal to various columns of $\mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^T\mathbf{X}_{\mathcal{A}})^{-1}$
- Carrying out this conditional inference turns out to be quite a bit easier than one would expect due to a remarkable result known as the *polyhedral lemma*, which we will state on the next slide without proof

Polyhedral lemma

Theorem: The conditional distribution of $\boldsymbol{\eta}^T\mathbf{y}|\{\mathbf{Ay} \leq \mathbf{b}\}$ is equivalent to distribution of $\boldsymbol{\eta}^T\mathbf{y}$ given

$$\begin{aligned}\mathcal{V}^-(\mathbf{y}) \leq \boldsymbol{\eta}^T\mathbf{y} \leq \mathcal{V}^+(\mathbf{y})\\ \mathcal{V}^0(\mathbf{y}) \geq 0,\end{aligned}$$

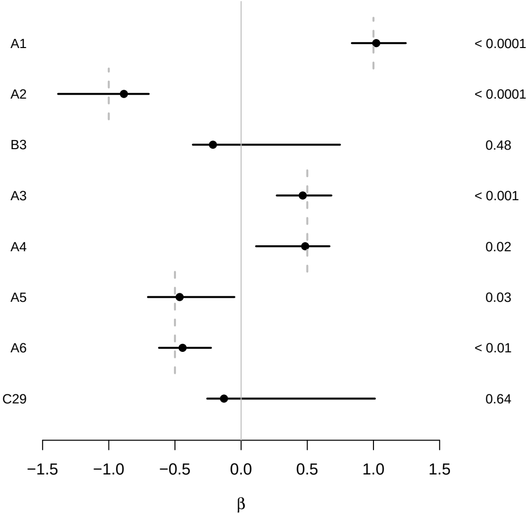
where

$$\begin{aligned}\boldsymbol{\alpha} &= \mathbf{A}\boldsymbol{\eta}/\|\boldsymbol{\eta}\|^2\\ \mathcal{V}^-(\mathbf{y}) &= \max_{j:\alpha_j<0} \frac{b_j - (\mathbf{Ay})_j + \alpha_j\boldsymbol{\eta}^T\mathbf{y}}{\alpha_j}\\ \mathcal{V}^+(\mathbf{y}) &= \max_{j:\alpha_j>0} \frac{b_j - (\mathbf{Ay})_j + \alpha_j\boldsymbol{\eta}^T\mathbf{y}}{\alpha_j}\\ \mathcal{V}^0(\mathbf{y}) &= \min_{j:\alpha_j=0} \{b_j - (\mathbf{Ay})_j\}\end{aligned}$$

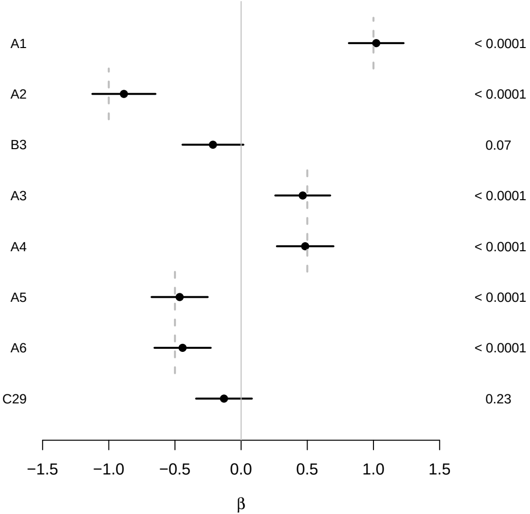
Using the lemma for inference

- The polyhedral lemma appears complex, but its upshot is actually quite simple: $\boldsymbol{\eta}^T\mathbf{y}$ would ordinarily follow a normal distribution, but conditional on the model we have selected, it follows a truncated normal distribution with support determined by $\mathbf{A}$ and $\mathbf{b}$
- Thus, letting F denote the CDF of this truncated normal distribution, we can carry out hypothesis tests of $\theta = \boldsymbol{\eta}^T\boldsymbol{\mu} = 0$ by calculating tail areas with respect to this distribution
- Likewise, we can construct confidence intervals by inverting the above tests, searching for values of θ satisfying $F(\theta) = \alpha/2$ and $F(\theta) = 1 - \alpha/2$

Application to example data



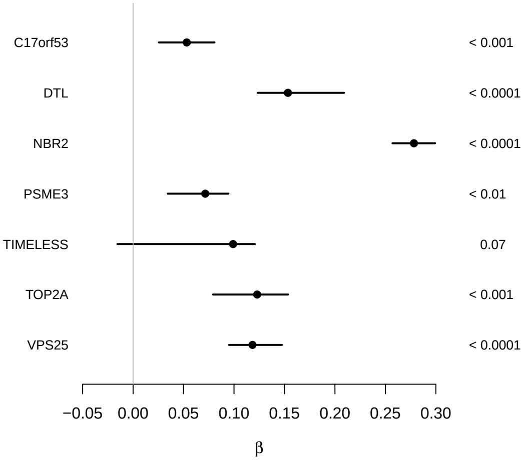
Lasso-OLS hybrid intervals



covTest: TCGA data

	T_c	p
C17orf53	126.0942	< 0.0001
VPS25	12.4525	< 0.0001
NBR2	8.8638	0.0001
DTL	22.4862	< 0.0001
PSME3	2.1363	0.1181
TOP2A	8.0346	0.0003
TIMELESS	4.5714	0.0103
CDC25C	1.2952	0.2738
CCDC56	2.1602	0.1153
CENPK	0.3062	0.7363

selectiveInference: TCGA data



Final remarks

- The covariance test lies somewhere in between the classical and "false inclusion" definitions with respect to the error rate it controls
- The covariance test approach scales up well to high dimensions, but how powerful it is has not been extensively studied
- Selective inference is a very promising approach to inference that appears to work very well in the $n > p$ case
- How well it works in the $p > n$ case is something of an unanswered question (the method is very new at this point); for the TCGA data, once the 8th predictor enters the model, all confidence intervals become infinitely wide