9

Introduction to Matrices and Linear Models

We have already encountered several examples of models in which response variables are linear functions of two or more explanatory (or predictor) variables. For example, we have been routinely expressing an individual's phenotypic value as the sum of genotypic and environmental values. A more complicated example is the use of linear regression to decompose an individual's genotypic value into average effects of individual alleles and residual contributions due to interactions between alleles (Chapters 4 and 5). Such **linear models** not only form the backbone of parameter estimation in quantitative genetics (Chapters 21–31), but are also the basis for incorporating marker (and other genomic) information into modern quantitative-genetic inference (Chapters 20, 30, and 31).

This chapter provides the foundational tools for the analysis of linear models, which are developed more fully, and extended to the powerful mixed model, in Chapter 10. We start by introducing multiple regression, wherein two or more variables are used to make predictions about a response variable. A review of elementary matrix algebra then follows, starting with matrix notation and building up to matrix multiplication and solutions of simultaneous equations using matrix inversion. We next use these results to develop tools for statistical analysis, considering the expectations and covariance matrices of transformed random vectors. We then introduce the multivariate normal distribution, which is by far the most important distribution in quantitative-genetics theory, and conclude with the analysis of the geometry (eigenvalues and eigenvectors) of variance-covariance matrices. Those with strong statistical backgrounds will find little new in this chapter, other than perhaps some immediate contact with quantitative genetics in the examples and familiarization with our notation. Additional background material is given in Appendix 3.

MULTIPLE REGRESSION

As a point of departure, consider the multiple regression

$$y = \alpha + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_n z_n + e \tag{9.1a}$$

where y is the **response variable**, and the z_i are the **predictor** (or **explanatory**) **variables** used to predict the value of the response variable. This multivariate equation is similar to the expression for a simple linear regression (Equation 3.12a) except that y is now a function of n predictor variables, rather than just one. The variables $y, z_1, \ldots, z_n$ represent observed measures, whereas α and $\beta_1, \ldots, \beta_n$ are constants to be estimated using some best-fit criterion. As in the case of simple linear regression, e (the **residual error**) is the deviation between the observed (y) and predicted (or fitted) value ($\widehat{y}$) of the response variable,

$$y = \widehat{y} + e, \quad \text{where} \quad \widehat{y} = \alpha + \sum_{i=1}^{n} \beta_i z_i \tag{9.1b}$$

Recall that the use of a linear model involves no assumptions regarding the true form of relationship between y and $z_1, \ldots, z_n$, nor is any assumption about the residuals being normally distributed required. It simply gives the best linear approximation. Many statistical

techniques, including path analysis (Appendix 2) and analysis of variance (Chapter 22), are based on versions of Equation 9.1.

The terms $\beta_1, \ldots, \beta_n$ are known as **partial regression coefficients**. The interpretation of β_i is the expected change in y given a unit change in z_i while all other predictor values are *held constant*. It is important to note that the partial regression coefficient associated with predictor variable z_i often differs from the regression coefficient, β_i', that is obtain in a univariate regression based solely on z_i, viz., $y = \alpha + \beta_i' z_i + e$ (Example 9.3). Suppose, for example, that a simple regression of y on z_1 has a slope of zero. This might lead to the suggestion that there is no relationship between z_1 and y. However, it is conceivable that z_1 actually has a strong positive effect on y that is obscured by positive correlations of z_1 with other variables that have negative influences on y. A multiple regression that included the appropriate variables would clarify this situation by yielding a positive value of β_1.

Because it is usually impossible for biologists to evaluate partial regression coefficients by empirically imposing constancy on all extraneous variables, we require a more indirect approach to the problem. From Chapter 3, the covariance of y and a predictor variable is

$$\begin{aligned}\sigma(y, z_i) &= \sigma\left[(\alpha + \beta_1 z_1 + \beta_2 z_2 + \cdots + \beta_n z_n + e), z_i\right] \\ &= \beta_1\sigma(z_1, z_i) + \beta_2\sigma(z_2, z_i) + \cdots + \beta_n\sigma(z_n, z_i) + \sigma(e, z_i)\end{aligned} \tag{9.2a}$$

The term $\sigma(\alpha, z_i)$ has dropped out because the covariance of z_i with a constant (α) is zero. By applying Equation 9.2a to each predictor variable, we obtain a set of n equations in n unknowns $(\beta_1, \ldots, \beta_n)$,

$$\begin{aligned}
\sigma(y, z_1) &= \beta_1\sigma^2(z_1) &&+ \beta_2\sigma(z_1, z_2) + \cdots + \beta_n\sigma(z_1, z_n) + \sigma(z_1, e) \\
\sigma(y, z_2) &= \beta_1\sigma(z_1, z_2) + &&\beta_2\sigma^2(z_2) \quad + \cdots + \beta_n\sigma(z_2, z_n) + \sigma(z_2, e) \\
&\vdots \qquad\quad \vdots \qquad\quad \vdots \qquad\quad \ddots \qquad\quad \vdots \qquad\quad \vdots \\
\sigma(y, z_n) &= \beta_1\sigma(z_1, z_n) + \beta_2\sigma(z_2, z_n) + \cdots &&+ \beta_n\sigma^2(z_n) \quad + \sigma(z_n, e)
\end{aligned} \tag{9.2b}$$

As in univariate regression, our task is to find the set of constants (α and the partial regression coefficients, β_i) that gives the best linear fit of the conditional expectation of y given $z_1, \cdots, z_n$. Again, the criterion we choose for "best" relies on the **least-squares** approach, which minimizes the squared differences between observed and expected values (i.e., the squared residuals). Thus, our task is to find that set of $\alpha, \beta_1, \cdots, \beta_n$ giving $\widehat{y} = \alpha + \sum \beta_i z_i$ such that $E[(y - \widehat{y})^2 | z_1, \cdots, z_n] = E[\, e^2\,]$ is minimized. Taking derivatives of this expectation with respect to α and the β_i and setting each equal to zero, it can be shown that the set of equations given by Equation 9.2b is, in fact, the least-squares solution to Equation 9.1 (WL Example A6.4). If the appropriate variances and covariances are known (i.e., we know their true population values), the β_i can be obtained exactly. If these are unknown, as is usually the case, the least-squares estimates b_i are obtained from Equation 9.2b by substituting the observed (estimated) variances and covariances, Var(z_i) and Cov(z_i, z_j), for their (unknown) population values, $\sigma^2(z_i)$ and $\sigma(z_i, z_j)$.

Finally, recall from Example 3.4 that one can also express the solutions of a least-squared regression entirely in terms of sums of squares and sums of cross-products (Equation 3.15d). This is the standard solution in most statistic textbooks, as they treat the predictor variables (the observed data, z_i) as fixed. Conversely, predictor variables in this book are very often treated as random because this is what is usually most relevant in most quantitative-genetic applications, wherein we are attempting to make inferences on the nature of some underlying true least-square regression based on a sample. In this setting, expressing the solutions in terms of variances and covariances is most appropriate. In other settings, such as regressing on sex or age, it may be more intuitive to think of sums of squares/cross-products than in terms of variances and covariances. These two approaches (sums of squares/cross-products and variances/covariances) are equivalent if we simply substitute (co)variance components by their sample estimates, which are expressed as sums of squares and cross-products (Example 3.4).

The properties of least-squares multiple regression are analogous to those for simple regression (Chapter 3). First, the procedure yields a solution such that the average deviation of y from its predicted value $\widehat{y}$, $E[e]$, is zero. Hence $E[\,y\,] = E[\,\widehat{y}\,]$, implying

$$\bar{y} = a + b_1\bar{z}_1 + \cdots + b_n\bar{z}_n \tag{9.3a}$$

Thus, once the fitted values $b_1, \ldots, b_n$ are obtained from Equation 9.2b, the intercept is obtained by $a = \bar{y} - \sum_i^n b_i\bar{z}_i$. Using Eqution 9.3a, we can rewrite Equation 9.1a as

$$y - \overline{y} = \sum_{i=1}^{n} \beta_i(z_i - \overline{z}_i) + e \tag{9.3b}$$

implying that

$$y = \overline{y} + \sum_{i=1}^{n} \beta_i(z_i - \overline{z}_i) + e \tag{9.3c}$$

Second, least-squares analysis gives a solution in which the residual errors are uncorrelated with the predictor variables. Thus, the terms $\sigma(e, z_i)$ can be dropped from Equation 9.2b. Third, the partial regression coefficients are entirely defined by variances and covariances. However, unlike simple regression coefficients, which depend on only a single variance and covariance, each partial regression coefficient is a function of the variances and covariances of all measured variables. Notice that if $n = 1$, then $\sigma(y, z_1) = \beta_1\sigma^2(z_1)$, and we recover the univariate solution, $\beta_1 = \sigma(y, z_1)/\sigma^2(z_1)$.

A simple pattern exists in each of the n equations in 9.2b. The *i*th equation defines the covariance of y and z_i as the sum of two types of quantities: a single term, which is the product of the *i*th partial regression coefficient and the variance of z_i, and a set of $(n - 1)$ terms, each of which is the product of a partial regression coefficient and the covariance of z_i with the corresponding predictor variable. This general pattern suggests an alternative way of writing Equation 9.2b,

$$\begin{pmatrix} \sigma^2(z_1) & \sigma(z_1, z_2) & \ldots & \sigma(z_1, z_n) \\ \sigma(z_1, z_2) & \sigma^2(z_2) & \ldots & \sigma(z_2, z_n) \\ \vdots & \vdots & \ddots & \vdots \\ \sigma(z_1, z_n) & \sigma(z_2, z_n) & \ldots & \sigma^2(z_n) \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix} = \begin{pmatrix} \sigma(y, z_1) \\ \sigma(y, z_2) \\ \vdots \\ \sigma(y, z_n) \end{pmatrix} \tag{9.4a}$$

The table of variances and covariances on the left is referred to as a **matrix**, while the columns of partial regression coefficients and of covariances involving y are called **vectors**. If these matrices and vectors are abbreviated, respectively, as $\mathbf{V}$, $\boldsymbol{\beta}$, and $\mathbf{c}$, then Equation 9.4a can be written even more compactly as

$$\mathbf{V}_{n\times n}\boldsymbol{\beta}_{n\times 1} = \mathbf{c}_{n\times 1} \tag{9.4b}$$

where **c** denotes the vector of **c**ovariances between the predictor and response variables.

The standard procedure of denoting matrices as bold capital letters and vectors as bold lowercase letters is adhered to in this book. Notice that $\mathbf{V}$, which is generally called a **covariance matrix**, is symmetrical about the main diagonal. As we shall see shortly, the *i*th equation in 9.2b can be recovered from Equation 9.4a by multiplying the elements in $\boldsymbol{\beta}$ by the corresponding elements in the *i*th horizontal row of the matrix $\mathbf{V}$, i.e., $\sum \beta_j V_{ij}$. Although a great deal of notational simplicity has been gained by condensing the system of Equations 9.2b to matrix form, this does not alter the fact that the solution of a large system of simultaneous equations is a tedious task if performed by hand. Fortunately, such solutions are rapidly accomplished on computers. Before considering matrix methods in more detail, we present an application of Equation 9.1 to quantitative genetics.

An Application to Multivariate Selection

Karl Pearson developed the technique of multiple regression in 1896, although some of the fundamentals can be traced to his predecessors (Pearson 1920; Stigler 1986). Pearson is perhaps best known as one of the founders of statistical methodology, but his intense interest in evolution may have been the primary motivating force underlying many of his theoretical endeavors. Almost all of his major papers, including the one of 1896, contain rigorous analyses of data gathered by his contemporaries on matters such as resemblance between relatives, natural selection, correlation between characters, and assortative mating (recall Example 7.7). The foresight of these studies is remarkable considering that they were performed prior to the existence of a genetic interpretation for the expression and inheritance of polygenic traits.

Pearson's (1896, 1903) invention of multiple regression developed out of the need for a technique to decompose the observed directional selection on a character into its direct and various indirect components. In Chapter 3 we defined the selection differential S (the within-generation change in the mean phenotype due to selection) as a measure of the total directional selection on a character. However, S cannot be considered to be a measure of the *direct forces* of selection on a character unless that character is uncorrelated with all other selected traits. An unselected character can appear to be under selection if other characters with which it is correlated are under directional selection. Alternatively, a character under strong directional selection may exhibit a negligible selection differential if the indirect effects of selection on correlated traits are sufficiently compensatory.

As a hypothetical example, consider fitness measured by the number of visits to a flower by a pollinator. It is observed that larger flowers obtain more pollinators (S for flower size is positive), but also that flower size and nectar volume are positively correlated. Hence, pollinators might be visiting larger flowers simply because they have more nectar. A multiple regression of pollinator visit number on both flower size and nectar volume can resolve which trait is the actual target of selection (provided that neither is correlated to other, unmeasured, targets of selection).

Because he did not employ matrix notation, some of the mathematics in Pearson's papers can be rather difficult to follow. Lande and Arnold (1983) did a great service by extending this work and rephrasing it in matrix notation. Suppose that a large number of individuals in a population have been measured for n characters and for fitness. Individual fitness can then be approximated by the linear model

$$w = \alpha + \beta_1 z_1 + \cdots + \beta_n z_n + e \tag{9.5a}$$

where w is relative fitness (observed fitness divided by the mean fitness in the population, i.e., $w = W/\overline{W}$), and $z_1, \ldots, z_n$ are the phenotypic measures of the n characters. The interpretation of β_i is the expected change in w given a unit change in trait i while all other traits are held constant. Recall from Chapter 3 that the selection differential for the ith trait is defined as the covariance between phenotype and relative fitness, $S_i = \sigma(z_i, w)$. Thus, we have

$$\begin{aligned} S_i = \sigma(z_i, w) &= \sigma(z_i, \alpha + \beta_1 z_1 + \cdots + \beta_n z_n + e) \\ &= \beta_1 \sigma(z_i, z_1) + \cdots + \beta_n \sigma(z_i, z_n) + \sigma(z_i, e) \end{aligned} \tag{9.5b}$$

Note that this expression is of the same form as Equation 9.2b, so that by taking the β_i to be the partial regression coefficients we have $\sigma(z_i, e) = 0$. This expression can also be compactly written as $\mathbf{s} = \mathbf{V}\boldsymbol{\beta}$ (Equation 9.4b), where the vector of covariances ($\mathbf{s}$) has its i element given by S_i. Finally, note that the selection differential of any trait may be partitioned into a component estimating the **direct selection** on the character and the sum of components from **indirect selection** on all correlated characters,

$$S_i = \beta_i \sigma^2(z_i) + \sum_{j \neq i}^{n} \beta_j \sigma(z_i, z_j) \tag{9.5c}$$

It is important to realize that the labels "direct" and "indirect" apply strictly to the specific set of characters included in the analysis; the partial regression coefficients are subject to change if a new analysis includes additional correlated characters that are under selection. Returning to our flower example, suppose that β for size is negative and β for nectar volume positive. This suggests that the pollinators favor smaller flowers with more nectar. The positive correlation between size and nectar volume obscured this relationship, resulting (Equation 9.5c) in a positive value of S for flower size.

Example 9.1. A morphological analysis of a pentatomid bug (*Euschistus variolarius*) population performed by Lande and Arnold (1983) provides a good example of the insight that can be gained from a multivariate approach. The bugs were collected along the shore of Lake Michigan after a storm. Of the 94 individuals that were recovered, 39 were alive. All individuals were measured for four characters: head and thorax width, and scutellum and forewing length. The data were then logarithmically transformed to more closely approximate normality (Chapter 14). All surviving bugs were assumed to have equal fitness ($W = 1$), and all dead bugs to have zero fitness ($W = 0$). Hence, mean fitness is the fraction (p) of individuals that survived, giving **relative fitnesses**, $w = W/\overline{W}$, as

$$w = \begin{cases} 1/p & \text{if the individual survived} \\ 0 & \text{if the individual did not survive} \end{cases}$$

The selection differential for each of the characters is simply the difference between the mean phenotype of the 39 survivors and the mean of the entire sample. These are reported in units of phenotypic standard deviations in the following table, along with the partial regression coefficients of relative fitness on the four morphological characters. Here * and ** indicate significance at the 5% and 1% levels. All of the phenotypic correlations were highly significant.

Character	Selection Differential	Partial Regression Coef. of Fitness	Phenotypic Correlations			
z_i	S_i	β_i	H	T	S	F
Head (H)	–0.11	–0.7	1.00	0.72	0.50	0.60
Thorax (T)	–0.06	11.6**		1.00	0.59	0.71
Scutellum (S)	–0.28*	–2.8			1.00	0.62
Forewing (F)	–0.43**	–16.6**				1.00

The estimates of the partial regression coefficients nicely illustrate two points discussed earlier. First, despite the strong directional selection (β) operating directly on thorax size, the selection differential (S) for thorax size is negligible. This lack of apparent selection results because the positive correlation between thorax width and wing length is coupled with negative forces of selection on the latter character. Second, there is a significant negative selection differential on scutellum length even though there is no significant direct selection on the character. The negative selection differential is largely an indirect consequence of the strong selection for smaller wing length. WL Chapters 29 and 30 examine Lande-Arnold fitness estimation in considerable detail. We note in passing that given the 0,1 nature of the response variable, logistic regression (Chapter 14; WL Chapters 14 and 29) is a more appropriate analysis of these data than a linear model.

ELEMENTARY MATRIX ALGEBRA

The solutions of systems of linear equations, such as those introduced above, generally involve the use of matrices and vectors of variables. For those with little familiarity with such constructs and their manipulations, the remainder of the chapter provides an overview

of the basic tools of matrix algebra, with a focus on useful results for the analysis of linear models.

Basic Notation

A matrix is simply a rectangular array of numbers. Some examples are:

$$\mathbf{a} = \begin{pmatrix} 12 \\ 13 \\ 47 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} 2 & 0 & 5 & 21 \end{pmatrix} \qquad \mathbf{C} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} \qquad \mathbf{D} = \begin{pmatrix} 0 & 1 \\ 3 & 4 \\ 2 & 9 \end{pmatrix}$$

A matrix with r rows and c columns is said to have **dimensionality** $r \times c$ (a useful mnemonic for remembering this order is **r**ailroad **c**ar). In the examples above, $\mathbf{D}$ has three rows and two columns, and is thus a 3×2 matrix. An $r \times 1$ matrix, such as $\mathbf{a}$, is a **column vector** ($c = 1$), while a $1 \times c$ matrix, such as $\mathbf{b}$, is a **row vector** ($r = 1$). A matrix in which the number of rows equals the number of columns, such as $\mathbf{C}$, is called a **square matrix**. Numbers are also matrices (of dimensionality 1×1) and are often referred to as **scalars**.

A matrix is completely specified by the **elements** that comprise it, with M_{ij} denoting the element in the *i*th row and *j*th column of matrix $\mathbf{M}$. Using the sample matrices above, $C_{23} = 4$ is the element in the second row and third column of $\mathbf{C}$. Likewise, $C_{32} = 1$ is the element in the third row and second column. Two matrices are equal if and only if all of their corresponding elements are equal. Dimensionality is important, as operations on matrices (such as addition or multiplication) are only defined when matrix dimensions agree in the appropriate manner (as is discussed below).

Partitioned Matrices

It is often useful to work with **partitioned matrices** wherein each element in a matrix is itself a matrix. There are several ways to partition a matrix. For example, we could write the matrix $\mathbf{C}$ above as

$$\mathbf{C} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 3 & \vdots & 1 & 2 \\ \cdots & \cdots & \cdots & \cdots \\ 2 & \vdots & 5 & 4 \\ 1 & \vdots & 1 & 2 \end{pmatrix} = \begin{pmatrix} \mathbf{a} & \mathbf{b} \\ \mathbf{d} & \mathbf{B} \end{pmatrix}$$

where

$$\mathbf{a} = \begin{pmatrix} 3 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 & 2 \end{pmatrix}, \quad \mathbf{d} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}, \quad \mathbf{B} = \begin{pmatrix} 5 & 4 \\ 1 & 2 \end{pmatrix}$$

Alternatively, we could partition $\mathbf{C}$ into a single row vector whose elements are themselves column vectors,

$$\mathbf{C} = \begin{pmatrix} \mathbf{c_1} & \mathbf{c_2} & \mathbf{c_3} \end{pmatrix} \quad \text{where} \quad \mathbf{c_1} = \begin{pmatrix} 3 \\ 2 \\ 1 \end{pmatrix}, \quad \mathbf{c_2} = \begin{pmatrix} 1 \\ 5 \\ 1 \end{pmatrix}, \quad \mathbf{c_3} = \begin{pmatrix} 2 \\ 4 \\ 2 \end{pmatrix}$$

or as a column vector whose elements are row vectors,

$$\mathbf{C} = \begin{pmatrix} \mathbf{b_1} \\ \mathbf{b_2} \\ \mathbf{b_3} \end{pmatrix} \quad \text{where} \quad \mathbf{b_1} = \begin{pmatrix} 3 & 1 & 2 \end{pmatrix}, \quad \mathbf{b_2} = \begin{pmatrix} 2 & 5 & 4 \end{pmatrix}, \quad \mathbf{b_3} = \begin{pmatrix} 1 & 1 & 2 \end{pmatrix}$$

As we will shortly see, this partition of a matrix as either a set of row or column vectors forms the basis of matrix multiplication.

Addition and Subtraction

Addition and subtraction of matrices is straightforward. To form a new matrix $\mathbf{A} + \mathbf{B} = \mathbf{C}$, $\mathbf{A}$ and $\mathbf{B}$ must have the same dimensionality ($\mathbf{A}$ and $\mathbf{B}$ have the same number of columns and the same number of rows), so that they have corresponding elements. One then simply adds these corresponding elements, $C_{ij} = A_{ij} + B_{ij}$. Subtraction is defined similarly. For example, if

$$\mathbf{A} = \begin{pmatrix} 3 & 0 \\ 1 & 2 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{B} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \\ 1 & 0 \end{pmatrix}$$

then

$$\mathbf{C} = \mathbf{A} + \mathbf{B} = \begin{pmatrix} 4 & 2 \\ 3 & 3 \\ 1 & 1 \end{pmatrix} \quad \text{and} \quad \mathbf{D} = \mathbf{A} - \mathbf{B} = \begin{pmatrix} 2 & -2 \\ -1 & 1 \\ -1 & 1 \end{pmatrix}$$

Multiplication

Multiplying a matrix by a **scalar** (a 1×1 matrix) is also straightforward. If $\mathbf{M} = a\mathbf{N}$, where a is a scalar, then $M_{ij} = aN_{ij}$. Each element of $\mathbf{N}$ is simply multiplied by the scalar. For example,

$$(-2)\begin{pmatrix} 1 & 0 \\ 3 & 1 \end{pmatrix} = \begin{pmatrix} -2 & 0 \\ -6 & -2 \end{pmatrix}$$

Matrix multiplication is a little more involved. We start by considering the **dot product** of two vectors, as this forms the basic operation of matrix multiplication. Letting $\mathbf{a}$ and $\mathbf{b}$ be two n-dimensional vectors, their dot product $\mathbf{a} \cdot \mathbf{b}$ is a scalar given by

$$\mathbf{a} \cdot \mathbf{b} = \sum_{i=1}^{n} a_i b_i$$

For example, for the two vectors

$$\mathbf{a} = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix} \quad \text{and} \quad \mathbf{b} = \begin{pmatrix} 4 & 5 & 7 & 9 \end{pmatrix}$$

the dot product is $\mathbf{a} \cdot \mathbf{b} = (1 \times 4) + (2 \times 5) + (3 \times 7) + (4 \times 9) = 71$. The dot product ignores whether the vectors are row, column, or mixed, but is not defined if the vectors have different lengths. As we will see, this restriction determines whether the product of two matrices is defined.

The dot product operator allows us to express systems of equations compactly in matrix form. Consider the following system of three equations and three unknowns,

$$\begin{aligned} x_1 + 2x_2 + x_3 &= 3 \\ 2x_1 - 2x_2 - 4x_3 &= 6 \\ 8x_1 - 4x_2 + 3x_3 &= 9 \end{aligned}$$

This can be written in matrix form as $\mathbf{Ax} = \mathbf{c}$, with

$$\begin{pmatrix} 1 & 2 & 1 \\ 2 & -2 & -4 \\ 8 & -8 & 3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} 3 \\ 6 \\ 9 \end{pmatrix}$$

This matrix representation recovers the system of equations using dot products. The dot product of the first row of $\mathbf{A}$ with the column vector $\mathbf{x}$ (which is defined because each vector

has three elements) recovers the first equation. The last two equations similarly follow as the dot products of the second and third rows, respectively, of $\mathbf{A}$ on $\mathbf{x}$.

Now consider the matrix $\mathbf{L}_{r\times b} = \mathbf{M}_{r\times c}\mathbf{N}_{c\times b}$ produced by multiplying the $r \times c$ matrix $\mathbf{M}$ by the $c \times b$ matrix $\mathbf{N}$. It is important to note the matching c subscripts, with the number of columns of $\mathbf{M}$ matching the number of rows of $\mathbf{N}$. Partitioning $\mathbf{M}$ as a column vector of r row vectors,

$$\mathbf{M} = \begin{pmatrix} \mathbf{m_1} \\ \mathbf{m_2} \\ \vdots \\ \mathbf{m_r} \end{pmatrix} \qquad \text{where} \qquad \mathbf{m_i} = \begin{pmatrix} M_{i1} & M_{i2} & \cdots & M_{ic} \end{pmatrix}$$

and $\mathbf{N}$ as a row vector of b column vectors,

$$\mathbf{N} = \begin{pmatrix} \mathbf{n_1} & \mathbf{n_2} & \cdots & \mathbf{n_b} \end{pmatrix} \qquad \text{where} \qquad \mathbf{n_j} = \begin{pmatrix} N_{1j} \\ N_{2j} \\ \vdots \\ N_{cj} \end{pmatrix}$$

the *ij*th element of $\mathbf{L}$ is given by the dot product

$$L_{ij} = \mathbf{m_i} \cdot \mathbf{n_j} = \sum_{k=1}^{c} M_{ik}N_{kj} \tag{9.6a}$$

Recall that this dot product is only defined if the vectors $\mathbf{m_i}$ and $\mathbf{n_j}$ have the same number of elements, which requires that the number of columns in $\mathbf{M}$ must equal the number of rows in $\mathbf{N}$. The resulting matrix $\mathbf{L}$ is of dimension $r \times b$ with

$$\mathbf{L} = \begin{pmatrix} \mathbf{m_1}\cdot\mathbf{n_1} & \mathbf{m_1}\cdot\mathbf{n_2} & \cdots & \mathbf{m_1}\cdot\mathbf{n_b} \\ \mathbf{m_2}\cdot\mathbf{n_1} & \mathbf{m_2}\cdot\mathbf{n_2} & \cdots & \mathbf{m_2}\cdot\mathbf{n_b} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{m_r}\cdot\mathbf{n_1} & \mathbf{m_r}\cdot\mathbf{n_2} & \cdots & \mathbf{m_r}\cdot\mathbf{n_b} \end{pmatrix} \tag{9.6b}$$

Note that using this definition, the matrix product given by Equation 9.4a recovers the set of equations given by Equation 9.2b.

Example 9.2. Compute the product $\mathbf{L} = \mathbf{MN}$ where

$$\mathbf{M} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} \qquad \text{and} \qquad \mathbf{N} = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 1 & 3 \\ 3 & 2 & 2 \end{pmatrix}$$

Writing $\mathbf{M} = \begin{pmatrix} \mathbf{m_1} \\ \mathbf{m_2} \\ \mathbf{m_3} \end{pmatrix}$ and $\mathbf{N} = \begin{pmatrix} \mathbf{n_1} & \mathbf{n_2} & \mathbf{n_3} \end{pmatrix}$, we have

$$\mathbf{m_1} = \begin{pmatrix} 3 & 1 & 2 \end{pmatrix}, \quad \mathbf{m_2} = \begin{pmatrix} 2 & 5 & 4 \end{pmatrix}, \quad \mathbf{m_3} = \begin{pmatrix} 1 & 1 & 2 \end{pmatrix}$$

and

$$\mathbf{n_1} = \begin{pmatrix} 4 \\ 1 \\ 3 \end{pmatrix}, \quad \mathbf{n_2} = \begin{pmatrix} 1 \\ 1 \\ 2 \end{pmatrix}, \quad \mathbf{n_3} = \begin{pmatrix} 0 \\ 3 \\ 2 \end{pmatrix}$$

The resulting matrix $\mathbf{L}$ is 3×3. Applying Equation 9.6b, the element in the first row and first column of $\mathbf{L}$ is the dot product of the first row vector of $\mathbf{M}$ with the first column vector of $\mathbf{N}$,

$$L_{11} = \mathbf{m_1} \cdot \mathbf{n_1} = \begin{pmatrix} 3 & 1 & 2 \end{pmatrix} \begin{pmatrix} 4 \\ 1 \\ 3 \end{pmatrix} = \sum_{k=1}^{3} M_{1k} N_{k1}$$

$$= M_{11}N_{11} + M_{12}N_{21} + M_{13}N_{31} = (3 \times 4) + (1 \times 1) + (2 \times 3) = 19$$

Computing the other elements yields

$$\mathbf{L} = \begin{pmatrix} \mathbf{m_1} \cdot \mathbf{n_1} & \mathbf{m_1} \cdot \mathbf{n_2} & \mathbf{m_1} \cdot \mathbf{n_3} \\ \mathbf{m_2} \cdot \mathbf{n_1} & \mathbf{m_2} \cdot \mathbf{n_2} & \mathbf{m_2} \cdot \mathbf{n_3} \\ \mathbf{m_3} \cdot \mathbf{n_1} & \mathbf{m_3} \cdot \mathbf{n_2} & \mathbf{m_3} \cdot \mathbf{n_3} \end{pmatrix} = \begin{pmatrix} 19 & 8 & 7 \\ 25 & 15 & 23 \\ 11 & 6 & 7 \end{pmatrix}$$

These straightforward, but tedious, calculations for each element in the new matrix are easy performed on a computer. Indeed, most statistical and math packages have all the matrix operations introduced in this chapter as built-in functions.

As suggested above, certain dimensional properties must be satisfied when two matrices are to be multiplied. Specifically, because the dot product is defined only for vectors of the same length, for the matrix product $\mathbf{MN}$ to be defined, the number of columns in $\mathbf{M}$ must equal the number of rows in $\mathbf{N}$. Matrices satisfying this row-column restriction are said to **conform.** Thus, while

$$\begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix}_{2\times 2} \begin{pmatrix} 4 \\ 3 \end{pmatrix}_{2\times 1} = \begin{pmatrix} 12 \\ 10 \end{pmatrix}_{2\times 1}, \qquad \begin{pmatrix} 4 \\ 3 \end{pmatrix}_{2\times 1} \begin{pmatrix} 3 & 0 \\ 1 & 2 \end{pmatrix}_{2\times 2} \quad \text{is undefined.}$$

Writing

$$\mathbf{M}_{r\times c}\mathbf{N}_{c\times b} = \mathbf{L}_{r\times b}$$

shows that the ***inner indices must match***, while the outer indices (r and b) give the number of rows and columns, respectively, of the resulting matrix. A second key point is that ***the order in which matrices are multiplied is critical***. In general, $\mathbf{AB}$ is not equal to $\mathbf{BA}$. Indeed, even if the matricies conform in one order, they many not in the opposite order. In particular, unless one matrix is $r \times c$ and the other is $c \times r$, the two orders of multiplication will not conform (note that two square matrices, $r = c$, of the same dimension conform in either order).

For example, when the order of the matrices in Example 9.2 is reversed,

$$\mathbf{NM} = \begin{pmatrix} 4 & 1 & 0 \\ 1 & 1 & 3 \\ 3 & 2 & 2 \end{pmatrix} \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} = \begin{pmatrix} 14 & 9 & 12 \\ 8 & 9 & 12 \\ 15 & 15 & 18 \end{pmatrix}$$

which differs from $\mathbf{MN}$. Because order is important in matrix multiplication, it has specific terminology. For the product $\mathbf{AB}$, we say that matrix $\mathbf{B}$ is **premultiplied** by the matrix $\mathbf{A}$, or that matrix $\mathbf{A}$ is **postmultiplied** by the matrix $\mathbf{B}$.

Transposition

Another useful matrix operation is **transposition**. The transpose of a matrix $\mathbf{A}$ is written $\mathbf{A}^T$ (while not used in this book, the notation $\mathbf{A}'$ is also widely used), and is obtained simply by switching rows and columns of the original matrix, with $A^T_{ij} = A_{ji}$. If $\mathbf{A}$ is $r \times c$, then $\mathbf{A}^T$ is $c \times r$. As an example,

$$\begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix}^T = \begin{pmatrix} 3 & 2 & 1 \\ 1 & 5 & 1 \\ 2 & 4 & 2 \end{pmatrix}$$

$$\begin{pmatrix} 7 & 4 & 5 \end{pmatrix}^T = \begin{pmatrix} 7 \\ 4 \\ 5 \end{pmatrix}$$

A **symmetric matrix** satisfies $\mathbf{A} = \mathbf{A}^T$, and is necessarily square. An important example of a square, symmetric matrix is a covariance matrix (Equation 9.4a). Matrix expressions involving transposes often arise as a result of matrix derivatives (Equations A3.28 and A3.29). For example, the derivative (with respect to a vector, $\mathbf{x}$) of $\mathbf{Ax}$ is the matrix $\mathbf{A}^T$.

A useful identity (when $\mathbf{A}$ and $\mathbf{B}$ conform) for transposition is that

$$(\mathbf{AB})^T = \mathbf{B}^T\mathbf{A}^T \tag{9.7a}$$

which holds for any number of conformable matrices, e.g.,

$$(\mathbf{ABC})^T = \mathbf{C}^T\mathbf{B}^T\mathbf{A}^T \tag{9.7b}$$

Vectors in statistics and quantitative genetics are generally written as *column vectors* and we follow this convention by using lowercase bold letters, e.g., $\mathbf{a}$, for a column vector and $\mathbf{a}^T$ for the corresponding row vector. With this convention, we distinguish between two vector products, the **inner product** which yields a scalar and the **outer product** which yields a matrix. For the two n-dimensional column vectors $\mathbf{a}$ and $\mathbf{b}$,

$$\mathbf{a} = \begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

their inner product is given by

$$\begin{pmatrix} a_1 & \cdots & a_n \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix} = \mathbf{a}^T\mathbf{b} = \sum_{i=1}^{n} a_i b_i \tag{9.8a}$$

while their outer product yields the $n \times n$ matrix

$$\begin{pmatrix} a_1 \\ \vdots \\ a_n \end{pmatrix} \begin{pmatrix} b_1 & \cdots & b_n \end{pmatrix} = \mathbf{a}\mathbf{b}^T = \begin{pmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_n \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \cdots & a_nb_n \end{pmatrix} \tag{9.8b}$$

Inner products frequently appear in statistics and quantitative genetics, as they represent *weighted sums*. For example, the regression given by Equation 9.1 can be expressed as

$$y = \alpha + \boldsymbol{\beta}^T\mathbf{z} + e, \quad \text{where} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{pmatrix}, \quad \mathbf{z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix}$$

Another example would be a weighted marker score (occasionally called a **polygenic score**), $S = \boldsymbol{\beta}^T\mathbf{z}$, for an individual, where $\mathbf{z}$ is a vector of marker values (e.g., z_j = 0, 1, or 2, respectively, for biallelic marker genotypes m_jm_j, M_jm_j, and M_jM_j) for that individual and β_j is the weighted assigned to the jth marker (i.e., each copy of M_j adds an amount β_j to the score).

Outer products appear in covariance matrices. Consider a $(n \times 1)$ vector $\mathbf{e}$ of random variables, each with mean zero. The resulting $(n \times n)$ covariance matrix, Cov($\mathbf{e}$), has its ij element as $E[\,e_ie_j\,] = \sigma_{ij}$, where

$$\begin{aligned} \text{Cov}(\mathbf{e}) = E[\,\mathbf{e}^T\mathbf{e}\,] &= E\left[\begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix} \begin{pmatrix} e_1 & \cdots & e_n \end{pmatrix} \right] \\ &= \begin{pmatrix} E[e_1e_1] & E[e_1e_2] & \cdots & E[e_1e_n] \\ E[e_2e_1] & E[e_2e_2] & \cdots & E[e_2e_n] \\ \vdots & \vdots & \ddots & \vdots \\ E[e_ne_1] & E[e_ne_2] & \cdots & E[e_ne_n] \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1n} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \cdots & \sigma_n^2 \end{pmatrix} \end{aligned}$$

Inverses and Solutions to Systems of Equations

While matrix multiplication provides a compact way of writing systems of equations, we also need a compact notation for expressing the *solutions* of such systems. This is provided by the **inverse** of a matrix, an operation analogous to scalar division. The importance of matrix inversion can be noted by first considering the solution of the simple scalar equation $ax = b$ for x. Multiplying both sides by a^{-1}, we have $(a^{-1}a)x = 1 \cdot x = x = a^{-1}b$. Now consider a square matrix $\mathbf{A}$. The **inverse of A**, denoted $\mathbf{A}^{-1}$, satisfies $\mathbf{A}^{-1}\mathbf{A} = \mathbf{I} = \mathbf{A}\mathbf{A}^{-1}$, where $\mathbf{I}$, the **identity matrix**, is a square matrix with diagonal elements equal to one and all other elements equal to zero. The identity matrix serves the role that 1 plays in scalar multiplication. Just as $1 \times a = a \times 1 = a$ in scalar multiplication, for any matrix $\mathbf{A} = \mathbf{I}\mathbf{A} = \mathbf{A}\mathbf{I}$. A matrix is called **nonsingular** if its inverse exists. Conditions under which this occurs are discussed in the next section. A useful property of inverses is that if the matrix product $\mathbf{AB}$ is a square matrix (where $\mathbf{A}$ and $\mathbf{B}$ are square and both of their inverses exist), then

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{9.9}$$

The fundamental relationship between the inverse of a matrix and the solution of systems of linear equations can be seen as follows. For a square nonsingular matrix $\mathbf{A}$, the unique solution for $\mathbf{x}$ in the matrix equation $\mathbf{Ax} = \mathbf{c}$ is obtained by premultiplying by $\mathbf{A}^{-1}$,

$$\mathbf{A}^{-1}\mathbf{Ax} = \mathbf{x} = \mathbf{A}^{-1}\mathbf{c} \tag{9.10a}$$

If $\mathbf{A}^{-1}$ does not exist ($\mathbf{A}$ is said to be **singular**), then there are either *no solutions* (the set of equations is **inconsistent**), or there are *infinitely-many solutions*. Consider the follow two sets of equations,

$$\text{Set one:} \quad \begin{aligned} x_1 + 2x_2 &= 3 \\ 2x_1 + 4x_2 &= 6 \end{aligned} \qquad\qquad \text{Set two:} \quad \begin{aligned} x_1 + 2x_2 &= 3 \\ 2x_1 + 4x_2 &= 3 \end{aligned}$$

For both sets, the left-hand side of the second equation is just twice the left-hand set of the first equation. Set one is **consistent**, with a line of solutions, $x_1 = 3 - 2x_2$. More generally, the solution set of a consistent system could be a plane or hyperplane (whereas it is a point when the coefficient matrix is nonsingular). Set two is inconsistent, as no values of x_1 and x_2 can satisfy both equations.

When $\mathbf{A}$ is either singular or nonsquare, solutions for $\mathbf{x}$ (for a consistent system of equations) can still be obtained using **generalized inverses** (denoted by $\mathbf{A}^{-}$) in place of $\mathbf{A}^{-1}$ (Appendix 3). As we have seen, the solutions returned in such cases are certain linear combinations of the elements of $\mathbf{x}$, rather than a unique value for $\mathbf{x}$ (see Appendix 3 for details.) Recalling Equation 9.4b, the solution of the multiple regression equation can be expressed as

$$\boldsymbol{\beta} = \mathbf{V}^{-1}\mathbf{c} \tag{9.10b}$$

Likewise, for the Pearson-Lande-Arnold regression giving the best linear predictor of fitness,

$$\boldsymbol{\beta} = \mathbf{P}^{-1}\mathbf{s} \tag{9.10c}$$

where $\mathbf{P}$ is the covariance matrix for phenotypic measures $z_1, \ldots, z_n$, and $\mathbf{s}$ is the vector of selection differentials for the n characters.

Before developing the formal method for inverting a matrix, we consider two extreme (but very useful) cases that lead to simple expressions for the inverse. First, if the matrix is **diagonal** (all off-diagonal elements are zero), then the matrix inverse is also diagonal, with $\mathrm{A}_{ii}^{-1} = 1/A_{ii}$. For example,

$$\text{for} \quad \mathbf{A} = \begin{pmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{pmatrix}, \qquad \text{then} \qquad \mathbf{A}^{-1} = \begin{pmatrix} a^{-1} & 0 & 0 \\ 0 & b^{-1} & 0 \\ 0 & 0 & c^{-1} \end{pmatrix}$$

Note that if any of the diagonal elements of $\mathbf{A}$ are zero, $\mathbf{A}^{-1}$ is not defined, as $1/0$ is undefined. Second, for any 2×2 matrix,

$$\mathbf{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \qquad \text{then} \qquad \mathbf{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \tag{9.11}$$

To check this result, note that

$$\begin{aligned} \mathbf{A}\mathbf{A}^{-1} &= \frac{1}{ad - bc} \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} \\ &= \frac{1}{ad - bc} \begin{pmatrix} ad - bc & 0 \\ 0 & ad - bc \end{pmatrix} = \mathbf{I} \end{aligned}$$

If $ad = bc$, the inverse does not exist, as division by zero is undefined.

Example 9.3. Consider the multiple regression of y on two predictor variables, z_1 and z_2, so that $y = \alpha + \beta_1 z_1 + \beta_2 z_2 + e$. We solve for the β_i, as the estimate of α follows as $\overline{y} - \beta_1 \overline{z}_1 - \beta_2 \overline{z}_2$. In the notation of Equation 9.4b, we have

$$\mathbf{c} = \begin{pmatrix} \sigma(y, z_1) \\ \\ \sigma(y, z_2) \end{pmatrix} \qquad \mathbf{V} = \begin{pmatrix} \sigma^2(z_1) & \sigma(z_1, z_2) \\ \\ \sigma(z_1, z_2) & \sigma^2(z_2) \end{pmatrix}$$

Recalling that $\sigma(z_1, z_2) = \rho_{12}\, \sigma(z_1)\sigma(z_2)$, Equation 9.11 gives

$$\mathbf{V}^{-1} = \frac{1}{\sigma^2(z_1)\sigma^2(z_2)\,(1 - \rho_{12}^2)} \begin{pmatrix} \sigma^2(z_2) & -\sigma(z_1, z_2) \\ \\ -\sigma(z_1, z_2) & \sigma^2(z_1) \end{pmatrix}$$

The inverse exists provided both characters have nonzero variance and are not completely correlated ($|\rho_{12}| \neq 1$). Recalling Equation 9.10b, the partial regression coefficients are given by $\boldsymbol{\beta} = \mathbf{V}^{-1}\mathbf{c}$, or

$$\begin{pmatrix} \beta_1 \\ \\ \beta_2 \end{pmatrix} = \frac{1}{\sigma^2(z_1)\sigma^2(z_2)\,(1 - \rho_{12}^2)} \begin{pmatrix} \sigma^2(z_2) & -\sigma(z_1, z_2) \\ \\ -\sigma(z_1, z_2) & \sigma^2(z_1) \end{pmatrix} \begin{pmatrix} \sigma(y, z_1) \\ \\ \sigma(y, z_2) \end{pmatrix}$$

Again using $\sigma(z_1, z_2) = \rho_{12}\, \sigma(z_1)\sigma(z_2)$, this equation reduces to

$$\beta_1 = \frac{1}{1 - \rho_{12}^2} \left[\beta_1' - \rho_{12}\, \frac{\sigma(y, z_2)}{\sigma(z_1)\sigma(z_2)} \right]$$

and

$$\beta_2 = \frac{1}{1 - \rho_{12}^2} \left[\beta_2' - \rho_{12}\, \frac{\sigma(y, z_1)}{\sigma(z_1)\sigma(z_2)} \right]$$

where

$$\beta_1' = \frac{\sigma(y, z_1)}{\sigma^2(z_1)} \quad \text{and} \quad \beta_2' = \frac{\sigma(y, z_2)}{\sigma^2(z_2)}$$

are the univariate regression slopes ($y = \alpha' + \beta_i' z_i + e$; Equation 3.14b). Note that only when the predictor variables are uncorrelated ($\rho_{12} = 0$), do the partial regression coefficients β_1 and β_2 reduce to the univariate regression slopes, β_1' and β_2'.

For example, consider our earlier hypothetical example of pollinator visits (y) as a function of flower size (z_1) and nectar volume (z_2). The univariate regression $y = \mu + \beta_1' z_1$ of

number of visits as a function of just flower size has a regression slope of β_1', while the regression coefficients on flower size when both size and nectar volume are included in multiple regression ($y = \mu + \beta_1 z_1 + \beta_2 z_2$) is β_1. β_1 and β_1' are only equal when there is no correlation between size and volume.

Determinants and Minors

For a 2×2 matrix, the quantity

$$|\mathbf{A}| = A_{11}A_{22} - A_{12}A_{21} \tag{9.12a}$$

is called the **determinant**, which more generally is denoted by $\det(\mathbf{A})$ or $|\mathbf{A}|$. As with the 2- dimensional case, $\mathbf{A}^{-1}$ exists for a square matrix $\mathbf{A}$ (of any dimensionality) if and only if $\det(\mathbf{A}) \neq 0$. For square matrices with dimensionality greater than two, the determinant is obtained recursively from the general expression

$$|\mathbf{A}| = \sum_{j=1}^{n} A_{ij}(-1)^{i+j}|\mathbf{A}_{ij}| \tag{9.12b}$$

where i is any fixed row of the matrix $\mathbf{A}$ and $\mathbf{A}_{ij}$ is the $(n-1) \times (n-1)$ submatrix obtained by deleting the ith row and jth column from $\mathbf{A}$. Such a submatrix is known as a **minor.** In words, each of the n quantities in this equation is the product of three components: the element in the row around which one is working, -1 to the $(i+j)$th power, and the determinant of the ijth minor. In applying Equation 9.12b, one starts with the original $n \times n$ matrix and works down until the minors are reduced to 2×2 matrices whose determinants are scalars of the form $A_{11}A_{22} - A_{12}A_{21}$. A useful result is that the determinant of a diagonal matrix is the product of the diagonal elements of that matrix, so that if

$$A_{ij} = \begin{cases} a_i & i = j \\ 0 & i \neq j \end{cases} \qquad \text{then} \qquad |\,\mathbf{A}\,| = \prod_{i=1}^{n} a_i \tag{9.12c}$$

The next section shows how determinants are used in the computation of a matrix inverse.

Example 9.4. Compute the determinant of

$$\mathbf{A} = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 3 & 2 \\ 1 & 2 & 1 \end{pmatrix}$$

Letting $i = 1$, i.e., using the elements in the first row of $\mathbf{A}$,

$$|\mathbf{A}| = 1 \cdot (-1)^{1+1} \begin{vmatrix} 3 & 2 \\ 2 & 1 \end{vmatrix} + 1 \cdot (-1)^{1+2} \begin{vmatrix} 1 & 2 \\ 1 & 1 \end{vmatrix} + 1 \cdot (-1)^{1+3} \begin{vmatrix} 1 & 3 \\ 1 & 2 \end{vmatrix}$$

Using Equation 9.12a to obtain the determinants of the 2×2 matrices, this simplifies to

$$|\mathbf{A}| = [1 \times (3-4)] - [1 \times (1-2)] + [1 \times (2-3)] = -1$$

The same answer is obtained regardless of which row is used, and expanding around a column, instead of a row, produces the same result. Thus, in order to reduce the number of computations required to obtain a determinant, it is useful to expand using the row or column that contains the most zeros. As with all other matrix operations presented in this chapter, these calculations are almost always performed using the build-in matrix functions in most computer packages.

Example 9.5. To see further connections between the determinant and the solution to a set of equations, consider the following two systems of equations:

$$\text{Set one:} \quad \begin{matrix} x_1 + x_2 = 1 \\ 2x_1 + 2x_2 = 2 \end{matrix} \qquad \text{Set two:} \quad \begin{matrix} 0.9999 \cdot x_1 + x_2 = 1 \\ 2x_1 + 2x_2 = 2 \end{matrix}$$

The determinant for the coefficient matrix associated with set one is zero, and there is no unique solution, rather a line of solutions, $x_1 = 1 - x_2$. In contrast, the determinant for the matrix associated with set two is nonzero, hence its inverse exists and there is a unique solution. However, the determinant nearly zero, 0.0002. Such a matrix is said to be **nearly singular**, meaning that although the two sets of equations are distinct, they overlap so closely that there is little additional information from one (or more) of the equations. For this set of equations,

$$\mathbf{A}^{-1} = \begin{pmatrix} -10,000 & 5000 \\ -10,000 & -4999.5 \end{pmatrix}, \quad \mathbf{x} = \begin{pmatrix} -3.63 \times 10^{-12} \\ 1 \end{pmatrix} \simeq \begin{pmatrix} 0 \\ 1 \end{pmatrix}$$

While there *technically* is a unique solution, it is *extremely* sensitive to the coefficients in the set of equations, and a very small change (such as through measurement error) can dramatically change the solution. For example, replacing the first equation by $x_1 + 0.9999 \cdot x_2 = 1$, yields the solution of $x_1 = 1, x_2 \simeq 0$.

Computing Inverses

The general solution of a matrix inverse is

$$A^{-1}_{ji} = \frac{(-1)^{i+j}|\mathbf{A}_{ij}|}{|\mathbf{A}|} \tag{9.13}$$

where A^{-1}_{ji} denotes the jith element of $\mathbf{A}^{-1}$, and $\mathbf{A}_{ij}$ denotes the ijth minor of $\mathbf{A}$. The reversed subscripts (ji versus ij) in the left and right expressions arise because the right-hand side computes an element in the transpose of the inverse (see Example 9.6). It can be seen from Equation 9.13 that ***a matrix can only be inverted if it has a nonzero determinant***. Thus, ***a matrix is singular if its determinant is zero***. This occurs whenever a matrix contains a row (or column) that can be written as a weighted sum of the other rows (or columns). In the context of a linear model, this happens if one of the n equations can be written as a combination of the others, a situation that is equivalent to there being n unknowns but less than n independent equations.

Example 9.6. Compute the inverse of

$$\mathbf{A} = \begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix}$$

First, find the determinants of the minors,

$$|\mathbf{A}_{11}| = \begin{vmatrix} 5 & 4 \\ 1 & 2 \end{vmatrix} = 6 \qquad |\mathbf{A}_{23}| = \begin{vmatrix} 3 & 1 \\ 1 & 1 \end{vmatrix} = 2$$

$$|\mathbf{A}_{12}| = \begin{vmatrix} 2 & 4 \\ 1 & 2 \end{vmatrix} = 0 \qquad |\mathbf{A}_{31}| = \begin{vmatrix} 1 & 2 \\ 5 & 4 \end{vmatrix} = -6$$

$$|\mathbf{A}_{13}| = \begin{vmatrix} 2 & 5 \\ 1 & 1 \end{vmatrix} = -3 \qquad |\mathbf{A}_{32}| = \begin{vmatrix} 3 & 2 \\ 2 & 4 \end{vmatrix} = 8$$

$$|\mathbf{A}_{21}| = \begin{vmatrix} 1 & 2 \\ 1 & 2 \end{vmatrix} = 0 \qquad |\mathbf{A}_{33}| = \begin{vmatrix} 3 & 1 \\ 2 & 5 \end{vmatrix} = 13$$

$$|\mathbf{A}_{22}| = \begin{vmatrix} 3 & 2 \\ 1 & 2 \end{vmatrix} = 4$$

Using Equation 9.12b and expanding using the first row of $\mathbf{A}$ gives

$$|\mathbf{A}| = 3|\mathbf{A}_{11}| - |\mathbf{A}_{12}| + 2|\mathbf{A}_{13}| = 12$$

Returning to the matrix in brackets in Equation 9.13, we obtain

$$\frac{1}{12}\begin{pmatrix} 1\times 6 & -1\times 0 & 1\times -3 \\ -1\times 0 & 1\times 4 & -1\times 2 \\ 1\times -6 & -1\times 8 & 1\times 13 \end{pmatrix} = \frac{1}{12}\begin{pmatrix} 6 & 0 & -3 \\ 0 & 4 & -2 \\ -6 & -8 & 13 \end{pmatrix}$$

and then taking the transpose,

$$\mathbf{A}^{-1} = \frac{1}{12}\begin{pmatrix} 6 & 0 & -6 \\ 0 & 4 & -8 \\ -3 & -2 & 13 \end{pmatrix}$$

To verify that this is indeed the inverse of $\mathbf{A}$, multiply $\mathbf{A}^{-1}$ by $\mathbf{A}$,

$$\frac{1}{12}\begin{pmatrix} 6 & 0 & -6 \\ 0 & 4 & -8 \\ -3 & -2 & 13 \end{pmatrix}\begin{pmatrix} 3 & 1 & 2 \\ 2 & 5 & 4 \\ 1 & 1 & 2 \end{pmatrix} = \frac{1}{12}\begin{pmatrix} 12 & 0 & 0 \\ 0 & 12 & 0 \\ 0 & 0 & 12 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Again, these tedious calculations are performed using matrix inversion routines available in standard packages.

EXPECTATIONS OF RANDOM VECTORS AND MATRICES

Matrix algebra provides a powerful approach for analyzing linear combinations of random variables. Let $\mathbf{x}$ be a column vector containing n random variables, $\mathbf{x} = (x_1, x_2, \cdots, x_n)^T$. We may wish to construct a new univariate (scalar) random variable y by taking some linear combination of the elements of $\mathbf{x}$,

$$y = \sum_{i=1}^{n} a_i x_i = \mathbf{a}^T\mathbf{x}$$

where $\mathbf{a} = (a_1, a_2, \cdots, a_n)^T$ is a column vector of constants. Likewise, we can construct a new k-dimensional vector $\mathbf{y}$ by premultiplying $\mathbf{x}$ by a $k \times n$ matrix $\mathbf{A}$ of constants, $\mathbf{y}_{k\times 1} = \mathbf{A}_{k\times n}\mathbf{x}_{n\times 1}$. Here $\mathbf{y}$ is a vector of k weighted sums, the jth of which is

$$y_j = \sum_{i=1}^{n} A_{ji} x_i = \mathbf{a}_j^T\mathbf{x}$$

where $\mathbf{a}_j$ denotes the jth row of $\mathbf{A}$. More generally, an $(n \times k)$ matrix $\mathbf{X}$ of random variables can be transformed into a new $m \times \ell$ dimensional matrix, $\mathbf{Y}$, of elements consisting of linear combinations of the elements of $\mathbf{X}$ by

$$\mathbf{Y}_{m\times\ell} = \mathbf{A}_{m\times n}\mathbf{X}_{n\times k}\mathbf{B}_{k\times\ell} \tag{9.14}$$

where the matrices $\mathbf{A}$ and $\mathbf{B}$ are constants with dimensions as subscripted.

If $\mathbf{X}$ is a matrix whose elements are random variables, then the expected value of $\mathbf{X}$ is a matrix $E[\mathbf{X}]$ containing the expected value of each element of $\mathbf{X}$. If $\mathbf{X}$ and $\mathbf{Z}$ are matrices of the same dimension, then

$$E[\mathbf{X} + \mathbf{Z}] = E[\mathbf{X}] + E[\mathbf{Z}] \tag{9.15}$$

This easily follows because the ijth element of $E[\mathbf{X} + \mathbf{Z}]$ is $E[x_{ij} + z_{ij}] = E[x_{ij}] + E[z_{ij}]$. Similarly, the expectation of $\mathbf{Y}$ as defined in Equation 9.14 is

$$E[\mathbf{Y}] = E[\mathbf{AXB}] = \mathbf{A}E[\mathbf{X}]\mathbf{B} \tag{9.16a}$$

For example, for $\mathbf{y} = \mathbf{Xb}$ where $\mathbf{b}$ is an $n \times 1$ column vector,

$$E[\mathbf{y}] = E[\mathbf{Xb}] = E[\mathbf{X}]\mathbf{b} \tag{9.16b}$$

If $\mathbf{X}$ is a matrix of fixed constants, then $E[\mathbf{y}] = \mathbf{Xb}$. Likewise, for $y = \mathbf{a}^T\mathbf{x} = \sum_i^n a_i x_i$,

$$E[y] = E[\mathbf{a}^T\mathbf{x}] = \mathbf{a}^T E[\mathbf{x}] \tag{9.16c}$$

COVARIANCE MATRICES OF TRANSFORMED VECTORS

To develop expressions for variances and covariances of linear combinations of random variables, we must first introduce the concept of quadratic forms. Consider an $n \times n$ square matrix $\mathbf{A}$ and an $n \times 1$ column vector $\mathbf{x}$. From the rules of matrix multiplication,

$$\mathbf{x}^T\mathbf{Ax} = \sum_{i=1}^{n}\sum_{j=1}^{n} A_{ij}x_i x_j \tag{9.17a}$$

Expressions of this form are called **quadratic forms** (or **quadratic products**) as they involve squares (x_i^2) and cross-products ($x_i x_j$), and yield a scalar. A generalization of a quadratic form is the **bilinear form**, $\mathbf{b}^T\mathbf{Aa}$, where $\mathbf{b}$ and $\mathbf{a}$ are, respectively, $n \times 1$ and $m \times 1$ column vectors and $\mathbf{A}$ is an $n \times m$ matrix. Indexing the matrices and vectors in this expression by their dimensions, $\mathbf{b}_{1\times n}^T\mathbf{A}_{n\times m}\mathbf{a}_{m\times 1}$, shows that the resulting matrix product defined and yields a 1×1 matrix; in other words, a scalar. As scalars, bilinear forms equal their transposes (as the transpose of a scalar simply returns that scalar), giving the useful identity

$$\mathbf{b}^T\mathbf{Aa} = \left(\mathbf{b}^T\mathbf{Aa}\right)^T = \mathbf{a}^T\mathbf{A}^T\mathbf{b} \tag{9.17b}$$

Again let $\mathbf{x}$ be a column vector of n random variables. A compact way to express the n variances and $n(n-1)/2$ covariances associated with the elements of $\mathbf{x}$ is the $n \times n$ matrix $\mathbf{V}$, where $V_{ij} = \sigma(x_i, x_j)$ is the covariance between the random variables x_i and x_j. We will generally refer to $\mathbf{V}$ as a **covariance matrix**, noting that the *diagonal elements represent the variances* and *off-diagonal elements the covariances*. The $\mathbf{V}$ matrix is symmetric ($\mathbf{V} = \mathbf{V}^T$), as

$$V_{ij} = \sigma(x_i, x_j) = \sigma(x_j, x_i) = V_{ji} \tag{9.18a}$$

Note that we can write the covariance matrix as an outer product. Letting $E[\mathbf{x}] = \boldsymbol{\mu}$, then

$$\mathbf{V} = E\left[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\right] \tag{9.18b}$$

This follows as Equation 9.8b gives the ijth element of Equation 9.18b as $E[(x_i - \mu_i)(x_j - \mu_j)] = \sigma(x_i, x_j)$.

Now consider a univariate random variable, $y = \sum c_k\, x_k$, generated from a linear combination of the elements of $\mathbf{x}$. In matrix notation, $y = \mathbf{c}^T\mathbf{x}$, where $\mathbf{c}$ is a column vector of constants. The variance of y can be expressed as a quadratic form involving the covariance matrix $\mathbf{V}$ for the elements of $\mathbf{x}$,

$$\begin{aligned}
\sigma^2\left(\mathbf{c}^T\mathbf{x}\right) &= \sigma^2\left(\sum_{i=1}^{n} c_i x_i\right) = \sigma\left(\sum_{i=1}^{n} c_i\, x_i\,, \sum_{j=1}^{n} c_j\, x_j\right) \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n} \sigma\left(c_i\, x_i, c_j\, x_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} c_i\, c_j\, \sigma\left(x_i, x_j\right) \\
&= \mathbf{c}^T\mathbf{V}\,\mathbf{c}
\end{aligned} \tag{9.19}$$

Note that if $\mathbf{V}$ is a proper covariance matrix, then $\mathbf{c}^T\mathbf{V}\mathbf{c} \geq 0$ for all $\mathbf{c}$, as this quadratic product represents the variance (and hence ≥ 0) of some index of the elements of $\mathbf{x}$.

Similarly, the covariance between two univariate random variables created from different linear combinations of $\mathbf{x}$ is given by the bilinear form

$$\sigma(\mathbf{a}^T\mathbf{x}, \mathbf{b}^T\mathbf{x}) = \mathbf{a}^T\mathbf{V}\,\mathbf{b} \tag{9.20}$$

If we transform $\mathbf{x}$ to two new vectors, $\mathbf{y}_{\ell\times 1} = \mathbf{A}_{\ell\times n}\mathbf{x}_{n\times 1}$ and $\mathbf{z}_{m\times 1} = \mathbf{B}_{m\times n}\mathbf{x}_{n\times 1}$, then instead of a single covariance we have an $\ell \times m$ dimensional matrix of covariances, denoted $\boldsymbol{\sigma}(\mathbf{y}, \mathbf{z})$, whose ijth element is $\sigma(y_i, z_j)$. Letting $\boldsymbol{\mu}_\mathbf{y} = \mathbf{A}\boldsymbol{\mu}$ and $\boldsymbol{\mu}_\mathbf{z} = \mathbf{B}\boldsymbol{\mu}$, with $E(\mathbf{x}) = \boldsymbol{\mu}$, then $\boldsymbol{\sigma}(\mathbf{y}, \mathbf{z})$ can be expressed in terms of $\mathbf{V}$, the covariance matrix of $\mathbf{x}$,

$$\begin{aligned}
\boldsymbol{\sigma}(\mathbf{y}, \mathbf{z}) &= \boldsymbol{\sigma}(\mathbf{A}\mathbf{x}, \mathbf{B}\mathbf{x}) \\
&= E\left[(\mathbf{y} - \boldsymbol{\mu}_\mathbf{y})(\mathbf{z} - \boldsymbol{\mu}_\mathbf{z})^T\right] \\
&= E\left[\mathbf{A}(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^T\mathbf{B}^T\right] \\
&= \mathbf{A}\mathbf{V}\,\mathbf{B}^T
\end{aligned} \tag{9.21a}$$

In particular, the covariance matrix for $\mathbf{y} = \mathbf{A}\mathbf{x}$ is

$$\boldsymbol{\sigma}(\mathbf{y}, \mathbf{y}) = \mathbf{A}\mathbf{V}\,\mathbf{A}^T \tag{9.21b}$$

so that the covariance between y_i and y_j is given by the ijth element of the matrix product $\mathbf{A}\mathbf{V}\mathbf{A}^T$.

Finally, note that if $\mathbf{x}$ is a vector of random variables with expected value $\boldsymbol{\mu}$, then the expected value of the (scalar) quadratic product $\mathbf{x}^T\mathbf{A}\mathbf{x}$ is

$$E(\mathbf{x}^T\mathbf{A}\mathbf{x}) = \text{tr}(\mathbf{A}\mathbf{V}) + \boldsymbol{\mu}^T\mathbf{A}\boldsymbol{\mu} \tag{9.22}$$

where $\mathbf{V}$ is the covariance matrix for the elements of $\mathbf{x}$, and the **trace** of a square matrix, $\text{tr}(\mathbf{M}) = \sum M_{ii}$, is the sum of its diagonal values (Searle 1971). Note that when x is a scalar and $\mathbf{A} = (1)$, Equation 9.22 collapses to $E[x^2] = \sigma^2 + \mu^2$. More generally, $E[x_i x_j] = \sigma(x_i, x_j) + \mu_i\mu_j$.

Example 9.7 Consider three traits with the following covariance structure: $\sigma^2(x_1) = 10$, $\sigma^2(x_2) = 20, \sigma^2(x_3) = 30, \sigma(x_1, x_2) = -5, \sigma(x_1, x_3) = 10$, and $\sigma(x_2, x_3) = 0$, yielding the covariance matrix

$$\mathbf{V} = \begin{pmatrix} 10 & -5 & 10 \\ -5 & 20 & 0 \\ 10 & 0 & 30 \end{pmatrix}$$

Further, assume the vector of means for these variables is $\boldsymbol{\mu}^T = (10, 4, -3)$. Consider two new indices constructed from these variables, with $y_1 = 2x_1 - 3x_2 + 4x_3$ and $y_2 = x_2 - 2x_3$. We can express these as inner products, $y_i = \mathbf{c}_i^T \mathbf{x}$, where

$$\mathbf{c}_1 = \begin{pmatrix} 2 \\ -3 \\ 4 \end{pmatrix}, \qquad \mathbf{c}_2 = \begin{pmatrix} 0 \\ 1 \\ -2 \end{pmatrix}$$

From Equation 9.19, the resulting variances become $\sigma^2(y_1) = \mathbf{c}_1{}^T \mathbf{V}\, \mathbf{c}_1 = 920$ and $\sigma^2(y_2) = \mathbf{c}_2{}^T \mathbf{V}\, \mathbf{c}_2 = 140$. Applying Equation 9.20, the covariance between these two indices is $\sigma(y_1, y_2) = \mathbf{c}_1{}^T \mathbf{V}\, \mathbf{c}_2 = -350$, yielding their correlation as

$$\rho(y_1, y_2) = \frac{\sigma(y_1, y_2)}{\sqrt{\sigma^2(y_1) \cdot \sigma^2(y_2)}} = \frac{-350}{\sqrt{920 \cdot 140}} = 0.975$$

Finally, consider the following quadratic function of $\mathbf{x}$,

$$y = x_1^2 - 2x_1x_2 + 4x_2x_3 + 3x_2^2 - 4x_3^2$$

We can write this in matrix form as $\mathbf{x}^T \mathbf{A} \mathbf{x}$, where

$$\mathbf{A} = \begin{pmatrix} 1 & -1 & 0 \\ -1 & 3 & 2 \\ 0 & 2 & -4 \end{pmatrix}$$

The expected value of y follows from Equation 9.22, with $E[y] = \boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} + \text{tr}(\mathbf{AV})$, where $\boldsymbol{\mu}^T \mathbf{A} \boldsymbol{\mu} = -16$ and

$$\mathbf{AV} = \begin{pmatrix} 15 & -25 & 10 \\ -5 & 65 & 50 \\ -50 & 40 & -120 \end{pmatrix}$$

which has a trace of $15 + 65 - 120 = -40$, yielding an expected value of $E[y] = -16 - 40 = -56$.

THE MULTIVARIATE NORMAL DISTRIBUTION

As we have seen above, matrix notation provides a compact way to express vectors of random variables. We now consider the most commonly assumed distribution for such vectors, the multivariate analog of the normal distribution discussed in Chapter 2. Much of the machinery of quantitative genetics is based on this distribution, which we hereafter denote as the **MVN**.

Consider the probability density function for n independent normal random variables, where x_i is normally distributed with mean μ_i and variance σ_i^2, which we denote as $x_i \sim$

$\mathrm{N}(\mu_i, \sigma_i^2)$, with $\mathbf{x} = (x_1 \cdots, x_n)^T$. In this case, because the variables are independent, the joint probability density function is simply the product of each univariate density,

$$\begin{aligned} p(\mathbf{x}) &= \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} (2\pi)^{-1/2} \sigma_i^{-1} \exp\left(-\frac{(x_i - \mu_i)^2}{2\,\sigma_i^2} \right) \\ &= (2\pi)^{-n/2} \left(\prod_{i=1}^{n} \sigma_i \right)^{-1} \exp\left(-\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{2\,\sigma_i^2} \right) \end{aligned} \tag{9.23}$$

We can express this equation more compactly in matrix form by defining the matrices

$$\mathbf{V} = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & \sigma_n^2 \end{pmatrix} \qquad \text{and} \qquad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{pmatrix}$$

Because $\mathbf{V}$ is diagonal, its determinant is simply the product of the diagonal elements

$$|\mathbf{V}| = \prod_{i=1}^{n} \sigma_i^2$$

Likewise, $\mathbf{V}^{-1}$ is also diagonal, with ith diagonal element $1/\sigma_i^2$. Hence, using quadratic products, we have

$$\sum_{i=1}^{n} \frac{(x_i - \mu_i)^2}{\sigma_i^2} = (\mathbf{x} - \boldsymbol{\mu})^T \, \mathbf{V}^{-1} \, (\mathbf{x} - \boldsymbol{\mu})$$

Substituting in these expressions, Equation 9.23 can be rewritten as

$$p(\mathbf{x}) = (2\pi)^{-n/2} \, |\mathbf{V}|^{-1/2} \, \exp\left[-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \, \mathbf{V}^{-1} \, (\mathbf{x} - \boldsymbol{\mu}) \right] \tag{9.24}$$

We will also write this density as $p(\mathbf{x}, \boldsymbol{\mu}, \mathbf{V})$ when we wish to stress that it is a function of the mean vector $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{V}$.

More generally, when the elements of $\mathbf{x}$ are correlated ($\mathbf{V}$ is not diagonal), Equation 9.24 gives the probability density function for a vector of multivariate normally distributed random variables, with mean vector $\boldsymbol{\mu}$ and covariance matrix $\mathbf{V}$. We denote this by

$$\mathbf{x} \sim \mathrm{MVN}_n(\boldsymbol{\mu}, \mathbf{V})$$

where the subscript indicating the dimensionality of $\mathbf{x}$ is usually omitted. The multivariate normal distribution is also referred to as the **Gaussian distribution**. We restrict our attention to those situations where $\mathbf{V}$ is nonsingular, in which case it is **positive-definite** (namely, $\mathbf{c}^T \mathbf{V} \mathbf{c} > 0$ for all vectors $\mathbf{c} \neq 0$; WL Appendix 5). When $\mathbf{V}$ is singular, some of the elements of $\mathbf{x}$ are linear functions of other elements in $\mathbf{x}$, and hence we can construct a reduced vector of variables whose covariance matrix is now nonsingular.

Properties of the MVN

As in the case of its univariate counterpart, the MVN is expected to arise naturally when the quantities of interest result from a large number of underlying variables. Because this condition seems (at least at first glance) to describe many biological systems, the MVN is a natural starting point in biometrical analysis. Further details on the wide variety of applications of the MVN to multivariate statistics can be found in the introductory texts by Morrison (1976) and Johnson and Wichern (2002) and in the more advanced treatment by Anderson (2003). The MVN has a number of useful properties, which we summarize below.

1. *If* $\mathbf{x} \sim$ MVN, *then the distribution of any subset of the variables in* $\mathbf{x}$ *is also MVN*. For example, each x_i is normally distributed and each pair (x_i, x_j) is bivariate normally distributed.

2. *If* $\mathbf{x} \sim$ MVN, *then any linear combination of the elements of* $\mathbf{x}$ *is also MVN*. Specifically, if $\mathbf{x} \sim \text{MVN}_n(\boldsymbol{\mu}, \mathbf{V})$, $\mathbf{a}$ is a vector of constants, and $\mathbf{A}_{m\times n}$ is a matrix of constants, then

$$\text{for} \quad \mathbf{y} = \mathbf{x} + \mathbf{a}, \qquad \mathbf{y} \sim \text{MVN}_n(\boldsymbol{\mu} + \mathbf{a}, \mathbf{V}) \tag{9.25a}$$

$$\text{for} \quad y = \mathbf{a}^T\mathbf{x} = \sum_{k=1}^{n} a_i x_i, \qquad y \sim \text{N}(\mathbf{a}^T\boldsymbol{\mu}, \mathbf{a}^T\mathbf{V}\,\mathbf{a}) \tag{9.25b}$$

$$\text{for} \quad \mathbf{y}_{m\times 1} = \mathbf{A}_{m\times n}\mathbf{x}_{n\times 1}, \qquad \mathbf{y} \sim \text{MVN}_m\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}^T\mathbf{V}\mathbf{A}\right) \tag{9.25c}$$

3. *Conditional distributions associated with the MVN are also multivariate normal*. Consider the partitioning of $\mathbf{x}$ into two components, an $(m \times 1)$ column vector $\mathbf{x_1}$ and an $[(n-m)\times 1]$ column vector $\mathbf{x_2}$ of the remaining variables, e.g.,

$$\mathbf{x} = \begin{pmatrix} \mathbf{x_1} \\ \mathbf{x_2} \end{pmatrix}$$

The mean vector and covariance matrix can be partitioned similarly as

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu_1} \\ \boldsymbol{\mu_2} \end{pmatrix} \qquad \text{and} \qquad \mathbf{V} = \begin{pmatrix} \mathbf{V_{x_1x_1}} & \mathbf{V_{x_1x_2}} \\ \mathbf{V}^T_{\mathbf{x_1x_2}} & \mathbf{V_{x_2x_2}} \end{pmatrix} \tag{9.26}$$

where the $m\times m$ and $(n-m)\times(n-m)$ matrices $\mathbf{V_{x_1x_1}}$ and $\mathbf{V_{x_2x_2}}$ are, respectively, the covariance matrices for $\mathbf{x_1}$ and $\mathbf{x_2}$, while the $m \times (n-m)$ matrix $\mathbf{V_{x_1x_2}}$ is the matrix of covariances between the elements of $\mathbf{x_1}$ and $\mathbf{x_2}$. If we condition on $\mathbf{x_2}$, the resulting conditional random variable, $\mathbf{x_1}|\mathbf{x_2}$, is MVN with $(m \times 1)$ mean vector

$$\boldsymbol{\mu}_{\mathbf{x_1}|\mathbf{x_2}} = \boldsymbol{\mu_1} + \mathbf{V_{x_1x_2}}\mathbf{V}^{-1}_{\mathbf{x_2x_2}}(\mathbf{x_2} - \boldsymbol{\mu_2}) \tag{9.27}$$

and $(m \times m)$ covariance matrix

$$\mathbf{V}_{\mathbf{x_1}|\mathbf{x_2}} = \mathbf{V_{x_1x_1}} - \mathbf{V_{x_1x_2}}\mathbf{V}^{-1}_{\mathbf{x_2x_2}}\mathbf{V}^T_{\mathbf{x_1x_2}} \tag{9.28}$$

A proof can be found in most multivariate statistics texts, e.g., Morrison (1976).

4. *If* $\mathbf{x} \sim$ MVN, *the regression of any subset of* $\mathbf{x}$ *on another subset is linear and homoscedastic*. Rewriting Equation 9.27 in terms of a regression of the predicted value of the vector $\mathbf{x_1}$ given an observed value of the vector $\mathbf{x_2}$, we have

$$\mathbf{x_1} = \boldsymbol{\mu_1} + \mathbf{V_{x_1x_2}}\mathbf{V}^{-1}_{\mathbf{x_2x_2}}(\mathbf{x_2} - \boldsymbol{\mu_2}) + \mathbf{e} \tag{9.29a}$$

where

$$\mathbf{e} \sim \text{MVN}_m\left(\mathbf{0}, \mathbf{V}_{\mathbf{x_1}|\mathbf{x_2}}\right) \tag{9.29b}$$

Extending the terminology of a univariate regression (Chapter 3) to this multivariate setting, one can think of $\mathbf{x_1}$ as the response variable (a vector in this case) and $\mathbf{x_2}$ as the predictor variable (also now a vector). Hence, $\mathbf{V_{x_1x_2}}$ is the covariance between the response and predictor variances, while $\mathbf{V_{x_2x_2}}$ is the (co)variance structure of the predictor variables. This generalizes the univariate slope of $\sigma(y,x)\sigma^{-2}(x)$ to an $m \times (n-m)$ matrix of slopes, $\mathbf{V_{x_1x_2}}\mathbf{V}^{-1}_{\mathbf{x_2x_2}}$, in a multivariate setting.

Example 9.8. Consider the regression of the phenotypic value of an offspring (z_o) on that of its parents (z_s and z_d for sire and dam, respectively). Assume that the joint distribution of z_o, z_s, and z_d is multivariate normal. For the simplest case of noninbred and unrelated parents, no epistasis or genotype-environment correlation, the covariance matrix can be obtained from the theory of correlation between relatives (Chapter 7), giving the joint distribution as

$$\begin{pmatrix} z_o \\ z_s \\ z_d \end{pmatrix} \sim \text{MVN}\left[\begin{pmatrix} \mu_o \\ \mu_s \\ \mu_d \end{pmatrix}, \sigma_z^2 \cdot \begin{pmatrix} 1 & h^2/2 & h^2/2 \\ h^2/2 & 1 & 0 \\ h^2/2 & 0 & 1 \end{pmatrix} \right] \tag{9.30a}$$

where the off-diagonal elements, $\sigma_A^2/2 = (h^2/2)\sigma_z^2$, follow from the parent-offspring covariance. Let

$$\mathbf{x_1} = (\, z_o \,), \quad \mathbf{x_2} = \begin{pmatrix} z_s \\ z_d \end{pmatrix}$$

giving

$$\mathbf{V}_{\mathbf{x_1},\mathbf{x_1}} = \sigma_z^2, \quad \mathbf{V}_{\mathbf{x_1},\mathbf{x_2}} = \frac{h^2\sigma_z^2}{2}(\,1 \quad 1\,), \quad \mathbf{V}_{\mathbf{x_2},\mathbf{x_2}} = \sigma_z^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

From Equation 9.29a, the regression of offspring value on parental values is linear and homoscedastic with

$$\begin{aligned} z_o &= \mu_o + \frac{h^2\sigma_z^2}{2}(\,1 \quad 1\,)\,\sigma_z^{-2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} z_s - \mu_s \\ z_d - \mu_d \end{pmatrix} + e \\ &= \mu_o + \frac{h^2}{2}(z_s - \mu_s) + \frac{h^2}{2}(z_d - \mu_d) + e \end{aligned} \tag{9.30b}$$

where, from Equations 9.28 and 9.29b, the residual error is normally distributed with mean zero and variance

$$\begin{aligned} \sigma_e^2 &= \sigma_z^2 - \frac{h^2\sigma_z^2}{2}(\,1 \quad 1\,)\,\sigma_z^{-2} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \frac{h^2\sigma_z^2}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \\ &= \sigma_z^2 \left(1 - \frac{h^4}{2}\right) \end{aligned} \tag{9.30c}$$

This same approach allows one to consider more complex situations. For example, suppose that the parents assortatively mate (with phenotypic correlation ρ_z). From Table 7.5, the parent-offspring covariance now becomes $\sigma_A^2(1+\rho_z)/2$, and the resulting covariance matrix becomes

$$\sigma_z^2 \cdot \begin{pmatrix} 1 & h^2(1+\rho_z)/2 & h^2(1+\rho_z)/2 \\ h^2(1+\rho_z)/2 & 1 & \rho_z \\ h^2(1+\rho_z)/2 & \rho_z & 1 \end{pmatrix} \tag{9.30d}$$

Similarly, when the parent are inbred and/or related, Equation 7.4b gives the covariances between an offspring and its sire and dam as $(2\sigma_A^2)(\Theta_{ss} + \Theta_{sd})/2$ and $(2\sigma_A^2)(\Theta_{dd} + \Theta_{sd})/2$, respectively, while Equation 7.11a gives the covariance between sire and dam as $\sigma_A(s,d) = 2\Theta_{sd}\sigma_A^2$. Finally, the phenotypic variance of an inbred individual is $\sigma_A^2(1+f) + \sigma_E^2 = \sigma_z^2 + \sigma_A^2 f = \sigma_z^2(1+fh^2)$. Incorporating these expressions, the resulting covariance matrix becomes

$$\sigma_z^2 \cdot \begin{pmatrix} 1 + f_0 h^2 & h^2(\Theta_{ss} + \Theta_{sd}) & h^2(\Theta_{dd} + \Theta_{sd}) \\ h^2(\Theta_{ss} + \Theta_{sd}) & 1 + f_s h^2 & 2h^2\Theta_{sd} \\ h^2(\Theta_{dd} + \Theta_{sd}) & 2h^2\Theta_{sd} & 1 + f_d h^2 \end{pmatrix} \tag{9.30d}$$

Recalling Equation 7.3b allows us to entirely write this matrix in terms of the three coefficients of coancestry associated with the parents (Θ_{ss}, Θ_{dd}, and Θ_{sd}), as $f_x = 2\Theta_{xx} - 1$, while

$f_o = \Theta_{sd}$. Equation 9.11 easily allows one to invert the 2 × 2 matrix $\mathbf{V}_{\mathbf{X_2},\mathbf{X_2}}$ associated with the parents, and using the approach leading to Equationa 9.30b and 9.30c yields the regression and associated residual variance under these more complex cases.

Example 9.9. The previous example dealt with the prediction of the phenotypic value of an offspring given its parental phenotypic values. The same approach can be used to predict an offspring's additive genetic (or breeding) value (A_o) given knowledge of the parental values (A_s, A_d). Again assuming that the joint distribution is multivariate normal and that the parents are unrelated and noninbred, the joint distribution can be written as

$$\begin{pmatrix} A_o \\ A_s \\ A_d \end{pmatrix} \sim \text{MVN}\left[\begin{pmatrix} \mu_o \\ \mu_s \\ \mu_d \end{pmatrix}, \sigma_A^2 \begin{pmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 0 \\ 1/2 & 0 & 1 \end{pmatrix} \right] \tag{9.31a}$$

Proceeding in the same fashion as in Example 9.8, the conditional distribution of offspring additive genetic values, given the parental values, is normal, so that the regression of offspring additive genetic value on parental value is linear and homoscedastic with

$$A_o = \mu_o + \frac{A_s - \mu_s}{2} + \frac{A_d - \mu_d}{2} + e \tag{9.31b}$$

and

$$e \sim \text{N}(0, \sigma_A^2/2) \tag{9.31c}$$

Finally, an important merger of concepts from this and the previous example is the prediction of an offspring's *breeding value* (A_o) given the *phenotypes* (z_s, z_d) of its parents. Assuming multivariate normality,

$$\begin{pmatrix} A_o \\ z_s \\ z_d \end{pmatrix} \sim \text{MVN}\left[\begin{pmatrix} 0 \\ \mu_s \\ \mu_d \end{pmatrix}, \sigma_z^2 \begin{pmatrix} h^2 & h^2/2 & h^2/2 \\ h^2/2 & 1 & 0 \\ h^2/2 & 0 & 1 \end{pmatrix} \right] \tag{9.31d}$$

Elements that differ from Equation 9.30a are that the expected value of A_o is zero (from the definition of breeding values), and $\sigma^2(A_o) = \sigma_A^2 = h^2\sigma_z^2$. Using the same approach as in Example 9.8 yields

$$A_o = \frac{h^2}{2}(z_s - \mu_s) + \frac{h^2}{2}(z_d - \mu_d) + e, \qquad \sigma_e^2 = \sigma_A^2\left(1 - \frac{h^4}{2}\right) \tag{9.31e}$$

Under more general settings (such as inbred and/or related parents), the covariance matrix in Equation 9.31d is replaced by Equation 9.30d, subject to the minor change that the 1,1 element $(1 + f_0 h^2)$ is replaced by $h^2(1 + f_o)$. Notice that the prediction of breeding values from phenotypic information under very general types of relationships is a function of h^2 and the coefficients of coancestry of the measured (i.e., phenotyped) relatives. This serves as a lead-in to the very general method of BLUP for predicting breeding values given a set of phenotypic values on a known group of relatives (Chapter 30).

MATRIX GEOMETRY

Eigenvalues and Eigenvectors

A vector can be thought of as an arrow, corresponding to a direction in space, as it has a length and an orientation. Similarly, a matrix can be thought of as describing a vector transformation, so that when one multiplies a vector by that matrix, it generates a new

vector. Such a transformation generally has the effect that the resulting vector is both *rotated* in direction and *scaled* (shrinking or expanding its length) relative to the original vector. Hence, the new vector $\mathbf{b} = \mathbf{A}\mathbf{c}$ usually points in a different direction than $\mathbf{c}$ as well as usually having a different length. However, a set of vectors exists for any square matrix that satisfy

$$\mathbf{A}\mathbf{y} = \lambda\mathbf{y} \tag{9.32}$$

Namely, the new vector ($\lambda\mathbf{y}$) is still in the same direction (as multiplying a vector by a constant does not change its orientation, except for reflecting it about the origin when $\lambda < 0$), although it is scaled so that its new length is an amount λ of the original length. Vectors that satisfy Equation 9.32 are called the **eigenvectors** of that matrix, with λ as the corresponding **eigenvalue** for a given eigenvector. For such vectors, the only action of the matrix transformation is to scale their length by some amount, λ. These vectors thus represent the ***inherent axes associated with the vector transformation given by*** $\mathbf{A}$, and the set of all such vectors, along with their corresponding scalar multipliers, completely describes the geometry of this transformation. Note that if $\mathbf{y}$ is an eigenvector, then so is $a\mathbf{y}$, as $\mathbf{A}(a\mathbf{y}) = a(\mathbf{A}\mathbf{y}) = \lambda(a\mathbf{y})$, while the associated eigenvalue, λ, remains unchanged. Hence, we typically scale eigenvectors to be of unit length to yield **unit** or **normalized eigenvectors**, which are typically denoted by $\mathbf{e}$.

The resulting collection of eigenvectors and their associated scaling eigenvalues are called the **eigenstructure** of a matrix. This structure provides powerful insight into the geometric aspects of a matrix, such as the major axes of variation in a covariance matrix. Appendix 3 examines the **singular value decomposition** (**SVD**) that generalizes this concept to nonsquare matrices, while WL Appendix 5 examines eigenstructure in more detail.

Notice that if a matrix has an eigenvalue of zero, then from Equation 9.32, there is some vector that satisfies $\mathbf{A}\mathbf{e} = \mathbf{0}$, namely that one can write at least one of the rows of $\mathbf{A}$ as a linear combination of the other rows. Put another way, if one has a set of n equations, one or more of these equations is redundant, as it is simply a linear combination of other equations, and hence there is no unique solution to the system of equations. Recall that the determinant of a matrix is zero in such cases, and hence we have an important result that ***the determinant of a matrix is zero when it has one (or more) zero eigenvalues***.

The eigenvalues of an n-dimensional square matrix, $\mathbf{A}$, are solutions of Equation 9.32, which can be written as $(\mathbf{A} - \lambda\,\mathbf{I})\mathbf{y} = \mathbf{0}$. This implies that the determinant of $(\mathbf{A} - \lambda\,\mathbf{I})$ must equal zero, which gives rise to the **characteristic equation** for λ,

$$|\mathbf{A} - \lambda\mathbf{I}| = 0 \tag{9.33a}$$

whose solution yields the eigenvalues of $\mathbf{A}$. This equation can be also be expressed using the **Laplace expansion**,

$$|\mathbf{A} - \lambda\mathbf{I}| = (-\lambda)^n + S_1(-\lambda)^{n-1} + \cdots + S_{n-1}(-\lambda)^1 + S_n = 0 \tag{9.33b}$$

where S_i is the sum of all **principal minors** (minors including diagonal elements of the original matrix) of order i. Finding the eigenvalues thus requires solving a polynomial equation of order n, implying that there are exactly n eigenvalues (some of which may be identical, i.e., **repeated**). In practice, for $n > 2$ this is accomplished numerically, and most analysis packages offer routines to accomplish this task.

Two of these principal minors are easily obtained and provide information on the nature of the eigenvalues. The only principal minor having the same order of the matrix is the full matrix itself, which means that $S_n = |\mathbf{A}|$, the determinant of $\mathbf{A}$. S_1 is also related to an important matrix quantity, the trace,

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} A_{ii} \tag{9.34a}$$

Observe that $S_1 = \text{tr}(\mathbf{A})$, as the only principal minors of order one are the diagonal elements themselves, the sum of which equals the trace. Both the trace and determinant can be expressed as functions of the eigenvalues, with

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} \lambda_i \qquad \text{and} \qquad |\mathbf{A}| = \prod_{i=1}^{n} \lambda_i \tag{9.34b}$$

Hence, $\mathbf{A}$ ***is singular*** ($|\mathbf{A}| = 0$) ***if, and only if, at least one eigenvalue is zero***. As we will see, if $\mathbf{A}$ is a covariance matrix, then its trace (the sum of its eigenvalues) measures its total amount of variation, as the eigenvalues of a covariance matrix are nonnegative ($\lambda_i \geq 0$). Another useful result is that ***the diagonal elements in a diagonal matrix correspond to its eigenvalues***.

A final useful concept is the idea of the **rank** of a square matrix, which its number of nonzero eigenvalues. Essentially, this is the amount of information (number of independent equations) that are specified by a matrix. A square $n \times n$ matrix of **full rank** was n nonzero eigenvalues and hence is nonsingular (a unique inverse exists).

Example 9.10. Consider the following matrix,

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}. \qquad \text{Hence,} \quad \det(\mathbf{A} - \lambda\mathbf{I}) = \det\begin{pmatrix} 1-\lambda & 2 \\ 2 & 1-\lambda \end{pmatrix}$$

Using Equation 9.12a to compute the determinant and solving Equation 9.33a yields $(1-\lambda)^2 - 4 = 0$, or the quadratic equation $\lambda^2 - 2\lambda - 3 = 0$, which solutions of 3 and -1. Noting that $S_1 = \text{tr}(\mathbf{A}) = 2$ and $S_2 = \det(\mathbf{A}) = -3$, we also recover this equation using Equation 9.33b, $|\mathbf{A} - \lambda\mathbf{I}| = (-\lambda)^2 + S_1(-\lambda)^1 + S_2 = \lambda^2 - 2\lambda - 3$.

Next, note that the vectors $\mathbf{y}_1^T = (1, 1)$ and $\mathbf{y}_2^T = (1, -1)$ correspond to the eigenvectors of $\mathbf{A}$, as

$$\mathbf{A}\mathbf{y}_1 = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = 3 \cdot \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad \text{and} \quad \mathbf{A}\mathbf{y}_2 = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ -1 \end{pmatrix} = (-1) \cdot \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

Hence, $\mathbf{y}_1$ is an eigenvector associated with $\lambda_1 = 3$, while $\mathbf{y}_2$ is an eigenvector associated with $\lambda_2 = -1$. Recalling that the length of a vector $\mathbf{x}$ is given by $||\mathbf{x}|| = \sqrt{\mathbf{x}^T\mathbf{x}}$ (WL Appendix 5), the lengths of both $\mathbf{y}_1$ and $\mathbf{y}_2$ are $\sqrt{2}$. Hence, the two unit eigenvectors are given by $\mathbf{e}_i = \mathbf{y}_i/\sqrt{2}$. Note that $\mathbf{y}_1^T\mathbf{y}_2 = 1 \cdot 1 + (-1) \cdot 1 = 0$, which implies (WL Appendix 5) that these two vectors are **orthogonal** (at right angles) to each other.

Example 9.11. Insight into the role of zero eigenvalues can be gained by considering two systems of linear equations, where in both cases the corresponding matrix of coefficients is singular. This implies that some of the equations are redundant with each other (i.e., are linear combinations of the others). However, the determinant is a very coarse measure of the *amount* of redundancy, as it simply indicates that *at least one* equation is redundant. The number of zero eigenvalues goes much further, giving the *number* of redundant equations. Consider the following two sets of equations:

$$\text{Set one:} \quad \begin{aligned} x_1 + x_2 + x_1 &= 1 \\ 2x_1 + 2x_2 + 2x_3 &= 2 \\ 3x_1 + 3x_2 + 3x_3 &= 3 \end{aligned} \qquad\qquad \text{Set two:} \quad \begin{aligned} x_1 + x_2 + x_3 &= 1 \\ 2x_1 + 2x_2 + 2x_3 &= 2 \\ x_1 - x_2 - 2x_3 &= 3 \end{aligned}$$

Clearly, set one consists of three redundant equations (all are multiples of each other), while set two contains at least one redundant equation, as the first and second equations are multiples of each other. The corresponding coefficient matrices for these sets become

$$\mathbf{A}_1 = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 3 & 3 & 3 \end{pmatrix}, \qquad \mathbf{A}_2 = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 2 & 2 \\ 1 & -1 & -2 \end{pmatrix}$$

As expected, both of these matrices have determinants of zero. $\mathbf{A}_1$ has eigenvalues of 6, 0, and 0, while the eigenvalues for $\mathbf{A}_2$ are 2.79, -1.79, and 0. Note in both cases that the trace (6 and 1, respectively) equals the sum of the eigenvalues. Because $\mathbf{A}_1$ contains only one nonzero eigenvalue (it has rank 1), only one relationship can be estimated from set one: the solution is the plane defined as all points satisfying $x_1 = 1 - x_2 - x_3$. Set two, by virtue of having two nonzero eigenvalues (it has rank 2), yields two relationships ($x_2 = x_1 + 5$ and $x_3 = 2x_1 - 4$). This importance the number of nonzero eigenvalues will reappear next chapter when we examine the number of fixed effects in a linear model that we can uniquely estimate.

Principal Components of the Variance-covariance Matrix

An important application of eigenstructure is **principal component analysis** (**PCA**), the eigenanalysis of a covariance matrix. Consider a random vector, $\mathbf{x}$, with an associated covariance matrix, $\mathbf{V}$. We are often interested in how the variance of the elements of $\mathbf{x}$ can be decomposed into independent components. For example, even though we may be measuring n variables, only one or two of these may account for the majority of the variation. If this is the case, we may wish to exclude those variables contributing very little variation from further analysis. More generally, if random variables are correlated, then certain linear combinations of the elements of $\mathbf{x}$ may account for most of the variance. PCA extracts these combinations by decomposing the variance of $\mathbf{x}$ into the contributions from a series of orthogonal vectors, the first (PC1) of which explains the most variation possible for any single vector, the second (PC2) the next possible amount, and so on until we account for the entire variance of $\mathbf{x}$. These vectors (or **axes**) correspond to the eigenvectors of $\mathbf{V}$ associated with the largest, second largest, and so on, eigenvalues.

Our starting point for PCA is to note that the eigenvalues of a covariance matrix are never negative, and are all positive if $\mathbf{V}$ is nonsingular. A matrix, $\mathbf{V}$, with all positive eigenvalues is said to be **positive definite**, and this implies that $\mathbf{c}^T\mathbf{V}\mathbf{c} > 0$ for values of $\mathbf{c}$ (other than the trivial case $\mathbf{c} = \mathbf{0}$). Recall (Equation 9.19) that this quadratic product is non-negative as it corresponds to the variance of the linear combination $\mathbf{c}^T\mathbf{x}$. Because all of the eigenvalues of $\mathbf{V}$ are non-negative, their sum represents the **total variance** implicit in the elements of $\mathbf{x}$. From Equation 9.34b, this sum is simply the trace of $\mathbf{V}$, tr($\mathbf{V}$). A matrix $\mathbf{V}$ is said to be **non-negative definite** if $\mathbf{c}^T\mathbf{V}\mathbf{c} \geq 0$, which happens when all its eigenvalues are non-negative, so that a covariance matrix, even if singular, is always non-negative definite, and a nonsingular covariance matrix is always positive definite.

Suppose that $\mathbf{V}$ is an n-dimensional covariance matrix, and we order its eigenvalues from largest to smallest, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, with their associated (unit-length) eigenvectors denoted by $\mathbf{e}_1, \mathbf{e}_2, \cdots, \mathbf{e}_n$, respectively. λ_1 is referred to as the **leading eigenvalue**, with $\mathbf{e}_1$ the **leading eigenvector**. It can be shown (WL Appendix 5) that the maximum variance for any linear combination of the elements of $\mathbf{x}$ ($y = \mathbf{c}_1^T\mathbf{x}$, subject to the constraint that $||\mathbf{c}_1|| = 1$), is

$$\max\left[\sigma^2(y)\right] = \max_{||\mathbf{c}_1||=1}\left[\sigma^2\left(\mathbf{c}_1^T\mathbf{x}\right)\right] = \max_{||\mathbf{c}_1||=1}\left[\mathbf{c}_1^T\mathbf{V}\mathbf{c}_1\right] = \lambda_1$$

which occurs when $\mathbf{c}_1 = \mathbf{e}_1$. This vector is the **first principal component** (often abbreviated as **PC1**), and accounts for a fraction $\lambda_1/\text{tr}(\mathbf{V})$ of the total variation in $\mathbf{x}$. We can partition the remaining variance in $\mathbf{x}$ after the removal of PC1 in a similar fashion. For example, the vector $\mathbf{c}_2$, that is orthogonal to PC1 ($\mathbf{c}_2^T\mathbf{c}_1 = 0$) and maximizes the remaining variance can be shown to be $\mathbf{e}_2$, which accounts for a fraction $\lambda_2/\text{tr}(\mathbf{V})$ of the total variation in $\mathbf{x}$. By proceeding in this fashion, we can see that the ith PC is given by $\mathbf{e}_i$, and that the amount of variation it accounts for is

$$\lambda_i \Big/ \sum_{k=1}^{n}\lambda_k = \frac{\lambda_i}{\text{tr}(\mathbf{V})} \tag{9.35a}$$

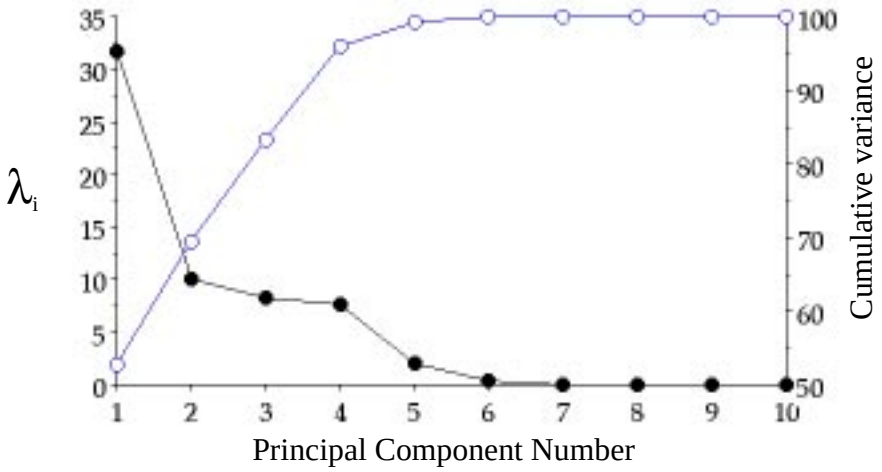

Figure 9.1. A joint scree and cumulative variance plot. The horizontal axis gives the eigenvalue number (λ_i corresponds to PCi), while the vertical axis jointly displays a scree plot (filled circles, the corresponding eigenvalue, λ_i) and the cumulative variance associated with the first i PCs (open circles; Equation 9.35b).

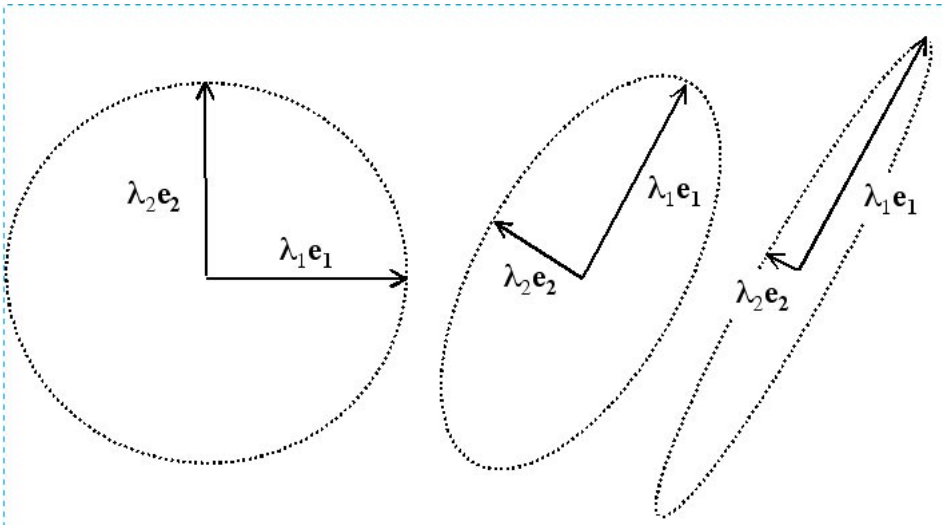

Figure 9.2. The impact of increasing the difference between the major and minor eigenvalues $(\lambda_1 - \lambda_2)$ while their sum is held constant. When the two eigenvalues are equal, the spread of $\mathbf{x}$ values about their mean is spherical, while it becomes increasingly elongated as the difference between the eigenvalues increases. Any tilt in this elongated distribution is generated by correlations between elements of $\mathbf{x}$ (the orientation being given by the direction of the associated eigenvectors, $\mathbf{e}_1$ and $\mathbf{e}_2$).

Put another way, $\lambda_i/\text{tr}(\mathbf{V})$ is the fraction of that total variance explained by the linear combination $\mathbf{e}_i^T\mathbf{x}$. It follows that the fraction of total variation accounted for by the first k PCs is

$$\frac{\lambda_1 + \cdots + \lambda_k}{\text{tr}(\mathbf{V})} \tag{9.35b}$$

A graph of Equation 9.35a as a function of k is called a **cumulative variance plot** (Figure 9.1).

Another useful visual display of the eigenstructure is a **scree plot**, which graphs the

eigenvalues ranked from largest to smallest (Figure 9.1). The term *scree* refers to the loose pile of rocks that comprise the steep slope of a mountain, as most scree plots display a rapid falloff, akin to what one would see in a scree field. Suppose the eigenvalues of $\mathbf{V}$ are roughly similar in magnitude (a relatively flat scree plot). For three dimensions this implies that the distribution of $\mathbf{x}$ is roughly spherical (i.e., a 3D plot of the elements of $\mathbf{x}$ corresponds to a soccer ball) and hence has little structure. As the eigenvalues become increasingly dissimilar, the scree plots starts to show a dramatic falloff in values, and the distribution of values of $\mathbf{x}$ becomes stretched and elongated, generating some axes with larger, and others with smaller, variances. Figure 9.2 shows the impact in two dimension when the trace of a matrix (the sum of its eigenvalues) is held constant, while the difference between λ_1 and λ_2 increases.

A nearly flat scree plot indicates very little structure in $\mathbf{V}$, while a typical scree plot (a rapid decline in eigenvalues) indicates that much of the variance is concentrated in a few directions (or **major axes**). Another way to state this is that a small variance in the eigenvalues, $\sigma^2(\lambda)$, implies little structure (roughly equal variance in all directions) in the structure of $\mathbf{x}$, while the distribution of $\mathbf{x}$ becomes increasing concentrated in a smaller number of directions as the variance in the eigenvalues increases.

Example 9.12 Here we perform PCA for the covariance matrix given in Example 9.7,

$$\mathbf{V} = \begin{pmatrix} 10 & -5 & 10 \\ -5 & 20 & 0 \\ 10 & 0 & 30 \end{pmatrix}$$

The eigenvalues and their associated eigenvectors are found to be $\lambda_1 = 34.41, \lambda_2 = 21.12$, and $\lambda_3 = 4.47$, with

$$\mathbf{e}_1 = \begin{pmatrix} 0.400 \\ -0.139 \\ 0.906 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0.218 \\ -0.948 \\ -0.238 \end{pmatrix}, \quad \mathbf{e}_3 = \begin{pmatrix} 0.892 \\ 0.287 \\ -0.349 \end{pmatrix}$$

Hence, PC1 accounts for $\lambda_1/\text{tr}(\mathbf{V})$ = 34.41/60 = 57% of the total variation of $\mathbf{V}$. There are several ways to interpret PC1. The first is as the direction of the maximal axis of variation (e.g., Figure 9.2). A second is that the weighted index $y_i = \mathbf{e}_1^T\mathbf{z}$ (a new composite variable),

$$y_1 = 0.400z_1 - 0.139z_2 + 0.906z_3$$

accounts for 57% of the total variation by itself.

Similarly, PC2 accounts for $\lambda_2/\text{tr}(\mathbf{V})$ = 21.12/60 = 35% of the total variance, and gives the direction of the most variation orthogonal to PC1 ($\mathbf{e}_i^T\mathbf{e}_2 = 0$). The weighted index corresponding to PC2 is

$$y_2 = 0.218z_1 - 0.948z_2 - 0.238z_3$$

Using the two dimensional vector $\mathbf{y}^T = (y_1, y_2)$ accounts for (34.31+21.12)/60 or 93% of the variation of $\mathbf{x}$. This illustrates one use of PCA, which is for **dimensional reduction**, extracting a set of weighted indices of much lower dimension than the original vector as a proxy for the variation in $\mathbf{x}$.

Example 9.13 As we will see in Chapter 20, an important application of PCA is in controlling for population structure in genome-wide association studies (GWAS). The basic idea of a GWAS is to search for marker-trait associations using densely-packed SNPs. For the kth SNP marker (assumed to be biallelic), individuals are grouped into three genotype classes (M_kM_k, M_km_k, and m_km_k), trait means are computed for each class, and a marker-effect examined

using ANOVA (i.e., an among-group difference in trait means). One simple linear model to test for an effect from SNP k would be the regression

$$z = \mu + \beta_k n_k + e \tag{9.36a}$$

where z is the trait value, μ is the population mean, n_k denotes the number of copies of allele M_k in an indvidual (with values of 0, 1, or 2), and $2\beta_k$ is the difference in average trait value between the two different SNPr homozygotes. A significant value of β_k indicates a marker-trait association.

Marker-trait associations can arise from linkage disequilibrium (LD) between the marker and a very closely-linked QTL (Chapter 5). However, they can also arise from population structure. Suppose our GWAS sample, unbeknownst to the investigator, consists of two populations, with population one trending to be taller than population two. Further, because of population structure, some marker allele frequencies differ between the populations. A marker that is predictive of group membership (e.g., an allele very common in population one but very rare in population two) will show a marker-trait association even when it is unlinked to any QTLs for height.

This complication from population structure arises when subpopulations in the sample differ in mean trait value (the subpopulation mean $\mu^* \neq \mu$). If one could first adjust for any subpopulation-specific differences, then any remaining marker-trait associations are likely due to LD with nearby QTLs. In a typical GWAS, a very large number of markers are scored, and these provide information on any population structure. To adjust for structure, the investigator first constructs a marker covariance matrix using markers *outside* of those on the chromosome being tested. A PCA analysis of this covariance matrix would look for the presence of structure by examining either a scree or a cumulative variance plot, and choose the first p PCs by some criteria. Let $\mathbf{m}_i$ denote the marker vector for individual i for these scored markers, for example with the value for element j being 0, 1, or 2, depending on the number of copies of allele M_j in individual i. The idea is to predict the mean trait value in a subpopulation (μ^*) by regression on these PCs,

$$z = \mu + \sum_{\ell=1}^{p} \gamma_\ell z_{\ell,i} + \beta_k n_k + e \tag{9.36b}$$

$$= \mu^* + \beta_k n_k + e \tag{9.36c}$$

Here, $z_{\ell,i} = \mathbf{e}_\ell^T \mathbf{m}_i$ is the value for individual i in the index of marker information given by PC ℓ. The γ_ℓ are the best fit predictors (partial regression coefficients) of how PC ℓ influences the overall mean (which are fit along with μ and β_k by least-squares). Hence, $\mu^* = \mu + \sum \gamma_\ell z_{\ell,i}$ is the predicted mean given the population from which individual i is drawn, leaving $\beta_k n_k$ as any residual effect from marker k (which was *not* used in the population structure correction).

Example 9.14 A related issue to the population structure correction in a GWAS is accessing the amount of shared relatedness among all pairwise combinations over a collection of n individuals (Chapter 20). This is done by constructing an $n \times n$ matrix whose ijth element is an estimate of $2\Theta_{ij}$, twice the coefficient of coancestry (Chapters 7 and 8). Note that 2Θ is the expected fraction of the genome shared by two individuals and is also the coefficient on the amount of additive genetic variance that contributes to their phenotypic correlation (Equation 7.11a). When estimated from pedigree data alone, this is called the **numerator relationship matrix**, and is denoted by $\mathbf{A}$. When estimated solely from marker data (Chapter 8), this is usually denoted by $\mathbf{G}$ and called the **genomic relationship matrix** (Chapters 30 and 31). One delicate issue with marker data is that there is no guarantee the matrix constructed from all pairwise estimators will be non-negative definite (and hence a proper covariance matrix). Ways to both ensure this, and also calculate $\mathbf{G}$ in a single matrix operation, were provided by VanRaden (2007, 2008).

The starting point is the **marker information matrix**, $\mathbf{M}$. For a set of m markers scored over n individuals, $\mathbf{M}$ is $n \times m$, with the ith row corresponding to the marker genotypes for individual i, while the jth column shows the genotypes for marker j over all scored individuals. At each SNP, we count the number of copies of the so-called reference allele (one

allele at the SNP is set to value 1, the alternative to value 0; Chapter 8), with M_{ij} denoting the number of reference alleles at locus i in individual j, which takes on values of 0, 1, and 2 for, respectively, SNP genotypes at locus i of 00, 10, and 11 (VanRanden 2007). Alternatively, SNP data is often coded by subtracting one from each category, giving scores of $-1, 0$, and 1 (VanRanden 2008). As a toy example, suppose that four SNPs are scored in each of three individuals, resulting in the 3×4 matrix (using the VanRanden 2007 coding)

$$\mathbf{M} = \begin{pmatrix} 1 & 0 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ -1 & 0 & 0 & 0 \end{pmatrix}$$

This indicates that individual one has genotypes of 11, 10, 11, and 11 at the four markers, individual two has genotypes of 11, 00, 11, and 00, and individual three has genotypes of 00, 10, 10, and 10. The 3×3 and 4×4 matrices, $\mathbf{M}\mathbf{M}^T$ and $\mathbf{M}^T\mathbf{M}$ respectively, become

$$\mathbf{M}\mathbf{M}^T = \begin{pmatrix} 3 & 1 & -1 \\ 1 & 4 & -1 \\ -1 & -1 & 1 \end{pmatrix} \qquad \mathbf{M}^T\mathbf{M} = \begin{pmatrix} 3 & -1 & 2 & 0 \\ -1 & 1 & -1 & 1 \\ 2 & -1 & 2 & 0 \\ 0 & 1 & 0 & 2 \end{pmatrix}$$

$\mathbf{M}_{n\times m}\mathbf{M}^T_{m\times n}$ corresponds to the covariance matrix for marker scores (genotypes) among individuals, and as will be shortly demonstrated, can be used to construct $\mathbf{G}$. Diagonal elements correspond to the number of homozygotes for individuals, showing that individuals one, two, and three have, respectively, 3, 4, and 1 homozygotes among their four scored markers. Off diagonal elements correspond to the similarity (in marker genotypes) of the two individuals. Individuals one and two are more similar accross the markers, while one and three, as well as two and three, are more dissimilar (negative values). More generally, the dimensionality of this square matrix is the number of individuals, n.

Conversely, the $m \times m$ matrix, $\mathbf{M}^T_{m\times n}\mathbf{M}_{n\times m}$, quantifies the covariance over markers, with negative values showing indicating negative associations between alleles at two different markers and positive value indicating positive associations. This is the matrix from which PCs are extracted to correct for population structure (Example 9.13). The dimensionality of this square matrix is the number of markers, m, and often $m \gg n$, so that this matrix is expected to be singular (the maximum number of positive eigenvalues is the smaller of n and m),

VanRaden (2007, 2008) proposed two methods (Equations 8.15b and 8.16b) for using $\mathbf{M}$ to generate an estimate of $\mathbf{G}$. Define the allele-frequency matrix $\mathbf{P}$ whose ith column is given by $2p_i\mathbf{1}$ when the VanRanden (2007) marker coding scheme (0, 1, 2) is used. Namely, a vector whose values are all $2p_i$, the frequency of the reference allele at locus i. When markers are scored by the VanRandem (2008) coding $(-1, 0, 1)$, then the ith column is given by $2(p_i - 0.5)\mathbf{1}$. Finally, define $\mathbf{Z} = \mathbf{M} - \mathbf{P}$. With these definitions, Van Randen's first method is

$$\widehat{\mathbf{G}}_{VR1} = \frac{\mathbf{Z}\mathbf{Z}^T}{2\sum_{i=1}^{m} p_i(1-p_i)} \tag{9.37a}$$

When using the VanRaden (2007) coding, this recovers Equation 8.16a in matrix form. This estimate weights all marker loci equally.

VanRanden's second method weights the information from each locus, by defining the diagonal matrix $\mathbf{D}$ whose iith element is given by

$$D_{ii} = \frac{1}{m\,2\,p_i(1-p_i)} \tag{9.37b}$$

with

$$\widehat{\mathbf{G}}_{VR2} = \mathbf{Z}\mathbf{D}\mathbf{Z}^T \tag{9.37c}$$

which, when using VanRanden's (2007) coding recovers Equation 8.15b. This estimator places more weight on loci with rare alleles.

To demonstrate that Equations 9.37a and 9.37c generate non-negative definite matrices, we need to show that $\mathbf{c}^T\mathbf{G}\mathbf{c} \geq 0$ for all vectors $\mathbf{c}$. Ignoring the positive constant in the denominator of Equation 9.37a, we have

$$\mathbf{c}^T\widehat{\mathbf{G}}_{VR1}\,\mathbf{c} = \mathbf{c}^T\mathbf{Z}\mathbf{Z}^T\mathbf{c} = \mathbf{y}^T\mathbf{y} = \sum y_i^2 \geq 0$$

with the vector $\mathbf{y} = \mathbf{Z}^T\mathbf{c}$. Similarly,

$$\mathbf{c}^T\widehat{\mathbf{G}}_{VR2}\,\mathbf{c} = \mathbf{c}^T\mathbf{Z}\mathbf{D}\mathbf{Z}^T\mathbf{c} = \mathbf{z}^T\mathbf{z} = \sum z_i^2 \geq 0$$

where the vector $\mathbf{z} = \mathbf{D}^{1/2}\,\mathbf{Z}^T\mathbf{c}$, with the square root matrix defined by Equation A3.7a.

Literature Cited

Anderson, T. W. 2003. *An introduction to multivariate statistical analysis.* 3rd Ed. John Wiley & Sons, New York, NY. [9]

Johnson, R. A., and D. W. Wichern. 2002. *Applied multivariate statistical analysis.* 5th Ed. Prentice-Hall, Upper Saddle River, NJ. [9]

Lande, R., and S. J. Arnold. 1983. The measurement of selection on correlated characters. *Evolution* 37: 1210–1226. [9]

Morrison, D. F. 1976. *Multivariate statistical methods.* McGraw-Hill, New York, NY. [9]

Pearson, K. 1896. Contributions to the mathematical theory of evolution. III. Regression, heredity and panmixia. *Phil. Trans. Royal Soc. Lond.* A 187: 253–318. [9]

Pearson, K. 1903. Mathematical contributions to the theory of evolution. XI. On the influence of natural selection on the variability and correlation of organs. *Phil. Trans. Royal Soc. Lond.* A 200: 1–66. [9]

Pearson, K. 1920. Notes on the history of correlation. *Biometrika* 13: 25–45. [9]

Searle, S. R. 1971. *Linear models*. John Wiley & Sons, New York, NY. [9]

Stigler, S. M. 1986. *The history of statistics.* Harvard Univ. Press, Cambridge, MA. [9]

VanRaden, P. M. 2007. Genomic measures of relationship and inbreeding. *INTERBULL Bull.* 37: 33–36. [9]

VanRaden, P. M. 2008. Efficient methods to compute genomic predictions. *J. Dairy Sci.* 91: 4414–4423. [9]