MA477: Data Science
Lesson 9 Outline — 29 January 2026
United States Military Academy, West Point
Instructor: MAJ Patrick Kuiper

---

1 Administrative

- Student review
- Model Review Questions
- Model Review Lecture

2 Shrinkage Lesson Objectives

- Understand the mathematical formulation of shrinkage methods and be able to compare/contrast them.
- Employ shrinkage methods in regression settings

3 Student Review

4 Shrinkage Methods Lecture

Shrinkage Methods (Ridge Regression and the Lasso)

Discussion Questions

We will organize this lecture around the following five questions:

1. Why might shrinking regression coefficients improve prediction, even if it makes the model less flexible?
2. What do the regression penalties do to the coefficients as the tuning parameter λ increases?
3. Why must predictors be standardized before applying shrinkage methods?
4. How do shrinkage methods' penalties differ, and why does that difference lead to variable selection?
5. Any thoughts on "smart" ways to select λ? Why might these fail?
6. How is the tuning parameter λ chosen in practice, and what is being optimized?

1. Why Can Shrinkage Improve Prediction?

Least squares regression estimates coefficients by minimizing the residual sum of squares (RSS). While this produces an unbiased estimator, it can have very high variance when predictors are highly correlated or when the number of predictors is large relative to the number of observations.

Shrinkage methods intentionally pull coefficient estimates toward zero. This introduces some bias, but it can dramatically reduce variance. If the reduction in variance outweighs the increase in bias, the overall test mean squared error (MSE) improves. This trade-off between bias and variance explains why shrinkage methods can outperform least squares in prediction.

2. What Does Ridge Regression Do as λ Increases?

Ridge regression minimizes the objective function

$$\text{RSS} + \lambda \sum_{j=1}^{p} \beta_j^2.$$

The tuning parameter $\lambda \geq 0$ controls the strength of the penalty on the squared coefficients.

- When $\lambda = 0$, ridge regression reduces to least squares.
- As λ increases, all coefficients are shrunk toward zero.
- As $\lambda \to \infty$, the coefficient estimates approach zero, yielding the null model.

Ridge regression shrinks coefficients smoothly but generally does not set them exactly equal to zero.

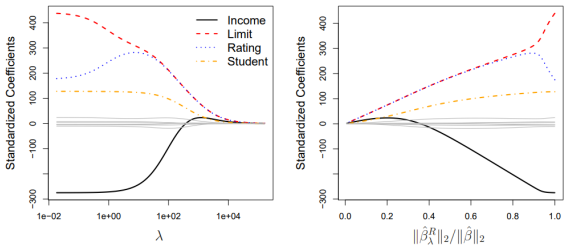

FIGURE 6.4. *The standardized ridge regression coefficients are displayed for the* `Credit` *data set, as a function of* λ *and* $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$.

Figure 1

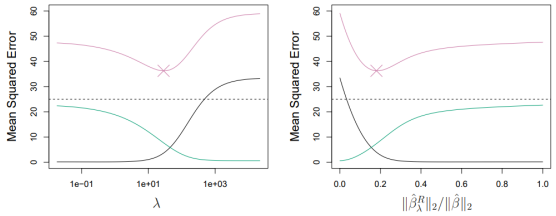

FIGURE 6.5. *Squared bias (black), variance (green), and test mean squared error (purple) for the ridge regression predictions on a simulated data set, as a function of* λ *and* $\|\hat{\beta}_\lambda^R\|_2/\|\hat{\beta}\|_2$. *The horizontal dashed lines indicate the minimum possible MSE. The purple crosses indicate the ridge regression models for which the MSE is smallest.*

Figure 2

3. Why Do We Standardize Predictors?

The ridge and lasso penalties depend directly on the magnitude of the regression coefficients. If predictors are measured on different scales, the penalty will affect them unequally.

To avoid this issue, predictors are standardized by subtracting their mean and dividing by their standard deviation. After standardization, all predictors have mean zero and standard deviation one. This ensures that the penalty treats all predictors fairly and that the results do not depend on the units of measurement.

4. How Does the Lasso Differ from Ridge Regression?

The lasso minimizes the objective function

$$\text{RSS} + \lambda \sum_{j=1}^{p} |\beta_j|.$$

This penalty uses the ℓ_1 norm of the coefficient vector, rather than the ℓ_2 norm used by ridge regression.

As a result:

- Ridge regression shrinks all coefficients toward zero but keeps them nonzero.
- The lasso can shrink some coefficients exactly to zero when λ is sufficiently large.

Because coefficients that equal zero correspond to excluded predictors, the lasso performs variable selection automatically and produces sparse, more interpretable models.

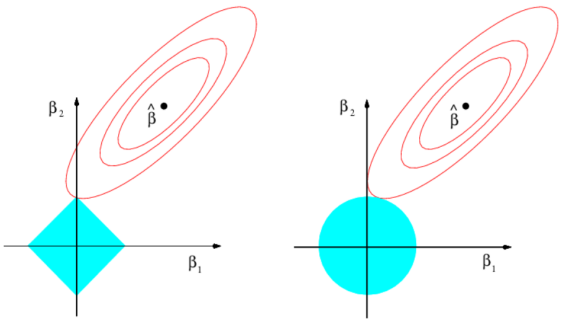

FIGURE 6.7. *Contours of the error and constraint functions for the lasso* (left) *and ridge regression* (right). *The solid blue areas are the constraint regions,* $|\beta_1| + |\beta_2| \le s$ *and* $\beta_1^2 + \beta_2^2 \le s$, *while the red ellipses are the contours of the RSS.*

Figure 3

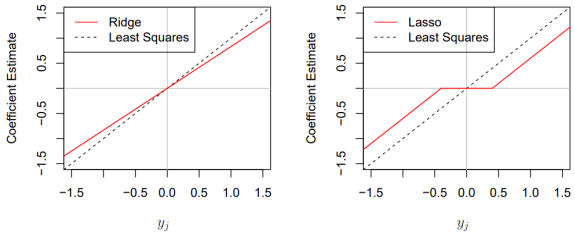

FIGURE 6.10. *The ridge regression and lasso coefficient estimates for a simple setting with* $n = p$ *and* $\mathbf{X}$ *a diagonal matrix with* 1*'s on the diagonal.* Left: *The ridge regression coefficient estimates are shrunken proportionally towards zero, relative to the least squares estimates.* Right: *The lasso coefficient estimates are soft-thresholded towards zero.*

Figure 4

Geometric Interpretation and the Role of Lagrange Multipliers

- Regularized regression is often written in constrained form, but the constraint level is *not* determined by the data.

- Without fixing the constraint, the problem is ill posed:

$$\min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta}) \quad \Rightarrow \quad \hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{\text{LS}}.$$

- The constraint must be imposed externally:

$$\min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta}) \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_2^2 \leq s \quad \text{(ridge)},$$

$$\min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta}) \quad \text{s.t.} \quad \|\boldsymbol{\beta}\|_1 \leq s \quad \text{(lasso)}.$$

- The tuning parameter λ is the Lagrange multiplier enforcing the chosen constraint:

$$\min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta}) + \lambda \mathcal{P}(\boldsymbol{\beta}).$$

- Larger λ corresponds to a tighter constraint; smaller λ to a looser constraint.
- The solution occurs where RSS contours first touch the constraint set.
- The ℓ_1 constraint has corners aligned with coordinate axes $\Rightarrow$ exact zeros.
- The ℓ_2 constraint is smooth $\Rightarrow$ coefficients are shrunk but rarely zero.
- In contrast, when constraints are hard and data-determined (e.g. $\min_x f(x)$ s.t. $g(x) = 0$), the Lagrange multiplier λ is solved jointly with x via the KKT conditions rather than chosen externally.

Optimization: Gradient Descent for Ridge and the Lasso

- Ridge and lasso differ not only in their penalties, but in how their optimization problems are solved.
- Ridge regression minimizes a smooth, convex objective:

$$\min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_2^2,$$

  which is differentiable for all $\boldsymbol{\beta}$.
- Ridge admits a closed-form solution:

$$\hat{\boldsymbol{\beta}} = (X^\top X + \lambda I)^{-1} X^\top y,$$

  and can also be solved efficiently via gradient descent.
- The lasso minimizes a convex but nonsmooth objective:

$$\min_{\boldsymbol{\beta}} \text{RSS}(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1,$$

  which is not differentiable at $\beta_j = 0$.
- Standard gradient descent is not directly applicable due to the nondifferentiability of $|\beta_j|$ at zero.
- The lasso is instead solved using methods such as coordinate descent or proximal gradient algorithms.
- These methods enforce the KKT conditions via soft-thresholding, allowing coefficients to be set exactly to zero.

5. How Is λ Chosen in Practice?

The tuning parameter λ is typically selected using cross-validation.

1. A grid of λ values is chosen.
2. Cross-validation error is computed for each value.
3. The value of λ that minimizes the validation error is selected.
4. The model is refit using all available data and the chosen λ.

The goal of this process is to minimize test-set prediction error, usually measured by mean squared error.

Summary

Ridge regression and the lasso both reduce variance by shrinking coefficients, but they behave differently:

- Ridge regression works well when many predictors have small or moderate effects.
- The lasso works well when only a small number of predictors have substantial effects.

Since the true structure is unknown in practice, cross-validation is used to determine which approach provides better predictive performance on a given data set.