Theoretical results: Classical setting

Patrick Breheny

March 25, 2025

Notation

- Our next topic will cover some theoretical results for the lasso, MCP, and SCAD
- There is a large body of literature on these results, which could easily fill an entire course on its own – we will just spend two lectures on this topic and focus on some important main results
- Notation:
  - Let $\boldsymbol{\beta}^*$ denote the (unknown) true value of $\boldsymbol{\beta}$
  - Let $\mathcal{S} = \{j : \beta_j^* \neq 0\}$ denote the set of nonzero coefficients (i.e., the *sparse set*), with $\boldsymbol{\beta}_\mathcal{S}$ and $\mathbf{X}_\mathcal{S}$ the corresponding subvector and submatrix
  - Let $\mathcal{N} = \{j : \beta_j^* = 0\}$ denote the set of "null" features.

Theoretical property #1: Estimation

- There are three main categories of theoretical results, concerning three desirable qualities we would like our estimator $\widehat{\boldsymbol{\beta}}$ to possess
- The first is that obviously, we would like our estimator to be close to the true value of $\boldsymbol{\beta}$; this is typically measured by mean squared (estimation) error:

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

- This may take the form of an asymptotic result such as $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 \to 0$, or in the form of a bound such as $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2^2 < B$, where B will typically depend on n, p, etc.

Theoretical property #2: Prediction

- A separate desirable property is that we would like our model to produce accurate predictions
- This is typically measured by mean squared prediction error:

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

- It is worth noting that although $\widehat{\boldsymbol{\beta}} \approx \boldsymbol{\beta}^* \implies \mathbf{X}\widehat{\boldsymbol{\beta}} \approx \mathbf{X}\boldsymbol{\beta}^*$, the converse is not true; thus, typically prediction consistency can occur under weaker conditions than estimation consistency

Theoretical property #3: Variable selection

- Finally, for a sparse model, we might also be interested in its properties as a variable selection method
- This can be measured a few different ways; one of them is sign consistency:

$$\mathrm{sign}(\widehat{\beta}_j) = \mathrm{sign}(\beta_j^*)$$

  with high probability
- This is the most challenging property to achieve, since $\widehat{\beta}_j$ and β_j^* may be very close, but if one of them is zero and the other is a small nonzero quantity, then they do not have the same sign

Asymptotic vs non-asymptotic settings

- Generally speaking, there are two sorts of theoretical results for high-dimensional regression models:
  - Classical/asymptotic results, in which p is fixed
  - Modern/non-asymptotic results, in which p increases with n, or in which finite-sample bounds are obtained
- The classical form of analysis, in which we treat the parameter as fixed (i.e., $\boldsymbol{\beta}^*$ is fixed), offers a number of interesting insights into the methods we have introduced so far, and is the setup we will be using today

Asymptotic setup: $p > n$

- However, these results also have the potential to be misleading, in that, if n increases while $\boldsymbol{\beta}$ remains fixed, in the limit we are always looking at $n \gg p$ situations; is this really relevant to $p \gg n$?
- For this reason, it is also worth considering theoretical analysis in which p is allowed to increase with n
- Typically, this involves assuming that the size of the sparse set, $|\mathcal{S}|$, stays fixed, and it is only the size of the null set that increases, so that $|\mathcal{S}| \ll n$ and $|\mathcal{N}| \gg n$; we will discuss this more next time

Sparsity regimes

- The setup we have been describing is sometimes referred to as "hard sparsity", in which $\boldsymbol{\beta}$ has a fixed, finite number of nonzero entries
- An alternative setup is to assume that most elements of $\boldsymbol{\beta}$ are small, but not necessarily exactly zero; i.e., assume something along the lines of letting $m = \max\{|\beta_j^*| : j \in \mathcal{N}\}$
- Yet another setup is to assume that $\boldsymbol{\beta}$ is not necessarily sparse, but is limited in size in the sense that $\sum_j |\beta_j^*| \leq R$ (i.e., within an ℓ_1 "ball" of radius R about $\mathbf{0}$)
- We will focus on the hard sparsity setting; many of the results are applicable to the other settings as well, however

Orthonormal case: Introduction

- We will begin our examination of the theoretical properties of the lasso by considering the special case of an orthonormal design: $\mathbf{X}^\top\mathbf{X}/n = \mathbf{I}$ for all n, with $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$ and $\varepsilon_i \overset{\perp\!\!\!\perp}{\sim} \mathrm{N}(0, \sigma^2)$
- For the sake of brevity, I'll refer to these assumptions in what follows as (O1)
- This might seem like an incredibly special case, but many of the important theoretical results carry over to the general design case provided some additional regularity conditions are met
- Once we show the basic results for the lasso, it is straightforward to extend them to MCP and SCAD

Eliminating all the variables in $\mathcal{N}$

- Let us begin by considering the question: how large must λ be in order to ensure that all the coefficients in $\mathcal{N}$ are eliminated?
- **Theorem:** Under (O1),

$$\mathbb{P}(\exists j \in \mathcal{N} : \widehat{\beta}_j \neq 0) \leq 2 \exp\left\{-\frac{n\lambda^2}{2\sigma^2} + \log p\right\}$$

Corollary

- So how large must λ be in order to accomplish this with probability 1?
- **Corollary:** Under (O1), if $\sqrt{n}\lambda \to \infty$, then

$$\mathbb{P}(\widehat{\beta}_j = 0 \, \forall j \in \mathcal{N}) \to 1$$

- Note that if instead $\sqrt{n}\lambda \to c$, where c is some constant, then $\mathbb{P}(\widehat{\beta}_j = 0 \, \forall j \in \mathcal{N}) \to 1 - \epsilon$, where $\epsilon > 0$
- In other words, if $\sqrt{n}\lambda$ is not large enough, there remains the possibility that the lasso will select variables from $\mathcal{N}$

A glimpse of $p \gg n$ theory

- Nevertheless, if $\lambda = O(\sigma\sqrt{n^{-1}\log p})$, then there is at least a chance of completely eliminating all variables in $\mathcal{N}$; setting λ to something of this order will come up often in our next lecture
- For now, we can note that unless p is growing exponentially fast with n, the ratio $\log(p)/n$ can still go to zero even if $p > n$, giving some insight into how high-dimensional regression is possible

Selecting all the variables in $\mathcal{S}$

- The previous theorem considered eliminating all of the variables in $\mathcal{N}$
- Likewise, we can ask: what is required in order for the lasso to select all of the variables in $\mathcal{S}$?
- **Theorem:** Under (O1), if $\lambda \to 0$ as $n \to \infty$, then

$$\mathbb{P}\{\text{sign}(\widehat{\beta}_j) = \text{sign}(\beta_j^*) \,\forall j \in \mathcal{S}\} \to 1$$

Selection consistency

- Putting these two theorems together, we obtain the asymptotic conditions necessary for selection consistency as $n \to \infty$
- For the lasso to be selection consistent (select the correct model with probability tending to 1), we require:
  - $\lambda \to 0$
  - $\sqrt{n}\lambda \to \infty$

Estimation consistency

- Let us now consider estimation consistency
- It is trivial to show that under (O1), $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}^*$ if $\lambda \to 0$: if $\lambda \to 0$, $\widehat{\boldsymbol{\beta}}$ converges to the OLS, which is consistent
- A more interesting condition is $\sqrt{n}$-consistency
- **Theorem:** Under (O1), $\widehat{\boldsymbol{\beta}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}^*$ if $\sqrt{n}\lambda \to c$, with $c < \infty$

Remarks

- **Corollary:** Suppose $\exists j : \beta_j^* \neq 0$. Then under (O1), $\widehat{\boldsymbol{\beta}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}^*$ if and only if $\sqrt{n}\lambda \to c$, with $c < \infty$
- In this case, $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)$ will contain a bias term on the order of $\sqrt{n}\lambda$, which will blow up unless λ rapidly goes to zero

Remarks (cont'd)

- It is possible for the lasso to be $\sqrt{n}$-consistent
- It is also possible for the lasso to be selection consistent
- However, it is not possible for the lasso to achieve both goals *at the same time*
- Specifically, we require $\sqrt{n}\lambda \to \infty$ for selection consistency, but $\sqrt{n}\lambda \to c < \infty$ for $\sqrt{n}$-estimation consistency
- As we will see soon, this is one of the main theoretical shortcomings of the lasso that methods such as MCP and SCAD aim to correct

Prediction and estimation in the orthonormal case

- In the orthonormal case, note that

$$\frac{1}{n}\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|^2 = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2$$

- Thus, since $\sqrt{n}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) = O_p(1)$ by our previous theory, we have the immediate corollary that if $\sqrt{n}\lambda \to c$, the prediction error is $O_p(n^{-1})$
- Prediction and estimation are not necessarily equivalent when features are correlated, however

Remarks

- Still, we see the connection between prediction and estimation – this suggests that if we use a prediction-based criterion such as cross-validation to choose λ, we emphasize estimation accuracy over selection accuracy
- In other words, cross-validation will tend to choose small values of λ; recall that if $\sqrt{n}\lambda \to c < \infty$,
  - All $\beta_j : j \in \mathcal{S}$ will be selected
  - Some $\beta_j : j \in \mathcal{N}$ will also be selected

Screening property

- This result (lasso with cross-validation selects all the true features, but also selects null features) is true in general, not just the orthonormal case
- This means that the lasso is not ideal if one desires a low false positive rate among the features selected by a model
- However, the lasso can be very useful for purposes of a screening tool to recover the important variables as the first step in an analysis such as the adaptive lasso

Extension to MCP and SCAD

- The lasso cannot simultaneously achieve both $\sqrt{n}$-consistency and selection consistency; MCP and SCAD, however, *can*
- In fact, they can achieve an even stronger result called the *oracle property*
- Let $\widehat{\beta}^*$ denote the oracle estimator:
  - $\widehat{\boldsymbol{\beta}}^*_{\mathcal{N}} = \mathbf{0}$
  - $\widehat{\boldsymbol{\beta}}^*_{\mathcal{S}}$ minimizes $\|\mathbf{y} - \mathbf{X}_{\mathcal{S}}\boldsymbol{\beta}_{\mathcal{S}}\|_2^2$
- **Theorem:** Under (O1), suppose $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$. Then $\widehat{\beta} = \widehat{\beta}^*$ with probability tending to 1, where $\widehat{\beta}$ is either the MCP or SCAD estimate.

More on the oracle property

- The oracle property is usually defined as: $\widehat{\boldsymbol{\beta}}$ must satisfy
  - $\widehat{\boldsymbol{\beta}}_{\mathcal{N}} = \mathbf{0}$ with probability tending to 1
  - $\widehat{\boldsymbol{\beta}}_{\mathcal{S}}$ is $\sqrt{n}$-consistent for $\boldsymbol{\beta}^*_{\mathcal{S}}$
- This broader definition encompasses the adaptive lasso as well
  - The adaptive lasso would never be exactly equal to the oracle estimator $\widehat{\boldsymbol{\beta}}^*$
  - However, with a consistent initial estimator, the bias term goes to zero, giving $\sqrt{n}$-consistency

General case: Introduction

- The essence of these results carries over to the case of a general design matrix, although we will need some new conditions regarding eigenvalues
- In what follows, I will refer to the following set of assumptions as (C1):
  - $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
  - $\varepsilon_i \overset{\perp\!\!\!\perp}{\sim} \mathrm{N}(0, \sigma^2)$
  - $\frac{1}{n}\mathbf{X}^\top\mathbf{X} = \boldsymbol{\Sigma}_n$, with $\boldsymbol{\Sigma}_n \to \boldsymbol{\Sigma}$
  - $\boldsymbol{\Sigma}$ has minimum eigenvalue ξ_* and maximum eigenvalue ξ^*

General case: $\sqrt{n}$-consistency

- For technical reasons, we must start our discussions of the general case with estimation (later proofs require the consistency result)
- **Theorem:** Under (C1), the lasso estimator $\widehat{\boldsymbol{\beta}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}^*$ if (i) $\sqrt{n}\lambda \to c$, with $c < \infty$ and (ii) $\xi_* > 0$.
- As in the orthonormal case, note that if $\sqrt{n}\lambda \to \infty$, the result no longer holds

General case: Prediction accuracy

- **Theorem:** Under (C1), if (i) $\sqrt{n}\lambda \to c$, with $c < \infty$ and (ii) $\xi_* > 0$, we have

$$\frac{1}{n}\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}^*\|^2 = O_p(n^{-1})$$

- You may be wondering: do we actually need $\xi_* > 0$ for prediction accuracy?
- Turns out the answer is no, you don't, although the prediction accuracy isn't quite as good if $\mathbf{X}$ is not full rank; we'll return to this point next time

MCP and SCAD in the general case: Consistency

- For MCP and SCAD, we can prove some stronger results
- First, we provide a corresponding consistency theorem; note the weaker condition on λ
- **Theorem:** Under (C1), $\widehat{\boldsymbol{\beta}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}^*$ if (i) $\lambda \to 0$ and (ii) $\xi_* > 0$, where $\widehat{\boldsymbol{\beta}}$ is an MCP or SCAD estimator
- Note: I say "an" estimator rather than "the" estimator since what we're actually proving is that there exists a local minimizer of the MCP/SCAD objective with $\sqrt{n}$-consistency

MCP and SCAD in the general case: Oracle property

- Based on this result, we can also prove that MCP and SCAD enjoy the oracle property in the general case:
- **Theorem:** Under (C1), if (i) $\lambda \to 0$, (ii) $\sqrt{n}\lambda \to \infty$, and (iii) $\xi_* > 0$, then $\widehat{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}}^*$ with probability tending to 1, where $\widehat{\boldsymbol{\beta}}$ is an MCP or SCAD estimator