# Multivariate analysis

Lecture 7 ▪ Quantitative Genetics

Grum Gebreyesus ▪ ASAP-Bio Knowledge Hub

asap-bio.github.io/lec-qg-multivariate.html

## Contents



## 1 Intended learning outcomes

After this lecture you will be able to:

- *apply* the multivariate normal distribution to analyse multi-trait models
- *explain* general properties of the univariate and multivariate normal distributions
- *apply* concepts for multivariate distributions to analyse problems in quantitative genetics

What is covered: the multivariate normal distribution, the two-trait model, genetic and environmental correlation, and genotype-environment interaction.

## 2 Motivation

- Different traits, such as height, weight or disease susceptibility, exhibit characteristic distributions. Studying these distributions helps quantify and characterise trait variability.
- The frequency distribution of most metric traits resembles the normal curve, so we can use the properties of the normal distribution to make inferences about the trait.

Lecture 1 explained *why* that bell shape arises, many small independent gene effects summing under the Central Limit Theorem. This lecture takes the next step: what happens when we consider **several traits, or several animals, at once**.

## 3 The univariate normal distribution

A reminder of the one-dimensional case, written $X \sim N\left(\mu, \sigma^2\right)$:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \qquad -\infty < x < \infty$$

with $\mathbb{E}(X) = \mu$ and $\mathbb{V}(X) = \sigma^2$. The distribution is completely specified by two numbers, its mean and its variance. That economy is what makes it so useful, and it is exactly what generalises next.

## 4 The multivariate normal distribution

Now consider $k$ variables together. If all are normally distributed marginally, and jointly they form a multivariate normal distribution, we write $\boldsymbol{X} \sim N\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$ with density

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^k \left|\boldsymbol{\Sigma}\right|}} \exp\left[-\tfrac{1}{2}\left(\boldsymbol{x}-\boldsymbol{\mu}\right)' \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{x}-\boldsymbol{\mu}\right)\right]$$

The two parameters are now a *vector* of means and a *matrix* of variances and covariances:

$$\boldsymbol{X} = \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_k \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_k \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1k} \\ \sigma_{12} & \sigma_2^2 & \cdots & \sigma_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1k} & \sigma_{2k} & \cdots & \sigma_k^2 \end{bmatrix}$$

**READING THE COVARIANCE MATRIX**

- The **diagonal** holds the variances of each variable.
- The **off-diagonal** holds the covariances between pairs.
- $\boldsymbol{\Sigma}$ is symmetric, since $\sigma_{12} = \sigma_{21}$.

### 4.1 Covariance or correlation

Covariance measures association, but its units are awkward, so we usually standardise it:

$$\rho_{XZ} = \frac{\mathbb{C}(X,Z)}{\sqrt{\mathbb{V}(X)\,\mathbb{V}(Z)}} = \frac{\sigma_{XZ}}{\sqrt{\sigma_X^2 \sigma_Z^2}} \qquad \Longleftrightarrow \qquad \sigma_{XZ} = \rho_{XZ}\,\sigma_X \sigma_Z$$

which lets the covariance matrix be written entirely in terms of correlations and standard deviations. If two traits are **independent** they are uncorrelated, so $\sigma_{XZ} = \rho_{XZ}\sigma_X\sigma_Z = 0$ and the corresponding off-diagonal entries vanish. A diagonal $\boldsymbol{\Sigma}$ therefore means mutually independent traits.

## 5 Why this matters in animal science

- It is embedded in the **infinitesimal model** assumptions.
- Selection of livestock is usually based on a **combination of several traits**.
- Those traits may be phenotypically *and* genetically correlated.
- Animals are ranked and selected by combining these correlated traits.

In other words, single-trait theory is the exception in practice. As soon as you have an index over several traits, or you want to predict a breeding value from relatives' records, you are working inside a multivariate normal distribution whether you name it or not.

## 6 Conditional distributions

The single most useful property of the multivariate normal is that its **conditional distributions are also normal**, with a mean that is a linear function of what you condition on. Partition the vector, mean and covariance matrix as

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

**CONDITIONAL MEAN AND VARIANCE**

$$\boldsymbol{X}_1 \mid \boldsymbol{X}_2 = \boldsymbol{x}_2 \;\sim\; N\Big(\boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left(\boldsymbol{x}_2 - \boldsymbol{\mu}_2\right),\; \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\Big)$$

The conditional mean is the multivariate generalisation of the regression formula from Lecture 2, and the conditional variance is *smaller* than the unconditional one: observing $\boldsymbol{X}_2$ removes uncertainty about $\boldsymbol{X}_1$.

## 7 Predicting genetic values from phenotypes

This is where the machinery pays off. Everything below is one application of the conditional formula.

### 7.1 One individual

Consider the two variables $g$ (breeding value) and $y$ (phenotype) for a single animal:

- $\mathbb{E}(y) = \mu$, $\mathbb{E}(g) = 0$
- $\mathbb{V}(y) = \sigma_y^2$, $\mathbb{V}(g) = \sigma_g^2 = h^2\sigma_y^2$
- $\mathbb{C}(y, g) = \mathbb{C}(g + e, g) = \mathbb{C}(g, g) = \sigma_g^2 = h^2\sigma_y^2$

$$\begin{bmatrix} g \\ y \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} h^2\sigma_y^2 & h^2\sigma_y^2 \\ h^2\sigma_y^2 & \sigma_y^2 \end{bmatrix}\right)$$

**Worked example: predicting one animal's breeding value**

Suppose $y = 10$, $\mu = 8$, $\sigma_y^2 = 4$ and $h^2 = 0.25$.
Apply the conditional mean with $\Sigma_{12} = h^2\sigma_y^2 = 1$ and $\Sigma_{22} = \sigma_y^2 = 4$:

$$\hat{g} = E(g \mid y) = 0 + \frac{h^2\sigma_y^2}{\sigma_y^2}(y - \mu) = h^2(y - \mu) = 0.25 \times 2 = 0.5$$

The regression coefficient is exactly $h^2$, which is the result met in Lecture 6 when deriving the breeder's equation. The conditional variance, $\sigma_g^2 - \left(h^2\sigma_y^2\right)^2/\sigma_y^2 = h^2\sigma_y^2(1 - h^2) = 0.75$, quantifies how much uncertainty remains.

### 7.2 Two individuals with independent environments

Now stack two related animals. With relationship $F_{ij}$ between them and independent environments, the joint distribution of both breeding values and both phenotypes is

$$\begin{bmatrix} g_1 \\ g_2 \\ y_1 \\ y_2 \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \\ \mu \\ \mu \end{bmatrix}, \begin{bmatrix} h^2\sigma_y^2 & F_{ij}h^2\sigma_y^2 & h^2\sigma_y^2 & F_{ij}h^2\sigma_y^2 \\ F_{ij}h^2\sigma_y^2 & h^2\sigma_y^2 & F_{ij}h^2\sigma_y^2 & h^2\sigma_y^2 \\ h^2\sigma_y^2 & F_{ij}h^2\sigma_y^2 & \sigma_y^2 & F_{ij}h^2\sigma_y^2 \\ F_{ij}h^2\sigma_y^2 & h^2\sigma_y^2 & F_{ij}h^2\sigma_y^2 & \sigma_y^2 \end{bmatrix}\right)$$

**INTERPRETATION**

Conditioning the top block (the breeding values) on the bottom block (the observed phenotypes) predicts *both* animals' genetic values from *both* records, weighting each by relationship. Notice that $F_{ij}$, from the relationship lecture, is what couples the two animals. This is BLUP in miniature: the numerator relationship matrix $\mathbf{A}$ simply extends this $2 \times 2$ block to the whole pedigree.

## 8 The two-trait model

Real examples of trait relationships:

- Animals with higher growth rate tend to have higher fat accumulation.
- Animals with higher weaning weight tend to have higher birth weight.
- Animals with lower body weight tend to have smaller litter sizes.

Starting from the general model $y = \mu + \sum_{j=1}^{L} g_j + e$, write it once per trait:

$$\begin{aligned} y_1 &= \mu_1 + g_1 + e_1 \\ y_2 &= \mu_2 + g_2 + e_2 \end{aligned} \quad \Longrightarrow \quad \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \sim MVN\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 \\ \rho_{21}\sigma_2\sigma_1 & \sigma_2^2 \end{bmatrix}\right)$$

## 9 Decomposing the phenotypic correlation

Where does a phenotypic correlation between two traits come from? Decompose it.

**From covariance to correlation**

1. By definition $\rho_{12} = \dfrac{\mathbb{C}(y_1, y_2)}{\sqrt{\mathbb{V}(y_1)\mathbb{V}(y_2)}}$.
2. Expand the covariance: $\mathbb{C}(y_1, y_2) = \mathbb{C}\left[(\mu_1 + g_1 + e_1), (\mu_2 + g_2 + e_2)\right]$, which produces nine terms.
3. Means are constants so contribute nothing, and assuming $\mathbb{C}(g_1, e_2) = \mathbb{C}(g_2, e_1) = 0$, only two terms survive: $\mathbb{C}(y_1, y_2) = \mathbb{C}(g_1, g_2) + \mathbb{C}(e_1, e_2)$.
4. Writing each in correlation form, $\mathbb{C}(y_1, y_2) = \rho_{g_{12}}\sqrt{\mathbb{V}(g_1)\mathbb{V}(g_2)} + \rho_{e_{12}}\sqrt{\mathbb{V}(e_1)\mathbb{V}(e_2)}$.
5. Dividing through by $\sqrt{\mathbb{V}(y_1)\mathbb{V}(y_2)}$ and recognising $\mathbb{V}(g)/\mathbb{V}(y) = h^2$ and $\mathbb{V}(e)/\mathbb{V}(y) = 1 - h^2$ gives the result.

**THE DECOMPOSITION**

$$\rho_{12} = \rho_{g_{12}}\sqrt{h_1^2 h_2^2} + \rho_{e_{12}}\sqrt{(1 - h_1^2)\,(1 - h_2^2)}$$

There are therefore **three** correlations between any two traits: phenotypic $\rho_{12}$, genetic $\rho_{g_{12}}$ and environmental $\rho_{e_{12}}$.

**PRACTICAL IMPLICATION**

The phenotypic correlation you can measure directly is a *blend* of the genetic and environmental correlations, weighted by the heritabilities. The two components can even have opposite signs, so a near-zero phenotypic correlation does not mean the traits are genetically unrelated. Only the genetic correlation $\rho_{g_{12}}$ predicts correlated response to selection, which is why the correlated-response formula in Lecture 6 contains $r_g$ and not $\rho_{12}$.

# 10 Genotype-environment correlation and interaction

The decomposition above rested on the assumption that $\mathbb{C}(g_1, e_2) = \mathbb{C}(g_2, e_1) = 0$. It is worth asking what happens when that fails, and the two ways it can fail are different phenomena.

## 10.1 Genotype-environment correlation

**DEFINITION**

Genotype-environment correlation occurs when individuals with high genetic values for a trait also experience environments with high values for that trait, that is, when certain genotypes "select" certain environments.

When it is present the cross terms no longer vanish and additional covariance components enter the expression for $\mathbb{C}(y_1, y_2)$. Practically, it inflates apparent genetic effects: the best-bred animals may also be the best-fed ones, and the phenotype cannot tell the difference.

## 10.2 Genotype-environment interaction

**DEFINITION**

- Both genes and environment contribute to the trait.
- The relationship between the environment and the phenotype *depends on the genotype*, or equivalently the relationship between genotype and phenotype depends on the environment.

- Genotypes are affected **differently** by different environments.

The distinction matters. A genotype-environment *correlation* means genotypes are non-randomly distributed across environments; an *interaction* means their ranking changes between environments. Interaction is the reason a sire that is elite in a temperate intensive system may be mediocre in a tropical extensive one, and it is why ASAP-Bio's concern with locally relevant evaluation is a genetic argument, not merely a political one.

## 11 Take-home messages

- The multivariate normal distribution is a useful tool for analysing multi-trait models in quantitative genetics.
- Genetic values can be predicted using the phenotypic information of relatives.
- The covariance matrix is the key parameter of the multivariate normal distribution, describing the degree of correlation between different traits or individuals.
- Genotype-environment interaction occurs when genotypes respond differently to different environments.

## 12 Exercises

**Exercise 7.1**

**Covariances from correlations.** Consider three variables with

- $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1.0$
- $\mu_1 = -2$, $\mu_2 = 3$, $\mu_3 = 5$
- $\rho_{12} = 0.3$, $\rho_{13} = -0.9$, $\rho_{23} = 0.0$

1. What are the covariances between the traits?
2. Simulate and plot the variables against each other, 1 vs 2, 1 vs 3, and 2 vs 3.

**1.** Since $\sigma_{ij} = \rho_{ij}\sigma_i\sigma_j$ and all standard deviations are 1, the covariances equal the correlations numerically: $\sigma_{12} = 0.3$, $\sigma_{13} = -0.9$, $\sigma_{23} = 0.0$. The covariance matrix is

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0.3 & -0.9 \\ 0.3 & 1 & 0 \\ -0.9 & 0 & 1 \end{bmatrix}$$

**2.** In R:

```
 install.packages("mvtnorm"); library(mvtnorm)
mu <- c(-2, 3, 5)
sigma <- matrix(c(1,0.3,-0.9, 0.3,1,0, -0.9,0,1), 3, 3)
x <- rmvnorm(n = 1000, mean = mu, sigma = sigma)
par(mfrow = c(3,3))
for (i in 1:3) for (j in 1:3)
plot(x[,i], x[,j], xlab = paste0("Var", i), ylab = paste0("Var", j))
```

Var1 against Var3 forms a tight negatively sloping cloud ($\rho = -0.9$); Var1 against Var2 is a weak positive cloud; Var2 against Var3 is a circular blob with no tilt ($\rho = 0$).

**Caution.** Var2 and Var3 are uncorrelated, yet each is correlated with Var1. Marginal independence of a pair does not survive conditioning: if you condition on Var1, Var2 and Var3 become correlated.

This is exactly why multi-trait evaluation cannot be replaced by a series of single-trait analyses.

## Exercise 7.2

**Predicting genetic values by conditioning.** Using the two-block partition, complete the R script below to obtain predicted genetic values from observed phenotypes. Block 1 is the genetic values, block 2 the phenotypes.

```
 mu1 <- c(......)
mu2 <- c(......)
V11 = V12 = V21 <- matrix(c(.......), ncol = 2)
V22 <- matrix(c(.......), ncol = 2)
x2 <- c(......)
x1 <- mu1 + V12 %*% solve(V22) %*% (x2 - mu2)
print(x1)
```

Take two half sibs ($F_{ij} = 0.25$) with $h^2 = 0.3$, $\sigma_y^2 = 100$, $\mu = 50$, and observed phenotypes $y_1 = 62$, $y_2 = 45$.

With $h^2\sigma_y^2 = 30$ and $F_{ij}h^2\sigma_y^2 = 0.25 \times 30 = 7.5$:

```
 mu1 <- c(0, 0) # genetic values have mean 0
mu2 <- c(50, 50) # phenotypes have mean mu
V11 = V12 = V21 <- matrix(c(30, 7.5, 7.5, 30), ncol = 2)
V22 <- matrix(c(100, 7.5, 7.5, 100), ncol = 2)
x2 <- c(62, 45)
x1 <- mu1 + V12 %*% solve(V22) %*% (x2 - mu2)
print(x1)
```

Note that $V_{12}$ and $V_{22}$ differ only in their diagonals: the genetic covariance between the two animals is $7.5$ in both, but the phenotypic variance on the diagonal of $V_{22}$ is $100$ whereas the genetic variance in $V_{12}$ is $30$.

Evaluating gives approximately $\hat{g}_1 \approx 3.5$ and $\hat{g}_2 \approx -0.6$. Two features are worth noting. First, the predictions are *regressed* towards zero relative to the phenotypic deviations ($+12$ and $-5$), because $h^2 < 1$. Second, each animal's prediction is influenced by its half sib's record through the off-diagonal $7.5$, which is precisely the information-sharing that makes BLUP more accurate than individual phenotypes alone.

## Exercise 7.3

**Decomposing a phenotypic correlation.** Two traits have $h_1^2 = 0.25$ and $h_2^2 = 0.49$. The genetic correlation is $\rho_{g_{12}} = -0.50$ and the environmental correlation is $\rho_{e_{12}} = +0.40$.

1. Compute the phenotypic correlation.
2. Comment on what a breeder who only measured the phenotypic correlation would conclude, and why that would be a mistake.

**1.** Apply the decomposition:

$$\rho_{12} = \rho_{g_{12}}\sqrt{h_1^2 h_2^2} + \rho_{e_{12}}\sqrt{(1-h_1^2)\,(1-h_2^2)}$$

$$= -0.50\sqrt{0.25 \times 0.49} + 0.40\sqrt{0.75 \times 0.51} = -0.50(0.35) + 0.40(0.6185)$$

$$= -0.175 + 0.247 = +0.072$$

**2.** The phenotypic correlation is small and *positive*, about $+0.07$. A breeder seeing only this would conclude the two traits are essentially unrelated, and that selecting on trait 1 is harmless for trait 2.
That conclusion is incorrect, and costly. The genetic correlation is $-0.50$, strongly **antagonistic**. Correlated response depends on $\rho_g$, not $\rho_{12}$, so selection on trait 1 will steadily degrade trait 2. The positive environmental correlation has masked the antagonism at the phenotypic level. This is the quantitative reason that breeding programmes must estimate genetic parameters rather than relying on observed phenotypic associations.