Lecture 21
Theory of the Lasso II

02 December 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

Yale

Class Notes

- Midterm II - Available now, due next Monday
- Problem Set 7 - Available now, due December 11th (grace period through the 16th)

Last time

Last class, we started investigating the theory of the lasso estimator.

For the case of $X^t X$ equal to the identity matrix, we were able to quickly establish bounds on the prediction error, estimation of β, and the reconstruction of the support of β.

For an arbitrary X matrix we were able to calculate a bound on $||X(\widehat{\beta} - \beta)||_2^2$.

Today's goal is to establish a bound on $||\widehat{\beta} - \beta||_2^2$

The basic starting point from last time was the following decomposition, which had no assumptions beyond linearity of the true model:

$$||X(\beta - b)||_2^2 \leq 2\epsilon^t X(b - \beta) + \lambda \cdot (||\beta||_1 - ||b||_1)$$

Where can think of this decomposition as the loss to be minimized, the empirical part, and the penalty term.

I then defined the set

$$\mathcal{A} = \left\{ 2||\epsilon^t X||_\infty \leq \lambda \right\}$$

And showed that for any $A > 1$ we have $\mathbb{P}\mathcal{A} \geq 1 - A^{-1}$ whenever

$$\lambda \geq A \cdot \sqrt{8 \log(2p)\sigma^2}.$$

Today we will motivate a stronger assumption on the model and use these two results to establish bounds on the prediction of β.

Also, it will be helpful to write the set $\mathcal{A}$ as being parameterized by the value of λ_0:

$$\mathcal{A}(\lambda_0) = \left\{2||\epsilon^t X||_\infty \le \lambda_0\right\}$$

Bounds on estimation error

We already know that on $\mathcal{A}(\lambda_0)$ and with $\lambda > 2 \cdot \lambda_0$, we have:

$$\begin{aligned} ||X(b-\beta)||_2^2 + \lambda \cdot ||b||_1 &\leq 2\epsilon^t X(b-\beta) + \lambda \cdot ||\beta||_1 \\ &\leq \lambda_0 ||b-\beta||_1 + \lambda \cdot ||\beta||_1 \end{aligned}$$

Now, multiplying by two gives:

$$2||X(b-\beta)||_2^2 + 2\lambda \cdot ||b||_1 \leq \lambda ||b-\beta||_1 + 2\lambda \cdot ||\beta||_1$$

Recall that we defined the notation: $S = \{j : \beta_j \neq 0\}$, s is the size of the set S, and v_S is the vector v which has components not in S set to zero.

Notice that:

$$\begin{aligned} ||b||_1 &= ||b_S||_1 + ||b_{S^c}||_1 \\ &\geq ||\beta||_1 - ||b_S - \beta||_1 + ||b_{S^c}||_1 \end{aligned}$$

Using the (reverse) triangle inequality and the fact that β_{S^c} is zero by definition.

Similarly, we have:

$$||b - \beta||_1 = ||b_S - \beta_S||_1 + ||b_{S^c}||_1$$

Where clearly β_S is redundant, but useful to keep the notation straight.

Plugging these in, we now get:

$$\begin{aligned} 2||X(b-\beta)||_2^2 + 2\lambda \cdot ||\beta_S||_1 - 2\lambda \cdot ||b_S - \beta||_1 + 2\lambda \cdot ||b_{S^c}||_1 \\ \leq \lambda ||b-\beta||_1 + \lambda \cdot ||b_S - \beta_S||_1 + 2\lambda \cdot ||b_{S^c}||_1 \end{aligned}$$

Which cancels out as:

$$2||X(b-\beta)||_2^2 + \lambda ||b_{S^c}||_1 \leq 3 \cdot \lambda \cdot ||b_S - \beta_S||_1$$

This result now actually gives two sub-results, as all three terms are positive and therefore each component of the left hand side is individually bounded by the right hand side.

In particular, we have:

$$||b_{S^c}||_1 \leq 3 \cdot ||b_S - \beta_S||_1$$

Which implies that the amount of error in b can not be too highly concentrated on S^c.

The other sub-result gives:

$$2||X(b-\beta)||_2^2 \leq 3\lambda \cdot ||b_S - \beta_S||_1$$

If σ_{min} is the minimum singular value of X, then the left hand side can be bounded below by:

$$2\sigma_{min}^2||b-\beta||_2^2 \leq 3\lambda \cdot ||b_S - \beta_S||_1$$

Using the Cauchy-Schwarz inequality, this becomes:

$$\begin{aligned} 2\sigma_{min}^2||b-\beta||_2^2 &\leq 3\lambda \cdot \sqrt{s}||b_S - \beta_S||_2 \\ ||b-\beta||_2 &\leq \frac{3\lambda\sqrt{s}}{2\sigma_{min}^2} \end{aligned}$$

Which gives a bound on the error of estimating β, which is exactly what we wanted to establish.

Why is this not sufficient for us? Well, in the high dimensional case $p > n$, we will always have σ_{min} equal to 0.

We can get around this problem by defining a modified version of the minimum eigenvector (or squared singular value) by only considering $b - \beta$ such that:

$$||b_{S^c}||_1 \leq 3 \cdot ||b_S - \beta_S||_1$$

The (minimum) restricted eigenvalue ϕ_S on the set S is defined as:

$$\phi_S = \arg\min_{v \in \mathcal{V}_S} \frac{||Xb||_2}{||b||_2}$$

Where:

$$\mathcal{V}_S = \{v \in \mathbb{R}^p \text{ s.t. } ||v_{S^c}||_1 \leq 3 \cdot ||v_S||_1\}$$

Because we do not know S, it is impossible to calculate ϕ_S in practice. In theoretical work, often one considers **the** restricted eigenvalue ϕ defined as the smallest ϕ_S for all sets S with size bounded by some predefined s_0.

Now, we can bound the following using our prior result:

$$\begin{aligned} 2||X(b-\beta)||_2^2 + \lambda \cdot ||b-\beta||_1 \\ &= 2||X(b-\beta)||_2^2 + \lambda \cdot ||b_S - \beta_S||_1 + \lambda \cdot ||b_{S^c}||_1 \\ &= 4\lambda \cdot ||b_S - \beta_S||_1 \end{aligned}$$

Using Cauchy-Schartz again, we can change the ℓ_1-norm to an ℓ_2-norm at the cost of a factor of $\sqrt{s}$:

$$2||X(b-\beta)||_2^2 + \lambda \cdot ||b-\beta||_1 \leq 4\lambda \cdot \sqrt{s} \cdot ||b_S - \beta_S||_2$$

Finally, we now use the restricted eigenvalue ϕ to convert from β space to $X\beta$ space:

$$2||X(b-\beta)||_2^2 + \lambda \cdot ||b-\beta||_1 \leq 4\lambda \cdot \sqrt{s} \cdot ||X(b_S - \beta_S)||_2/\phi$$

I am now going to use an inequality trick that is often useful in theoretical statistics derivations. For any u and v, notice that $4uv \leq u^2 + 4v^2$.

For a proof, notice that it is trivially true at zero and negative values of u and v. Then look at the derivatives and notice that the right hand side grows faster than the left hand side in the directions of both u and v.

Setting $u = ||X(b_S - \beta_S)||_2$, we then have:

$$\begin{aligned} 2||X(b-\beta)||_2^2 + \lambda \cdot ||b-\beta||_1 &\leq ||X(b_S - \beta_S)||_2 + 4\lambda^2 \cdot s \cdot /\phi^2 \\ &\leq ||X(b-\beta)||_2^2 + 4\lambda^2 \cdot s \cdot /\phi^2 \end{aligned}$$

And when canceling one factor of $||X(b-\beta)||_2$:

$$||X(b-\beta)||_2^2 + \lambda \cdot ||b-\beta||_1 \leq 4\lambda^2 \cdot s \cdot /\phi^2$$

Which holds on the entire set $\mathcal{A}(\lambda_0)$.

This establishes two simultaneous bounds:

$$||X(b-\beta)||_2^2 \leq 4\lambda^2 \cdot s \cdot /\phi^2$$
$$||b-\beta||_1 \leq 4\lambda \cdot s \cdot /\phi^2$$

Though the first is slightly less satisfying than our result in last class as it relies on ϕ^2, though it no longer requires the norm of β.

Asymptotic analysis

As before, we can convert a more natural re-scaled problem by dividing all of the λ parameters by $\sqrt{n}$

Also, remember that for some $A > 1$, we have $\mathbb{P}\mathcal{A}(\lambda_0) \geq 1 - A^{-1}$ for all $\lambda > A \cdot \sqrt{16 n^{-1} \log(2p) \sigma^2}$.

Therefore, we have:

$$\begin{aligned}||b-\beta||_1 &\le 4\lambda \cdot s \cdot /\phi^2 \\ &\le 8 \cdot A\sigma^2/\phi^2 \cdot \frac{s_n^2 \log(2p_n)}{n}\end{aligned}$$

Which is the same result as from the Bickel, Ritov, Tsybakov paper.

To establish consistency of the estimator under constant noise and restricted eigenvalues ϕ^2, we need the following limit to go to zero:

$$\lim_{n \to \infty} \frac{s_n^2 \log(2p_n)}{n} = 0$$

Which can happen with a number of different scalings, such as a constant number of non-zero terms but an exponential number of non-zero terms. Or, s_n growing like $n^{1/3}$ and p_n growing linearly with s_n.

Some (personal) closing thoughts on the application of the lasso theory to data analysis:

1. The theory is useful for establishing a rough rule of thumb for how large p_n and s_n can be to have a reasonable chance of reconstructing β or $X\beta$
2. The theory also helps guide where to start looking for the optimal λ
3. We still generally need some form of cross validation however, as the theoretical values tend to overestimate λ in practice; we also do not know σ^2 and in theory need to use an over-estimate for the convergence results to hold
4. Bounds on $||X(\beta - \widehat{\beta})||_2^2$ are nice to have, however the theoretical bounds on $||\beta - \widehat{\beta}||_2^2$ are difficult to use in practice due to the near-impossible to calculate restricted eigenvalue assumption
5. I have always been skeptical of the asymptotic results for the same reason; ϕ likely depends on n, p_n and s_n in complex ways that are not accounted for

For our next (and last) week we will:

1. use the lasso to encode more complex forms of linear sparsity (e.g., outlier detection and the fused lasso)
2. give an alternative approach to solving for the lasso solution at a particular value of λ