Theoretical results for lasso, MCP, and SCAD

Patrick Breheny

March 2

Introduction

- There is an enormous body of literature concerning theoretical results for high-dimensional penalized regression
- Our goal for today is to get an introduction to these results, focusing on proving some interesting, relevant results in relatively simple cases
- Time permitting, we may return to this topic later in the course and cover some additional extensions

Notation

- In today's lecture, we will let $\widehat{\boldsymbol{\beta}}$ denote the estimator in question and $\boldsymbol{\beta}_0$ denote the (unknown) true value of $\boldsymbol{\beta}$
- We will let $\mathcal{S} = \{j : \boldsymbol{\beta}_{0j} \neq 0\}$ 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
- Similarly, we will let $\mathcal{N} = \{j : \boldsymbol{\beta}_{0j} = 0\}$ denote the set of "null" coefficients equal to zero

Types of results

There are three main categories of theoretical results, depending on various qualities we would like our estimator to possess:

- *Prediction* The mean squared prediction error is small:

$$\frac{1}{n}\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_0\|^2$$

- *Estimation* The mean squared error is small:

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$$

- *Selection* The probability that $\operatorname{sign}(\widehat{\beta}_j) = \operatorname{sign}(\beta_{0j})$ for all j is large

Asymptotic setup: Fixed p

- As often in statistics, closed-form results for finite sample sizes are typically difficult to obtain, so we focus on asymptotic results as $n \to \infty$
- Classically, we would treat $\boldsymbol{\beta}$ as fixed and consider the behavior of $\widehat{\boldsymbol{\beta}}$ as n grows
- This offers a number of interesting insights, and is the setup we will mainly be sticking to 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, many researchers prefer instead to consider the high-dimensional case where 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 $|\mathcal{N}| \gg n$

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 $m = \max\{|\beta_{0j}| : j \in \mathcal{N}\}$
- Yet another setup is to assume that $\boldsymbol{\beta}$ is not necessarily sparse, but is limited in size in the sense that $\sum_j |\beta_{0j}| \leq R$ (i.e., within an ℓ_1 "ball" of radius R about $\mathbf{0}$)

Orthonormal case: Introduction

- We will begin our examination of the theoretical properties of the lasso by considering the special case of an orthonormal design: $\mathbf{X}^T\mathbf{X}/n = \mathbf{I}$ for all n, with $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon$ and $\varepsilon_i \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2)$
- 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

Theorem: Correct sparsity

- In this setting, it would seem possible for the lasso to set λ high enough that all the coefficients in $\mathcal{N}$ are eliminated
- How large must λ be in order to accomplish this?
- **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:** Assume $O1$. 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, even with an infinite amount of data, there is still the possibility that the lasso will select some variables from the null set $\mathcal{N}$

A glimpse of $p \gg n$ theory

- It is worth mentioning that 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 comes up very often in extending theoretical results to the case where p is allowed to grow with n, and gives us a glimpse of how it is possible to carry out statistical analyses in this setting
- Specifically, unless p is growing exponentially fast with n, the ratio $\log(p)/n$ can still go to zero even if $p > n$

Selection consistency

- Likewise, we can ask: what is required in order for the lasso to select all of the variables in $\mathcal{S}$?
- **Theorem:** Suppose $O1$ and $\lambda \to 0$ as $n \to \infty$. Then

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

- Note that it is possible to satisfy $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$ simultaneously; i.e., for the lasso to be selection consistent (select the correct model with probability tending to 1)

Estimation consistency

- Let us now consider estimation consistency
- **Theorem:** Under $O1$, $\widehat{\boldsymbol{\beta}}$ is a consistent estimator of $\boldsymbol{\beta}_0$ if $\lambda \to 0$.
- A more demanding condition is $\sqrt{n}$-consistency
- **Theorem:** Under $O1$, $\widehat{\boldsymbol{\beta}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}_0$ if and only if $\sqrt{n}\lambda \to c$, with $c < \infty$

Remarks

- It is possible for the lasso to be both selection consistent and $\sqrt{n}$-consistent for estimation
- However, it is not possible to achieve both goals *at the same time*
- Specifically, we require $\sqrt{n}\lambda \to \infty$ to correctly select the model with probability 1, but we require $\lambda = O(n^{-1/2})$ for $\sqrt{n}$-consistency

Lasso recovery and variable screening

- Note that in the orthonormal case,

$$\frac{1}{n}\|\mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}_0\|^2 = \|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|^2$$

- The tendency, then, if use a prediction-based criterion such as cross-validation to choose λ is that we emphasize estimation accuracy and select λ values for which the probability of allowing null coefficients into the model is high (this is the case for non-orthonormal $\mathbf{X}$ as well)
- This means that lasso models tend not to be as sparse as the ideal model would be, although it does make the lasso useful for variable screening (as in the adaptive lasso and other procedures), as it recovers the true variables with high probability

Extension to MCP and SCAD

- It *is* possible, however, to achieve both $\sqrt{n}$-consistency and selction consistency simultaneously with MCP and SCAD, however
- **Theorem:** Under $O1$, $\widehat{\boldsymbol{\beta}}$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}_0$ if $\lambda \to 0$, where $\widehat{\boldsymbol{\beta}}$ is either the MCP or SCAD estimate
- As we previously noted, it is possible to satisfy $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$ simultaneously
- A related result can also be shown for the adaptive lasso, although we will not prove it in class

General case: Introduction

- The essence of these results carries over to the case of a general design matrix and a general likelihood, although additional regularity conditions are required
- Generally speaking, these are the basic regularity conditions required to ensure asymptotic normality of the MLE: common support, identifiability, the Fisher information $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})$ is positive definite at $\boldsymbol{\beta}_0$, and all third derivatives of the log-likelihood are bounded
- It is worth mentioning that these regularity conditions need to be revised substantially if we allow $p > n$, since $\boldsymbol{\mathcal{I}}(\boldsymbol{\beta})$ cannot be positive definite in that case
- In what follows, I will refer to this set of assumptions as $G1$

Notation

- We will present and prove some results from Fan & Li's 2001 paper introducing the SCAD estimator, which concern general likelihoods and general penalties (i.e., the same theorem will apply to lasso, SCAD, and MCP)
- Let $\mathbf{v} = (p'(|\beta_{0j}|)\text{sign}(\beta_{0j}))_{j=1}^{p}$, with $v = \max_{j \in \mathcal{S}} |v_j|$, where p is the penalty function
- Likewise, let $\mathbf{A} = \text{diag}\{p''(|\beta_{0j}|)\}_{j=1}^{p}$, with $a = \max_{j \in \mathcal{S}} |a_{jj}|$

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

- Fan & Li prove three key theorems in their seminal paper; the first concerns $\sqrt{n}$-consistency
- **Theorem 1:** Under $G1$, suppose that $\lambda \to 0$ and $a \to 0$. Then there exists a local maximizer of the objective function Q such that

$$\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(n^{-1/2} + v)$$

General case: Sparsity

- Their second theorem concerns the sparsity of $\widehat{\boldsymbol{\beta}}$
- **Theorem 2:** Suppose the conditions of Theorem 1 are met, with $\lambda \to 0$, $\sqrt{n}\lambda \to \infty$, and $\lim_{\theta \to 0+} p'(\theta) = \lambda$. Then with probability tending to 1, $\widehat{\boldsymbol{\beta}}_{\mathcal{N}} = \mathbf{0}$ is a minimizer of $Q(\boldsymbol{\beta})$

General case: Asymptotic normality

- Their final result concerns the asymptotic normality of $\widehat{\boldsymbol{\beta}}_{\mathcal{S}}$
- **Theorem 3:** Suppose that the conditions of Theorem 2 are met, with $\lambda \to 0$ and $\sqrt{n}\lambda \to \infty$. Then

$$\sqrt{n}(\boldsymbol{\mathcal{I}}_{\mathcal{S}} + \mathbf{A}_{\mathcal{S}})(\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{0\mathcal{S}}) + \sqrt{n}\mathbf{v}_{\mathcal{S}} \xrightarrow{\text{d}} \text{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}_{\mathcal{S}}),$$

where $\boldsymbol{\mathcal{I}}_{\mathcal{S}}$ is the Fisher information for $\boldsymbol{\beta}_{\mathcal{S}}$ knowing that $\boldsymbol{\beta}_{\mathcal{N}} = 0$

Corollary: MCP and SCAD

- Note that for MCP and SCAD, $\mathbf{A}_{\mathcal{S}} \to \mathbf{0}$ and $\mathbf{v}_{\mathcal{S}} \to \mathbf{0}$ as $\lambda \to 0$
- Thus, for MCP and SCAD, the result of Theorem 3 simplifies to

$$\sqrt{n}\boldsymbol{\mathcal{I}}_{\mathcal{S}}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{0\mathcal{S}}) \xrightarrow{\text{d}} \text{N}(\mathbf{0}, \boldsymbol{\mathcal{I}}_{\mathcal{S}})$$

- Note that this is the same asymptotic result we have for the "oracle estimator", in which we know in advance which coefficients are zero and which ones are not, and maximum likelihood is applied using only the nonzero variables

Oracle property and corollary for lasso

- This property, in which asymptotically, an estimator performs as well as the oracle MLE, is known as the *oracle property*
- Note that the lasso does not have the oracle property:
  - For the lasso, $\mathbf{v}_{\mathcal{S}} = \lambda \mathbf{s}_{\mathcal{S}}$, where $\mathbf{s}_{\mathcal{S}} = (\text{sign}(\widehat{\beta}_j))_{j \in \mathcal{S}}$
  - Thus, if $\sqrt{n}\lambda \to \infty$, the $\sqrt{n}\mathbf{v}_{\mathcal{S}}$ term in the final theorem goes to infinity and $\sqrt{n}(\widehat{\boldsymbol{\beta}}_{\mathcal{S}} - \boldsymbol{\beta}_{0\mathcal{S}})$ no longer converges to a normal distribution