Regularization and Maximum a Posteriori Estimation

Gioele La Manno

École Polytechnique Fédérale de Lausanne (EPFL)

School of Life Science (SV)

April 2025

EPFL - BMI - UPLAMANNO

Contents

The Limits of Maximum Likelihood and Shrinkage Methods

The James-Stein Phenomenon: A Theoretical Foundation

The James-Stein phenomenon showed that maximum likelihood estimation is not always optimal, especially when estimating multiple parameters simultaneously.

Consider estimating multiple means of a multivariate normal distribution:

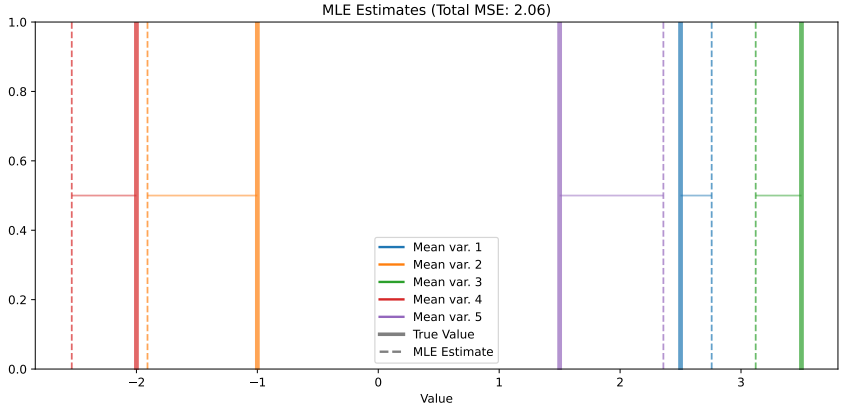
$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$$

James and Stein proved that for $p \geq 3$, we can achieve better estimation with:

$$\hat{\mu}_i^{JS} = \left(1 - \frac{(p-2)}{\sum_{j=1}^{p} Y_j^2}\right) \cdot Y_i$$

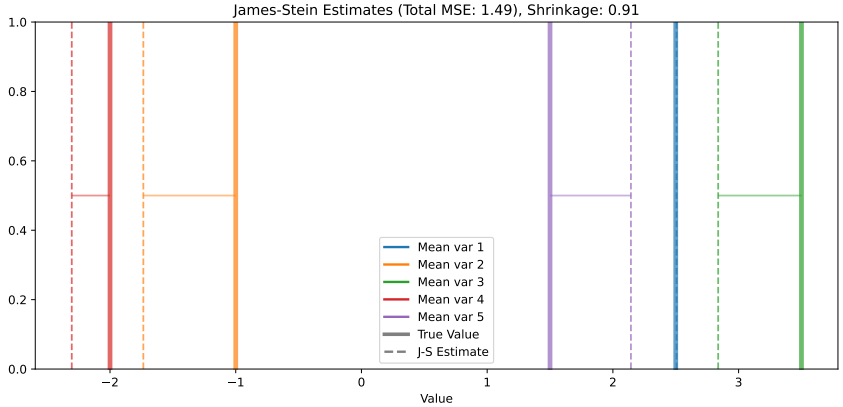
This "shrinks" each estimate toward zero, reducing overall error.

James-Stein Phenomenon: Visual Comparison



MLE estimates (dashed lines) vs. ground truth (solid lines). The MLE uses each observation directly as the estimate of its mean.

James-Stein Phenomenon: Visual Comparison



James-Stein estimates (dashed lines) vs. ground truth (solid lines). The shrinkage estimator produces values closer to the true means on average.

Underdetermination in Regression

In high-dimensional settings where parameters approach or exceed observations, MLE tends to capture noise rather than underlying patterns.

Consider a gene expression study with:

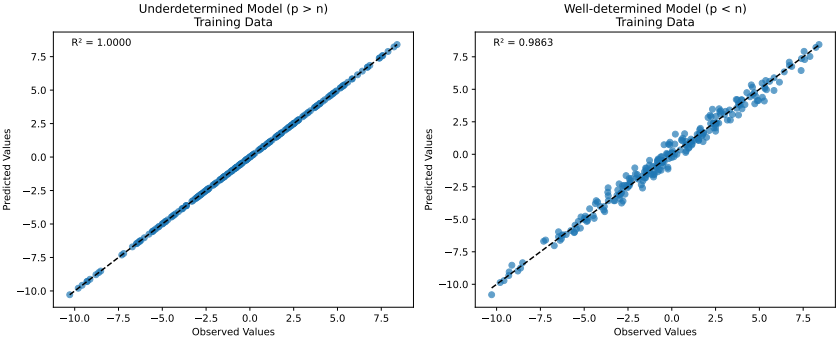
$$\mathbf{y} \in \mathbb{R}^n \quad \text{(clinical measurements)} \tag{1}$$

$$\mathbf{X} \in \mathbb{R}^{n \times p} \quad \text{(gene expression values)} \tag{2}$$

$$\boldsymbol{\beta} \in \mathbb{R}^p \quad \text{(gene effects)} \tag{3}$$

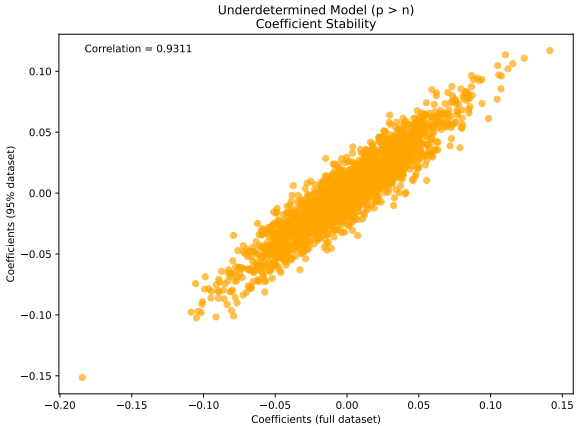
When $n \ll p$, our system becomes underdetermined, allowing infinitely many solutions that fit the training data perfectly.

Underdetermined vs. Well-Determined Systems



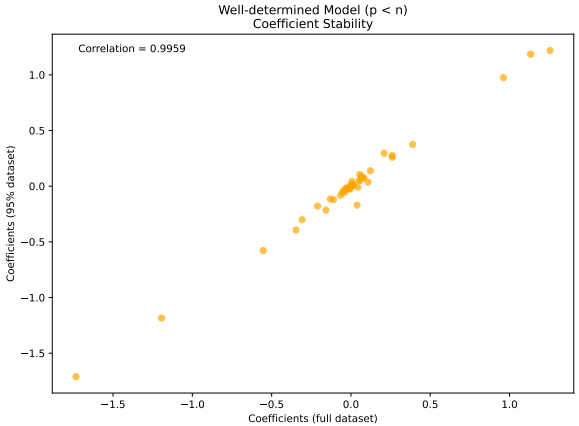
In an underdetermined system ($p > n$), we can achieve perfect fit to training data (points fall exactly on diagonal line), but this often leads to poor generalization.

Coefficient Stability in Regression Models



In underdetermined systems, coefficient estimates become highly unstable. Small changes in the training data lead to drastically different coefficient values.

Coefficient Stability in Well-Determined Systems



In contrast, well-determined systems ($n > p$) yield stable coefficient estimates that remain similar when small changes are made to the training data.

Overfitting and Generalization

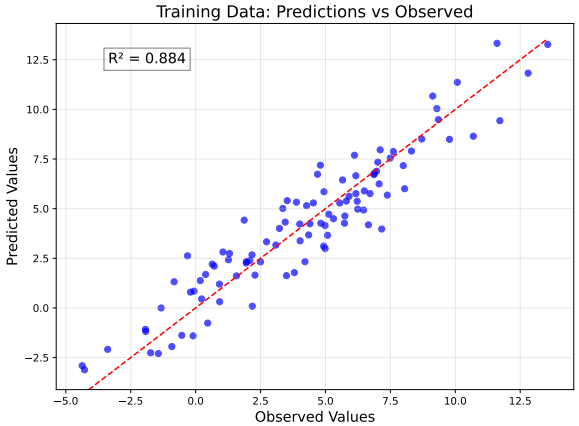
Overfitting occurs when our model captures random noise rather than underlying patterns, compromising its ability to generalize to new data.

Definition (Overfitting)

Overfitting occurs when a statistical model captures random noise or fluctuations in the training data that do not represent the underlying relationship, leading to poor generalization performance on unseen data.

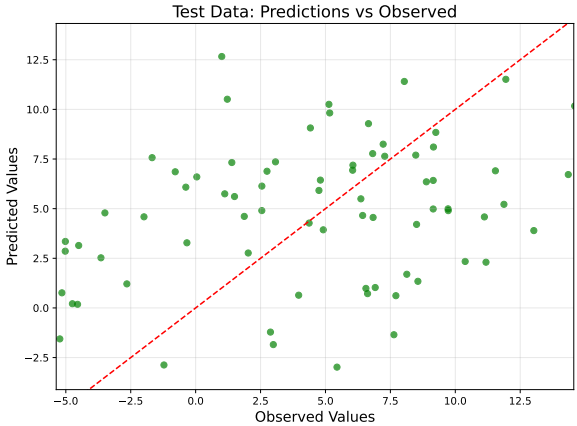
The primary goal of most biological modeling is to discover patterns that extend beyond our specific samples to the broader biological phenomenon.

Overfitting: Training vs. Test Performance



An overfit model performs deceptively well on training data (points align along diagonal), potentially misleading researchers about its predictive value.

Overfitting: Poor Generalization



The same model performs poorly on new test data, revealing that it captured noise rather than the underlying biological relationship.

Cross-Validation: A Robust Approach to Model Assessment

Cross-validation offers a systematic approach to assess model generalization by repeatedly partitioning data:

In K-fold cross-validation:

- Data is divided into K equally sized subsets or "folds"
- Model is trained K times, each using a different fold as validation
- Performance is averaged across all K iterations

Advantages:

- Uses all data points for both training and validation
- Reduces impact of random partitioning
- Provides measure of performance variability

Ridge Regression: Controlling Coefficient Magnitude

Inspired by the James-Stein phenomenon, ridge regression modifies the standard loss function by adding a penalty on coefficient magnitudes:

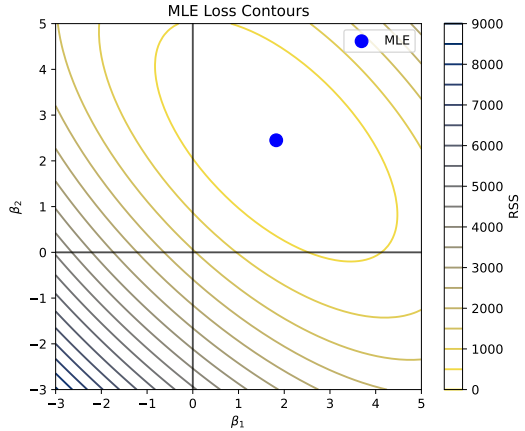
$$\text{Minimize: } \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

The regularization parameter $\lambda \geq 0$ controls the trade-off between:

- Fitting the data well (first term)
- Keeping coefficients small (second term)

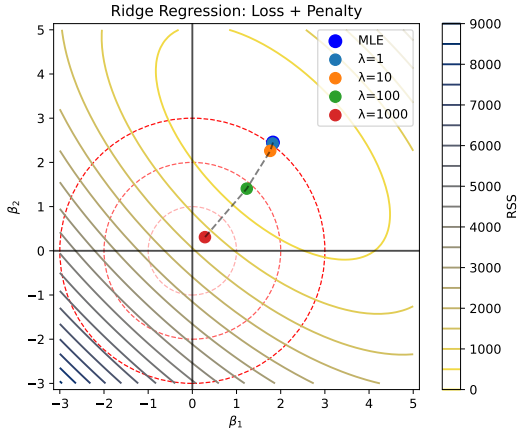
This approach introduces bias toward smaller coefficient values but can substantially reduce variance.

Ridge Regression as Constrained Optimization



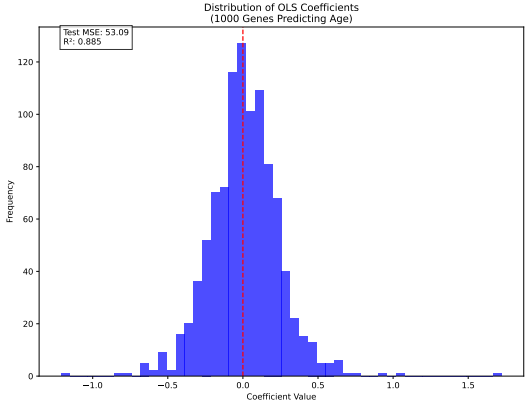
Contours of the likelihood function for two coefficients (β_1 and β_2). Without ridge regularization, the MLE is at the center of the contours.

Ridge Regression as Constrained Optimization



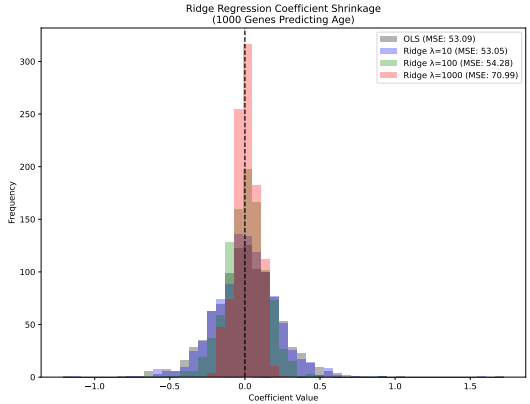
With ridge regularization, we add a penalty represented by concentric circles. The solution (marked with X) is where these circles meet the likelihood contours, shrinking

Ridge Regression: Coefficient Shrinkage



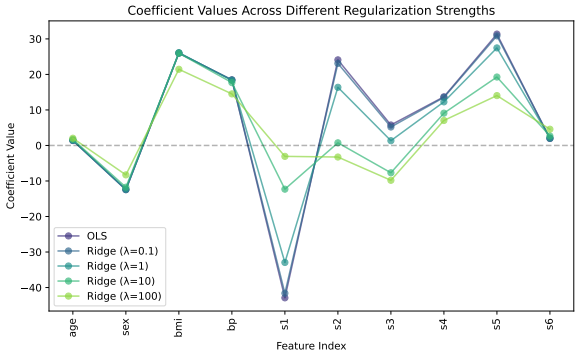
Histogram of OLS coefficients for a gene expression dataset with 1000 genes. Without regularization, many coefficients take extreme values.

Ridge Regression: Effect of Regularization



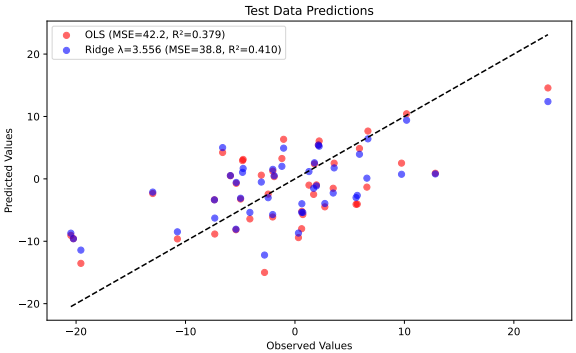
Ridge regression with different λ values. As regularization strength increases, coefficients are progressively shrunk toward zero, reducing model complexity.

Ridge Regression in Biological Applications



Ridge regression produces more stable coefficient estimates, preventing any single gene from having excessive influence on predictions.

Ridge Regression: Improved Generalization



Ridge regression typically improves prediction accuracy on new data by reducing overfitting, leading to better alignment between predicted and observed values.

Bias-Variance Trade-off: The Theoretical Foundation

Any predictor's mean squared error can be decomposed into three components:

$$\text{MSE}(\hat{f}(x)) = \text{Bias}^2(\hat{f}(x)) + \text{Variance}(\hat{f}(x)) + \text{Irreducible Error}$$

Where:

- **Bias**: Systematic deviation from the true parameter value
- **Variance**: Sensitivity to random fluctuations in training data
- **Irreducible Error**: Inherent randomness in the data-generating process

Regularization methods like ridge regression introduce some bias but can substantially reduce variance, improving overall prediction performance.

Derivation of the Bias-Variance Decomposition

Starting with the Mean Squared Error at a fixed point x:

$$\text{MSE}(\hat{f}(x)) = \mathbb{E}[(Y - \hat{f}(x))^2] = \mathbb{E}[(f(x) + \varepsilon - \hat{f}(x))^2] \tag{4}$$

Since $\mathbb{E}[\varepsilon] = 0$ and ε is independent of $\hat{f}(x)$:

$$\text{MSE}(\hat{f}(x)) = \mathbb{E}[(f(x) - \hat{f}(x))^2] + \sigma_\varepsilon^2 \tag{5}$$

Adding and subtracting $\mathbb{E}[\hat{f}(x)]$:

$$\mathbb{E}[(f(x) - \mathbb{E}[\hat{f}(x)] + \mathbb{E}[\hat{f}(x)] - \hat{f}(x))^2] + \sigma_\varepsilon^2 \tag{6}$$

$$= \mathbb{E}[(f(x) - \mathbb{E}[\hat{f}(x)])^2 + 2(f(x) - \mathbb{E}[\hat{f}(x)])(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)) + (\mathbb{E}[\hat{f}(x)] - \hat{f}(x))^2] \tag{7}$$

$$= (f(x) - \mathbb{E}[\hat{f}(x)])^2 + \mathbb{E}[(\mathbb{E}[\hat{f}(x)] - \hat{f}(x))^2] + \sigma_\varepsilon^2 \tag{8}$$

The Geometry of Regression and Rank Deficiency

In linear regression, we seek parameters $\boldsymbol{\beta}$ that minimize:

$$\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

Leading to the normal equations:

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y}$$

When $\mathbf{X}^T\mathbf{X}$ is invertible, the unique solution is:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

When p approaches or exceeds n, $\mathbf{X}$ becomes rank deficient, making the solution unstable or non-unique.

Ridge Regression: Mathematical Solution

Ridge regression modifies the normal equations by adding a diagonal matrix:

$$\text{Ridge objective} : \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \tag{9}$$

Setting the derivative to zero and solving:

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y} \tag{10}$$

$$\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \tag{11}$$

This ensures invertibility even when $p \geq n$ or predictors are collinear, stabilizing the solution.

Choosing the Regularization Parameter: The Lambda Dilemma

The regularization parameter λ controls the trade-off between:

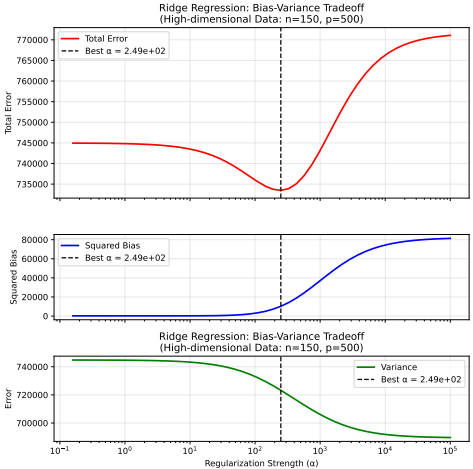
- **Low** λ: Minimal shrinkage, closer to OLS, potential overfitting
- **High** λ: Strong shrinkage, potentially oversimplified model

Cross-validation offers a principled approach to selecting λ:

1. Split data into K folds
2. For each candidate λ value:
   - Train ridge model on K-1 folds
   - Evaluate performance on held-out fold
3. Average performance across all K iterations
4. Select λ with lowest cross-validation error

This produces a characteristic U-shaped error curve that reveals the optimal regularization strength.

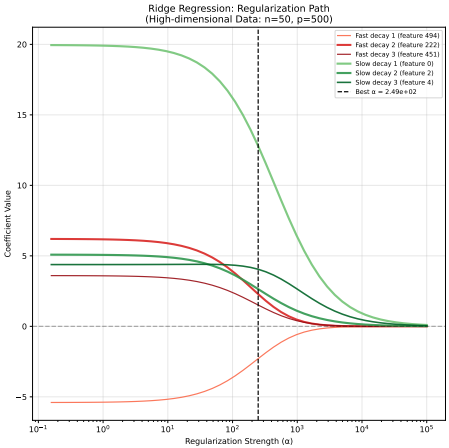
Choosing the Regularization Parameter: Visual Representation



The U-shaped curve illustrates how prediction error initially decreases as λ increases (reducing overfitting), then increases again as excessive shrinkage introduces bias.

This visualization helps identify the optimal regularization strength that balances the bias-variance tradeoff for your specific biological dataset.

Ridge Regression: Regularization Path



The regularization path illustrates how coefficients change as λ varies. Each line represents a coefficient's magnitude, revealing which features are most resistant to shrinkage (likely the most important predictors).

Bayesian Perspective and Maximum a Posteriori Estimation

The Bayesian Framework: Incorporating Prior Knowledge

The Bayesian approach treats parameters as random variables with probability distributions, not fixed but unknown quantities.

At the heart of Bayesian statistics lies Bayes' rule:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)}$$

In parameter estimation, this becomes:

$$P(\boldsymbol{\theta}|\mathbf{y}) = \frac{P(\mathbf{y}|\boldsymbol{\theta}) \cdot P(\boldsymbol{\theta})}{P(\mathbf{y})}$$

Where:

- $P(\boldsymbol{\theta}|\mathbf{y})$: Posterior distribution
- $P(\mathbf{y}|\boldsymbol{\theta})$: Likelihood function
- $P(\boldsymbol{\theta})$: Prior distribution
- $P(\mathbf{y})$: Normalizing constant

Maximum a Posteriori Estimation: A Bridge Between Paradigms

MAP estimation finds parameter values that maximize the posterior probability:

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{\theta}|\mathbf{y}) = \arg\max_{\boldsymbol{\theta}} [P(\mathbf{y}|\boldsymbol{\theta}) \cdot P(\boldsymbol{\theta})]$$

Taking logarithms:

$$\hat{\boldsymbol{\theta}}_{MAP} = \arg\max_{\boldsymbol{\theta}} [\log P(\mathbf{y}|\boldsymbol{\theta}) + \log P(\boldsymbol{\theta})]$$

This reveals that MAP estimation equals maximum likelihood estimation plus a term representing the log-prior. The prior effectively serves as a regularization term, penalizing parameter values that are a priori unlikely.

Connection Between Ridge Regression and MAP Estimation

For regression coefficients, a common choice is a Gaussian prior centered at zero:

$$\beta_j \sim \mathcal{N}(0, \tau^2)$$

The log of this prior is:

$$\log P(\boldsymbol{\beta}) = -\frac{1}{2\tau^2} \sum_{j=1}^{p} \beta_j^2 + \text{constant}$$

Incorporating this into MAP estimation:

$$\hat{\boldsymbol{\beta}}_{MAP} = \arg\max_{\boldsymbol{\beta}} \left[\log P(\mathbf{y}|\boldsymbol{\beta}) - \frac{1}{2\tau^2} \sum_{j=1}^{p} \beta_j^2 \right]$$

This is mathematically equivalent to ridge regression with $\lambda = \frac{1}{2\tau^2}$.

Beyond Ridge: Lasso Regression for Feature Selection

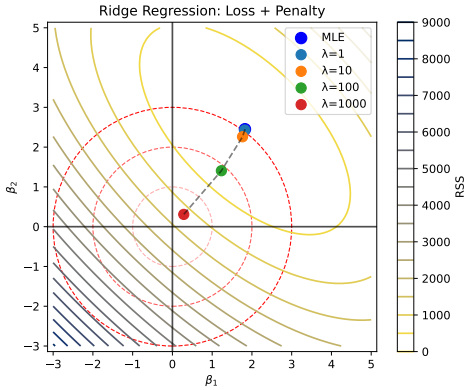
While ridge regression shrinks coefficients toward zero, it rarely sets any coefficient exactly to zero. Lasso addresses this limitation with a different penalty:

$$\text{Minimize: } \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

Where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j|$ is the L1 norm of the coefficient vector.

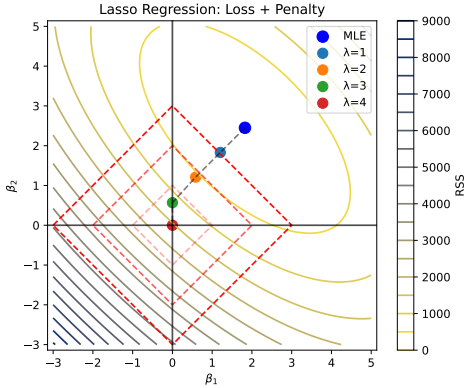
This seemingly minor change produces a fundamentally different effect: the lasso can shrink some coefficients exactly to zero, performing automatic feature selection.

Lasso vs. Ridge: Geometric Comparison



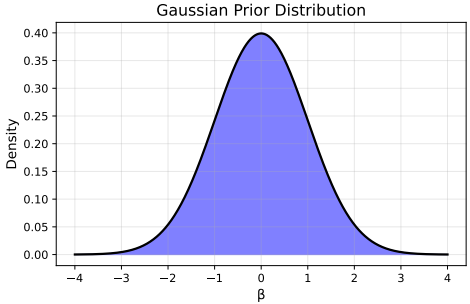
Ridge regression uses an L2 penalty, creating circular constraint regions that shrink coefficients proportionally but rarely to exactly zero.

Lasso vs. Ridge: Geometric Comparison



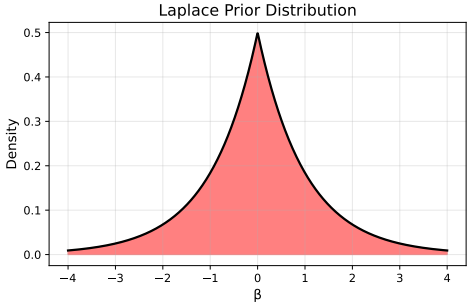
Lasso uses an L1 penalty, creating diamond-shaped constraint regions. When contours touch the corners of this region, some coefficients become exactly zero, enabling feature selection.

Bayesian Interpretation: Different Priors



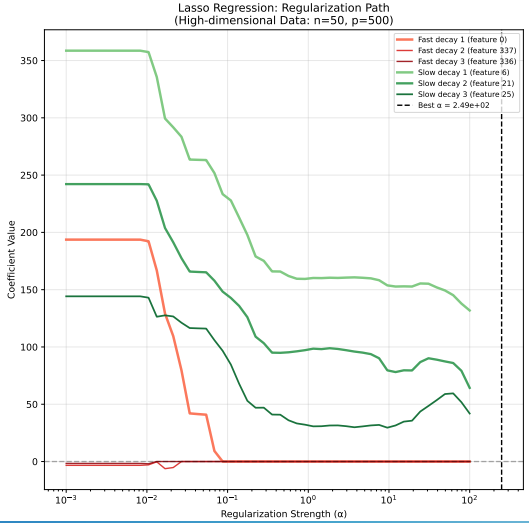
Ridge regression corresponds to MAP estimation with a Gaussian prior. The rounded peak reflects the belief that coefficients are likely small but rarely exactly zero.

Bayesian Interpretation: Different Priors



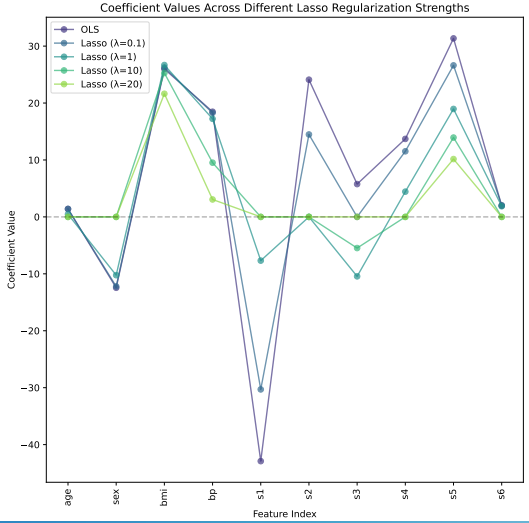
Lasso corresponds to MAP estimation with a Laplace (double-exponential) prior. The sharp peak at zero reflects the belief that many coefficients are likely exactly zero.

Lasso Regression: Variable Selection



Lasso's regularization path shows coefficients reaching exactly zero as λ increases, performing automatic variable selection by retaining only the most important predictors.

Lasso in Biological Applications



In genomics, lasso can identify a subset of genetic variants associated with disease risk from thousands of candidates, providing a focused set of targets for functional validation studies.