Ridge regression: Selection of λ and a case study

Patrick Breheny

February 10

Introduction

- As we discussed last time, the parameter λ controls the tradeoff between the penalty and the model fit, and therefore has a very large impact on the resulting estimate:
  - As $\lambda \to 0$, Q approaches L and $\widehat{\boldsymbol{\beta}}$ approaches the OLS estimate
  - On the other hand, as $\lambda \to \infty$, the penalty dominates the objective function and $\widehat{\boldsymbol{\beta}} \approx \mathbf{0}$
- Clearly, selection of λ is a very important practical aspect of fitting penalized regression models.

General regression framework

- In general, a reasonable approach to selecting λ in an objective manner is to choose the value of λ with the greatest predictive power: if $\lambda = 1$ can predict future observations better than $\lambda = 5$, this is a clear reason to prefer $\lambda = 1$
- Suppose that $\mathbb{E}(y_i) = f(\mathbf{x}_i)$, $\mathrm{Var}(y_i) = \sigma^2$, and that we have fit a model to obtain $\hat{f}$, an estimate of f, with fitted values $\{\hat{y}_i(\lambda)\}_{i=1}^n$
- It is clearly misleading to evaluate predictive accuracy by comparing $\hat{y}_i(\lambda)$ to y_i; the observed value y_i has already been used to calculate $\hat{y}_i(\lambda)$, and is therefore not a genuine prediction

Prediction error

- Simply calculating the residual sum of squares (RSS), then, will underestimate the true predictive accuracy of the model
- Instead, we must examine how well $\hat{y}_i(\lambda)$ predicts a new observation y_i^{new} generated from the underlying model:

$$y_i^{\text{new}} = f(\mathbf{x}_i) + \varepsilon_i^{\text{new}}$$

- Then the prediction error can be measured by

$$\text{PE}(\lambda) = \sum_{i=1}^{n} (y_i^{\text{new}} - \hat{y}_i(\lambda))^2$$

- To clarify, under this framework we are measuring new responses $\{y_i^{\text{new}}\}_{i=1}^n$, but at the original values of the predictors $\{\mathbf{x}_i\}_{i=1}^n$

Expected prediction error

- The model with the greatest predictive power, then, is the model that minimizes the expected prediction error

$$\mathbb{E}\,\mathrm{PE}(\lambda) = \mathbb{E}\sum_{i=1}^{n}\{y_i^{\mathrm{new}} - \hat{y}_i(\lambda)\}^2,$$

  where the expectation is taken over both the original observations $\{y_i\}_{i=1}^n$ as well as the new observations $\{y_i^{\mathrm{new}}\}_{i=1}^n$

- **Theorem:**

$$\mathbb{E}\,\mathrm{PE}(\lambda) = \mathbb{E}\sum_{i=1}^{n}\{y_i - \hat{y}_i(\lambda)\}^2 + 2\sum_{i=1}^{n}\mathrm{Cov}\{\hat{y}_i(\lambda), y_i\}$$

Remarks

- So the expected prediction error consists of two terms:
  - The first term is the within-sample fitting error
  - the second term is a bias correction factor that arises from the tendency of within-sample fitting error to underestimate out-of-sample prediction error, also known as the *optimism* of the model fit
- The second term can also be considered a measure of model complexity, or degrees of freedom:

$$\begin{aligned} \mathrm{df}(\lambda) &= \sum_{i=1}^{n} \frac{\mathrm{Cov}(\hat{y}_i(\lambda), y_i)}{\sigma^2} \\ &= \frac{\mathrm{tr}\{\mathrm{Cov}(\hat{\mathbf{y}}(\lambda), \mathbf{y})\}}{\sigma^2} \end{aligned}$$

Degrees of freedom: Linear regression

- For example, consider OLS regression, with $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- **Result:**

$$\mathrm{df} = \mathrm{rank}(\mathbf{X})$$

- Thus, our covariance-based definition agrees with the usual notion of degrees of freedom as the number of parameters (in an unpenalized model)

Degrees of freedom: Ridge regression

- A model fitting method is said to *linear* if we can write $\hat{\mathbf{y}} = \mathbf{S}\mathbf{y}$ for some matrix S
- **Result:** For any linear method,

$$\mathrm{df} = \mathrm{tr}(\mathbf{S})$$

- Ridge regression is a linear fitting method, with $\mathbf{S} = n^{-1}\mathbf{X}(n^{-1}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T$; thus,

$$\mathrm{df}(\lambda) = \mathrm{tr}(n^{-1}\mathbf{X}(n^{-1}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T) = \sum_{j=1}^{p} \frac{d_j}{d_j + \lambda}$$

where $d_1, \ldots, d_p$ are the eigenvalues of $n^{-1}\mathbf{X}^T\mathbf{X}$

Remarks

- This result illustrates that in penalized regression, model selection is continuous
- As we change λ, we gradually increase the complexity of the model, and small changes in λ result in small changes in estimation
- This is in sharp contrast to "best subsets" model selection, where complexity is added by discrete jumps as we introduce parameters, and adding just a single parameter can introduce large changes in model estimates

The C_p statistic

- Now that we've generalized the concept of degrees of freedom, I'll describe various model selection criteria that can be used to select λ
- This account will be brief, since you have likely encountered these criteria in other classes
- To begin, let us turn our attention back to $\mathbb{E}(\mathrm{PE})$; recall that it consisted of two terms, a within-sample error term and a model complexity term
- Using RSS/σ^2 for the first term and $\mathrm{df}(\lambda)$ for the second, we obtain a criterion known as the C_p statistic:

$$C_p = \frac{\mathrm{RSS}(\lambda)}{\sigma^2} + \mathrm{df}(\lambda)$$

Leave-one-out error

- An alternative approach is to consider leaving observations out of the fitting process and saving them to use for evaluating predictive accuracy; in general, this is known as cross-validation, which we will discuss later
- However, for linear fitting methods, there is an elegant closed-form solution to the leave-one-out cross-validation error that does not require actually refitting the model
- Letting $\hat{f}_{(-i)}$ denote the fitted model with observation i left out,

$$\sum_i \left\{ y_i - \hat{f}_{(-i)}(x_i) \right\}^2 = \sum_i \left(\frac{y_i - \hat{f}(x_i)}{1 - S_{ii}} \right)^2,$$

where S_{ii} is the ith diagonal element of $\mathbf{S}$

GCV

- Replacing S_{ii} by its average, $\mathrm{tr}(\mathbf{S})/n = \mathrm{df}(\lambda)/n$, we arrive at the *generalized cross validation* criterion:

$$\mathrm{GCV} = \frac{\mathrm{RSS}(\lambda)}{(1 - \mathrm{df}(\lambda)/n)^2}$$

- Like C_p, the GCV criterion combines $\mathrm{RSS}(\lambda)$ with a model complexity term, although in GCV it takes the form of an inflation factor $(1 - \mathrm{df}/n)^2$ multiplicative factor rather than an additive term
- One attractive aspect of GCV as opposed to C_p statistic is that it does not require an estimate of σ^2

AIC

- Both C_p and GCV are developed with the least squares objective in mind; the *Akaike information criterion* (AIC) is a generalization of the C_p idea to general maximum likelihood models
- Rather than consider the expected value of $\{y_i^{\text{new}} - \hat{y}_i(\hat{\theta})\}^2$, Akaike proposed estimating the expected value of $\log \mathbb{P}(y_i^{\text{new}}|\hat{\theta})$, where $\hat{\theta}$ denotes the estimated parameters of the model based on the original data $\{y_i\}_{i=1}^n$
- Asymptotically, it can be shown that for maximum likelihood estimation,

$$\text{AIC} = 2L(\hat{\theta}(\lambda)|\mathbf{X}, \mathbf{y}) + 2\text{df}(\lambda)$$

AIC (cont'd)

- For the normal distribution,

$$\text{AIC} = n \log \sigma^2 + \frac{\text{RSS}(\lambda)}{\sigma^2} + 2\text{df}(\lambda) + \text{constant}$$

- Thus, in the case of normally distributed errors with known variance σ^2, AIC and C_p are equivalent up to a constant

Bayesian model selection

- A rather different approach is to consider model selection from a Bayesian perspective
- Letting M_λ denote the model with regularization parameter λ, we would be interested in calculating the posterior probability of M_λ given the data, $\mathbb{P}(M_\lambda|\mathbf{X}, \mathbf{y})$
- If we assume a uniform prior across all models, then $\mathbb{P}(M_\lambda|\mathbf{X}, \mathbf{y}) \propto \mathbb{P}(\mathbf{y}|\mathbf{X}, M_\lambda)$

BIC

- In general, calculating this quantity involves numerical integration, but this integral can be approximated to yield

$$\log \mathbb{P}(\mathbf{y}|\mathbf{X}, M_\lambda) \approx -L(\hat{\theta}(\lambda)|\mathbf{X}, \mathbf{y}) - \frac{1}{2}\mathrm{df}(\lambda)\log(n)$$

- The *Bayesian information criterion* (BIC) is defined as -2 times this quantity:

$$\mathrm{BIC} = 2L(\hat{\theta}(\lambda)|\mathbf{X}, \mathbf{y}) + \mathrm{df}(\lambda)\log(n)$$

- Thus, choosing the model with the smallest BIC is (approximately) equivalent to choosing the model with the highest posterior probability

Remarks

- Note that, despite the very different derivations, the equations for AIC and BIC are surprisingly similar; the only difference is $\log(n)$ instead of 2 as the multiplicative factor for $\mathrm{df}(\lambda)$
- In practice, this means that BIC applies a heavier penalty to model complexity than does AIC (provided $n \geq 8$) and will therefore favor more parsimonious models

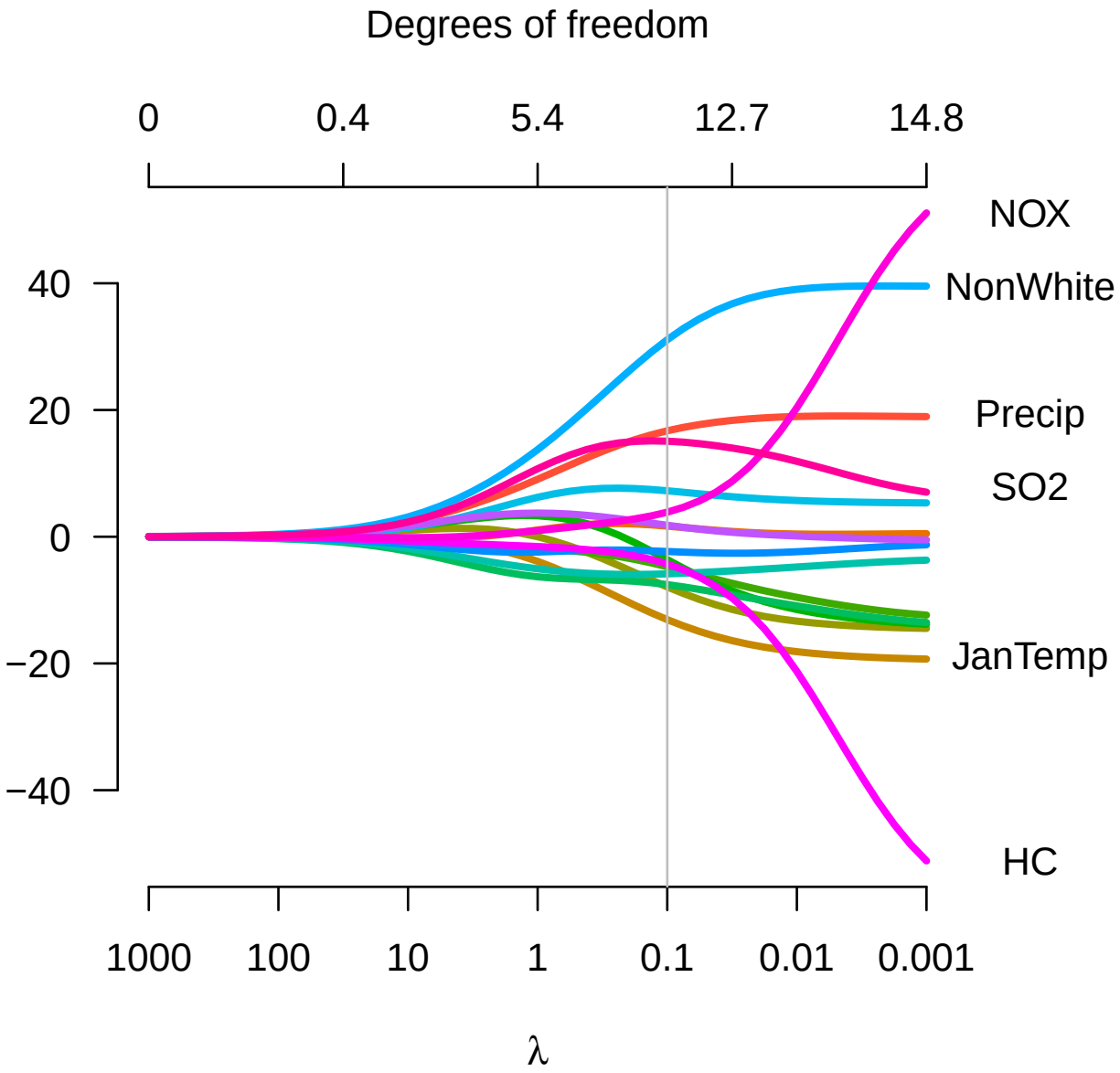
Pollution study

- To illustrate ridge regression in practice, we will now consider a study designed to estimate the relationship between pollution and mortality while adjusting for the potentially confounding effects of climate and socioeconomic conditions
- To quantify pollution, "relative pollution potential" was measured for three pollutants – hydrocarbons (HC), nitrogen oxides (NOX), and sulfur dioxide (SO_2) – in 60 Standard Metropolitan Statistical Areas in the United States between 1959-1961
- The outcome of interest is total age-adjusted mortality from all causes, in deaths per 100,000 population

Pollution study (cont'd)

- In total, there are $p = 15$ explanatory variables: the three pollution variables, 8 demographic/socioeconomic variables, and 4 climate variables
- Although few would consider $p = 15$ "high-dimensional", the full maximum likelihood model nevertheless struggles with a sample size of just 60 and strong correlation among several variables
- As we will see, this leaves it unable to provide a trustworthy answer to the primary question of the relationship between pollution and mortality

Ridge trace / coefficient path



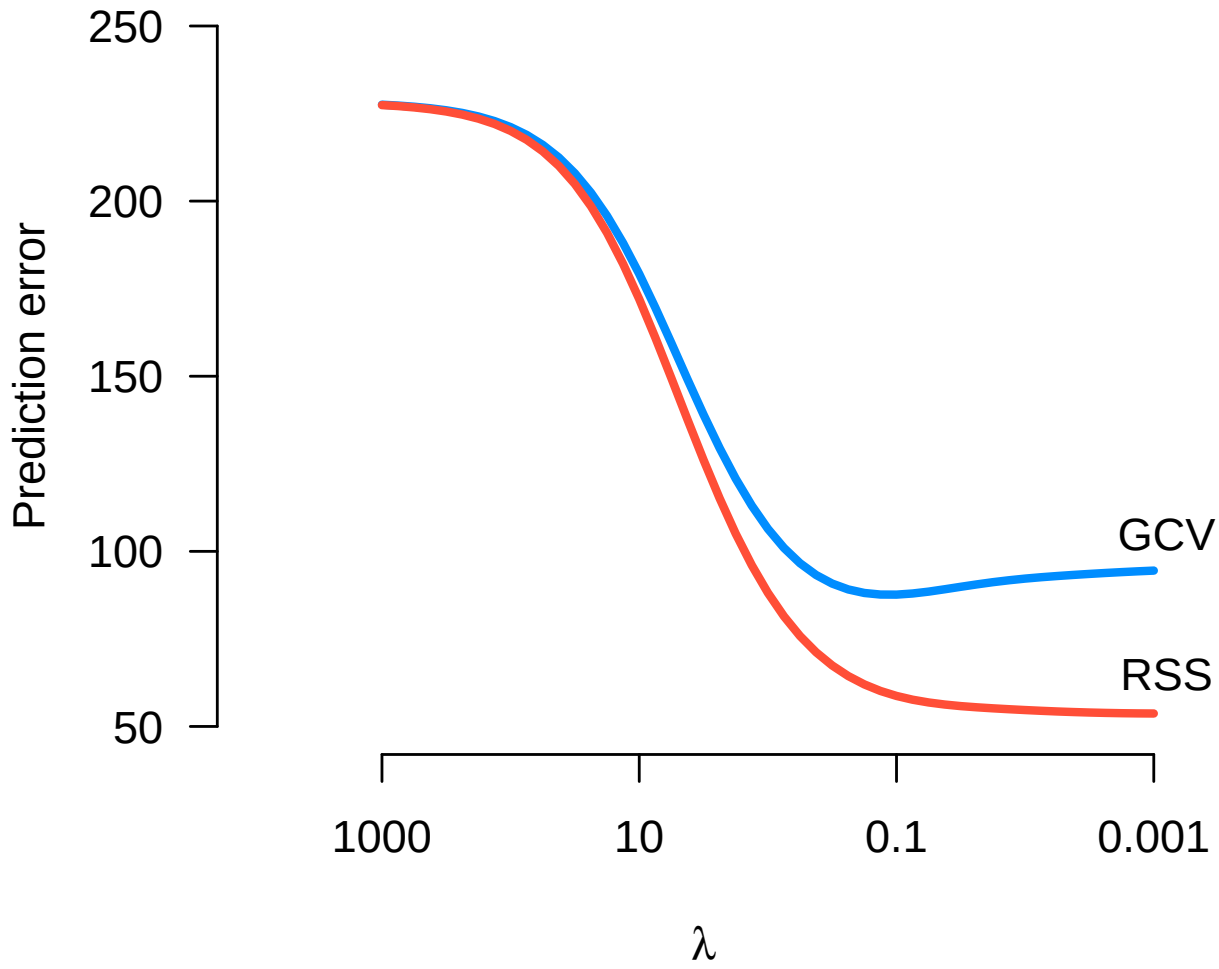
Remarks

- It is particularly instructive to look at the coefficient paths of the three pollution parameters, all of which are fairly highly correlated with each other
- At small λ values, the estimates indicate that NOX pollution has a very strong harmful effect, while HC pollution has a very strong protective effect
- This result is surprising, and indeed rather difficult to believe – increasing the amount of HC pollution should *save* 60 lives per 100,000?
- However, as we increase the ridge penalty, we see that the estimated effects for these two types of pollution quite rapidly drop to near zero

Remarks (cont'd)

- A parallel story is told by examining the SO_2 coefficient path
- SO_2 is correlated with HC and NOX (although not as highly correlated as HC and NOX are with each other), so its solution is affected by the estimated effects for the other two pollutants
- In particular, while most of the other coefficient estimates increase monotonically as λ decreases from ∞ to 0, the estimated effect of SO_2 goes up, then decreases
- As a result, depending on the value of λ one chooses, SO_2 pollution is either far more important, or far less important, than HC and NOX pollution

Fitting error and prediction error



t-statistics for OLS and ridge

Pollution terms:

	Ridge	OLS
SO2	2.78	0.59
NOX	0.37	1.35
HC	-0.41	-1.39

	Ridge	OLS
NonWhite	3.94	3.40
Precip	2.51	2.09
Density	1.45	0.92
Humidity	0.34	0.09
Poor	0.23	-0.05
WhiteCol	-0.39	-0.12
House	-0.46	-1.54
Over65	-0.60	-1.08
Sound	-0.89	-0.38
Educ	-1.04	-1.46
JulyTemp	-1.09	-1.65
JanTemp	-1.77	-1.77

Remarks

- For the t-statistics on the previous page, we take the standard error to be the square root of the diagonal elements of

$$\nabla^2_{\boldsymbol{\beta}} Q^{-1} = \frac{\sigma^2}{n}(n^{-1}\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1},$$

  using $\hat{\sigma}^2 = \mathrm{RSS}/(n - \mathrm{df})$ to estimate σ^2
- Note that some terms become more significant with an added ridge penalty, while others become less significant; although the *estimates* are shrunken towards zero, the fact that variance is reduced can cause the significance (i.e., the evidence against $\beta = 0$) to increase

Concluding remarks

- The major limitation of ridge regression is the fact that all of its coefficients are nonzero
- This poses two considerable problems for high-dimensional regression:
  - Solutions become very difficult to interpret
  - The computational burden becomes large
- It is desirable, then, to have models which allow for both shrinkage *and selection*; in other words, to retain the benefits of ridge regression while at the same time selecting a subset of important variables