Mathematical Analysis: The Impact of Multicollinearity on OLS Estimator Stability

Lecture Notes: Advanced Machine Learning

1 Introduction

In Ordinary Least Squares (OLS) regression, we seek to estimate the coefficients β for the linear model $y = X\beta + \epsilon$. While OLS is the Best Linear Unbiased Estimator (BLUE) under ideal conditions, the presence of **Multicollinearity** (high correlation between features) catastrophically destabilizes the model.

This report provides a rigorous derivation of this instability, tracing the error propagation from geometric flatness to eigenvalue singularity, and finally to the explosion of variance in the coefficient estimates.

2 Step 1: Deriving the Variance of the Estimator

We begin by establishing the variance of the OLS estimator $\hat{\beta}$ in terms of the data matrix X.

Consider the standard linear model:

$$y = X\beta + \epsilon \tag{1}$$

Where:

- $y \in \mathbb{R}^n$ is the target vector.
- $X \in \mathbb{R}^{n \times p}$ is the deterministic feature matrix.
- $\epsilon \in \mathbb{R}^n$ is the noise vector, with $E[\epsilon] = 0$ and $\text{Var}(\epsilon) = \sigma^2 I$.

The OLS estimator is given by the normal equation:

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{2}$$

To measure stability, we calculate the Variance-Covariance Matrix of $\hat{\beta}$. Using the property $\text{Var}(Ay) = A\text{Var}(y)A^T$:

$$\text{Var}(\hat{\beta}) = \text{Var}\left((X^T X)^{-1} X^T y\right) \tag{3}$$

$$= \left[(X^T X)^{-1} X^T\right] \text{Var}(y) \left[(X^T X)^{-1} X^T\right]^T \tag{4}$$

Since $\text{Var}(y) = \text{Var}(\epsilon) = \sigma^2 I$, and using the identity $(AB)^T = B^T A^T$:

$$\text{Var}(\hat{\beta}) = \left[(X^T X)^{-1} X^T\right] (\sigma^2 I) \left[X((X^T X)^{-1})^T\right] \tag{5}$$

$$= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \quad \text{(Since } (X^T X) \text{ is symmetric)} \tag{6}$$

$$= \sigma^2 (X^T X)^{-1} I \tag{7}$$

$$\boxed{\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}} \tag{8}$$

Checkpoint: The stability of our weights is entirely dependent on the inverse of the correlation matrix $(X^T X)^{-1}$.

3 Step 2: Spectral Decomposition

To understand the behavior of the inverse matrix $(X^TX)^{-1}$, we utilize Eigendecomposition. Since X^TX is a real symmetric matrix, it can be decomposed as:

$$X^TX = V\Lambda V^T \tag{9}$$

Where:

- V is the matrix of orthogonal eigenvectors ($V^T = V^{-1}$).
- Λ is the diagonal matrix of eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$.

The inverse is derived as:

$$(X^TX)^{-1} = (V\Lambda V^T)^{-1} = V\Lambda^{-1}V^T \tag{10}$$

We can express the variance of the coefficient vector as a summation:

$$\text{Var}(\hat{\beta}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}\mathbf{v}_j\mathbf{v}_j^T \tag{11}$$

Specifically, the variance of the k-th coefficient $\hat{\beta}_k$ is:

$$\text{Var}(\hat{\beta}_k) = \sigma^2 \sum_{j=1}^{p} \frac{v_{kj}^2}{\lambda_j} \tag{12}$$

4 Step 3: The Geometric Link (Why $\lambda \approx 0$)

Equation (12) shows that variance depends on $\frac{1}{\lambda}$. We must now prove mathematically why multicollinearity forces $\lambda \to 0$.

4.1 Geometric Flatness

Multicollinearity implies the data points lie on a lower-dimensional subspace (e.g., a line in 2D space, or a plane in 3D space). Let $\mathbf{v}$ be a unit direction vector ($||\mathbf{v}|| = 1$). The projection of the data matrix X onto direction $\mathbf{v}$ is the vector $X\mathbf{v}$.

The **variance** (or energy) of the data along direction $\mathbf{v}$ is the squared Euclidean norm of this projection:

$$\text{Variance along } \mathbf{v} = ||X\mathbf{v}||^2 \tag{13}$$

If the data is perfectly multicollinear, there exists a direction $\mathbf{v}$ perpendicular to the data distribution where the variance is zero (the data is "flat").

$$||X\mathbf{v}||^2 \approx 0 \tag{14}$$

4.2 Linking Variance to Eigenvalues

We expand the squared norm using vector algebra:

$$||X\mathbf{v}||^2 = (X\mathbf{v})^T(X\mathbf{v}) \tag{15}$$

$$= \mathbf{v}^TX^TX\mathbf{v} \tag{16}$$

Let $\mathbf{v}$ be an eigenvector of X^TX with eigenvalue λ. By definition: $(X^TX)\mathbf{v} = \lambda\mathbf{v}$. Substituting this back:

$$||X\mathbf{v}||^2 = \mathbf{v}^T(\lambda\mathbf{v}) \tag{17}$$
$$= \lambda(\mathbf{v}^T\mathbf{v}) \tag{18}$$
$$= \lambda \quad (\text{Since } ||\mathbf{v}|| = 1) \tag{19}$$

$$\boxed{\lambda = ||X\mathbf{v}||^2} \tag{20}$$

Conclusion:

1. **Geometry:** Multicollinearity implies the data is flat in some direction $\mathbf{v}$.
2. **Algebra:** Flatness implies the variance $||X\mathbf{v}||^2 \to 0$.
3. **Eigenvalue:** Therefore, the eigenvalue associated with that direction $\lambda \to 0$.

5 Step 4: The Explosion of Variance

We now substitute our finding back into the variance equation derived in Step 2.

The variance of the coefficient $\hat{\beta}_k$ is:

$$\text{Var}(\hat{\beta}_k) = \sigma^2 \left(\frac{v_{k1}^2}{\lambda_1} + \frac{v_{k2}^2}{\lambda_2} + \cdots + \frac{v_{kp}^2}{\lambda_{min}} \right) \tag{21}$$

Due to multicollinearity, we established that $\lambda_{min} \to 0$. Consequently, the term $\frac{1}{\lambda_{min}} \to \infty$.

$$\lim_{\lambda_{min} \to 0} \text{Var}(\hat{\beta}_k) = \infty \tag{22}$$

6 Summary

Mathematically, multicollinearity renders the matrix X^TX nearly singular (rank-deficient). This results in near-zero eigenvalues. Since the variance of the OLS estimator is proportional to the reciprocal of these eigenvalues, the variance explodes.

Practically, this means that even a microscopic change in the input noise ϵ results in macroscopic, chaotic swings in the estimated model weights $\hat{\beta}$.