Lecture 20
Theory of the Lasso I

30 November 2015

Taylor B. Arnold
Yale Statistics
STAT 312/612

Yale

Class Notes

- Midterm II - Posted this afternoon, due next Monday
- Problem Set 7 - Posted Wed., due December 11th (grace period through the 16th)

Types of bounds

Recall that the lasso regression replaces the ℓ_2 penalty with an ℓ_1 penalty, and looks deceptively similar to the ridge regression:

$$\widehat{\beta}_\lambda = \arg\min_b \left\{ ||y - Xb||_2^2 + \lambda ||b||_1 \right\}$$

Where the ℓ_1-norm is defined as the sum of the absolute values of the vector's components:

$$||\beta||_1 = \sum_i |\beta_i|$$

There are three types of errors that we commonly are concerned with in lasso regression.

There are three types of errors that we commonly are concerned with in lasso regression. The prediction loss:

$$||X(\beta - \widehat{\beta})||_2^2$$

There are three types of errors that we commonly are concerned with in lasso regression. The prediction loss:

$$||X(\beta - \widehat{\beta})||_2^2$$

Parameter estimation:

$$||\beta - \widehat{\beta}||_2^2$$

There are three types of errors that we commonly are concerned with in lasso regression. The prediction loss:

$$||X(\beta - \widehat{\beta})||_2^2$$

Parameter estimation:

$$||\beta - \widehat{\beta}||_2^2$$

And model selection:

$$\mathbb{P}\left\{\text{supp}(\beta) = \text{supp}(\widehat{\beta})\right\}$$

The type of results we want to establish may look different than you are used to seeing in more introductory courses.

The type of results we want to establish may look different than you are used to seeing in more introductory courses. What we will want to be able to construct is a set $\mathcal{A}$ such that:

$$\mathbb{P}\mathcal{A} = 1 - \epsilon$$

For some small $\epsilon > 0$, where we have bounds such as

$$||X(\beta - \widehat{\beta})||_2^2 \leq \delta$$

Conditioned on being in event $\mathcal{A}$.

The type of results we want to establish may look different than you are used to seeing in more introductory courses. What we will want to be able to construct is a set $\mathcal{A}$ such that:

$$\mathbb{P}\mathcal{A} = 1 - \epsilon$$

For some small $\epsilon > 0$, where we have bounds such as

$$||X(\beta - \widehat{\beta})||_2^2 \leq \delta$$

Conditioned on being in event $\mathcal{A}$.

Today I'll establish bounds on all three for a simple case where X^tX is the identity matrix and bounds on the first in the general case of an arbitrary X matrix.

Simple case

Simple case

Let's consider the simple case where X^tX is equal to the identity matrix.

Simple case

Let's consider the simple case where X^tX is equal to the identity matrix.

We know that the lasso solution can be written as (Lecture 17):

$$\widehat{\beta}_j^{\lambda} = \begin{cases} 0, & \lambda > 2 \cdot |x_j^t y| \\ x_j^t y - \text{sign}(x_j^t y) \cdot \lambda, & \lambda \leq 2 \cdot |x_j^t y| \end{cases}$$

Define the set $\mathcal{A}$ such that λ bounds all of the correlations of X with the noise vector ϵ:

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

We will develop a general result later showing what the probability of $\mathcal{A}$ occurring is.

We see that for any j:

$$\begin{aligned} 2 \cdot |x_j^t y| &= 2 \cdot |x_j^t (X\beta + \epsilon)| \\ &= 2 \cdot |(\beta_j + x_j^t \epsilon)| \end{aligned}$$

We see that for any j:

$$\begin{aligned} 2 \cdot |x_j^t y| &= 2 \cdot |x_j^t (X\beta + \epsilon)| \\ &= 2 \cdot |(\beta_j + x_j^t \epsilon)| \end{aligned}$$

If β_j is equal to 0, then on $\mathcal{A}$:

$$\begin{aligned} 2 \cdot |x_j^t y| &= 2 \cdot |x_j^t \epsilon| \\ &\leq \lambda \end{aligned}$$

And therefore $\widehat{\beta}_j$ will also be set exactly to zero.

What about j such that β_j is not equal to 0?

What about j such that β_j is not equal to 0? Then on $\mathcal{A}$:

$$\begin{aligned} 2 \cdot |x_j^t y| &= 2 \cdot |x_j^t (X\beta + \epsilon)| \\ &= 2 \cdot |\beta_j + x_j^t \epsilon| \\ &\geq 2 \cdot |\beta_j| - 2 \cdot |x_j^t \epsilon| \\ &\geq 2 \cdot |\beta_j| - \lambda \end{aligned}$$

What about j such that β_j is not equal to 0? Then on $\mathcal{A}$:

$$\begin{aligned} 2 \cdot |x_j^t y| &= 2 \cdot |x_j^t (X\beta + \epsilon)| \\ &= 2 \cdot |\beta_j + x_j^t \epsilon| \\ &\geq 2 \cdot |\beta_j| - 2 \cdot |x_j^t \epsilon| \\ &\geq 2 \cdot |\beta_j| - \lambda \end{aligned}$$

We then have that as long as the following holds:

$$\begin{aligned} \lambda &\leq 2 \cdot |\beta_j| - \lambda \\ \lambda &\leq |\beta_j| \end{aligned}$$

$\widehat{\beta}_j$ will be non-zero.

Therefore, on $\mathcal{A}$ the estimator $\widehat{\beta}_\lambda$ finds the correct support of β if:

$$2||X\epsilon||_2^2 \leq \lambda \leq \min_{j,\beta_j \neq 0} |\beta_j|$$

As one would expect, this forces there to be no particularly small elements β_j, as otherwise we could not differentiate between that and 0.

To establish a bound on the estimation error, we need a bit more notation. Let $S = \{j : \beta_j \neq 0\}$ and s be the size of the set S. Also, let v_S be the vector v which has components not in S set to zero.

To establish a bound on the estimation error, we need a bit more notation. Let $S = \{j : \beta_j \neq 0\}$ and s be the size of the set S. Also, let v_S be the vector v which has components not in S set to zero.

Further, let:

$$\widehat{\beta}^{\text{oracle}} = (X_S^t X_S)^{-1} X_S^t y$$

That is, the ordinary least squares estimator that knows that the correct support of β is S.

Now, notice that on $\mathcal{A}$, the estimator $\widehat{\beta}^{\lambda}$ will be zero on S^c. Therefore:

$$||\beta - \widehat{\beta}||_2 \quad = \quad ||\beta_S - \widehat{\beta}_S||_2$$

Now, notice that on $\mathcal{A}$, the estimator $\widehat{\beta}^{\lambda}$ will be zero on S^c. Therefore:

$$\begin{aligned} ||\beta - \widehat{\beta}||_2 &= ||\beta_S - \widehat{\beta}_S||_2 \\ &= ||\beta_S - \widehat{\beta}^{\text{oracle}} + \widehat{\beta}^{\text{oracle}} - \widehat{\beta}_S||_2 \end{aligned}$$

Now, notice that on $\mathcal{A}$, the estimator $\widehat{\beta}^{\lambda}$ will be zero on S^c. Therefore:

$$\begin{aligned} ||\beta - \widehat{\beta}||_2 &= ||\beta_S - \widehat{\beta}_S||_2 \\ &= ||\beta_S - \widehat{\beta}^{\text{oracle}} + \widehat{\beta}^{\text{oracle}} - \widehat{\beta}_S||_2 \\ &\leq ||\beta_S - \widehat{\beta}^{\text{oracle}}||_2 + ||\widehat{\beta}^{\text{oracle}} - \widehat{\beta}_S||_2 \end{aligned}$$

Now, notice that on $\mathcal{A}$, the estimator $\widehat{\beta}^{\lambda}$ will be zero on S^c. Therefore:

$$
\begin{aligned}
||\beta - \widehat{\beta}||_2 &= ||\beta_S - \widehat{\beta}_S||_2 \\
&= ||\beta_S - \widehat{\beta}^{\text{oracle}} + \widehat{\beta}^{\text{oracle}} - \widehat{\beta}_S||_2 \\
&\leq ||\beta_S - \widehat{\beta}^{\text{oracle}}||_2 + ||\widehat{\beta}^{\text{oracle}} - \widehat{\beta}_S||_2 \\
&= ||\beta_S - \widehat{\beta}^{\text{oracle}}||_2 + \sqrt{\sum_{j \in S} (x_j^t y - \text{sign}(x_j^t y)\lambda - x_j^t y)^2}
\end{aligned}
$$

Now, notice that on $\mathcal{A}$, the estimator $\widehat{\beta}^{\lambda}$ will be zero on S^c. Therefore:

$$
\begin{aligned}
||\beta - \widehat{\beta}||_2 &= ||\beta_S - \widehat{\beta}_S||_2 \\
&= ||\beta_S - \widehat{\beta}^{\text{oracle}} + \widehat{\beta}^{\text{oracle}} - \widehat{\beta}_S||_2 \\
&\leq ||\beta_S - \widehat{\beta}^{\text{oracle}}||_2 + ||\widehat{\beta}^{\text{oracle}} - \widehat{\beta}_S||_2 \\
&= ||\beta_S - \widehat{\beta}^{\text{oracle}}||_2 + \sqrt{\sum_{j \in S} (x_j^t y - \text{sign}(x_j^t y)\lambda - x_j^t y)^2} \\
&= ||\beta_S - \widehat{\beta}^{\text{oracle}}||_2 + \sqrt{\sum_{j \in S} (\text{sign}(x_j^t y)\lambda)^2}
\end{aligned}
$$

Now, notice that on $\mathcal{A}$, the estimator $\widehat{\beta}^{\lambda}$ will be zero on S^c. Therefore:

$$
\begin{aligned}
||\beta - \widehat{\beta}||_2 &= ||\beta_S - \widehat{\beta}_S||_2 \\
&= ||\beta_S - \widehat{\beta}^{\text{oracle}} + \widehat{\beta}^{\text{oracle}} - \widehat{\beta}_S||_2 \\
&\leq ||\beta_S - \widehat{\beta}^{\text{oracle}}||_2 + ||\widehat{\beta}^{\text{oracle}} - \widehat{\beta}_S||_2 \\
&= ||\beta_S - \widehat{\beta}^{\text{oracle}}||_2 + \sqrt{\sum_{j \in S} (x_j^t y - \text{sign}(x_j^t y)\lambda - x_j^t y)^2} \\
&= ||\beta_S - \widehat{\beta}^{\text{oracle}}||_2 + \sqrt{\sum_{j \in S} (\text{sign}(x_j^t y)\lambda)^2} \\
&= ||\beta_S - \widehat{\beta}^{\text{oracle}}||_2 + \lambda\sqrt{s}
\end{aligned}
$$

Now, notice that on $\mathcal{A}$, the estimator $\widehat{\beta}^{\lambda}$ will be zero on S^c. Therefore:

$$
\begin{aligned}
||\beta - \widehat{\beta}||_2 &= ||\beta_S - \widehat{\beta}_S||_2 \\
&= ||\beta_S - \widehat{\beta}^{\text{oracle}} + \widehat{\beta}^{\text{oracle}} - \widehat{\beta}_S||_2 \\
&\leq ||\beta_S - \widehat{\beta}^{\text{oracle}}||_2 + ||\widehat{\beta}^{\text{oracle}} - \widehat{\beta}_S||_2 \\
&= ||\beta_S - \widehat{\beta}^{\text{oracle}}||_2 + \sqrt{\sum_{j \in S} (x_j^t y - \text{sign}(x_j^t y)\lambda - x_j^t y)^2} \\
&= ||\beta_S - \widehat{\beta}^{\text{oracle}}||_2 + \sqrt{\sum_{j \in S} (\text{sign}(x_j^t y)\lambda)^2} \\
&= ||\beta_S - \widehat{\beta}^{\text{oracle}}||_2 + \lambda\sqrt{s}
\end{aligned}
$$

So the cost of not knowing S is an extra factor of $\lambda\sqrt{s}$ in the prediction error.

This result is a type of *oracle inequality*, which relates the error of not knowing some quantity to the error that can be attained when the quantity is known.

In the lasso literature, this almost always refers to comparing a penalized estimator on X to ordinary least squares fit on the set S.

Now, how far off will the prediction of β be? Notice that:

$$\begin{aligned} ||X(\beta - \widehat{\beta})||_2 &= \sqrt{(\beta - \widehat{\beta})^t X^t X (\beta - \widehat{\beta})} \\ &= \sqrt{(\beta - \widehat{\beta})^t (\beta - \widehat{\beta})} \\ &= ||\beta - \widehat{\beta}||_2 \end{aligned}$$

So we simultaneously established a prediction and estimation bound for the simple lasso regression.

So now on the set $\widehat{A}$ we have bounds on all three quantities of interest.

So now on the set $\widehat{A}$ we have bounds on all three quantities of interest.

Notice that none of our bounds so far depend on n, p, or σ^2, which may seem quite odd. These are actually bound up in the choice of λ and the scale of X (which we set to be 1).

Prediction error: general case

We want to establish some result about the general lasso solution.
We don't have an analytic form anymore, so where to begin?

We want to establish some result about the general lasso solution. We don't have an analytic form anymore, so where to begin?

Let's start with the one relationship we know to be true between $\widehat{\beta}$ (which I will call b to simplify the slides) and β:

$$||y - Xb||_2^2 + \lambda||b||_1 \leq ||y - X\beta||_2^2 + \lambda||\beta||_1$$

By expanding the ℓ_2-norm and writing y as $X\beta + \epsilon$, the right-hand side can be written as:

$$\begin{aligned} ||y - X\beta||_2^2 + \lambda||\beta||_1 &= y^t y + \beta^t X^t X\beta - 2y^t X\beta + \lambda||\beta||_1 \\ &= y^t y + \beta^t X^t X\beta - 2\beta^t X^t X\beta - 2\epsilon^t X\beta + \lambda||\beta||_1 \\ &= y^t y - \beta^t X^t X\beta - 2\epsilon^t X\beta + \lambda||\beta||_1 \end{aligned}$$

Similarly, the left hand side simply becomes:

$$\begin{aligned} ||y - Xb||_2^2 + \lambda||b||_1 &= y^t y + b^t X^t X b - 2y^t X b + \lambda||b||_1 \\ &= y^t y + b^t X^t X b - 2\beta^t X^t X b - 2\epsilon^t X b + \lambda||b||_1 \end{aligned}$$

Canceling the $y^t y$ on both sides and putting similar terms on each side of the inequality, this yields:

$$b^t X^t X b - 2\beta^t X^t X b - 2\epsilon^t X b + \lambda||b||_1 \leq -\beta^t X^t X \beta - 2\epsilon^t X \beta + \lambda||\beta||_1$$

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

Canceling the $y^t y$ on both sides and putting similar terms on each side of the inequality, this yields:

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

The left hand side can be written as an inner product, and we now have a basic inequality with three terms:

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

We see that this decomposes nicely into three distinct terms:

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

These are the the loss to be minimized, the empirical part, and the penalty term.

Bound on inner product $u^t v$

Let u and v be arbitrary vectors. Notice that:

$$\begin{aligned}
|u^t v| &= \left| \sum_i u_i v_i \right| \\
&\leq \sum_i |u_i v_i| \\
&\leq \sum_i |\max_i(u_i) v_i| \\
&= |\max_i(u_i)| \cdot \sum_i |v_i| \\
&= ||u||_\infty \cdot ||v||_1
\end{aligned}$$

This is a special case of Hölder's inequality.

We will use this trick to bound the empirical part of our inequality:

$$2|\epsilon^t X(b-\beta)| \leq 2||\epsilon^t X||_\infty \cdot ||b-\beta||_1$$

We will use this trick to bound the empirical part of our inequality:

$$2|\epsilon^t X(b-\beta)| \leq 2||\epsilon^t X||_\infty \cdot ||b-\beta||_1$$

We will then use the same definition of $\mathcal{A}$:

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

On this set, the random part is on the same order of magnitude as the penalty part. We will return in a bit to talk about what would make this quantity small.

On the set $\mathcal{A}$, we have:

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

On the set $\mathcal{A}$, we have:

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

And using the inequality $||\beta|| + ||b|| \geq ||\beta - b||$, we then have:

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

So, it seems that picking a smaller λ yields a tighter bound. However, the probability of being on $\mathcal{A}$ decreases as λ decreases.

There is ultimately a trade off to be made, and we will see that it is possible to parameterize λ in a nice way to show-off the various bounds.

Bounds on empirical process

Now, let's consider when the event $\mathcal{A}$ occurs. We will assume that the columns of X_j have a norm of 1.

Start by setting z_j equal to $\epsilon^t X_j$. These are distributed (not necessarily independent) as $\mathcal{N}(0, \sigma^2)$.

To bound these probabilities, recall this result on the maximum of p observations from identically distributed normal random variables:

$$\mathbb{E}\left[\max_j |z_j|\right] \leq \sigma\sqrt{2\log(2p)}$$

Using the Markov inequality, we get for any $a > 0$:

$$\begin{aligned}\mathbb{P}\left[\max_j |z_j| \geq a/2\right] &\leq \frac{\mathbb{E}\max_j |z_j|}{a/2} \\ &\leq \frac{\sigma\sqrt{8\log(2p)}}{a}\end{aligned}$$

Using the Markov inequality, we get for any $a > 0$:

$$\begin{aligned}\mathbb{P}\left[\max_j |z_j| \geq a/2\right] &\leq \frac{\mathbb{E}\max_j |z_j|}{a/2} \\ &\leq \frac{\sigma\sqrt{8\log(2p)}}{a}\end{aligned}$$

Now, for simplicity, set $a = A \cdot \sqrt{8\log(2p)\sigma^2}$ for any $A > 1$. Then $\mathbb{P}\mathcal{A} = 1 - A^{-1}$ whenever

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

Summary and consistency

So far, I have been solving the following optimization problem:

$$\widehat{\beta}_\lambda = \arg\min_b \left\{ ||y - Xb||_2^2 + \lambda ||b||_1 \right\}$$

It will be useful now to re-scale the problem.

Up until now we have assumed that $X^t X$ has 1's on the diagonal. A more natural assumption is that it was n's on the diagonal. Let $X' = \sqrt{n} \cdot X$ and $\beta' = n^{-1/2}\beta$. Then, the new criterion becomes:

$$||y - X'b'||_2^2 + \lambda\sqrt{n} \cdot ||b'||_1$$
$$||y - X'b'||_2^2 + \lambda' \cdot ||b'||_1$$

So, therefore we can solve this re-scaled problem by simply dividing all of the λ parameters by $\sqrt{n}$.

So we now can put all of this in a single theorem.

So we now can put all of this in a single theorem.

Theorem For some $A > 1$ and all $\lambda > A \cdot \sqrt{8n^{-1}\log(2p)\sigma^2}$, we have with probability $1 - A^{-1}$:

$$||X(\beta - b)||_2^2/n \leq 2\frac{\lambda}{\sqrt{n}}||\beta||_1$$

How do we get consistency out of this? We need both of these things to go to zero as n goes to infinity:

$$A^{-1} \to 0$$

$$A \cdot \sqrt{\frac{8 \log(2p)\sigma^2}{n}} ||\beta||_1 \to 0$$

How do we get consistency out of this? We need both of these things to go to zero as n goes to infinity:

$$A^{-1} \to 0$$

$$A \cdot \sqrt{\frac{8 \log(2p)\sigma^2}{n}} ||\beta||_1 \to 0$$

Notice that as long as:

$$\sqrt{\frac{8 \log(2p)\sigma^2}{n}} ||\beta||_1 \to 0$$

We can set A to decay very slowly, such as $A^{-1} = n^{-0.01}$.

So, assuming that σ^2 and the size of the true β stay fixed, we have consistency of the lasso estimator as long as p_n is dominated by e^n.

Simulations