Lecture 9 - **Regularization and Maximum a Posteriori Estimation**

BIOENG-210 Course Notes
Prof. Gioele La Manno

April 2025

1 The Limits of Maximum Likelihood and Shrinkage Methods

1.1 The James-Stein Phenomenon: A Theoretical Foundation

Before examining the limitations of maximum likelihood estimation, it is worth exploring a remarkable theoretical discovery that challenged it: the James-Stein phenomenon.

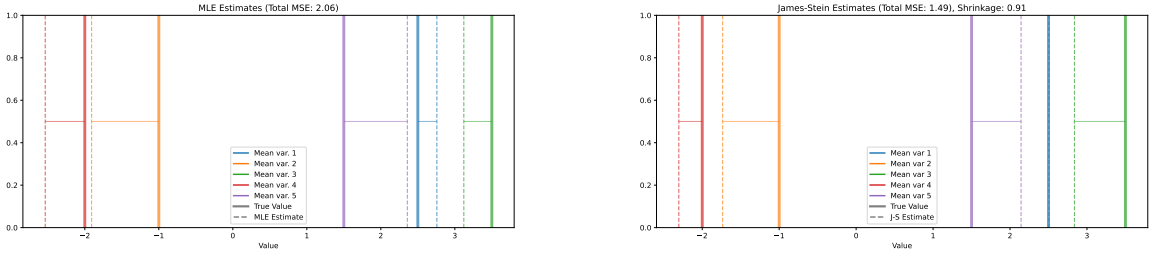
This statistical surprise revealed that MLE is not always optimal, especially when estimating multiple parameters simultaneously.

Consider a simple scenario: we observe a multivariate normal random vector $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I})$ with unknown mean vector $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$ and known diagonal covariance matrix $\boldsymbol{\Sigma} = \mathbf{I}$ (identity matrix with $\sigma^2 = 1$ along the diagonal). According to maximum likelihood theory, our best estimate for each mean would be its corresponding observation: $\hat{\mu}_i^{MLE} = Y_i$. This seems intuitive and uncontroversial.

Yet, in a groundbreaking paper in 1961, Charles Stein and later with Willard James, proved something counter-intuitive: when estimating three or more means simultaneously ($p \geq 3$), we can always achieve better performance by using:

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

This estimator "shrinks" each individual estimate toward zero by an amount that depends on the overall magnitude of all observations. The James-Stein estimator always achieves lower total mean squared error than the maximum likelihood estimator when $p \geq 3$.



(a) James-Stein phenomenon: MLE vs. G.T.

(b) James-Stein phenomenon: JS estimates vs. G.T.

Figure 1: **James-Stein phenomenon: MLE vs. JS estimates**

The surprising mathematical result that biased estimators can outperform unbiased ones has profound implications. An intuitive explanation of this phenomenon is that optimizing for the overall

mean-squared error of a combined estimator produces different results than optimizing individual parameters separately. When we care about the collective performance across all parameters, the combined shrinkage estimator performs better, even if the parameters are independent. However, if our goal is to estimate a single specific parameter with maximum accuracy, using the shrinkage estimator actually performs worse than the standard approach.

It suggests that properly constrained models—even if slightly biased—often generalize better to new data than unconstrained maximum likelihood approaches. The key insight from this phenomenon is that by introducing some bias (shrinking estimates toward zero), we can reduce variance enough to improve overall estimation performance. This insight forms the foundation for regularization methods like ridge regression. This "borrowing strength" across multiple parameters reveals a fundamental trade-off between bias and variance that underlies modern regularization methods. This insight fundamentally challenges our statistical intuition and provides the theoretical foundation for modern regularization techniques.

1.2 Underdetermination in Regression

With the James-Stein phenomenon as background, we can now examine further practical limitations of maximum likelihood estimation and the regression methods we have discussed so far.

One fundamental limitation arises when we encounter data with a (low) complexity that approaches the maximum capacity of our model. In other words, where the model has nearly enough parameters to perfectly fit the observed data, MLE tends to capture noise rather than underlying patterns. This for example happens in high-dimensional settings, where the number of parameters approaches or exceeds the number of observations.

This situation is increasingly common in modern biology, where we might collect thousands of measurements (gene expression levels, protein abundances, metabolite concentrations) but have relatively few samples due to cost, ethical considerations, or practical constraints.

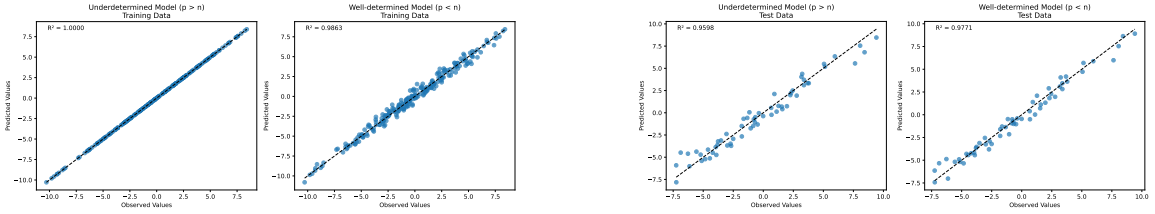
Consider a gene expression study where we measure the expression levels of $p = 20,000$ genes across only $n = 100$ patient samples. Mathematically, we have:

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

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

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

Our model takes the form $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2\mathbf{I})$. With $n \ll p$, the linear system is severely underdetermined, allowing infinitely many solutions $\hat{\boldsymbol{\beta}}$ that satisfy $\mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{y}$ exactly. The maximum likelihood estimate in this case suffers from rank deficiency of $\mathbf{X}^T\mathbf{X}$, making it unstable and non-unique. Specifically, since $\text{rank}(\mathbf{X}) \leq \min(n, p) = n = 100$, the column space of $\mathbf{X}$ can span the entire response space $\mathbb{R}^n$, enabling perfect interpolation of the training data with zero residuals: $\|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2 = 0$.



(a) Underdetermined regression: training data

(b) Underdetermined regression: training data

Figure 2: **Underdetermined regression: training data**

Definition 1.1 (Underdetermined System)**.** A linear system $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$ is underdetermined when the number of unknown parameters exceeds the number of independent equations (effective constraints).

Mathematically, this occurs when $\text{rank}(\mathbf{X}) < p$, where p is the number of parameters. In regression contexts, this typically happens when:

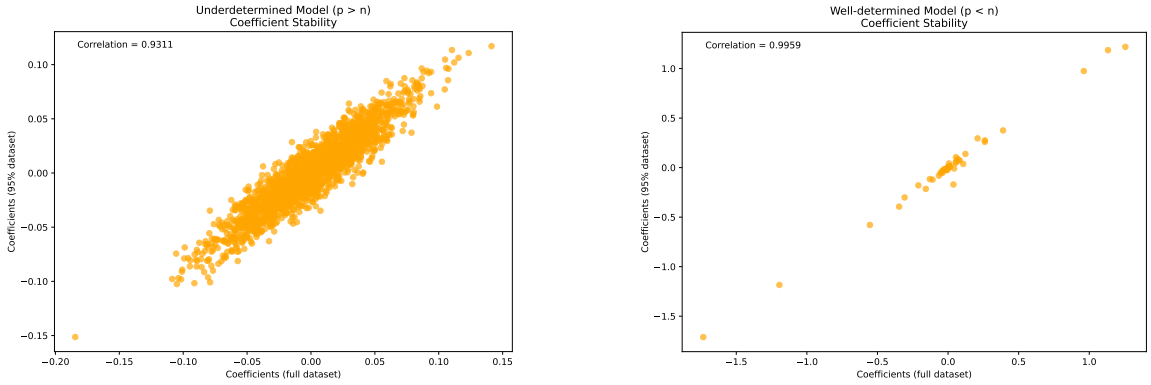
- The number of predictors exceeds the number of observations ($p > n$)
- The predictors contain perfect collinearities or near-collinearities

Underdetermined systems have infinitely many solutions, making parameter estimation unstable without additional constraints.

Definition 1.2 (Well-Determined System)**.** A linear system $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$ is well-determined when there are exactly as many independent equations as unknown parameters. Mathematically, this occurs when $\text{rank}(\mathbf{X}) = p$, where p is the number of parameters. In regression contexts, this typically requires:

- The number of observations exceeds the number of predictors ($n > p$)
- The design matrix $\mathbf{X}$ has full column rank (no perfect collinearities)

Well-determined systems have unique solutions, allowing for stable parameter estimation without regularization.



(a) Underdetermined regression: betas

(b) Not underdetermined regression: betas

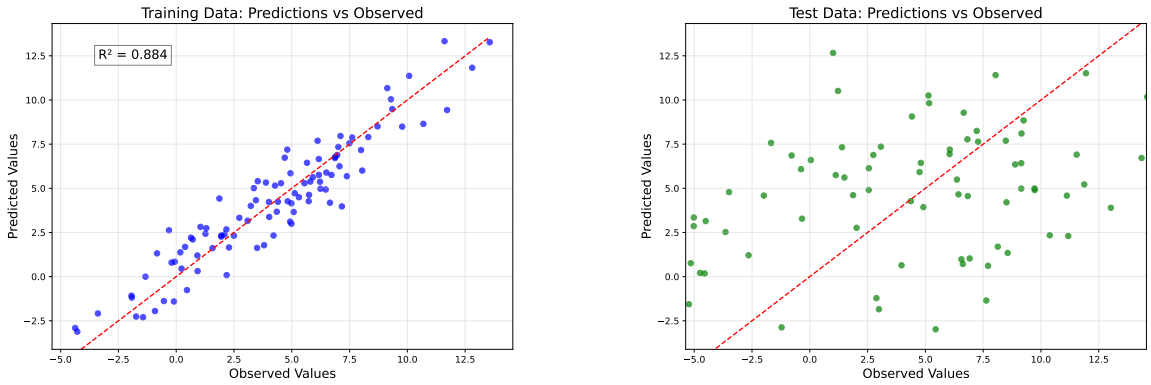
Figure 3: **Beta coefficient stability in underdetermined vs. well-determined regression**

1.3 Overfitting and Generalization: The Need for Testing

Overfitting represents a critical failure mode where our model captures random noise in the training data rather than the underlying biological signal. An overfit model performs well on the data used to train it but generalizes poorly to new, unseen data. This undermines the primary goal of most biological modeling: to discover patterns that extend beyond our specific samples to the broader biological phenomenon.

Definition 1.3 (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.

Consider a gene expression study examining how thousands of genes predict disease risk. With traditional maximum likelihood approaches, the model might identify seemingly strong associations that actually represent random patterns in the small sample rather than genuine biological relationships. When applied to new patients, such a model would likely perform poorly, potentially leading researchers toward spurious conclusions and wasted experimental effort.



(a) Overfitting in regression: training vs. test data

(b) Overfitting in regression: training vs. test data

Figure 4: **Overfitting in regression: training vs. test data**

To detect and prevent overfitting, we need methods to assess a model's generalization performance—its ability to make accurate predictions on new, unseen data. The most direct approach is to use a test set:

1. Split the available data into a training set (typically 70-80% of the data) and a test set (the remaining 20-30%)
2. Fit the model using only the training data
3. Evaluate the model's performance on the test set, which was not used during model fitting

The test set provides an unbiased estimate of the model's generalization performance because these data points were not used to fit the model. A large discrepancy between training and test performance indicates overfitting.

1.4 Cross-Validation: A More Robust Approach

The single train-test split approach has limitations, particularly with smaller datasets where the specific division might significantly influence results. Cross-validation offers a more robust alternative by repeatedly partitioning the data in different ways.

In K-fold cross-validation, the data is divided into K equally sized subsets or "folds" (typically K=5 or K=10). The model is then trained and evaluated K times, with each iteration using a different fold as the validation set and the remaining K-1 folds as the training set. The final performance metric is averaged across all K iterations, providing a more reliable estimate of the model's generalization performance.

This approach has several advantages:

- It uses all data points for both training and validation
- It reduces the impact of random partitioning on performance estimates
- It provides a measure of variability in performance across different data subsets

When data is particularly limited, leave-one-out cross-validation (LOOCV) represents an extreme case where K equals the number of observations. While computationally intensive, this maximizes the amount of training data available in each iteration.

1.5 Ridge Regression: Controlling Coefficient Magnitude

Inspired by the insights from the James-Stein phenomenon, let's take an empirical approach to address the limitations we have seen. What if we modify our standard loss function by adding a secondary term that penalizes coefficient values far from zero? This approach is simple but powerful:

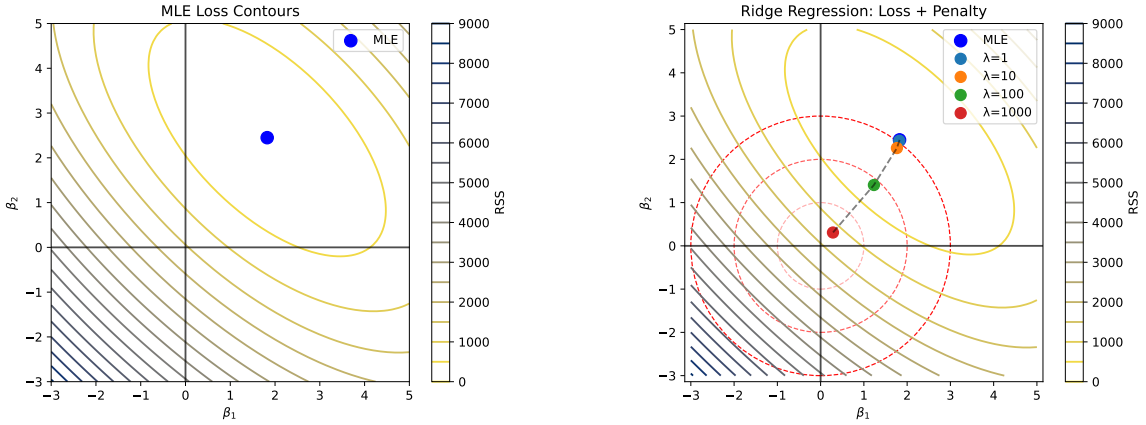
$$\text{Minimize: } \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}\beta_j^2$$

Or in matrix notation:

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

This modification introduces a compromise between two objectives: fitting the data well (first term) and keeping coefficients small (second term).

The parameter $\lambda \geq 0$ controls how much we prioritize small coefficients over fitting the data perfectly. Let's see what this achieves. By intentionally introducing some bias toward smaller coefficient values, we gain several important benefits:



(a) Contour plot of MLE without ridge

(b) Contour plot of ridge regression

Figure 5: **Contour plot of MLE without ridge vs. ridge regression**

This approach is called "Ridge Regression", "L2 Regularization" or "Shrinkage" regression.

It addresses several key limitations we saw before. First, it mitigates multicollinearity by ensuring that no coefficient becomes excessively large due to correlated predictors. When predictors are perfectly correlated, ordinary least squares produces unstable estimates, while ridge provides a unique, stable solution.

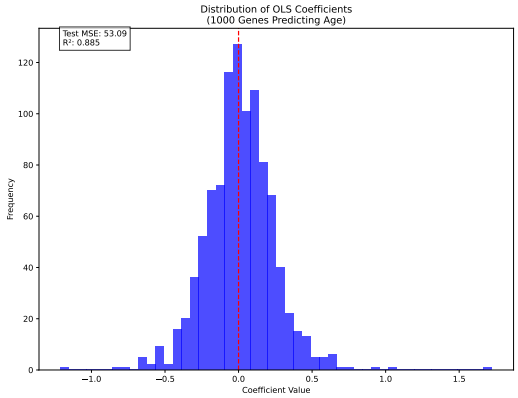
Second, by constraining coefficient magnitudes, ridge reduces model complexity and variance, helping prevent overfitting especially when the number of predictors approaches or exceeds the number of observations.

More in general adding a penalty term to the loss function is called regularization because: It regulates the complexity of the model by adding constraints on the coefficients. It regularizes the solution by making it more stable and less sensitive to small changes in the training data.

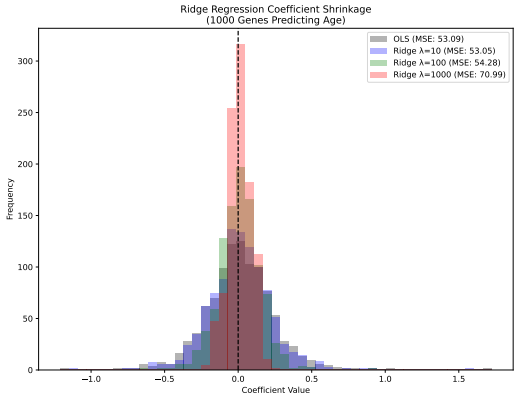
1.5.1 Ridge Regression in Biological Applications

Ridge regression also provides computational advantages for high-dimensional biological data.

This regularization is particularly valuable in biomarker studies where we might evaluate many potential predictors of disease status with limited samples. The ridge penalty prevents any single marker from having an excessive influence, producing a more stable and generalizable model.

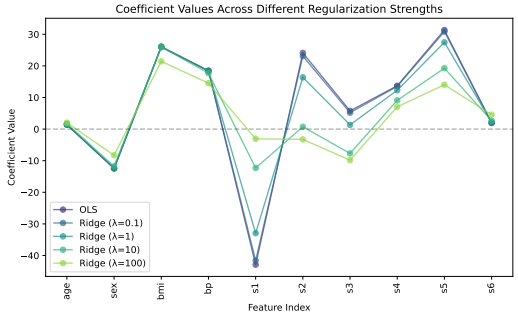


(a) Ridge regression coefficients

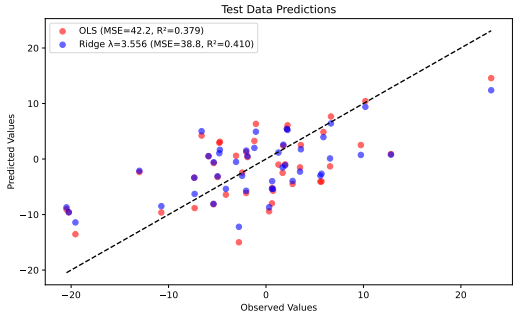


(b) Ridge regression coefficients

Figure 6: **Ridge regression coefficients**



(a) Ridge regression coefficients



(b) Prediction vs observed

Figure 7: **Ridge regression coefficients and predictions**

This makes ridge regression suitable for applications like genomic prediction, where we model phenotypes based on thousands or millions of genetic variants.

1.6 Bias-Variance Trade-off: The Theoretical Foundation

The efficacy of regularization methods lies in their ability to balance the bias-variance trade-off—a fundamental concept in statistical learning. This trade-off explains why introducing some bias through regularization can actually improve overall prediction performance. 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** represents the systematic deviation of an estimator from the true parameter value, reflecting how far the average estimate (across many potential datasets) is from the truth
- **Variance** captures the estimator's sensitivity to random fluctuations in the training data, measuring how much estimates would vary if we repeatedly sampled new datasets from the same population
- **Irreducible Error** stems from inherent randomness in the data-generating process that cannot be eliminated regardless of model quality, often associated with unmeasured factors or fundamental stochasticity

But ok, this might seem out of the blue, let's see how this is derived using the algebra of expectation.

To derive the bias-variance decomposition, let's consider the Mean Squared Error (MSE) of a predictor $\hat{f}(x)$ for a true function $f(x)$ at a fixed point x:

$$\text{MSE}(\hat{f}(x)) = \mathbb{E}[(Y - \hat{f}(x))^2] \tag{4}$$
$$\tag{5}$$

If we denote the true function as $f(x)$ and acknowledge that $Y = f(x) + \varepsilon$ where ε represents random noise with $\mathbb{E}[\varepsilon] = 0$ and $\text{Var}(\varepsilon) = \sigma_\varepsilon^2$, we can rewrite:

$$\begin{aligned}
\text{MSE}(\hat{f}(x)) &= \mathbb{E}[(f(x) + \varepsilon - \hat{f}(x))^2] && (6)\\
&= \mathbb{E}[((f(x) - \hat{f}(x)) + \varepsilon)^2] && (7)\\
&= \mathbb{E}[(f(x) - \hat{f}(x))^2 + 2\varepsilon(f(x) - \hat{f}(x)) + \varepsilon^2] && (8)\\
& && (9)
\end{aligned}$$

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

$$\begin{aligned}
\text{MSE}(\hat{f}(x)) &= \mathbb{E}[(f(x) - \hat{f}(x))^2] + \mathbb{E}[\varepsilon^2] && (10)\\
&= \mathbb{E}[(f(x) - \hat{f}(x))^2] + \sigma_\varepsilon^2 && (11)\\
& && (12)
\end{aligned}$$

Let's focus on the first term. We can add and subtract $\mathbb{E}[\hat{f}(x)]$:

$$\mathbb{E}[(f(x) - \hat{f}(x))^2] = \mathbb{E}[(f(x) - \mathbb{E}[\hat{f}(x)] + \mathbb{E}[\hat{f}(x)] - \hat{f}(x))^2] \tag{13}$$
$$= \mathbb{E}[((f(x) - \mathbb{E}[\hat{f}(x)]) + (\mathbb{E}[\hat{f}(x)] - \hat{f}(x)))^2] \tag{14}$$
$$\tag{15}$$

Expanding the square:

$$\mathbb{E}[(f(x) - \hat{f}(x))^2] = \mathbb{E}[(f(x) - \mathbb{E}[\hat{f}(x)])^2 + 2(f(x) - \mathbb{E}[\hat{f}(x)])(\mathbb{E}[\hat{f}(x)] - \hat{f}(x)) + (\mathbb{E}[\hat{f}(x)] - \hat{f}(x))^2] \tag{16}$$
$$\tag{17}$$

Note that $f(x) - \mathbb{E}[\hat{f}(x)]$ is constant with respect to the expectation, and:

$$\mathbb{E}[\mathbb{E}[\hat{f}(x)] - \hat{f}(x)] = \mathbb{E}[\hat{f}(x)] - \mathbb{E}[\hat{f}(x)] = 0 \tag{18}$$
$$\tag{19}$$

So the middle term vanishes, and we get:

$$\mathbb{E}[(f(x) - \hat{f}(x))^2] = (f(x) - \mathbb{E}[\hat{f}(x)])^2 + \mathbb{E}[(\mathbb{E}[\hat{f}(x)] - \hat{f}(x))^2] \tag{20}$$
$$= (f(x) - \mathbb{E}[\hat{f}(x)])^2 + \mathrm{Var}(\hat{f}(x)) \tag{21}$$
$$\tag{22}$$

The term $(f(x) - \mathbb{E}[\hat{f}(x)])^2$ is the squared bias of $\hat{f}(x)$, and $\mathrm{Var}(\hat{f}(x))$ is its variance. Substituting back into our MSE expression:

$$\mathrm{MSE}(\hat{f}(x)) = (f(x) - \mathbb{E}[\hat{f}(x)])^2 + \mathrm{Var}(\hat{f}(x)) + \sigma_\varepsilon^2 \tag{23}$$
$$= \mathrm{Bias}^2(\hat{f}(x)) + \mathrm{Variance}(\hat{f}(x)) + \text{Irreducible Error} \tag{24}$$
$$\tag{25}$$

Unregularized maximum likelihood estimators are typically unbiased but may have high variance, especially with limited data or many parameters. Regularization methods like ridge regression introduce some bias by shrinking estimates toward zero, but they can substantially reduce variance. When the reduction in variance exceeds the increase in squared bias, the overall mean squared error decreases—the key insight from the James-Stein phenomenon.

1.7 Tunable model complexity

In biological applications, this trade-off has practical implications. Complex models with many parameters might precisely fit training data (low bias) but generalize poorly to new observations (high variance). Conversely, overly simplistic models might generalize more reliably (low variance) but systematically miss important patterns (high bias). Regularization methods help navigate this trade-off by controlling model complexity based on the available data.

This framework explains why regularization becomes increasingly important as the number of predictors approaches or exceeds the number of observations—a common scenario in modern biological studies, for example using gene expression data to predict a biological outcome. With limited samples relative to predictors, unregularized estimates suffer extremely high variance, making some form of regularization essential for reliable inference and prediction.

1.8 The Geometry of Model Estimation and Rank Deficiency

A deeper understanding of estimation challenges in high-dimensional settings comes from examining the geometric interpretation of regression that we explored in previous lectures. Recall that in multiple linear regression, we can view the estimation process as projecting the response vector $\mathbf{y} \in \mathbb{R}^n$ (an n-dimensional vector of observations) onto the column space of the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ (an $n \times p$ matrix where columns represent the p predictors).

In linear regression, we seek to find parameters $\boldsymbol{\beta}$ that minimize the sum of squared residuals:

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

This leads 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:

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

substituting the beta expression one gets y hat

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

Geometrically, this represents the orthogonal projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$.

When the number of predictors p approaches or exceeds the number of observations n, the design matrix $\mathbf{X}$ becomes rank deficient or nearly so.

Geometrically, this means that:

- If the design matrix $\mathbf{X}$ has full rank equal to n (the number of observations), its column space spans the entire response space $\mathbb{R}^n$. This allows us to find a $\boldsymbol{\beta} \in \mathbb{R}^p$ such that $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$ exactly, resulting in zero training error. This typically happens when we have at least as many predictors as observations ($p \geq n$) and the predictors are not redundant.
- If $\mathbf{X}$ has rank less than p (the number of parameters), the system $\mathbf{X}\boldsymbol{\beta} = \mathbf{y}$ is underdetermined, and infinitely many parameter vectors $\boldsymbol{\beta} \in \mathbb{R}^p$ can perfectly fit the training data. This occurs either when $p > n$ or when the predictors contain collinearities.

These scenarios highlight a fundamental problem: when the model has enough flexibility to fit the training data perfectly, it captures not only the underlying pattern but also the random noise. This perfect fit to training data is precisely what constitutes overfitting.

I will show how regularization act on the problem by adding a diagonal matrix to the normal equations:

$$\text{Ridge objective} : \min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \tag{26}$$

Taking the derivative with respect to $\boldsymbol{\beta}$ and setting to zero:

$$\frac{\partial}{\partial \boldsymbol{\beta}}(\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2) = \mathbf{0} \tag{27}$$

$$-2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} = \mathbf{0} \tag{28}$$

$$-\mathbf{X}^T\mathbf{y} + \mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta} = \mathbf{0} \tag{29}$$

$$\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \lambda\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y} \tag{30}$$

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\boldsymbol{\beta} = \mathbf{X}^T\mathbf{y} \tag{31}$$

$$\tag{32}$$

This modification ensures invertibility by adding $\lambda\mathbf{I}$ to $\mathbf{X}^T\mathbf{X}$, ensuring that the matrix is invertible even when $p \geq n$ or when predictors are collinear, which stabilizes the solution.

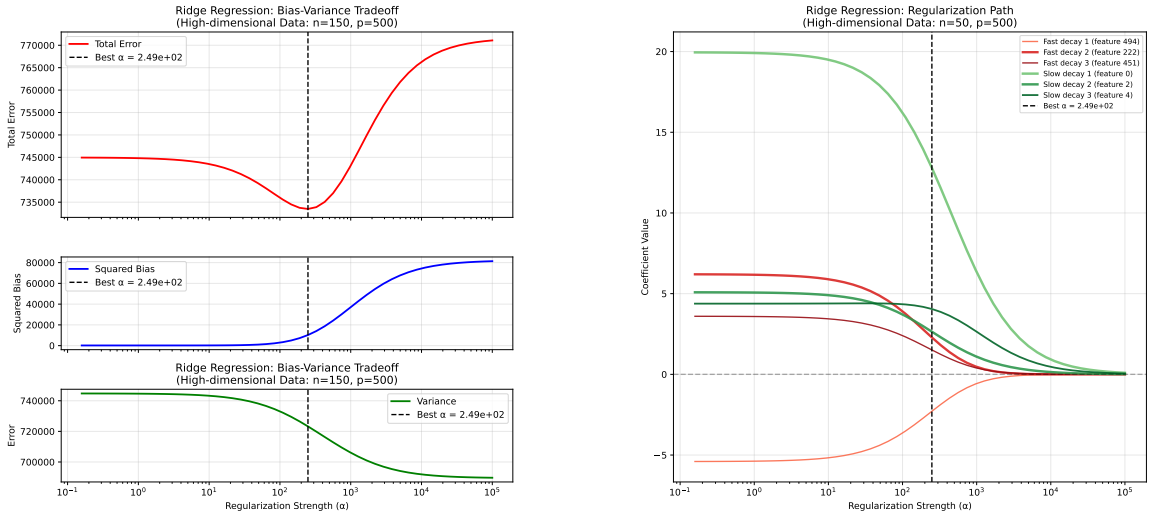
1.9 Choosing the Regularization Parameter

The choice of the regularization parameter λ critically affects model performance. Too small a value provides insufficient regularization, while too large a value overshrinks coefficients toward zero, potentially removing genuine biological signal.

Cross-validation provides a systematic approach to determine the optimal λ by fitting models with various penalty strengths and selecting the value that minimizes prediction error on held-out data. A typical implementation follows these steps:

1. Create a grid of λ values, usually on a logarithmic scale (e.g., $\lambda \in \{10^{-4}, 10^{-3}, ..., 10^{4}\}$)
2. For each λ value:
   (a) Divide the data into K folds (typically K=5 or K=10)
   (b) For each fold, fit a ridge regression model on the remaining K-1 folds using the current λ
   (c) Calculate the prediction error on the held-out fold
   (d) Average the prediction errors across all K folds to obtain the cross-validation error for this λ
3. Select the λ that minimizes the average cross-validation error

A useful visualization of regularization's effect is the "regularization path," which shows how coefficient values change as λ varies.



(a) Cross-validation error for different λ values

(b) Regularization path for ridge regression

Figure 8: **cross-validation for ridge regression and Regularization path**

In the regularization path, each line represents one coefficient's magnitude as λ increases (typically moving right to left). For ridge regression, coefficients shrink continuously toward zero but rarely reach exactly zero. The path reveals which features are most resistant to shrinkage (the features that decay the latest when λ increases, which are likely the most important predictors) and how the model complexity changes with regularization strength.

The cross-validation curve typically exhibits a U-shape: error decreases as λ increases from zero (reducing overfitting), reaches a minimum at the optimal regularization strength, then increases again as excessive shrinkage introduces bias. Some implementations select the largest λ within one standard error of the minimum (the "one-standard-error rule") to favor simpler models when the performance difference is statistically negligible.

In biological applications, examining the regularization path can provide insights into feature importance and sensitivity to regularization, while the cross-validation curve helps select an appropriate model complexity that balances bias and variance for the available data.

2 Bayesian Perspective and Maximum a Posteriori Estimation

2.1 The Bayesian Framework: Incorporating Prior Knowledge

While ridge regression addresses overfitting through coefficient constraints, we can develop a deeper understanding by adopting a Bayesian perspective. The Bayesian approach differs fundamentally from maximum likelihood by treating parameters themselves as random variables with probability distributions, not fixed but unknown quantities.

At the heart of Bayesian statistics lies Bayes' rule, which provides a framework for updating beliefs based on new evidence:

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

In the context of parameter estimation, we rewrite this as:

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

Here, $P(\boldsymbol{\theta}|\mathbf{y})$ represents the posterior distribution of parameters given the observed data, $P(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function we are already familiar with, $P(\boldsymbol{\theta})$ is the prior distribution encoding our beliefs about parameters before seeing the data, and $P(\mathbf{y})$ serves as a normalizing constant.

The Bayesian framework offers several advantages for biological data analysis. First, it naturally incorporates prior knowledge—perhaps from previous studies, published literature, or biological constraints—through the prior distribution. Second, rather than providing point estimates, it yields a full posterior distribution that quantifies uncertainty in parameter values. Third, it handles small sample sizes more gracefully than maximum likelihood by relying partially on prior information when data are limited.

In systems biology, where we might model complex networks with many parameters, the Bayesian approach allows integration of existing knowledge about pathway structure or protein interactions. Similarly, in clinical biomarker studies, we might incorporate prior information about which markers have shown promise in earlier investigations.

2.2 Maximum a Posteriori Estimation: A Bridge Between Paradigms

While full Bayesian analysis involves characterizing the entire posterior distribution—often requiring complex computational methods—maximum a posteriori (MAP) estimation provides a middle ground between Bayesian and frequentist approaches. MAP estimation finds the parameter values that maximize the posterior probability:

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

Since $P(\mathbf{y})$ doesn't depend on $\boldsymbol{\theta}$, we can ignore it during maximization. Taking logarithms, MAP estimation becomes:

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

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

2.3 Connection Between Ridge Regression and MAP Estimation

The connection between regularization methods and Bayesian statistics becomes clear when we consider specific prior distributions. For regression coefficients, a common choice is a Gaussian prior centered at zero:

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

This reflects the belief that most coefficients are likely to be small, with extreme values being rare. The log of this prior is:

$$\log P(\beta_j) = -\frac{\beta_j^2}{2\tau^2} + \text{constant}$$

For all coefficients together:

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

Incorporating this prior into MAP estimation gives:

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

Comparing this with ridge regression:

$$\text{Maximize: } \ell(\boldsymbol{\beta}) - \lambda \sum_{j=1}^{p} \beta_j^2$$

We see that ridge regression is mathematically equivalent to MAP estimation with a Gaussian prior, where $\lambda = \frac{1}{2\tau^2}$.

This Bayesian interpretation provides deeper insight into regularization. The regularization parameter λ corresponds to the precision of our prior belief about coefficient values. A large λ (small prior variance τ^2) represents strong prior conviction that coefficients should be close to zero, while a small λ (large prior variance) represents weak prior beliefs, giving more weight to the data.

In biological applications, this interpretation helps guide the choice of regularization strength based on existing knowledge. For example, in genomic studies where most genetic variants have small or negligible effects, a relatively strong regularization (informed by the expected distribution of effect sizes from previous studies) might be appropriate. Conversely, when modeling a small set of pre-selected biomarkers with established relevance, weaker regularization may be warranted.

2.4 Beyond Ridge: Lasso Regression for Feature Selection

While ridge regression effectively addresses overfitting by shrinking all coefficients toward zero, it rarely sets any coefficient exactly to zero. This limitation becomes significant in high-dimensional biological datasets where we often seek to identify a smaller subset of relevant features from thousands of candidates.

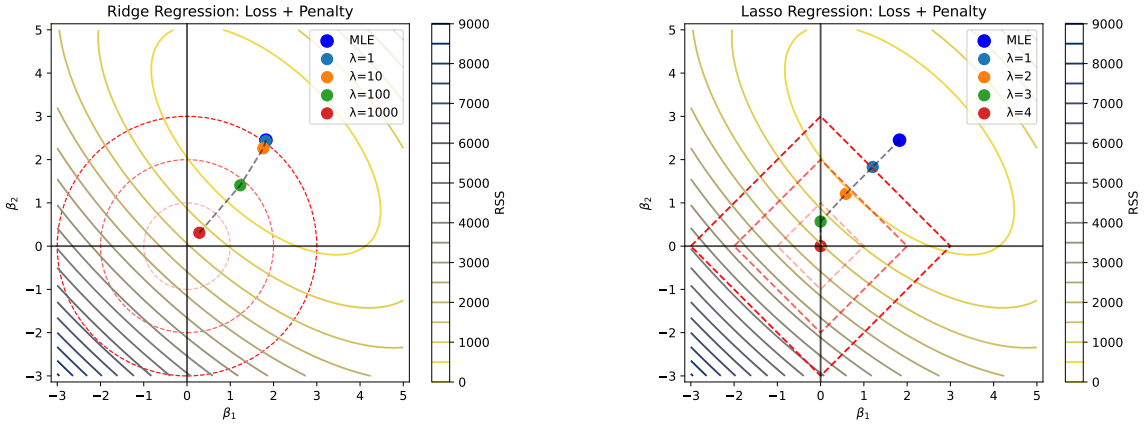
Lasso (Least Absolute Shrinkage and Selection Operator) regression addresses this need by using a different penalty term:

$$\text{Minimize: } \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

in matrix and norm notation:

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

where $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p}|\beta_j|$ is the L1 norm of the coefficient vector. The L1 norm do not produces concentric circles ("ridges") like the L2 norm, but rather produces a diamond shape.

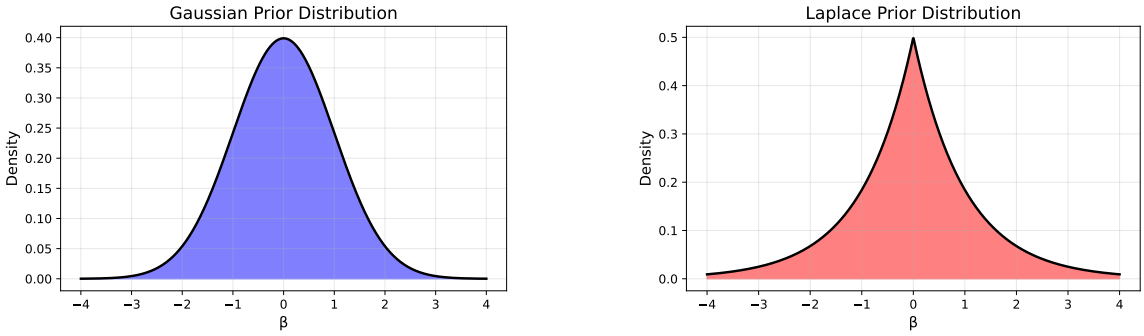


(a) Contour plot of Ridge regression

(b) Contour plot of Lasso regression

Figure 9: **Contour plot of MLE without ridge vs. ridge regression**

The lasso penalty encourages sparsity in the coefficient estimates, effectively performing variable selection by shrinking some coefficients exactly to zero.

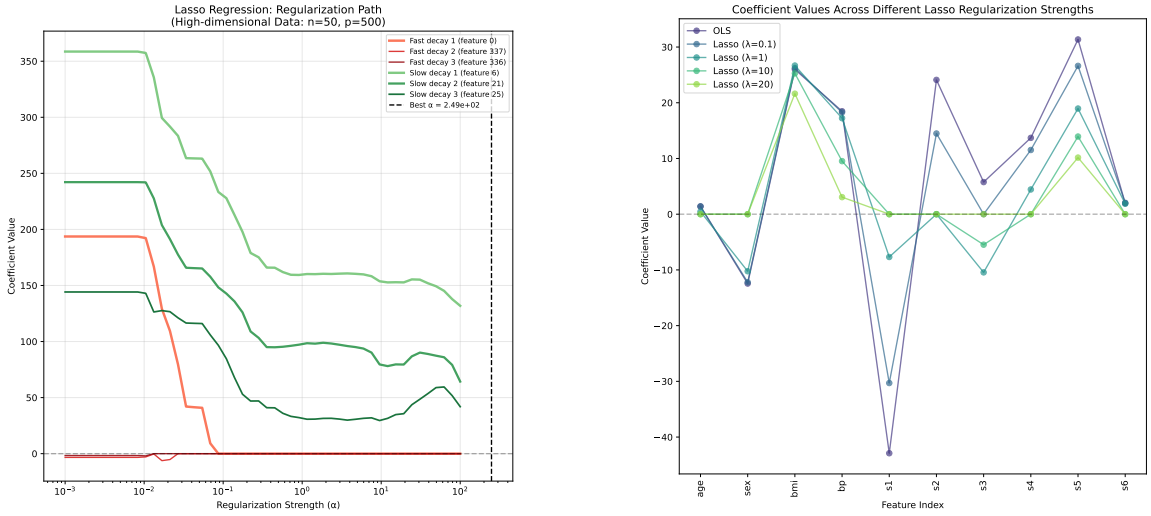


(a) Gaussian prior distribution

(b) Laplace prior distribution

Figure 10: **Comparison of Gaussian and Laplace priors**

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



(a) Regularization path for lasso regression

(b) Lasso regression coefficients

Figure 11: **Regularization path for lasso regression**

From a Bayesian perspective, the lasso corresponds to MAP estimation with a Laplace (double-exponential) prior on the coefficients:

$$P(\beta_j) = \frac{1}{2b} \exp\left(-\frac{|\beta_j|}{b}\right)$$

Where $b > 0$ is the scale parameter controlling the spread of the distribution. Comparing with the lasso penalty, we can explicitly see that $\lambda = \frac{1}{b}$, meaning the regularization strength is inversely proportional to the scale of our prior distribution.

The sharper peak of the Laplace distribution around zero (compared to the Gaussian prior's more rounded shape) explains the lasso's tendency to produce sparse solutions.

2.5 Lasso Regression in Biological Applications

Lasso regression has become particularly valuable in modern biological applications:

In genomics, researchers might use lasso to identify a subset of genetic variants associated with disease risk from hundreds of thousands of candidates, providing a more focused set of targets for functional validation studies.

Despite these advantages, lasso has limitations. When predictors are highly correlated—common in biological data—lasso tends to select one variable from each correlated group somewhat arbitrarily. This can make biological interpretation challenging when we know that multiple correlated features (e.g., genes in the same pathway) likely influence the outcome together.