Shrinkage Methods: Ridge Regression and LASSO

Saurabh

March 8, 2025

Contents

1 Introduction

Multiple Linear Regression (MLR) is among the most widely used statistical techniques for modeling the relationship between a continuous response variable y and one or more predictor variables $\mathbf{X} = (X_1, X_2, \ldots, X_p)$. The classical *Ordinary Least Squares* (OLS) approach estimates the coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$ by minimizing the sum of squared errors (SSE):

$$\text{SSE}(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2.$$

While OLS is straightforward and powerful, it is not without pitfalls. Two key challenges frequently arise:

1) **Multicollinearity:** When predictor variables are highly correlated, the matrix $\mathbf{X}^\top\mathbf{X}$ becomes nearly singular, making the OLS estimates unstable and prone to high variance.

2) **Overfitting:** In high-dimensional settings or when the number of predictors is large relative to the number of observations, the OLS solution may overfit the training data, leading to poor generalization.

Shrinkage methods, especially **Ridge Regression** and **LASSO**, offer systematic ways to handle these issues. By introducing a penalty on the size of the coefficients, these methods shrink the estimates toward zero, thereby reducing variance and often improving prediction accuracy.

2 Problems in Multiple Linear Regression

2.1 Problem 1: Multicollinearity

Multicollinearity refers to a situation where two or more predictors in the regression model are highly correlated. This leads to:

- A **near-singular $\mathbf{X}^\top\mathbf{X}$** matrix.
- **Instability** in coefficient estimates, where small changes in data can cause large changes in $\boldsymbol{\beta}$.
- **Wide confidence intervals**, complicating interpretation of individual coefficients.

Illustrative Example

Suppose you measure both the weight in pounds and the weight in kilograms of an object. Since

$$\text{weight in kilograms} \approx 0.4536 \times \text{weight in pounds},$$

the corresponding columns in $\mathbf{X}$ will be nearly linearly dependent. In OLS, $\mathbf{X}^\top\mathbf{X}$ will almost fail to invert, leading to extremely large and erratic coefficient estimates that can drastically change with minor data perturbations.

2.2 Problem 2: Overfitting

When the number of predictors p is large (or comparable to the number of observations n), the model may fit random noise rather than the true underlying signal. Overfitting symptoms include:

- **Low training error** but **high test error**.
- **Poor generalization** to new/unseen data.

3 Ridge Regression

Ridge Regression addresses multicollinearity and overfitting by introducing an ℓ_2 penalty on the coefficients.

3.1 The Ridge Objective Function

The Ridge objective function modifies the standard SSE with a penalty on the sum of squared coefficients:

$$J(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda \geq 0$ is the **regularization parameter** controlling the strength of the penalty. In vector form:

$$J(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + \lambda\, \boldsymbol{\beta}^\top \boldsymbol{\beta}.$$

3.2 Derivation of the Ridge Estimator

To minimize $J(\boldsymbol{\beta})$, take the gradient and set it to zero:

$$J(\boldsymbol{\beta}) = \mathbf{y}^\top \mathbf{y} - 2\,\boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{y} + \boldsymbol{\beta}^\top \mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} + \lambda\, \boldsymbol{\beta}^\top \boldsymbol{\beta}.$$

Differentiating with respect to $\boldsymbol{\beta}$:

$$\nabla_{\boldsymbol{\beta}} J(\boldsymbol{\beta}) = -2\,\mathbf{X}^\top \mathbf{y} + 2\,\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} + 2\,\lambda\,\boldsymbol{\beta} = \mathbf{0}.$$

Divide by 2 and rearrange:

$$\mathbf{X}^\top \mathbf{X}\boldsymbol{\beta} + \lambda\,\boldsymbol{\beta} = \mathbf{X}^\top \mathbf{y}.$$

Noting that $\lambda\boldsymbol{\beta} = \lambda\mathbf{I}\boldsymbol{\beta}$, we have:

$$\left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right) \boldsymbol{\beta} = \mathbf{X}^\top \mathbf{y}.$$

Provided $\lambda > 0$, $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}$ is invertible. Thus, the **Ridge estimator** is:

$$\boldsymbol{\beta}_{\text{ridge}} = \left(\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}\right)^{-1} \mathbf{X}^\top \mathbf{y}.$$

3.3 Why Ridge Mitigates Multicollinearity

By adding $\lambda\mathbf{I}$ to $\mathbf{X}^\top\mathbf{X}$, Ridge protects against near-zero eigenvalues that occur with strongly correlated predictors. This helps:

- **Stabilize** the inversion of $\mathbf{X}^\top\mathbf{X}$.
- **Shrink** coefficients to control variance.
- Trade a small increase in bias for a potentially large decrease in variance.

3.4 Practical Illustration of Ridge

Consider a dataset with 200 observations and 10 predictors, where two predictors X_1 and X_2 are correlated ($\rho \approx 0.95$):

a) **OLS Fit**: You may see extremely large magnitudes for $\hat{\beta}_1$ and $\hat{\beta}_2$, with wide confidence intervals due to collinearity.

b) **Ridge Fit**: By choosing a moderate λ (found via cross-validation), you shrink both β_1 and β_2 substantially, leading to more stable estimates.

4 LASSO

LASSO (Least Absolute Shrinkage and Selection Operator) is another shrinkage method but uses an ℓ_1 penalty. Its objective:

$$J(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

4.1 Geometric Perspective

While Ridge's L2 penalty is akin to constraining $\boldsymbol{\beta}$ to lie within a sphere (circle in 2D), LASSO's L1 penalty corresponds to a diamond shape in 2D (an octahedron in higher dimensions). When the elliptic contours of the SSE intersect this diamond boundary, they often touch at a vertex or axis, yielding a zero value for some coefficient(s).

4.2 Why Coefficients Go to Zero

Mathematically, the corners of the L1 penalty boundary enable some solutions where $\beta_j = 0$. LASSO thus induces **coefficient sparsity**, which can be extremely helpful for:

- **Feature Selection**: Unimportant predictors often get removed from the model (their coefficients become zero).
- **Interpretability**: Sparser models tend to be easier to explain and visualize.

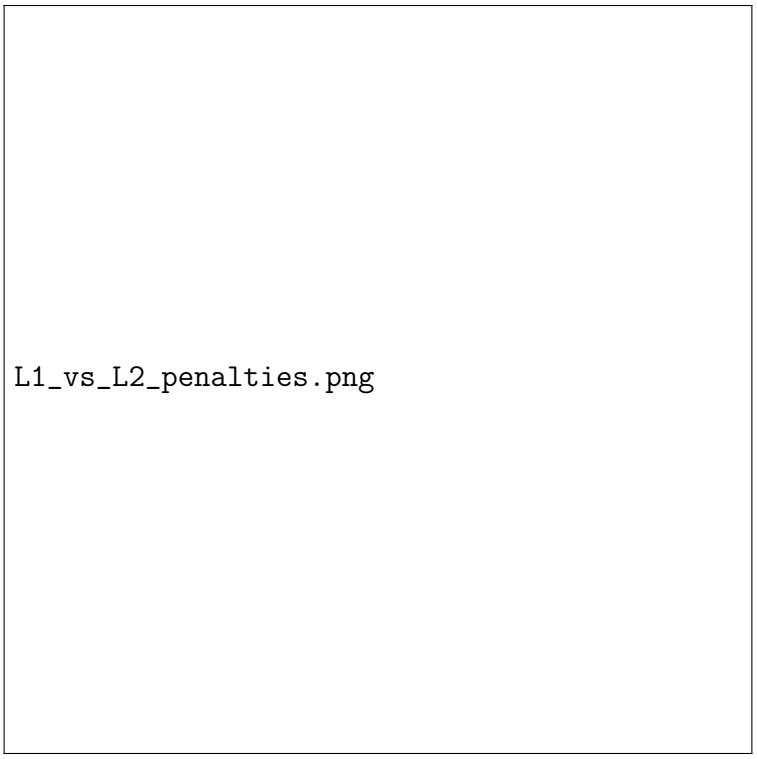

Figure 1: Geometric illustration of L1 vs. L2 penalties. The ellipse represents SSE contours, the circle represents an L2 (Ridge) penalty region, and the diamond represents an L1 (LASSO) penalty region.

4.3 Comparison with Ridge

- **Multicollinearity**: Ridge tends to handle correlated predictors smoothly by shrinking their coefficients but keeping them all nonzero.
- **Variable Selection**: LASSO sets some coefficients exactly to zero, making the model potentially much simpler.
- **Performance**: Both can outperform OLS in terms of prediction error, especially in high-dimensional or correlated settings. The best choice often depends on domain priorities (e.g., interpretability vs. strictly improved stability).

5 Choosing the Regularization Parameter λ

A critical aspect of applying Ridge or LASSO is selecting λ. Too small a λ gives minimal shrinkage, offering little advantage over OLS, while too large a λ over-penalizes coefficients, potentially causing underfitting.

5.1 Cross-Validation

Cross-validation (CV) is the standard procedure for tuning λ. In k-fold CV:

a) Partition the data into k roughly equal folds.

b) For a candidate λ, train on $k-1$ folds and validate on the remaining fold.

c) Repeat for all k folds and average the validation errors.

Vary λ over a reasonable range and select the one yielding the lowest average error.

5.2 Grid Search and Other Techniques

A simple **grid search** across a logarithmic scale of λ (e.g., $\{10^{-4}, 10^{-3}, \ldots, 10^{3}\}$) is common. For each λ, you perform the above CV process, track the mean validation error, and pick the optimal λ.

More advanced approaches like **random search** or **Bayesian optimization** can be used, especially for larger spaces or multiple hyperparameters (e.g., λ plus α in Elastic Net).

5.3 Practical Tips on λ

- **Standardization**: Always standardize or normalize predictors so the penalty treats all coefficients fairly.
- **Wide Range**: Start with a broad range of λ (spanning several orders of magnitude).
- **Coefficient Paths**: For LASSO, plot how coefficients evolve as λ changes to see which features drop out and when.

6 Detailed Illustrative Example

To demonstrate how Ridge and LASSO behave, consider a simulated dataset:

6.1 Data Generation

- Observations: $n = 100$.
- Predictors: $p = 5$.
- Let $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, with
$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.8 & 0 & 0 & 0 \\ 0.8 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$
making X_1 and X_2 strongly correlated ($\rho = 0.8$).
- True coefficients: $\boldsymbol{\beta}^* = (3, 3, 0, 0, 2)$.
- Response: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

6.2 Fitting Ridge and LASSO

a) **Standardize**: Scale each predictor to mean 0 and variance 1.

b) **Range of λ**: Use values such as $\{10^{-3}, 10^{-2}, \ldots, 10^{2}\}$.

c) **Cross-Validation**: Perform 10-fold CV for each λ.

d) **Optimal λ**: Pick the λ that yields the lowest mean CV error.

6.3 Observations

- **Ridge**: Both X_1 and X_2 remain in the model but are shrunk significantly, reflecting their correlation.
- **LASSO**: Possibly sets some coefficients to exactly zero. Given $\beta^* = 0$ for some predictors, LASSO can identify those irrelevant ones (e.g., X_3 or X_4 if they truly have no effect).
- **Predictive Performance**: Both Ridge and LASSO typically outperform OLS on a test set, especially under correlated or irrelevant predictors.

7 Extensions and Related Methods

7.1 Elastic Net

Elastic Net combines the L1 and L2 penalties:

$$J(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \alpha\lambda \sum_{j=1}^{p} |\beta_j| + (1-\alpha)\lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\alpha \in [0, 1]$. This approach is useful when you want both the feature selection property of LASSO and the stability offered by Ridge.

7.2 Regularization Beyond Linear Models

Shrinkage principles extend to other scenarios:

- **Logistic Regression**: Incorporate L1 or L2 penalties for classification tasks.
- **Neural Networks**: Weight decay (an L2 penalty) is popular for controlling overfitting in deep learning.

8 Summary and Best Practices

Key Points:

- **Multicollinearity Solution**: Ridge adds $\lambda \mathbf{I}$ to $\mathbf{X}^\top \mathbf{X}$, stabilizing coefficients for correlated predictors.
- **Coefficient Sparsity**: LASSO sets some coefficients to zero, aiding feature selection.
- **Choosing λ**: Tuning λ (via cross-validation) strikes a balance between bias and variance.

Best Practices:

a) **Standardize Predictors**: Ensures the penalty is applied uniformly across variables.

b) **Explore Broadly**: Use a wide range of λ values, typically on a log scale.

c) **Consider Elastic Net**: When you need both shrinkage of correlated predictors and feature selection.

d) **Interpret with Care**: Large λ can drastically reduce coefficients; domain expertise is essential to validate these effects.

9 Conclusion

Ridge Regression and **LASSO** are powerful techniques that help overcome the limitations of ordinary least squares by penalizing large coefficients. Ridge effectively combats multicollinearity, whereas LASSO promotes sparsity by driving some coefficients exactly to zero.

Finding an appropriate λ is paramount—too small yields minimal benefit; too large causes underfitting. Cross-validation, combined with a systematic search (grid or otherwise), is the recommended method to identify an optimal λ.

Ultimately, shrinkage methods have become cornerstone tools in modern data analysis and machine learning, allowing practitioners to build more robust, interpretable, and efficient predictive models even when predictors are highly correlated or abundant.

References

[1] Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1), 55–67.

[2] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.