Nonconvex penalties: Signal-to-noise ratio and algorithms

Patrick Breheny

March 2

Introduction

- In today's lecture, we will discuss the performance of nonconvex penalties with respect to the signal-to-noise ratio of the data-generating process, the most critical factor determining their success relative to the lasso
- We will then turn our attention to the details of model fitting, discussing algorithms for nonconvex penalties as well as the impact of nonconvexity on model-fitting

Signal to noise ratio

- For linear regression,

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{Var}(\mathrm{E}(Y|X)) + \mathrm{E}(\mathrm{Var}(Y|X)) \\ &= \boldsymbol{\beta}^\top \mathrm{Var}(X)\boldsymbol{\beta} + \sigma^2\end{aligned}$$

- The first term in the sum is known as the *signal* and the second term the *noise*
- Thus, we may define the *signal-to-noise ratio*

$$\mathrm{SNR} = \boldsymbol{\beta}^\top \mathrm{Var}(X)\boldsymbol{\beta}/\sigma^2$$

SNR and R^2

- Recall that we have seen this decomposition before, in calculating R^2, which is also a function of the signal and noise
- In particular, note that

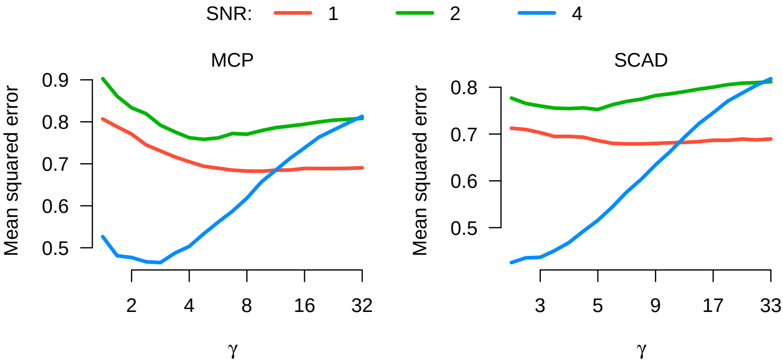
$$R^2 = \frac{\text{SNR}}{1 + \text{SNR}}$$

- As a general piece of advice, I strongly recommend considering the signal-to-noise ratio when designing simulations, and avoiding settings where SNR is, say, 50 ($R^2 = .98$); is this realistic?

Simulation: Setup

- To see the impact of SNR, let's set $n = 50$, $p = 100$, and let all features $\mathbf{x}_j$ follow independent, standard Gaussian distributions
- In the generating model, we set $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_6 \neq 0$ and $\beta_7 = \beta_8 = \cdots = \beta_{100} = 0$, varying the nonzero values of β_1 through β_6 to produce a range of signal to noise ratios
- For each data set, an independent data set of equal size was generated for the purposes of selecting the regularization parameter

Simulation: Results



Remarks

- The motivation of MCP/SCAD/etc. is to eliminate bias for large coefficients; it should not come as little surprise, then, that the advantage of these methods only becomes apparent when some nonzero coefficients are large
- It is also worth noting that $\gamma \approx 3$ is generally a reasonable choice for MCP – its performance was never far from the best
- Also note that the SCAD is somewhat less sensitive to the choice of γ, in the sense that many values of γ produce rather lasso-like estimates

Algorithm

Letting $\tilde{z} = n^{-1}\mathbf{x}_j^\top\tilde{\mathbf{r}}_j$, F is the firm-thresholding operator, and T_{SCAD} is the SCAD-thresholding operator, the CD algorithm for MCP/SCAD is

```
repeat
    for $j = 1, 2, \ldots, p$
        $\tilde{z}_j = n^{-1}\sum_{i=1}^n x_{ij}r_i + \widetilde{\beta}_j^{(s)}$
        $\widetilde{\beta}_j^{(s+1)} \leftarrow \begin{cases} F(\tilde{z}_j|\lambda,\gamma) & \text{for MCP, or} \\ T_{\text{SCAD}}(\tilde{z}_j|\lambda,\gamma) & \text{for SCAD} \end{cases}$
        $r_i \leftarrow r_i - (\widetilde{\beta}_j^{(s+1)} - \widetilde{\beta}_j^{(s)})x_{ij}$ for all $i$
until convergence
```

The algorithm is identical to our earlier algorithm for the lasso except for the step in which $\widetilde{\beta}_j$ is updated

Convergence

- Although the MCP and SCAD penalties are not convex functions, $Q(\beta_j|\boldsymbol{\beta}_{-j})$ is still convex
- As a result, the coordinate-wise updates are unique and always occur at the global minimum with respect to that coordinate
- **Proposition:** Let $\{\boldsymbol{\beta}^{(s)}\}$ denote the sequence of coefficients produced at each iteration of the coordinate descent algorithms for SCAD and MCP. For all $s = 0, 1, 2, \ldots$,

$$Q(\boldsymbol{\beta}^{(s+1)}) \leq Q(\boldsymbol{\beta}^{(s)}).$$

Furthermore, the sequence is guaranteed to converge to a local minimum of $Q(\boldsymbol{\beta})$.

Local linear approximation

- For MCP and SCAD, one can obtain closed-form coordinate-wise minima and use those solutions as updates
- An alternative approach, which is particularly useful in penalties that do not yield tidy closed-form solutions, is to construct a local approximation of the penalty about a point $\widetilde{\beta}$:

$$P(|\beta|) \approx P(|\widetilde{\beta}|) + \dot{P}(|\widetilde{\beta}|)(|\beta| - |\widetilde{\beta}|)$$

- Note that with this approximation, the penalty takes on the form of the lasso penalty (with $\dot{P}(|\widetilde{\beta}|)$ playing the role of the regularization parameter) plus a constant

LLA algorithm

- The approximation is applied in an iterative fashion: at the sth iteration, letting $\tilde{\lambda}_j = \dot{P}(|\beta_j^{(s-1)}|)$, the update is given by solving for the value minimizing

$$\frac{1}{2n}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \sum_{j=1}^{p} \tilde{\lambda}_j |\beta_j|$$

- Note that this equation is essentially identical to the one for the adaptive lasso; however, the adaptive lasso weights are assigned in a more or less ad hoc fashion based on an initial estimator, while the LLA modifications to λ are explicitly determined by the penalty function P

Remarks

- Like coordinate descent, the local linear approximation (LLA) algorithm is guaranteed to drive the objective function downhill with every iteration and to converge to a local minimum of $Q(\boldsymbol{\beta})$
- For MCP and SCAD, CD is more efficient, as it avoids the extra approximation introduced by LLA
- However, LLA is still quite efficient, and a valuable alternative when dealing with penalties without a simple solution in the one-dimensional case

Convexity challenges

- While the objective functions for SCAD and MCP are convex in each coordinate dimension, they are not convex over $\mathbb{R}^p$
- Thus, multiple minima may exist, each satisfying the KKT conditions
- Neither the CD or LLA algorithms are guaranteed to converge to the global minimum in such cases
- As we have discussed earlier, the existence of multiple minima poses considerable problems for MLE / penalized MLE methods, both numerically (convergence to an inferior solution) and statistically (increased variance as the solution jumps from one minima to another)

Global convexity

- We begin by noting that it is possible for the objective function Q to be convex with respect to $\boldsymbol{\beta}$ even though the penalty component is nonconvex
- Letting $c_{\min}$ denote the minimum eigenvalue of $\mathbf{X}^\top\mathbf{X}/n$, the MCP objective function is strictly convex if $\gamma > 1/c_{\min}$, while the SCAD objective function is strictly convex if $\gamma > 1 + 1/c_{\min}$
- In this case, the coordinate descent and LLA algorithms will converge to the unique global minimum of Q

Is global convexity desirable?

- However, obtaining strict convexity is not always possible or desirable; for example, in high-dimensional settings where $p > n$, $c_{\min} = 0$ and the MCP/SCAD objective functions cannot be globally convex
- Nevertheless, as we saw in the earlier simulations (where $p > n$), it is not true in general that convex penalties outperform nonconvex ones in such scenarios
- For low signal-to-noise ratios there was indeed some benefit to increasing γ in an effort to make the objective function more convex; however, for larger SNR values, this strategy diminished estimation accuracy

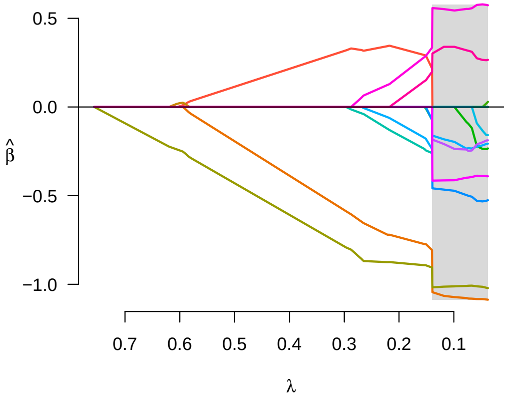
Local convexity

- One reason this happens is that the solutions are sparse
- Although $Q(\boldsymbol{\beta})$ may not be convex over the entire p-dimensional parameter space (i.e., *globally convex*), it is still convex on many lower-dimensional spaces
- If these lower-dimensional spaces contain the solution of interest, then the existence of other local minima in much higher dimensions may not be relevant
- This concept is known as *local convexity*

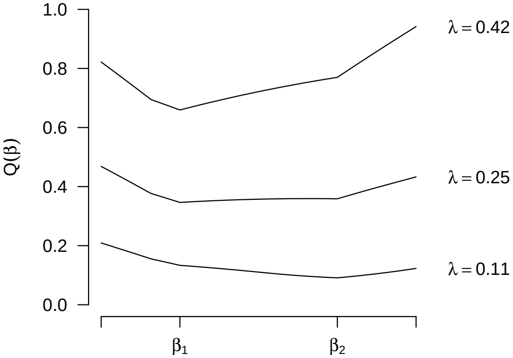
Local convexity: Details

- Recall the conditions for global convexity: γ must be greater than $1/c_{\min}$ for MCP and $1+1/c_{\min}$ for SCAD, where $c_{\min}$ denoted the minimum eigenvalue of $\mathbf{X}^\top\mathbf{X}/n$
- A straightforward modification is to include only the covariates with nonzero coefficients (the covariates which are "active" in the model) in the calculation of $c_{\min}$
- Note that neither γ nor $\mathbf{X}$ change with λ; what does vary with λ is the set of active covariates; generally speaking, this will increase as λ decreases
- Thus, local convexity of the objective function will not be an issue for large λ, but may cease to hold as λ is lowered past some critical value λ^*

Convexity diagnostic: Example (MCP)



Convexity diagnostic: Example (cont'd)



Remarks

- As the second figure indicates, when $\lambda = 0.42$, $\boldsymbol{\beta}_1$ clearly minimizes the objective function and when $\lambda = 0.11$, $\boldsymbol{\beta}_2$ clearly minimizes the objective function
- For $\lambda \approx 0.25$, however, the objective function is very broad and flat, indicating substantial uncertainty about which solution is preferable
- Calculation of the locally convex region (the unshaded region in the earlier figure) can be a useful diagnostic in practice to indicate which regions of the solution path may suffer from multiple local minima and discontinuous paths