Theoretical results: Non-asymptotic

Patrick Breheny

March 28

Introduction

- Last time we derived results from a classical perspective in which $\boldsymbol{\beta}^*$ was fixed as $n \to \infty$
- Today, we will consider things from a non-asymptotic perspective, obtaining bounds on estimation and prediction error while allowing $p > n$
- Although results along these lines can be shown for other penalized regression estimators as well, today's lecture will focus entirely on the lasso

A preliminary lemma

- We'll begin by discussing prediction, as we can prove results here without requiring any additional conditions
- First, let us prove the following lemma, from which several of our later results will derive
- **Lemma:** If $\lambda \geq \frac{2}{n}\|\mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty$, then the lasso prediction error satisfies

$$\frac{1}{n}\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2 \leq \lambda\|\boldsymbol{\delta}\|_1 + 2\lambda\|\boldsymbol{\beta}^*\|_1 - 2\lambda\|\boldsymbol{\delta} + \boldsymbol{\beta}^*\|_1,$$

where $\boldsymbol{\delta} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$

Prediction bound

- Based on this lemma, we have the following
- **Theorem:** If $\lambda \geq \frac{2}{n}\|\mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty$, then the lasso prediction error satisfies

$$\frac{1}{n}\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2 \leq 4\lambda\|\boldsymbol{\beta}^*\|_1$$

- **Corollary:** If $\lambda = 2\sigma\sqrt{c\log(p)/n}$ and $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ with $\varepsilon_i \overset{\perp\!\!\!\perp}{\sim} \mathrm{N}(0, \sigma^2)$, then the lasso prediction error satisfies

$$\frac{1}{n}\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2 \leq 8\sigma\|\boldsymbol{\beta}^*\|_1\sqrt{\frac{c\log p}{n}}$$

with probability at least $1 - 2\exp\{-\frac{1}{2}(c-2)\log p\}$

Remarks

- The prediction error increases with noise and dimension, and decreases with sample size – these dependencies are intuitive
- The dependence on $\|\beta^*\|$ is less obvious; it is worth noting, however, that up until this point, we have assumed nothing about β^* (or about $\mathbf{X}$)
- This prediction result differs from our previous results: previously, we had shown that prediction error was $O(n^{-1})$, whereas this result is $O(n^{-1/2})$

Eigenvalue conditions

- In the previous lecture, we introduced an eigenvalue condition: namely, that $\mathbf{X}^\top\mathbf{X}/n \to \boldsymbol{\Sigma}$, with the minimum eigenvalue of $\boldsymbol{\Sigma}$ bounded above 0
- Why is this important?
- We're finding the value $\widehat{\boldsymbol{\beta}}$ that minimizes $Q(\boldsymbol{\beta})$; but even if we can guarantee that $Q(\widehat{\boldsymbol{\beta}}) \approx Q(\boldsymbol{\beta}^*)$, if the function is flat, we have no guarantee that $\widehat{\boldsymbol{\beta}}$ is close to $\boldsymbol{\beta}^*$
- If $p > n$, however, it is clear that this condition can never be met

Restricting our eigenvalue conditions

- In other words, our previous condition was:

$$\frac{\frac{1}{n}\boldsymbol{\delta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2^2} > \tau$$

  for all $\boldsymbol{\delta} \neq \mathbf{0}$ and some $\tau > 0$
- However, what if this condition didn't have to be met for *all* $\boldsymbol{\delta} \in \mathbb{R}^p$, but only for *some* $\boldsymbol{\delta} \in \mathbb{R}^p$?
- For example, what if we only had to satisfy the condition for $\boldsymbol{\delta} \in \mathbb{R}^{\mathcal{S}}$?

A cone condition

- This is a step in the right direction, but not nearly strong enough: for example, suppose a variable in $\mathcal{N}$ was perfectly correlated with a variable in $\mathcal{S}$
- We will definitely need to involve $\mathcal{N}$ in our condition as well, but how to do so without running into dimensionality problems?
- The key here is to require the eigenvalue condition for only those $\boldsymbol{\delta}$ vectors that fall mostly, or at least partially, in the direction of $\boldsymbol{\beta}^*$
- **Theorem:** If $\lambda \geq \frac{2}{n}\|\mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty$, then

$$\|\boldsymbol{\delta}_{\mathcal{N}}\|_1 \leq 3\|\boldsymbol{\delta}_{\mathcal{S}}\|_1$$

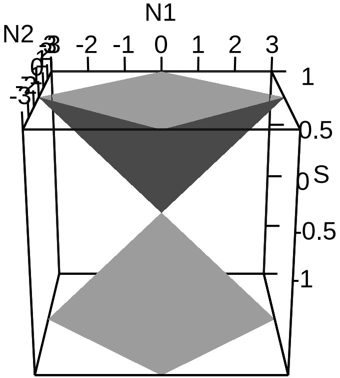
Examples

- For example, suppose $\mathbf{X}^\top\mathbf{X}/n$ looks like this:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix}$$

- We are in trouble if $\mathcal{S}$ contains either feature 2 or feature 3
- However, if $\mathcal{S} = \{1\}$ then there are no flat directions that lie within the lasso cones
- Second example: Suppose $\mathcal{S} = \{1\}$ and $\mathbf{x}_1 = \mathbf{x}_2 + \mathbf{x}_3 + \mathbf{x}_4$; then $L(\boldsymbol{\beta})$ would be perfectly flat in the direction $\boldsymbol{\delta} = (1, -1, -1, -1)$, with $\|\boldsymbol{\delta}_{\mathcal{N}}\|_1 \leq 3\|\boldsymbol{\delta}_{\mathcal{S}}\|_1$ satisfied – this kind of $\mathbf{X}$ must be ruled out also

Illustration



Restricted eigenvalue condition

- Let us now formally state the *restricted eigenvalue condition*, which I will denote RE(τ): There exists a constant $\tau > 0$ such that

$$\frac{\frac{1}{n}\boldsymbol{\delta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\delta}}{\|\boldsymbol{\delta}\|_2^2} \geq \tau$$

  for all nonzero $\boldsymbol{\delta} : \|\boldsymbol{\delta}_{\mathcal{N}}\|_1 \leq 3\|\boldsymbol{\delta}_{\mathcal{S}}\|_1$

- Note: This condition is specific to linear regression; the general condition is known as *restricted strong convexity* and would consist of replacing $\mathbf{X}^\top\mathbf{X}/n$ with $\nabla^2 L(\boldsymbol{\beta})$

Other conditions

This is certainly not the only condition that people have used to prove things in the high-dimensional setting; other similar conditions include

- Irrepresentable condition
- Restricted isometry property (RIP)
- Compatibility condition
- Coherence condition
- Sparse Riesz condition

All of these conditions require that $\mathbf{X}_{\mathcal{S}}$ is full rank as well as placing some sort of restriction on how strongly features in $\mathcal{S}$ can be correlated with features in $\mathcal{N}$

Estimation consistency

- With this condition in place, we're ready to prove the following theorem
- **Theorem:** Suppose $\mathbf{X}$ satisfies RE(τ) and $\lambda \geq \frac{2}{n}\|\mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty$; then

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{3}{\tau}\lambda\sqrt{|\mathcal{S}|}$$

- **Corollary:** Suppose $\mathbf{X}$ satisfies RE(τ), $\mathbf{y} = \mathbf{X}\beta^* + \boldsymbol{\varepsilon}$ with $\varepsilon_i \overset{\perp\!\!\!\perp}{\sim} \mathrm{N}(0, \sigma^2)$, and $\lambda = 2\sigma\sqrt{c\log(p)/n}$; then

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{6\sigma}{\tau}\sqrt{\frac{c\,|\mathcal{S}|\log p}{n}}$$

with probability $1 - 2\exp\{-\frac{1}{2}(c-2)\log p\}$

Remarks

- This rate makes a lot of sense:
  - The error of the oracle estimator is on the order $\sigma\sqrt{|\mathcal{S}|/n}$: no method can estimate $\mathcal{S}$ parameters based on n observations at a better rate than this
  - The $\log p$ term is the price we pay to search over p features in order to discover the sparse set $\mathcal{S}$
- Note also the dependence on the eigenvalue parameter τ; in particular, if the minimum eigenvalue is close to 0, the estimate rate will suffer significantly

Another look at prediction error

- Now that we've made some assumptions about $\mathbf{X}$ and β^*, does this affect our prediction accuracy?
- **Theorem:** Suppose $\mathbf{X}$ satisfies RE(τ) and $\lambda \geq \frac{2}{n}\|\mathbf{X}^\top \boldsymbol{\varepsilon}\|_\infty$; then

$$\frac{1}{n}\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2 \leq \frac{9}{\tau}\lambda^2 \, |\mathcal{S}|$$

- **Corollary:** Suppose $\mathbf{X}$ satisfies RE(τ), $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ with $\varepsilon_i \overset{\perp\!\!\!\perp}{\sim} \mathrm{N}(0, \sigma^2)$, and $\lambda = 2\sigma\sqrt{c\log(p)/n}$; then

$$\frac{1}{n}\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|_2^2 \leq 36c\frac{\sigma^2}{\tau}\frac{|\mathcal{S}|\log p}{n}$$

with probability $1 - 2\exp\{-\frac{1}{2}(c-2)\log p\}$

Remarks

- We have now derived two results concerning the prediction error of the lasso:
  - No assumptions on $\mathbf{X}$ or $\boldsymbol{\beta}^*$: $\text{MSPE} = O(n^{-1/2})$, the "slow rate"
  - $\boldsymbol{\beta}^*$ sparse, $\mathbf{X}$ satisfies RE(τ): $\text{MSPE} = O(n^{-1})$, the "fast rate"
- Further theoretical work has shown that these bounds are in fact tight: no method can achieve the fast rate without additional assumptions

Irrepresentable condition

- Finally, we'll take a look at the selection consistency of the lasso in high dimensions, although we're not going to have time to prove our result in class
- We begin by noting that our restricted eigenvalue condition is not enough to establish selection consistency; we need something stronger
- The feature matrix $\mathbf{X}$ satisfies the *irrepresentable condition*, which I will denote IR(τ), if there exists $\tau > 0$ such that

$$\max_{j \in \mathcal{N}} \|(\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}})^{-1} \mathbf{X}_{\mathcal{S}}^\top \mathbf{x}_j\|_1 \leq 1 - \tau$$

Remarks

- Note that this places an upper bound on the size of $(\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}}^\top \mathbf{x}_j$, the coefficient for regressing a null feature on the signal features
- In words, this is saying no noise feature can be highly "represented" by the true signal features; if this were the case, we might select the noise feature instead of the true signal
- For example, if $\mathbf{X}_{\mathcal{S}}$ and $\mathbf{X}_{\mathcal{N}}$ were orthogonal, then $\tau = 1$
- Note that
  - This is actually a fairly strong condition
  - IR(τ) requires $\mathbf{\Sigma}_{\mathcal{S}} = \frac{1}{n}\mathbf{X}_{\mathcal{S}}^\top \mathbf{X}_{\mathcal{S}}$ to be invertible; let ξ_* denote its minimum eigenvalue

Selection consistency theorem (Wainwright, 2009)

Theorem: Suppose that $\mathbf{X}$ satisfies IR(τ) and $\mathbf{y} = \mathbf{X}\beta^* + \varepsilon$ with $\varepsilon_i \overset{\perp\!\!\!\perp}{\sim} \mathrm{N}(0, \sigma^2)$; let

$$\lambda = \frac{8\sigma}{\tau}\sqrt{\frac{\log p}{n}}$$
$$B = \lambda\left(\frac{4\sigma}{\sqrt{\xi_*}} + \|\mathbf{\Sigma}_{\mathcal{S}}^{-1}\|_\infty\right)$$

Then with probability at least $1 - c_1 \exp\{-c_2 n\lambda^2\}$, the lasso solution $\widehat{\beta}$ has the following properties:

Selection consistency theorem (Wainwright, 2009) (cont'd)

- **Uniqueness:** $\widehat{\boldsymbol{\beta}}$ is unique
- **Estimation error bound:** $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_\infty \leq B$
- **No false inclusions:** $\hat{\mathcal{S}} \subseteq \mathcal{S}$
- **No false exclusions:** $\hat{\mathcal{S}}$ includes all indices j such that $|\beta_j^*| > B$ and is therefore selection consistent provided that all elements of $\boldsymbol{\beta}_{\mathcal{S}}^*$ are at least that large