Ridge and Lasso Regression
A Comprehensive Guide to Shrinkage Methods

YourName

March 7, 2025

Contents

1 Introduction

When dealing with linear regression models, one of the most common issues is **overfitting**, especially when the number of predictors is large or when predictors are highly correlated. **Ridge** and **Lasso** are two powerful techniques that address overfitting by introducing a *penalty* on the size of coefficients, effectively "shrinking" some coefficients toward zero. This document provides a detailed look into both methods, exploring their mathematical formulations, geometric interpretations, advantages, and disadvantages.

2 Brief Review: Ordinary Least Squares (OLS)

2.1 OLS Objective Function

Recall that in standard linear regression, we model the response y in terms of p predictors $x_1, \ldots, x_p$:

$$y \;=\; \beta_0 \;+\; \beta_1 x_1 \;+\; \cdots \;+\; \beta_p x_p \;+\; \epsilon,$$

where ϵ is an error term. The ordinary least squares method estimates the coefficients β_j by minimizing the **Residual Sum of Squares** (RSS):

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \text{RSS} = \min_{\boldsymbol{\beta}} \sum_{i=1}^{n} \bigl(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\bigr)^2.$$

2.2 Overfitting & High Variance

- If p is large (comparable to n, or even $p > n$), OLS can fit the training data well but generalize poorly to new data.

- Predictors may be highly correlated, leading to unstable (high variance) coefficient estimates.

Key Idea

Ridge and Lasso combat overfitting by adding a *penalty* on the size of coefficients, thus controlling their magnitude.

3 Ridge Regression

3.1 Definition and Objective

Ridge regression modifies the OLS objective by adding an L_2 (squared) penalty on coefficients:

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \left\{ \text{RSS}(\boldsymbol{\beta}) \; + \; \lambda \sum_{j=1}^{p} \beta_j^2 \right\}.$$

Typically, we do not penalize β_0 (the intercept) to keep the solution unbiased in terms of the mean response.

- $\lambda \geq 0$ is a tuning parameter that controls the amount of shrinkage.
- As λ increases, the coefficients become more heavily constrained and shrink toward 0.
- Setting $\lambda = 0$ recovers ordinary least squares.

3.2 Geometric Interpretation

- The *constraint region* from the penalty $\sum \beta_j^2 \leq c$ is a **circle** (or hypersphere in higher dimensions).
- Minimizing RSS subject to β_j^2 penalty pushes the solution to lie close to the origin in coefficient space.

3.3 Effect of Ridge Shrinkage

- Ridge regression **never fully zeroes out** any coefficient. It shrinks them but does not set them exactly to zero.
- This can improve prediction error in cases of high collinearity or when p is large.

3.4 Advantages

- **Stabilizes coefficients** in high correlation scenarios.
- Often improves **predictive performance** compared to OLS.
- Straightforward to compute via *modified* least squares formulas or algorithms.

3.5 Disadvantages

- **No true feature selection**: All variables remain in the model, albeit with smaller coefficients.
- May be less interpretable if you desire a sparse model where some coefficients are exactly zero.

4 The Lasso

4.1 Definition and Objective

Lasso regression changes the penalty to an L_1 norm (absolute values of coefficients):

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \left\{ \mathrm{RSS}(\boldsymbol{\beta}) \; + \; \lambda \sum_{j=1}^{p} |\beta_j| \right\}.$$

Similar to ridge, λ controls the amount of shrinkage.

4.2 Geometric Interpretation

- The L_1 penalty region $\sum |\beta_j| \leq c$ forms a **diamond** (or multidimensional cross-polytope).
- The corners of this diamond make it easier for the optimization to "lock onto" axes, setting some coefficients to zero.

Why Lasso Does Feature Selection

The diamond shape of the constraint region has corners that align with coordinate axes, so the solution often "hits" those corners, giving $\beta_j = 0$ for some j.

4.3 Advantages

- **Feature selection**: Some coefficients can become exactly zero, simplifying the model.
- Good predictive performance while also producing a *sparse* solution.

4.4 Disadvantages

- **If variables are highly correlated**, the lasso might randomly pick one from the group and ignore the others, leading to potentially unstable selection.
- The solution can be **harder to interpret** in scenarios of strong collinearity.

5 Choosing the Tuning Parameter λ

Both ridge and lasso depend critically on selecting λ. Common approaches:

- **Cross-Validation (CV):** For a grid of λ values, fit ridge/lasso models, compute CV error, and pick the λ that minimizes CV error.
- **Validation Set Approach**: If you have sufficient data, a single hold-out validation set can guide the choice of λ.

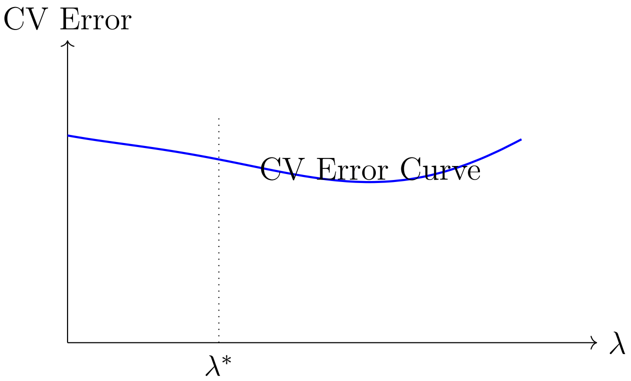

Figure 1: Selecting λ: We choose the value that minimizes cross-validation error (or near it).

6 Comparisons and Intuition

6.1 Ridge vs. Lasso: Key Differences

Aspect	Ridge	Lasso
Penalty Type	L_2 norm $(\sum \beta_j^2)$	L_1 norm $(\sum \lvert\beta_j\rvert)$
Feature Selection?	No (shrinks but doesn't set to zero)	Yes (some coefficients exactly zero)
Correlated Predictors	Spreads coefficients among them	Tends to pick one, discard others
Solution Path	Smooth	Piecewise linear
Computation	Easy to solve (analytical partial)	Requires iterative methods

Table 1: High-level comparison of Ridge vs. Lasso.

6.2 Elastic Net: A Middle Ground

Sometimes, a combination of L_1 and L_2 penalties (called **Elastic Net**) is used. It can benefit from both shrinkage and feature selection while addressing correlated predictors more smoothly.

7 Practical Tips

7.1 Preprocessing and Standardization

Because ridge and lasso penalize coefficient magnitudes, it's often recommended to **standardize or scale** each predictor before applying these methods. Otherwise, coefficients from variables with large scales might be shrunk more/less arbitrarily.

7.2 Interpretation

- Ridge might keep all variables in play, which can be useful if all have at least some partial predictive power.
- Lasso can simplify the model drastically by setting some coefficients to zero, but the specific choice of zero coefficients can vary if predictors are strongly correlated.

7.3 Model Tuning Workflow

1) **Split data** (train/validation or cross-validation).

2) **Pick a grid** of λ values (e.g., on a log scale).

3) **For each** λ, fit ridge or lasso, then compute CV error.

4) **Select the best** λ based on lowest CV error (or a one-standard-error rule).

5) **Refit** the model on the entire training set using that λ.

6) **Evaluate final performance** on a hold-out or test set if available.

8 Examples

8.1 House Price Prediction (with many features)

- **Dataset:** House prices, features like area, number of bedrooms, location, age, proximity to schools, etc.
- **Challenge:** Potentially dozens or hundreds of features, many correlated (e.g., living area vs. number of rooms).
- **Ridge:** Might distribute weights among correlated features, keeping them all but shrinking them.
- **Lasso:** Might pick only a few key features (area, location) and set the rest to zero, simplifying the model.

8.2 Gene Expression Analysis (High-Dimension)

- **Dataset:** Expression levels for thousands of genes, with only a few hundred samples.
- **Challenge:** $p \gg n$ scenario, prone to extreme overfitting if all genes are used in an OLS model.
- **Ridge:** Helps reduce variance by shrinking coefficients. All genes remain in the model but with smaller weights.
- **Lasso:** Potentially selects only a subset of genes as significant, facilitating biological interpretation (finding relevant biomarkers).

9 Conclusion

Key Takeaways

- **Ridge Regression** applies an L_2 penalty, shrinking coefficients toward zero but never setting them exactly to zero.
- **Lasso Regression** applies an L_1 penalty, promoting sparsity in coefficients. Some are set exactly to zero, yielding model simplification.
- Both methods require a **tuning parameter** λ that controls the amount of shrinkage, typically chosen via cross-validation.
- They are especially beneficial in high-dimensional or correlated predictor settings, improving model generalization and interpretability (especially for lasso).

Ridge and Lasso are cornerstones of modern regression practice when facing many predictors or risk of overfitting. With thoughtful tuning and interpretation, these methods often outperform plain OLS in predictive accuracy and can yield more stable, understandable models.