Elastic Net: Algorithms and case study

Patrick Breheny

March 13

Introduction

- Today's lecture will finish our discussion of the elastic net and its nonconvex analogs: algorithms, R code, and two high-dimensional case studies
- The coordinate descent algorithms for all of the elastic net-type methods from the previous lecture (Lasso + ridge, SCAD + ridge, MCP + ridge) are very similar to the coordinate descent algorithms we have previously described
- The only step that differs is the updating of $\widetilde{\beta}_j$, which uses the orthonormal solutions we derived previously

Convexity considerations

- Before moving on, however, it is worth revisiting the issue of convexity for the ridge-stabilized versions of MCP and SCAD
- In the orthogonal case, the objective function is strictly convex if

$$\text{MCP:} \qquad \gamma > \frac{1}{1+\lambda_2}$$

$$\text{SCAD:} \qquad \gamma > 1 + \frac{1}{1+\lambda_2}$$

- As we increase the ridge penalty regularization parameter λ_2, the objective function becomes increasingly convex
- Or, to put it differently, by increasing λ_2 we increase the range of γ values over which the objective function remains convex

Convexity considerations: General case

- The corresponding equations for convexity in the general (non-orthogonal) case are:

$$\text{MCP:} \qquad \gamma > \frac{1}{c_{\min} + \lambda_2}$$

$$\text{SCAD:} \qquad \gamma > 1 + \frac{1}{c_{\min} + \lambda_2}$$

- Last week, we discussed increasing γ to maintain the stability of the objective function and prevent discontinuous jumps between local minima along the solution path
- Here, we see that another way to accomplish that same goal is by introducing a ridge component; we will explore this further in the upcoming case studies

BRCA gene expression study

- To illustrate the performance of ridge-stabilizing penalties in practice, as well as how to fit them using available software, we begin by revisiting our running example involving breast cancer gene expression data
- In both `glmnet` and `ncvreg`, there is an `alpha` option that can be used to control the balance between lasso and ridge penalties, as in the reparameterization introduced in the previous lecture
- In what follows, we will compare elastic net and Mnet models in terms of their predictive accuracy and number of features selected for $\alpha \in \{1, 0.75, 0.5, 0.25\}$

Model fitting: R code

```
# Elastic net
cvfit1 <- cv.glmnet(X, y)
cvfit2 <- cv.glmnet(X, y, alpha=0.75)
cvfit3 <- cv.glmnet(X, y, alpha=0.5)
cvfit4 <- cv.glmnet(X, y, alpha=0.25)

# Mnet
cvfit5 <- cv.ncvreg(X, y)
cvfit6 <- cv.ncvreg(X, y, alpha=0.75)
cvfit7 <- cv.ncvreg(X, y, alpha=0.5)
cvfit8 <- cv.ncvreg(X, y, alpha=0.25)
```

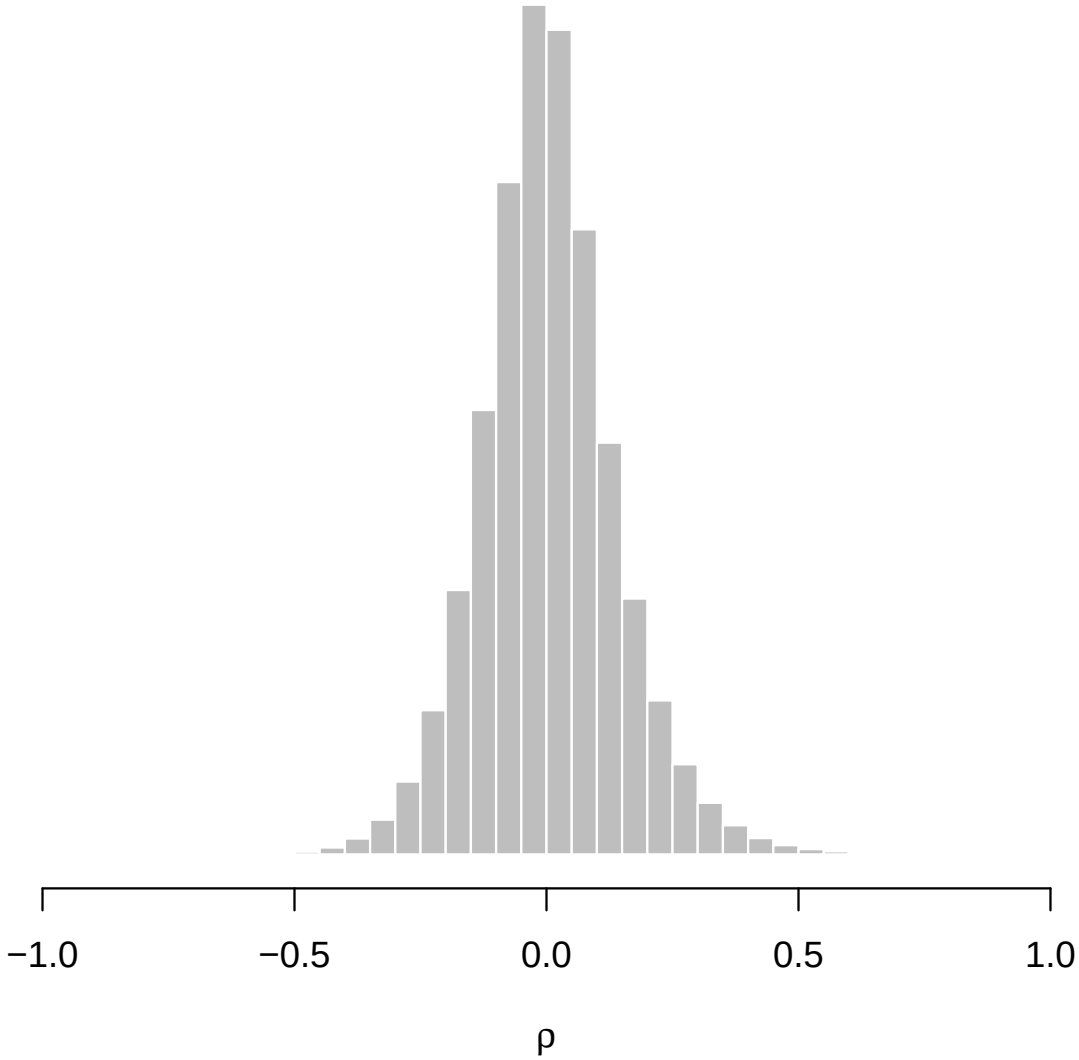
Results

	$\hat{R}^2$	Variables selected
Elastic Net		
$\alpha = 1$	0.60	49
$\alpha = 0.75$	0.60	57
$\alpha = 0.5$	0.60	63
$\alpha = 0.25$	0.60	82
Mnet		
$\alpha = 1$	0.55	28
$\alpha = 0.75$	0.56	27
$\alpha = 0.5$	0.57	37
$\alpha = 0.25$	0.58	35

Remarks

- In this example, the overall predictive accuracy for each approach is virtually identical across all the values of α considered here
- The solutions themselves, however, are quite different: by increasing the proportion of the penalty allocated to the ridge component, the number of variables selected went up by 67% for the elastic net as we dropped α from 1 to 0.25
- A similar trend holds for Mnet, although not as pronounced

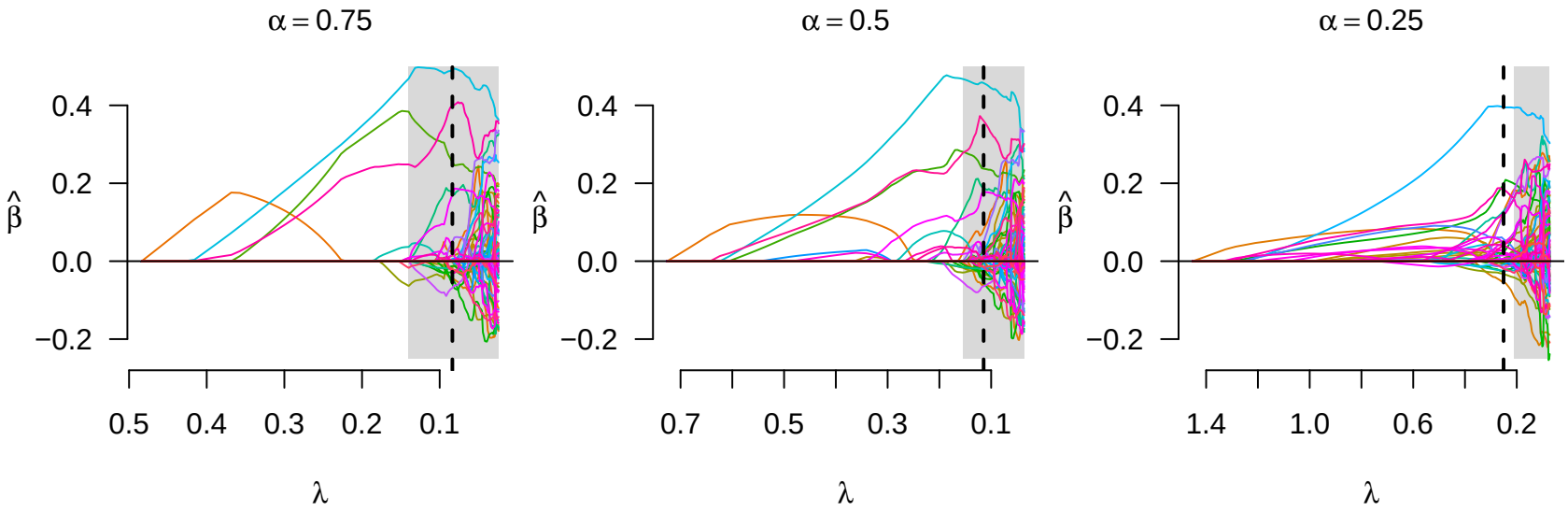
bcTCGA: Correlation



Remarks (cont'd)

- These results are essentially consistent with the simulation study of the previous lecture, in which the overall estimation accuracy of the lasso and elastic net were seen to be similar in the absence of strong correlation
  - For the breast cancer data, 99% of the pairwise correlations between genes were less than 0.4 in absolute value.
- Nevertheless, it is worth noting that the ridge component stabilizes the Mnet solutions in terms of reducing concerns about local minima

Mnet stability



The effect is similar to increasing γ; increasing α is often more effective, although in this case there isn't much difference

Rat eye data: Introduction

- The breast cancer data from the previous section was not particularly highly correlated, nor did it suggest highly sparse solutions (≈ 50 or more selected coefficients)
- As a contrast, let's look at a different study of gene expression data, this time gathered from the eye tissue of 120 twelve-week-old male rats
- The goal of the study was to detect genes whose expression patterns are related to that of the gene TRIM32, a gene known to be linked to a genetic disorder called Bardet-Biedl Syndrome (which, among other symptoms, leads to a number of problems with vision and proper formation of the retina)

Variable screening

- In the study, attention was restricted to the 5,000 genes with the largest variances in expression (on the log scale)
- It is worth taking a moment to discuss this strategy of variable screening
- Computationally, of course, we can certainly fit a model with all 18,975 genes
- From a statistical standpoint, however, it is often advantageous to reduce p – provided that the screening is effective at removing null features, we have increased power to detect the important features

Variable screening (cont'd)

- A variety of criteria for screening variables may be used in practice:
  - Variance (or, for categorical variables, frequency of cases)
  - Abundance: we may wish to screen out genes that are not expressed at high levels in eye tissue
  - Prior knowledge: e.g., genes that have been found to be associated with other eye diseases
  - Correlation: we may wish to screen out features that are highly correlated, retaining just one out of a correlated group
- It is also worth mentioning one criteria that often causes problems: screening based on univariate associations with the response, which typically invalidates downstream inferences

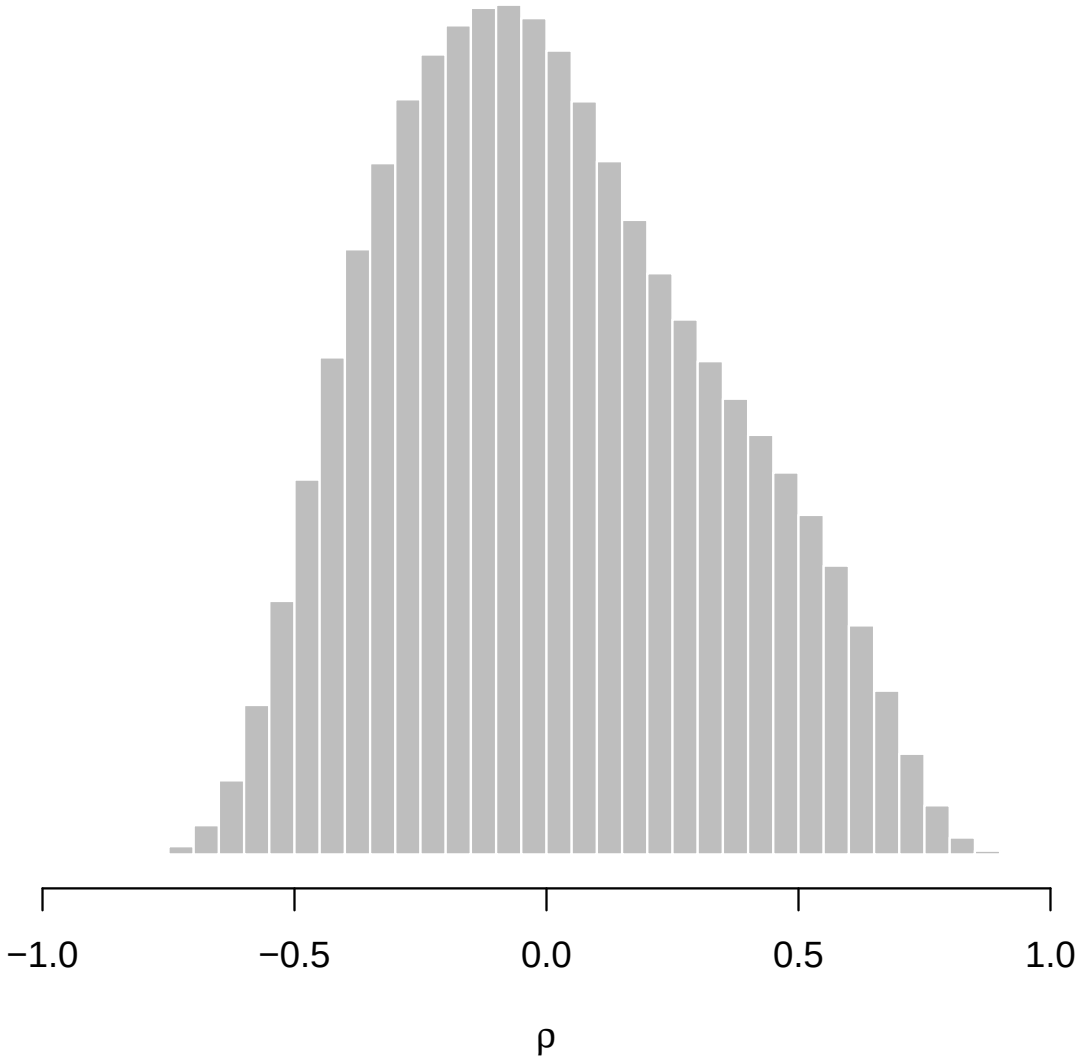
TRIM32 analysis for rat eye samples

- Returning to the present study, after screening we have $n = 120$ and $p = 5,000$
- Let's apply the same 8 models from the previous section to this data

Results

	$\hat{R}^2$	Variables selected
Elastic Net		
$\alpha = 1$	0.58	14
$\alpha = 0.75$	0.57	18
$\alpha = 0.5$	0.56	28
$\alpha = 0.25$	0.56	46
Mnet		
$\alpha = 1$	0.46	9
$\alpha = 0.75$	0.47	12
$\alpha = 0.5$	0.50	13
$\alpha = 0.25$	0.61	15

Scheetz2006: Correlation



Remarks

- This data differs from the breast cancer data in two important ways:
  - The variables are considerably more highly correlated with each other: only 76% of the pairwise correlations are below 0.4 in absolute value, and 8% of the correlations are above 0.6
  - We are able to identify accurate predictive models that include only a rather small number of features – perhaps as few as 9
- As a consequence, the incorporation of a ridge penalty has a larger impact in this setting than it did in the previous one, at least for MCP

Remarks (cont'd)

- Although MCP had inferior predictive accuracy compared to the lasso ($\hat{R}^2 = 0.46$ versus $\hat{R}^2 = 0.58$), lowering α substantially increased the predictive accuracy to $\hat{R}^2 = 0.61$
- The incorporation of a ridge penalty did not seem to benefit the lasso, although as usual it does affect the estimates and produce a more dense (less sparse) model
- The Mnet estimator with $\alpha = 0.25$ is particularly attractive here, as it achieve the best prediction accuracy out of all models considered, and does so using only 15 features (out of 5,000)

Theory for penalized regression

- 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 $\mathcal{S} = \{j : \boldsymbol{\beta}_j^* \neq 0\}$ 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 $\mathcal{N} = \{j : \boldsymbol{\beta}_j^* = 0\}$ 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:

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

- This may take the form of an asymptotic result such as $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2 \to 0$, or in the form of a bound such as $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2 < B$, 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:

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

- 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:

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

  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