Stability and the elastic net

Patrick Breheny

March 15

Introduction

- Our last several lectures have concentrated on methods for reducing the bias of lasso estimates
- This week, we will discuss methods for doing the opposite: introducing ridge penalties in order to reduce the variance of lasso estimates at the cost of further increasing their bias
- As we saw when discussing the ridge penalty itself, there is typically some degree of shrinkage we can introduce for which the gains of variance reduction outweigh the cost of increased bias to produce more accurate estimates

Elastic net penalty

- Recall that lasso solutions are not always unique; for coordinate descent in particular, this means that we will obtain different estimates depending on how the features are ordered (rather undesirable)
- However, solutions to ridge regression are always unique
- Consider, then, the following penalty, known as the *elastic net* penalty:

$$P_\lambda(\boldsymbol{\beta}) = \lambda_1 \|\boldsymbol{\beta}\|_1 + \frac{\lambda_2}{2}\|\boldsymbol{\beta}\|_2^2,$$

  which consists of two terms, a lasso term plus a ridge term

Example (revisited)

- Because the ridge penalty is strictly convex, the elastic net solution $\widehat{\beta}$ is unique provided that $\lambda_2 > 0$
- To see how this works, let us revisit the example from our first lecture on the lasso, which consisted of two observations: $(y_1, x_{11}, x_{12}) = (1, 1, 1)$ and $(y_2, x_{21}, x_{22}) = (-1, -1, -1)$
- For $\lambda < 1$, the lasso admits infinitely many solutions along the line $\beta_1 + \beta_2 = 1 - \lambda$ in the $\beta_1 > 0, \beta_2 > 0$ quadrant

Example (revisited, continued)

- In contrast, the elastic net penalty always yields a unique solution:

$$\begin{cases} \widehat{\beta}_1 = \widehat{\beta}_2 = 0 & \text{if } \lambda_1 \geq 1, \\ \widehat{\beta}_1 = \widehat{\beta}_2 = \dfrac{1-\lambda_1}{2+\lambda_2} & \text{if } \lambda_1 < 1. \end{cases}$$

- Note that regardless of λ_1 and λ_2, $\widehat{\beta}_1$ is always equal to $\widehat{\beta}_2$; this is reasonable, given that $\mathbf{x}_1 = \mathbf{x}_2$
- Indeed, this is a general property of the elastic net: whenever $\mathbf{x}_j = \mathbf{x}_k$, $\widehat{\beta}_j = \widehat{\beta}_k$

Remarks

- The example also illustrates that the elastic net retains properties of both the lasso and ridge regression methods
- From the lasso, it inherits sparsity – in particular, $\boldsymbol{\beta} = \mathbf{0}$ if $\lambda_1 > 1$
- From ridge regression, the elastic net inherits the ability to always produce a unique solution as well as ridge regression's property of proportional shrinkage: $\widehat{\beta}_1 = \widehat{\beta}_2 = (1 - \lambda_1)/(2 + \lambda_2)$ for elastic net, compared to $\widehat{\beta}_1 = \widehat{\beta}_2 = (1 - \lambda_1)/2$ for (one possible solution of) the lasso

Reparameterization

- A common reparameterization of the elastic net is to express the regularization parameters in terms of λ, which controls the overall degree of regularization, and α, which controls the balance between the lasso and ridge penalties:

$$\begin{aligned} \lambda_1 &= \alpha\lambda \\ \lambda_2 &= (1-\alpha)\lambda \end{aligned}$$

- This reparameterization is useful in practice, as it allows one to fix α and then select a single tuning parameter λ, which is more straightforward than attempting to select λ_1 and λ_2 separately

Orthonormal solutions: Introduction

- As with several other penalties we have considered, the elastic net has a closed form solution in the orthonormal case
- Considering this special case lends considerable insight into the nature of the and in addition, proves useful for optimization via the coordinate descent algorithm

KKT conditions

- The KKT conditions, or penalized likelihood equations, are given by:

$$\begin{aligned} \frac{1}{n}\mathbf{x}_j^T(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}) - \lambda_2\widehat{\beta}_j &= \lambda_1 \text{sign}(\widehat{\beta}_j) & \qquad \widehat{\beta}_j &\neq 0 \\ \frac{1}{n}|\mathbf{x}_j^T(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})| &\leq \lambda_1 & \qquad \widehat{\beta}_j &= 0 \end{aligned}$$

- Simplifying these conditions for the orthonormal case yields

$$\begin{aligned} z_j - \widehat{\beta}_j - \lambda_2\widehat{\beta}_j &= \lambda_1 \text{sign}(\widehat{\beta}_j) & \qquad \widehat{\beta}_j &\neq 0 \\ |z_j| &\leq \lambda_1 & \qquad \widehat{\beta}_j &= 0, \end{aligned}$$

where $z_j = \mathbf{x}_j^T\mathbf{y}/n$

Orthonormal solutions

- These equations can be further simplified by writing them in terms of the soft-thresholding operator:

$$\widehat{\beta}_j = \frac{S(z_j|\lambda_1)}{1+\lambda_2}$$

- In the orthonormal case, then, the elastic net solutions are simply the lasso solutions divided by $1+\lambda_2$
- In other words, the additional ridge penalty has the same effect on the lasso as the ridge penalty itself has on ordinary least squares regression: it provides shrinkage

Remarks

- As with ridge regression itself, shrinking the coefficients towards zero increases bias, but reduces variance
- Since this involves drawbacks as well as advantages, adding a ridge penalty is not always universally beneficial, as the bias can dominate the variance
- Still, as with ridge regression itself, it is typically the case that a profitable compromise can be reached by incorporating some (possibly small) ridge term into the penalty

The grouping effect: Introduction

- Our earlier example is an extreme example of a property possessed by the elastic net known as the *grouping effect*
- The property states that highly correlated features will have similar estimated coefficients, which seems intuitively reasonable
- Even if a data set does not contain identical variables as in the toy example, many data sets – particularly high dimensional ones – contain highly correlated predictors
- The shrinkage and grouping effects produced by the elastic net are an effective way of dealing with these correlated predictors

Grouping effect

- The property can be described formally in terms of an upper bound on the difference between two coefficients as it relates to the correlation between the predictors:

  $$|\widehat{\beta}_j - \widehat{\beta}_k| \leq \frac{\|\mathbf{y}\|\sqrt{2(1-\rho_{jk})}}{\lambda_2\sqrt{n}}$$

  where ρ_{jk} is the sample correlation between $\mathbf{x}_j$ and $\mathbf{x}_k$
- Note, in particular, that as $\rho_{jk} \to 1$, the difference between $\widehat{\beta}_j$ and $\widehat{\beta}_k$ goes to zero

Simulation example: Setup

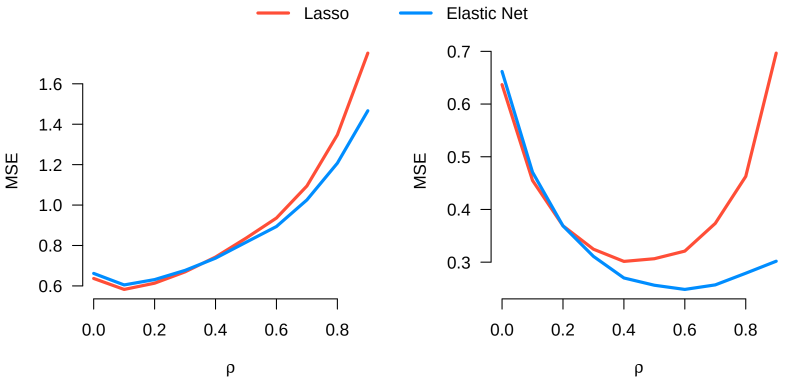
- To see the effect of the grouping property, we will carry out a simulation study with $n = 50$ and $p = 100$
- All features $\mathbf{x}_j$ will follow standard Gaussian distributions in the marginal sense, but we introduce correlation between the features in one of two ways:
  - Compound symmetric: All features have the same pairwise correlation ρ
  - Block diagonal: The 100 features are partitioned into blocks of 5 features each, with a pairwise correlation of ρ between features in a block, but features from separate blocks are independent
- In the generating model, we will set $\beta_1 = \beta_2 = \cdots = \beta_5 = 0.5$ and $\beta_6 = \beta_7 = \cdots = \beta_{100} = 0$

Simulation example: Setup (cont'd)

- Note that in the block diagonal case, this introduces a grouping property: correlated features have identical coefficients
- In the compound symmetric case, on the other hand, correlation between features does not tell us anything about their corresponding coefficients
- For the elastic net penalty, for the sake of simplicity we set $\lambda_1 = \lambda_2$ and select λ by independent validation

Simulation example: Results

Left: Compound symmetric; Right: Block diagonal



Remarks

- This simulation demonstrates that when the correlation between features is not large, there is often little difference between the lasso and elastic net estimators in terms of their estimation accuracy; indeed, when correlation is near zero, the lasso is often more accurate
- When the correlation between features is large, however, the elastic net has an advantage over the lasso
- This advantage is much more pronounced in the block diagonal case, where the coefficients have a grouping property

Commentary on the grouping property

- In practice, the grouping effect is often one of the strongest motivations for applying an elastic net penalty
- For example, in gene expression studies, genes that have similar functions, or that work together in a pathway to accomplish a certain function, are often correlated
- It is often reasonable to assume, then, that if the function is relevant to the response we are analyzing, the coefficients will be similar across the correlated group

Commentary on the grouping property (cont'd)

- It is worth pointing out, however, that grouping does not always hold
- For example, in a genetic association study, it is certainly quite possible for two nearby variants to be highly correlated in their inheritance patterns, but for one variant to be harmless and the other to be highly deleterious
- Nevertheless, in such a case, it is often quite difficult to determine which of two highly correlated features is the causative feature, and the elastic net, which splits the estimated signal between the correlated features, offers a reasonable compromise

Ridge + nonconvex penalty

- The motivation for adding a ridge penalty to the lasso penalty also applies to nonconvex penalties such as MCP and SCAD
- In fact, the motivation is perhaps even stronger in this case
- As we saw last week, the objective functions for MCP and SCAD may fail to be convex and present multiple local minima, which leads to difficulty in optimization and decreased numerical stability
- Adding a strictly convex ridge penalty can often substantially stabilize the problem

Orthonormal solutions for MCP

- The addition of a ridge penalty has a similar shrinkage effect on MCP and SCAD as it does on lasso-penalized models
- In particular, for MCP in the orthonormal case,

$$\widehat{\beta}_j = \begin{cases} \dfrac{z_j}{1+\lambda_2} & |z_j| > \gamma\lambda_1(1+\lambda_2) \\[2ex] \dfrac{S(z_j|\lambda_1)}{1-\frac{1}{\gamma}+\lambda_2} & |z_j| \le \gamma\lambda_1(1+\lambda_2). \end{cases}$$

- Similar, if somewhat more complicated, results are available for SCAD (equation 4.7 in the book)

Remarks

- From this solution, we can see that the shrinkage role played by λ_2 is, in a sense, the opposite of the bias reduction role played by γ
- While dividing by $1 - \gamma^{-1}$ inflates the value of $S(z_j|\lambda_1)$, dividing by $1 + \lambda_2$ shrinks it
- When both are present in the model, the orthonormal solution is the soft-thresholding solution divided by $1 - \gamma^{-1} + \lambda_2$, which could either shrink or inflate $S(z_j|\lambda_1)$ depending on the balance between γ and λ_2

Remarks

- It should be noted, however, that the terms are not entirely redundant; while they cancel each other out in the denominator of the orthonormal solution, they do not cancel out elsewhere
- In particular, they can have rather different effects in the presence of correlation among the features
- Finally, as in the elastic net, the regularization parameters for the ridge-stabilized versions of MCP and SCAD are often expressed in terms of λ and α

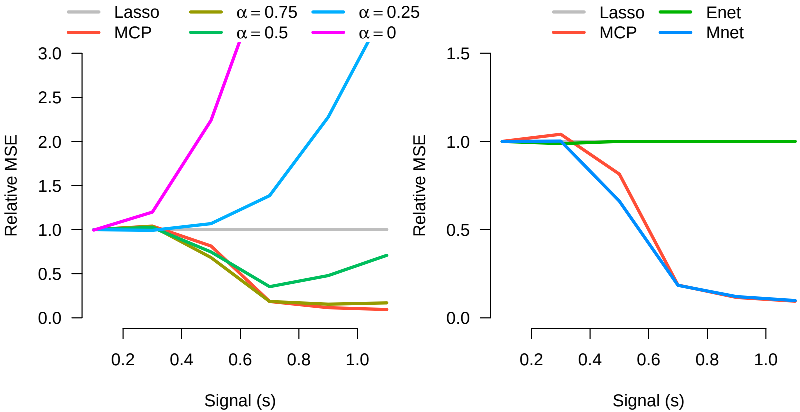
Simulation 1: Setup

- We close this lecture with two simulation studies comparing the estimation accuracy of lasso, MCP, the elastic net, and what we will abbreviate "MNet", the MCP version of the elastic net (i.e., a penalty that consists of MCP + Ridge)
- First, suppose all covariates $\{x_j\}$ follow independent standard Gaussian distributions, and that the outcome $\mathbf{y}$ equals $\mathbf{X}\boldsymbol{\beta}$ plus errors drawn from the standard Gaussian distribution
- For each independently generated set of data set, let $n = 100$ and $p = 500$, with 12 nonzero coefficients equal to s and the remaining 488 coefficients equal to zero; we will consider varying the signal strength s between 0.1 and 1.1

Simulation 1: Setup (cont'd)

- For all methods, tuning parameters are selected on the basis of mean-squared prediction error on an independent validation data set also of size $n = 100$
- For lasso and MCP, only one tuning parameter (λ) was selected (for MCP, $\gamma = 3$ was fixed)
- For the Enet and Mnet estimators, we consider both fixed-α estimators and estimators in which both λ and α were selected by external validation (i.e., prediction error was calculated over a two-dimensional grid and the best combination of λ and α was chosen)

Simulation 1: Results



Fixed-α remarks

- All methods behave rather similarly when s is small, as all models end up with estimates of $\widehat{\beta} \approx \mathbf{0}$ in these settings
- As one might expect, a modest ridge penalty is beneficial in the medium-signal settings, with $\alpha = 0.5$ achieving the highest accuracy when $s = 0.5$
- As signal increases, however, the downward bias of ridge and lasso play a larger role, and MCP becomes the most accurate estimator along with the $\alpha = 0.9$ Mnet estimator, which is similar to MCP

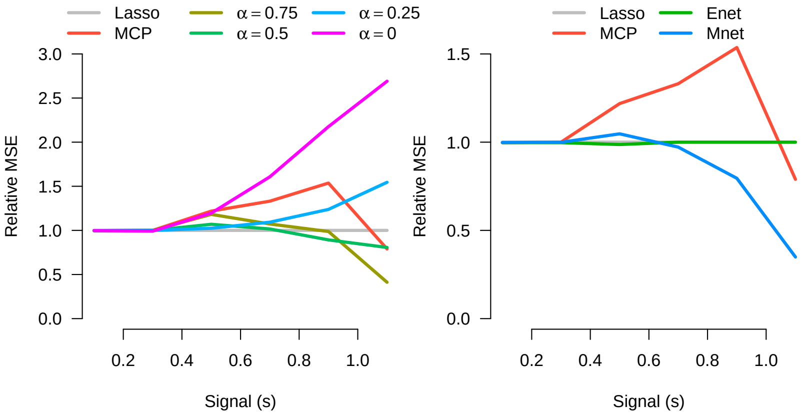
Variable-α remarks

- In the variable-α case, there is little difference between the lasso and elastic net estimators
- In particular, when s is large the two are virtually identical due, in part, to the fact that α is typically selected to be ≈ 1 for Enet when s is large
- MCP and Mnet are similar to lasso and Enet when s is small, but substantially outperform the lasso and elastic net when the signal is increased
- One can improve estimation accuracy by adding a ridge penalty, although the gains are not particularly dramatic when the features are independent

Simulation 2: Setup

- Let us modify the previous simulation to examine how correlation among the features affects the results
- In particular, all covariates $\{x_j\}$ follow a standard Gaussian distribution marginally, but are now correlated with a common (compound symmetric) correlation $\rho = 0.7$ between any two covariates
- This is a rather extreme amount of correlation, but helps to clearly illustrate the effect of correlation on the relative performance of the methods

Simulation 2: Results



Remarks

- The primary point illustrated by the figure is that the benefits of shrinkage are much more pronounced in the presence of correlation
- For example, while MCP was never far from the best estimator in the uncorrelated case, it is one of the worst methods for most signal strengths in the correlated case
- Meanwhile, although Mnet and MCP were generally similar in Simulation #1, here Mnet outperforms MCP rather dramatically

Conclusions

- The addition of a ridge penalty typically yields little improvement in the absence of correlation
- Benefits are more substantial in the presence of correlation, and very substantial in grouped scenarios
- Adding ridge penalties offers much more potential advantage for nonconvex penalties; indeed, adjusting α to stabilize MCP in this way is often a more fruitful approach than adjusting γ
- It is difficult to rely on any particular value of α; in practice, it is advisable to try out several values of α and use cross-validation to guide its selection