Linear Algebra in Quantitative Genetics

This repository explores key concepts from quantitative genetics using the language of linear algebra. Linear algebra provides the mathematical framework for modeling and predicting complex traits in populations.

Core Components

- **Vectors**: Used to represent data for a single trait across all individuals.
  - $\mathbf{y}$: The vector of phenotypes (observed traits).
  - $\mathbf{u}$: The vector of breeding values (genetic merit to be predicted).
- **Matrices**: Used to represent relationships and experimental designs.
  - **Genomic Relationship Matrix (G)**: A square matrix ($n \times n$) quantifying the genetic similarity between all pairs of individuals, calculated from SNP data.
  - **Incidence Matrices (X, Z)**: Sparse matrices that link observations (phenotypes) to fixed effects (like herd or year) and random effects (breeding values).

The Central Model

The workhorse of modern genetic prediction is the **Mixed Linear Model**, often called the "Animal Model":

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}$$

This is a system of linear equations where we solve for the vector of breeding values ($\mathbf{u}$). This is accomplished using **Henderson's Mixed Model Equations (MME)**, which simultaneously accounts for fixed environmental effects and random genetic effects to produce **Estimated Breeding Values (EBVs)** or **Genomic EBVs (GEBVs)**.

Data Analysis & Insights

- **Principal Component Analysis (PCA)**: By applying **eigen-decomposition** to the Genomic Relationship Matrix ($\mathbf{G}$), we can perform PCA.
  - The **eigenvectors** reveal patterns of population structure (e.g., clustering of related individuals or sub-populations).
  - The **eigenvalues** quantify the amount of genetic variance captured by each principal component.

Lecture 1

0.1 Matrix Algebra Basics

0.2 Special Matrices

Introduce some special matrices: Square matrix, Symmetric matrix, Upper triangular matrix, Diagonal matrix, Unit matrix, Zero matrix, Identity matrix.

0.3 Matrix Operations

Matrix Addition/Subtraction/Multiplication

Application: $\mathbf{Z} = (\mathbf{Y} - \mathbf{M}) \times \mathbf{P}$ standardize the phenotype or genotype matrix.

Example: Standardizing the Genotype Matrix

You standardize the **genotype matrix (M)** *before* you use it to calculate the **G** matrix. This standardization is crucial because it ensures that all genetic markers (SNPs) contribute fairly to the estimate of genomic relationships, regardless of their allele frequencies. Here are the key steps, following one of the most common approaches (VanRaden, 2008).

Step 1: Create and Center the Genotype Matrix First, you need a raw genotype matrix, let's call it **W**, where rows are individuals and columns are markers. Genotypes are typically coded as 0, 1, or 2 (counting the number of a specific allele).

1. **Calculate Allele Frequencies (p_i)**: For each marker i, calculate the frequency of the allele you are counting. The formula is:
$$p_i = \frac{\text{total count of allele at marker } i}{2 \times (\text{number of individuals})}$$

2. **Create the Frequency Matrix (P)**: Create a matrix **P** where every element in a column i is the value $2p_i$. This value represents the expected genotype score based on allele frequency.

3. **Center the Genotype Matrix (M)**: Subtract **P** from your raw genotype matrix **W**. This gives you the centered matrix **M**.
$$\mathbf{M} = \mathbf{W} - \mathbf{P}$$
This step adjusts the genotypes based on their allele frequencies, effectively giving a mean of zero to each marker.

Step 2: Scale the Centered Matrix and Calculate G Now you calculate the **Genomic Relationship Matrix (G)**.

1. **Calculate the Scaling Factor**: The denominator used for scaling is the sum of the variances of all the markers. This is calculated as:
$$k = 2\sum_{i=1}^{m} p_i(1 - p_i)$$
where m is the total number of markers. This factor ensures that the average diagonal element of the final G matrix is close to 1, making it analogous to a traditional pedigree-based relationship matrix.

2. **Calculate the G Matrix**: The final **G** matrix is calculated with the following formula:
$$\mathbf{G} = \frac{\mathbf{M}\mathbf{M}^T}{k} = \frac{\mathbf{M}\mathbf{M}^T}{2\sum_{i=1}^{m} p_i(1 - p_i)}$$

- **M**: Your centered genotype matrix from Step 1.
- $\mathbf{M}^T$: The transpose of the **M** matrix.
- $\mathbf{M}\mathbf{M}^T$: This matrix multiplication is what actually computes the genomic relationships between all pairs of individuals.
- k: The scaling factor you calculated above.

The resulting **G** matrix contains the estimated genomic relationships, which you can then plug directly into your GBLUP (Genomic Best Linear Unbiased Prediction) model equations, typically by inverting it ($\mathbf{G}^{-1}$).

0.4 Inner Product

If $\mathbf{a}^T\mathbf{b} = 0$, $\mathbf{a}$ and $\mathbf{b}$ are orthogonal.

- **Positive Inner Product**: When the angle θ is less than 90° (acute), $\cos(\theta)$ is positive. This means the vectors point in generally the same direction. The projection of $\mathbf{a}$ onto $\mathbf{b}$ is positive.
- **Zero Inner Product**: When the angle θ is exactly 90°, the vectors are orthogonal (perpendicular). Since $\cos(90^\circ) = 0$, the inner product is zero. There is no projection of one vector onto the other; they are completely independent in direction.
- **Negative Inner Product**: When the angle θ is greater than 90° (obtuse), $\cos(\theta)$ is negative. This means the vectors point in generally opposite directions. The projection is in the opposite direction of vector $\mathbf{b}$.

0.5 Transpose

$\mathbf{C}$ is the centered matrix. The covariance of the sample (sample covariance matrix ($\mathbf{S}$)):

$$\mathbf{S} = \frac{1}{n-1}\mathbf{C}^T\mathbf{C}$$

Note: For an $n \times p$ data matrix where rows are observations and columns are variables, this gives a $p \times p$ covariance matrix. The dimensions work as follows: If $\mathbf{C}$ is $n \times p$, then $\mathbf{C}^T$ is $p \times n$, and $\mathbf{C}^T\mathbf{C}$ is $p \times p$.

0.6 Trace

The trace of a square matrix, denoted as $\text{tr}(\mathbf{A})$, is the sum of the elements on the main diagonal. For an $n \times n$ matrix $\mathbf{A}$, the trace is calculated as:

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \cdots + a_{nn}$$

The trace has several useful properties, such as $\text{tr}(\mathbf{A} + \mathbf{B}) = \text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B})$ and $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$.

0.7 Quadratic Forms

A quadratic form is a scalar function of a vector $\mathbf{x}$ that can be expressed as:

$$\mathbf{x}^T\mathbf{A}\mathbf{x}$$

where $\mathbf{A}$ is a symmetric matrix. In simple terms, it's a polynomial where every term has a degree of two (e.g., x_1^2, x_1x_2). The result is a single number (a scalar) that is a sum of squared terms (like $a_{11}x_1^2$) and cross-product terms (like $a_{12}x_1x_2$).

While the final step of GBLUP involves solving a system of linear equations (the MME), the statistical foundation for setting up those equations relies heavily on quadratic forms.

- **Solving for Breeding Values (MME)**: This is a system of linear equations. Its goal is prediction. Estimate the variance components (σ_a^2 and σ_e^2).
- **Estimating the Variances (REML)**: This process relies on maximizing a function built from quadratic forms. Its goal is estimation of the parameters needed for prediction. Estimate the variance components (σ_a^2 and σ_e^2).

0.8 Determinant ($\det(\mathbf{A})$)

In any dimension (Laplace expansion):

- For a fixed **row** i: $\det(\mathbf{A}) = \sum_{j=1}^{n}(-1)^{i+j}a_{ij}M_{ij}$
- For a fixed **column** j: $\det(\mathbf{A}) = \sum_{i=1}^{n}(-1)^{i+j}a_{ij}M_{ij}$

where M_{ij} is the **minor** of a_{ij} (the determinant of the $(n-1)\times(n-1)$ submatrix obtained by removing row i and column j). A matrix is invertible if and only if its determinant is not zero. A zero determinant would imply that the system is singular. In a practical sense, this could mean that some effects are perfectly confounded (e.g., you can't separate a fixed effect from the average breeding value) or there are other linear dependencies in your model and data.

0.9 Inverse

If $\mathbf{A}$ is a square matrix, the inverse $\mathbf{A}^{-1}$ satisfies:

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

Higher Dimension Matrices

The inverse of a square matrix $\mathbf{A}$ is given by:

$$\mathbf{A}^{-1} = \frac{1}{\det(\mathbf{A})}\text{adj}(\mathbf{A})$$

- **Adjugate (adjoint) of A:** $\text{adj}(\mathbf{A}) = \mathbf{C}^T$, where $\mathbf{C}$ is the cofactor matrix.
- **Cofactor matrix C:** A matrix formed by all the cofactors of $\mathbf{A}$, where the cofactor $C_{ij} = (-1)^{i+j}M_{ij}$.

0.10 Rank

The rank of a matrix is the maximum number of linearly independent rows or columns in the matrix. It provides a measure of the dimensionality of the vector space spanned by its rows or columns.

0.11 Generalized Inverse

A matrix $\mathbf{A}^{-}$ is called a generalized inverse of $\mathbf{A}$ if it satisfies the condition:

$$\mathbf{A}\mathbf{A}^{-}\mathbf{A} = \mathbf{A}$$

Unlike the regular inverse ($\mathbf{A}^{-1}$), a generalized inverse exists for any matrix, including rectangular or singular (non-invertible) matrices. It is particularly useful for finding solutions to systems of linear equations that do not have a unique solution.

0.12 Linear System of Equations

0.13 Eigenvalues and Eigenvectors

An eigenvector must be a non-zero vector by definition. For a square matrix $\mathbf{C}$ and scalar λ, if there exists a non-zero vector $\mathbf{v}$ such that:

$$\mathbf{C}\mathbf{v} = \lambda\mathbf{v}$$

Then:

- λ is called an eigenvalue of $\mathbf{C}$
- $\mathbf{v}$ is called an eigenvector of $\mathbf{C}$ associated with eigenvalue λ

An $n \times n$ square matrix has exactly n eigenvalues. These eigenvalues are the roots of the characteristic polynomial, and while there are always n of them, some might be repeated or be complex numbers.

0.14 Cholesky Decomposition

Cholesky decomposition is a method for breaking down a special type of matrix into the product of a lower triangular matrix and its transpose. Think of it as a specialized and highly efficient way to find the "square root" of a matrix.

The Standard Method (Slow)

The most direct way to solve the system is to first compute the inverse of the matrix $\mathbf{A}$, and then multiply it by the vector $\mathbf{b}$:

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$$

For large matrices, such as those used in quantitative genetics (e.g., in the MME), calculating the inverse is computationally very expensive and can lead to an accumulation of numerical errors.

The Cholesky Method (Fast)

The Cholesky method is a two-step process that is significantly more efficient.

Step 1: Decompose the Matrix A First, we perform the Cholesky decomposition on our symmetric, positive-definite matrix $\mathbf{A}$ to find a lower triangular matrix $\mathbf{L}$ such that:

$$\mathbf{A} = \mathbf{L}\mathbf{L}^T$$

Step 2: Substitute and Solve Next, we substitute the decomposed form back into the original equation $\mathbf{A}\mathbf{x} = \mathbf{b}$:

$$(\mathbf{L}\mathbf{L}^T)\mathbf{x} = \mathbf{b}$$

We can split this into two separate, simpler problems by introducing an intermediate vector $\mathbf{y}$, where we define $\mathbf{L}^T\mathbf{x} = \mathbf{y}$.

Forward Substitution First, we solve the system for the intermediate vector $\mathbf{y}$:

$$\mathbf{L}\mathbf{y} = \mathbf{b}$$

Because $\mathbf{L}$ is a lower triangular matrix, this system is extremely fast to solve. The first element of $\mathbf{y}$ (y_1) can be solved for immediately. This value is then used to solve for y_2, and so on, in a process called **forward substitution**.

Backward Substitution Once the vector $\mathbf{y}$ is known, we solve the second system to find our final answer, $\mathbf{x}$:

$$\mathbf{L}^T\mathbf{x} = \mathbf{y}$$

Because $\mathbf{L}^T$ is an upper triangular matrix, this system is also very fast to solve. The last element of $\mathbf{x}$ (x_n) is solved for first, and its value is then used to solve for the second-to-last element, and so on, in a process called **backward substitution**.

Why the Cholesky Method is Better

- **Speed:** This two-step substitution process is significantly faster than calculating the full inverse of $\mathbf{A}$. For large systems, the time savings can be enormous, reducing a computation that might take hours to one that takes minutes or seconds.
- **Numerical Stability:** The Cholesky method is also more numerically stable. This means it is less susceptible to the small rounding errors that can occur during computation, leading to more accurate and reliable results.

Numerical Example

Let's solve the system $\mathbf{Ax} = \mathbf{b}$ where:

$$\mathbf{A} = \begin{bmatrix} 4 & 2 \\ 2 & 10 \end{bmatrix} \quad \text{and} \quad \mathbf{b} = \begin{bmatrix} 2 \\ -8 \end{bmatrix}$$

Step 1: Cholesky Decomposition of A First, we find the Cholesky decomposition of $\mathbf{A}$. For this matrix, the lower triangular matrix $\mathbf{L}$ is:

$$\mathbf{L} = \begin{bmatrix} 2 & 0 \\ 1 & 3 \end{bmatrix}$$

Verification that $\mathbf{LL}^T = \mathbf{A}$:

$$\mathbf{LL}^T = \begin{bmatrix} 2 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 2 & 10 \end{bmatrix} = \mathbf{A}$$

Let's verify each element:

- (1,1) element: $2 \times 2 + 0 \times 0 = 4$ ✓
- (1,2) element: $2 \times 1 + 0 \times 3 = 2$ ✓
- (2,1) element: $1 \times 2 + 3 \times 0 = 2$ ✓
- (2,2) element: $1 \times 1 + 3 \times 3 = 10$ ✓

Step 2: Forward Substitution (Solve $\mathbf{Ly} = \mathbf{b}$) We solve for the intermediate vector $\mathbf{y}$:

$$\begin{bmatrix} 2 & 0 \\ 1 & 3 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 2 \\ -8 \end{bmatrix}$$

This gives us two simple equations:

1. $2y_1 = 2 \implies y_1 = 1$
2. $y_1 + 3y_2 = -8 \implies 1 + 3y_2 = -8 \implies 3y_2 = -9 \implies y_2 = -3$

So, our intermediate vector is $\mathbf{y} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$.

Step 3: Backward Substitution (Solve $\mathbf{L}^T\mathbf{x} = \mathbf{y}$) Now we solve for our final answer $\mathbf{x}$ using $\mathbf{y}$:

$$\begin{bmatrix} 2 & 1 \\ 0 & 3 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ -3 \end{bmatrix}$$

This gives us two more simple equations, which we solve from the bottom up:

1. $3x_2 = -3 \implies x_2 = -1$
2. $2x_1 + x_2 = 1 \implies 2x_1 + (-1) = 1 \implies 2x_1 = 2 \implies x_1 = 1$

Final Solution The solution to the system is $\mathbf{x} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

A Complete PCA Example in Linear Algebra

This document provides a complete, step-by-step calculation for Principal Component Analysis (PCA) on a small dataset. It demonstrates how to use core linear algebra concepts—covariance matrices, eigenvalues, and eigenvectors—to identify the most important patterns in the data. An $n \times n$ square matrix has exactly n eigenvalues, which represent the variance of the principal components. For each eigenvalue, there is a corresponding non-zero eigenvector, which represents the direction of that component. Eigenvectors are required to be non-zero by definition to capture the unique, characteristic directions of the data's variance.

Step 1: The Raw Data

We start with a dataset for 5 students and 2 variables: **Study Hours** and **Exam Score**.

Student	Study Hours (x)	Exam Score (y)
A	2	3
B	3	5
C	5	6
D	7	8
E	8	8

Step 2: Center the Data

The first step in PCA is to center the data by subtracting the mean of each variable from all its observations. This ensures that the analysis focuses on the variance within the data.

- **Mean of Study Hours** $= (2+3+5+7+8)/5 = 5$
- **Mean of Exam Score** $= (3+5+6+8+8)/5 = 6$

This gives us the **centered data matrix (C)**:

$$\mathbf{C} = \begin{bmatrix} 2-5 & 3-6 \\ 3-5 & 5-6 \\ 5-5 & 6-6 \\ 7-5 & 8-6 \\ 8-5 & 8-6 \end{bmatrix} = \begin{bmatrix} -3 & -3 \\ -2 & -1 \\ 0 & 0 \\ 2 & 2 \\ 3 & 2 \end{bmatrix}$$

Step 3: Calculate the Covariance Matrix

With the data centered, we now compute the covariance matrix. This matrix describes the variance of each variable and how they vary together. For a data matrix with rows as observations and columns as variables, the covariance matrix is calculated as:

$$\text{Cov} = \frac{1}{n-1}\mathbf{C}^T\mathbf{C}$$

where $n = 5$ (number of observations).

1. **Calculate $\mathbf{C}^T\mathbf{C}$** (The sum of squares and cross-products):

$$\mathbf{C}^T\mathbf{C} = \begin{bmatrix} -3 & -2 & 0 & 2 & 3 \\ -3 & -1 & 0 & 2 & 2 \end{bmatrix} \times \begin{bmatrix} -3 & -3 \\ -2 & -1 \\ 0 & 0 \\ 2 & 2 \\ 3 & 2 \end{bmatrix} = \begin{bmatrix} 26 & 21 \\ 21 & 18 \end{bmatrix}$$

Verification of calculations:

- (1,1): $(-3)(-3) + (-2)(-2) + (0)(0) + (2)(2) + (3)(3) = 9 + 4 + 0 + 4 + 9 = 26$ ✓

- (1,2): $(-3)(-3) + (-2)(-1) + (0)(0) + (2)(2) + (3)(2) = 9 + 2 + 0 + 4 + 6 = 21$ ✓
- (2,1): Same as (1,2) = 21 ✓
- (2,2): $(-3)(-3) + (-1)(-1) + (0)(0) + (2)(2) + (2)(2) = 9 + 1 + 0 + 4 + 4 = 18$ ✓

2. **Divide by $(n-1)$** to get the sample covariance. We divide by $5 - 1 = 4$.

$$\text{Covariance Matrix } (\mathbf{A}) = \frac{1}{4}\begin{bmatrix} 26 & 21 \\ 21 & 18 \end{bmatrix} = \begin{bmatrix} 6.5 & 5.25 \\ 5.25 & 4.5 \end{bmatrix}$$

Step 4: Calculate the Eigenvalues (λ)

The eigenvalues of the covariance matrix represent the magnitude of the variance along each principal component. We find them by solving the characteristic equation $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$.

$$(6.5 - \lambda)(4.5 - \lambda) - (5.25)^2 = 0$$

This expands to the quadratic equation:

$$\lambda^2 - 11\lambda + 1.6875 = 0$$

Using the quadratic formula, we find the two eigenvalues:

- $\lambda_1 \approx 10.84$
- $\lambda_2 \approx 0.16$

Step 5: Calculate the Eigenvectors (Principal Components)

The eigenvectors of the covariance matrix give the direction of the principal components. We solve $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = \mathbf{0}$ for each eigenvalue.

For $\lambda_1 \approx 10.84$ (PC1)

1. **Set up the equation**: $(\mathbf{A} - 10.84\mathbf{I})\mathbf{v} = \mathbf{0}$ gives the system $-4.34v_1 + 5.25v_2 = 0$.
2. **Find the relationship**: This simplifies to $v_2 \approx 0.827v_1$.
3. **Normalize**: We choose a vector in this direction like $[1, 0.827]^T$ and normalize it to a length of 1. Its length is $\sqrt{1^2 + 0.827^2} \approx 1.298$.
   - **PC1** $= [1/1.298, 0.827/1.298]^T \approx [0.77, 0.64]^T$

For $\lambda_2 \approx 0.16$ (PC2)

1. **Set up the equation**: $(\mathbf{A} - 0.16\mathbf{I})\mathbf{v} = \mathbf{0}$ gives the system $6.34v_1 + 5.25v_2 = 0$.
2. **Find the relationship**: This simplifies to $v_2 \approx -1.208v_1$.
3. **Normalize**: We choose a vector in this direction like $[1, -1.208]^T$ and normalize it. Its length is $\sqrt{1^2 + (-1.208)^2} \approx 1.568$.
   - **PC2** $= [1/1.568, -1.208/1.568]^T \approx [0.64, -0.77]^T$

Step 6: Final Result Summary

The eigenvalues tell us how much variance is captured by each principal component.

	Eigenvalue (Variance)	% of Total Variance	Eigenvector (Direction)
PC1	**10.84**	98.5%	$[0.77, 0.64]^T$
PC2	**0.16**	1.5%	$[0.64, -0.77]^T$

This result shows that the first principal component (PC1) captures 98.5% of the total variance in the data. It represents the strong positive relationship between study hours and exam scores and can be used as a new, single dimension to represent the original two variables with minimal loss of information.