

# Appendix 5

## The Geometry of Vectors and Matrices: Eigenvalues and Eigenvectors

*Much of the presentation that follows is in matrix notation, and for this I offer no apology as this has rapidly become an essential tool of any serious student of animal breeding.*  
Henderson (1973)

The basic concepts of matrix algebra were introduced in LW Chapter 8 and LW Appendix A3, and we assume the reader has this level of understanding (which includes matrix multiplication, inverses, and determinants). If not, a quick review of LW Chapter 8 before proceeding will be helpful. A deeper understanding of multivariate issues in quantitative genetics requires an appreciation of matrix *geometry*. Our primary intent here is to introduce the reader to the idea of vectors and matrices as **geometric structures**, and thus viewing matrix operations as transformations converting one vector into another by a change in geometry (rotation and scaling), which is completely summarized by the **eigenvalues** (scaling), and their associated **eigenvectors** (rotation), of a matrix.

### THE GEOMETRY OF VECTORS AND MATRICES

As there are numerous excellent texts on matrix algebra, we made little effort to prove most of the results given below. For statistical applications, concise introductions can be found in the chapters on matrix methods in Johnson and Wichern (1988) and Morrison (1976), while Dhrymes (1978) and Searle (1982) provided more extended treatments. Wilf's (1978) short chapter on matrix methods provides a very nifty review of methods useful in applied mathematics. Franklin (1968), Horn and Johnson (1985), and Gantmacher (1960), respectively, presented increasingly sophisticated treatments of matrix analysis.

#### Comparing Vectors: Lengths and Angles

As Figure A5.1A shows, a vector,  $\mathbf{x}$ , can be treated as a geometric object, consisting of an arrow leading from the origin to an  $n$ -dimensional point whose coordinates are given by the elements of  $\mathbf{x}$ . By changing coordinate systems, we change the resulting vector, potentially changing both its direction (**rotating** the vector) and length (**scaling** the vector). This geometric interpretation suggests several ways for comparing vectors, such as the **angle** between two vectors and the **projection** of one vector onto another.

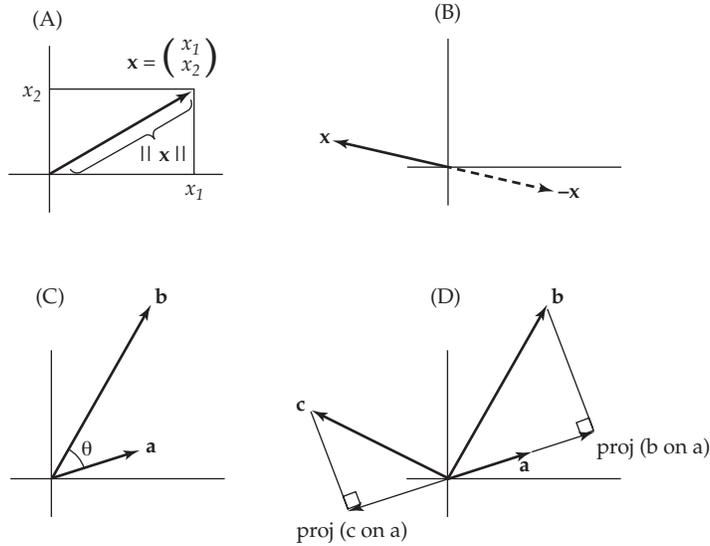
Consider first the length (or **norm**) of a vector. The most common measure of length is the Euclidean distance of the vector from the origin,  $\|\mathbf{x}\|$ , defined as

$$\|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} = \sqrt{\mathbf{x}^T \mathbf{x}} \quad (\text{A5.1a})$$

For any scalar  $a$ ,  $\|a\mathbf{x}\| = |a| \|\mathbf{x}\|$ . Similarly, the squared Euclidean distance between the vectors  $\mathbf{x}$  and  $\mathbf{y}$  is

$$\|\mathbf{x} - \mathbf{y}\|^2 = \sum_{i=1}^n (x_i - y_i)^2 = (\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y}) = (\mathbf{y} - \mathbf{x})^T (\mathbf{y} - \mathbf{x}) \quad (\text{A5.1b})$$

Vectors can differ by length, direction, or both. The angle,  $\theta$ , between two vectors ( $\mathbf{x}$  and  $\mathbf{y}$ ) provides a measure of how much they differ in direction (Figure A5.1C). If the vectors



**Figure A5.1** Some basic geometric concepts of vectors. While we use examples from two dimensions, these concepts easily extend to  $n$  dimensions. **A:** A vector  $\mathbf{x}$  can be thought of as an arrow from the origin to a point in space whose coordinates are given by the elements of  $\mathbf{x}$ . **B:** Multiplying a vector by  $-1$  results in a *reflection* about the origin. **C:** One measure of the difference in direction between two vectors is the angle ( $\theta$ ) between them. **D:**  $\text{Proj}(\mathbf{b}$  on  $\mathbf{a}$ ) is the vector resulting from the projection of  $\mathbf{b}$  onto  $\mathbf{a}$ . Note that the resulting projection vector is either in the same direction as  $\mathbf{a}$  or in the direction of the reflection of  $\mathbf{a}$ , as seen for  $\text{Proj}(\mathbf{c}$  on  $\mathbf{a}$ ).

satisfy  $a\mathbf{x} = \mathbf{y}$ , they both point in exactly the same direction ( $\theta = 0$ ; they are **codirectional**) when  $a > 0$ . If  $a < 0$ , they are exactly 180 degrees apart and differ in direction only by a **reflection** about the origin (Figure A5.1B). At the other extreme, two vectors can be at right angles to each other ( $\theta = 90^\circ$  or  $270^\circ$ ), in which case they are said to be **orthogonal**. Orthogonal vectors of unit length are further said to be **orthonormal**. For any two  $n$ -dimensional vectors,  $\theta$  satisfies

$$\cos(\theta) = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|} = \frac{\mathbf{y}^T \mathbf{x}}{\|\mathbf{x}\| \|\mathbf{y}\|} \tag{A5.2a}$$

Hence,

$$\theta = \cos^{-1} \left( \frac{\mathbf{y}^T \mathbf{x}}{\|\mathbf{x}\| \|\mathbf{y}\|} \right) \tag{A5.2b}$$

If both  $\mathbf{x}$  and  $\mathbf{y}$  are of unit length, then  $\theta = \cos^{-1}(\mathbf{y}^T \mathbf{x})$ , which reveals the close connection between vector angles and inner products. Note that because  $\cos(90^\circ) = \cos(270^\circ) = 0$ , **two vectors are orthogonal if, and only if, their inner product is zero**,  $\mathbf{x}^T \mathbf{y} = 0$ .

Another way to compare two vectors is to consider the **projection** vector of one onto the other.  $\text{Proj}(\mathbf{x}$  on  $\mathbf{y})$ , the projection of  $\mathbf{x}$  on  $\mathbf{y}$ , is a vector in the direction of  $\mathbf{y}$ , whose length is given by how much of the vector  $\mathbf{x}$  lies along the direction of  $\mathbf{y}$ . For any two  $n$ -dimensional vectors, the projection of  $\mathbf{x}$  on  $\mathbf{y}$  is defined by

$$\text{Proj}(\mathbf{x}$$
 on  $\mathbf{y}) = \frac{\mathbf{x}^T \mathbf{y}}{\mathbf{y}^T \mathbf{y}} \mathbf{y} = \frac{\mathbf{x}^T \mathbf{y}}{\|\mathbf{y}\|^2} \mathbf{y} = \left( \cos(\theta) \frac{\|\mathbf{x}\|}{\|\mathbf{y}\|} \right) \mathbf{y} \tag{A5.3a}$

The term in the parentheses (which follows from Equation A5.2a) is a scalar, representing the length that  $\mathbf{x}$  projects in the direction of  $\mathbf{y}$ , which means that  $\text{Proj}(\mathbf{x}$  on  $\mathbf{y})$  is a scaled version of the vector  $\mathbf{y}$  onto which we are projecting. If  $\|\mathbf{y}\| = 1$ , then

$$\text{Proj}(\mathbf{x}$$
 on  $\mathbf{y}) = (\mathbf{x}^T \mathbf{y}) \mathbf{y} = (\cos(\theta) \|\mathbf{x}\|) \mathbf{y} \tag{A5.3b}$

The vector resulting from the projection of  $\mathbf{x}$  on  $\mathbf{y}$  is in the same direction as  $\mathbf{y}$  unless  $90^\circ < \theta < 270^\circ$ , in which case  $\cos(\theta) < 0$  and the projection vector is in exactly the opposite direction (the reflection of  $\mathbf{y}$  about the origin). The length of the projection vector is

$$\|\text{Proj}(\mathbf{x} \text{ on } \mathbf{y})\| = |\cos(\theta)| \|\mathbf{x}\| \leq \|\mathbf{x}\| \tag{A5.3c}$$

If two vectors lie in exactly the same direction ( $\theta = 0$ ), the projection of one on the other simply recovers the vector (i.e.,  $\text{Proj}(\mathbf{x} \text{ on } \mathbf{y}) = \mathbf{x}$ ). Conversely, if two vectors are orthogonal, the projection of one on the other yields a vector of length zero.

An important property of projection vectors is that if  $\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n$  is any set of mutually orthogonal  $n$ -dimensional vectors, then any  $n$ -dimensional vector  $\mathbf{x}$  can be represented as the sum of projections of  $\mathbf{x}$  onto the members of this set, namely,

$$\mathbf{x} = \sum_{i=1}^n \text{Proj}(\mathbf{x} \text{ on } \mathbf{y}_i) \tag{A5.4}$$

One way to think about such a decomposition is as the transformation from one set of axes (or coordinates) into another (defined by the vectors,  $\mathbf{y}_i$ , that **span**, or completely cover, the vector space). We can also consider the projection of a vector into some **subspace** of a matrix (say  $\mathbf{y}_1, \dots, \mathbf{y}_k$ , where  $k < n$ ), namely, the projection onto some subset of the vectors that span the space of the original matrix. For example, one might consider the subspace of a covariance matrix imposed by (say) its three largest factors (eigenvalues). The notion of a subspace of the genetic covariance matrix  $\mathbf{G}$  will prove useful in describing the constraints caused by the genetic covariance structure (Volume 3).

**Matrices Describe Vector Transformations**

When we multiply a vector,  $\mathbf{x}$ , by a matrix,  $\mathbf{A}$ , to create a new vector,  $\mathbf{y} = \mathbf{A}\mathbf{x}$ ,  $\mathbf{A}$  *rotates* and *scales* the original vector,  $\mathbf{x}$ , into the new vector,  $\mathbf{y}$ .  $\mathbf{A}$  therefore describes a *transformation* of the original coordinate system of  $\mathbf{x}$  into a new coordinate system,  $\mathbf{y}$  (which has a different dimension from  $\mathbf{x}$  unless  $\mathbf{A}$  is square).

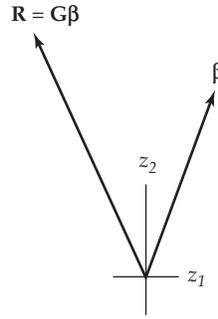
**Example A5.1.** Consider the Lande version of the multivariate breeder’s equation,  $\mathbf{R} = \mathbf{G}\boldsymbol{\beta}$  (Equation 13.26a). Here  $\mathbf{R}$  is the change in the vector of phenotypic means resulting from selection,  $\mathbf{G}$  is the covariance matrix of additive-genetic values (breeding values) of the characters, and  $\boldsymbol{\beta}$  is the directional selection gradient (the direction of change in character means that results in the greatest increase in mean population fitness; Chapters 13 and 30). Suppose

$$\mathbf{G} = \begin{pmatrix} 4 & -2 \\ -2 & 2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{\beta} = \begin{pmatrix} 1 \\ 3 \end{pmatrix}, \quad \text{yielding} \quad \mathbf{R} = \mathbf{G}\boldsymbol{\beta} = \begin{pmatrix} -2 \\ 4 \end{pmatrix}$$

The resulting direction of change in character means is different from that most favored by natural selection. Selection ( $\boldsymbol{\beta}$ ) favors an increase in trait one ( $z_1$ ), but when the genetic covariance structure is taken into account, the resulting change in the mean of  $z_1$  is negative. If we take the appropriate inner products, we find  $\|\boldsymbol{\beta}\| = \sqrt{10}$ ,  $\|\mathbf{R}\| = \sqrt{20}$ , and  $\boldsymbol{\beta}^T \mathbf{R} = 10$ . Equation A5.2a returns

$$\cos(\theta) = \frac{\boldsymbol{\beta}^T \mathbf{R}}{\|\mathbf{R}\| \|\boldsymbol{\beta}\|} = \frac{1}{\sqrt{2}}$$

The resulting angle between the selection gradient and response vector is  $\cos^{-1}(1/\sqrt{2}) = 45^\circ$ , implying that the constraints introduced by the genetic covariance matrix rotate the response vector considerably away from the direction most favored by natural selection (Figure A5.2).



**Figure A5.2** If we use the values of  $\beta$  and  $G$  from Example A5.1, observe that  $G$  translates the directional selection gradient vector ( $\beta$ ) into the response vector ( $R$ ) in a counterintuitive fashion. While  $\beta$  shows that fitness is maximized by increasing both traits 1 and 2, the resulting response vector,  $R$ , increases trait 2 but *decreases* trait 1. This behavior results from the strong negative additive-genetic covariance between  $z_1$  and  $z_2$ , as will become more obvious shortly, when we consider the eigenvectors of  $G$  (Figure A5.3). As shown in Example A5.1, the angle between the vectors  $\beta$  and  $R$  is 45 degrees.

**Orthonormal Matrices: Rigid Rotations**

A key building block on our way to the partitioning of a matrix into its rotational and scaling components is the idea of an **orthonormal matrix**. Writing a square  $n \times n$  matrix,  $U$ , as a row vector whose  $n$  elements are  $1 \times n$  column vectors,  $U = (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n)$ , then  $U$  is said to be orthonormal if

$$\mathbf{u}_i^T \mathbf{u}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Namely, each column of  $U$  is of unit length and is orthogonal to every other column. Matrices with this property are also referred to as **unitary** and satisfy

$$U^T U = U U^T = I \tag{A5.5a}$$

As a result, the inverse of a unitary matrix is simply its transpose,

$$U^T = U^{-1} \tag{A5.5b}$$

The coordinate transformation induced by an orthonormal matrix has a very simple geometric interpretation: it is a **rigid rotation** of the original coordinate system—axes of the original coordinates are all rotated by the same angle to create the new coordinate system. To see this, first note that *orthonormal matrices preserve all inner products*. Taking  $\mathbf{y}_1 = U\mathbf{x}_1$  and  $\mathbf{y}_2 = U\mathbf{x}_2$

$$\mathbf{y}_1^T \mathbf{y}_2 = \mathbf{x}_1^T (U^T U) \mathbf{x}_2 = \mathbf{x}_1^T \mathbf{x}_2 \tag{A5.5c}$$

Thus, orthonormal matrices do not change (scale) the length of vectors, as  $\|\mathbf{y}_1\| = \mathbf{y}_1^T \mathbf{y}_1 = \mathbf{x}_1^T \mathbf{x}_1 = \|\mathbf{x}_1\|$ . Using these results, note that if  $\theta$  is the angle between the vectors  $\mathbf{x}_1$  and  $\mathbf{x}_2$ , then following transformation by an orthonormal matrix

$$\cos(\theta | \mathbf{y}_1, \mathbf{y}_2) = \frac{\mathbf{y}_1^T \mathbf{y}_2}{\sqrt{\|\mathbf{y}_1\|} \sqrt{\|\mathbf{y}_2\|}} = \frac{\mathbf{x}_1^T \mathbf{x}_2}{\sqrt{\|\mathbf{x}_1\|} \sqrt{\|\mathbf{x}_2\|}} = \cos(\theta | \mathbf{x}_1, \mathbf{x}_2) \tag{A5.5d}$$

which shows that the angle between the two vectors remains unchanged following their transformation by the same orthonormal matrix.

**Eigenvalues and Eigenvectors**

The eigenvalues, and their associated eigenvectors, of a square matrix describe its transformational geometry. Eigenvalues describe how the original coordinate axes are *scaled* in the

new coordinate system that is described by the eigenvectors (i.e., how the original axes are rotated).

To more formally introduce eigenvalues and eigenvectors, suppose, for a square matrix  $\mathbf{A}$ , that the vector  $\mathbf{y}$  satisfies the matrix equation

$$\mathbf{A}\mathbf{y} = \lambda\mathbf{y} \tag{A5.6}$$

for some scalar value,  $\lambda$ . Geometrically, this means that the new vector resulting from transformation of  $\mathbf{y}$  by  $\mathbf{A}$  points in the same direction as  $\mathbf{y}$  (or is exactly reflected about the origin if  $\lambda < 0$ ). For such vectors, the only action of the matrix transformation is to scale them by some amount,  $\lambda$ . These vectors thus represent the *inherent axes associated with the 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. Vectors that satisfy Equation A5.6 are referred to as **eigenvectors**, and their associated scaling factors are **eigenvalues**, and together they jointly describe the **eigenstructure** (the intrinsic geometry) of the square matrix,  $\mathbf{A}$ . If  $\mathbf{y}$  is an eigenvector, then  $a\mathbf{y}$  is also an eigenvector, as  $\mathbf{A}(a\mathbf{y}) = a(\mathbf{A}\mathbf{y}) = \lambda(a\mathbf{y})$ . Note, however, that the associated eigenvalue,  $\lambda$ , remains unchanged. Hence, we typically scale eigenvectors to be of unit length to yield **unit** or **normalized** eigenvectors. In particular, if  $\mathbf{y}_i$  is any eigenvector associated with the  $i$ th eigenvalue, then the associated normalized eigenvector is  $\mathbf{e}_i = \mathbf{y}_i / \|\mathbf{y}_i\|$ .

The eigenvalues of an  $n$ -dimensional square matrix,  $\mathbf{A}$ , are solutions of Equation A5.6, 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**,  $|\mathbf{A} - \lambda\mathbf{I}| = 0$ , whose solution yields the eigenvalues of  $\mathbf{A}$ . This equation can be also be expressed using the **Laplace expansion**,

$$|\mathbf{A} - \lambda\mathbf{I}| = (-\lambda)^n + S_1(-\lambda)^{n-1} + \dots + S_{n-1}(-\lambda)^1 + S_n = 0 \tag{A5.7}$$

where  $|\mathbf{A}|$  denotes the determinant of  $\mathbf{A}$  and  $S_i$  is the sum of all **principal minors** (minors including diagonal elements of the original matrix) of order  $i$  (minors, which are subsets of the full matrix, were defined in LW Chapter 8). 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 statistical 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**. This is denoted by  $\text{tr}(\mathbf{A})$ , and is the sum of the diagonal elements of the matrix, namely,

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

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

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^n \lambda_i \quad \text{and} \quad |\mathbf{A}| = \prod_{i=1}^n \lambda_i \tag{A5.8}$$

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 ( $\lambda_i \geq 0$ ).

Let  $\mathbf{e}_i$  be the (unit-length) eigenvector associated with eigenvalue  $\lambda_i$ . If the eigenvectors of  $\mathbf{A}$  can be chosen to be mutually orthogonal, namely,  $\mathbf{e}_i^T \mathbf{e}_j = 0$  for  $i \neq j$ , then we can express  $\mathbf{A}$  as

$$\mathbf{A} = \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^T + \dots + \lambda_n \mathbf{e}_n \mathbf{e}_n^T \tag{A5.9a}$$

This is called the **spectral decomposition** of  $\mathbf{A}$ , and it is derived below in Equation A5.10d. Because  $\|\mathbf{e}_i\| = 1$ , Equation A5.3b gives the projection of  $\mathbf{x}$  on  $\mathbf{e}_i$  as  $(\mathbf{x}^T \mathbf{e}_i) \mathbf{e}_i$ . Note that  $\mathbf{e}_i (\mathbf{e}_i^T \mathbf{x}) = (\mathbf{e}_i^T \mathbf{x}) \mathbf{e}_i = (\mathbf{x}^T \mathbf{e}_i) \mathbf{e}_i$ , as  $\mathbf{e}_i^T \mathbf{x}$  is a scalar, which implies that  $\mathbf{e}_i^T \mathbf{x} = (\mathbf{e}_i^T \mathbf{x})^T = \mathbf{x}^T \mathbf{e}_i$ . Hence, from Equation A5.3b, we have

$$\begin{aligned} \mathbf{Ax} &= \lambda_1 \mathbf{e}_1 \mathbf{e}_1^T \mathbf{x} + \lambda_2 \mathbf{e}_2 \mathbf{e}_2^T \mathbf{x} + \cdots + \lambda_n \mathbf{e}_n \mathbf{e}_n^T \mathbf{x} \\ &= \lambda_1 (\mathbf{e}_1^T \mathbf{x}) \mathbf{e}_1 + \lambda_2 (\mathbf{e}_2^T \mathbf{x}) \mathbf{e}_2 + \cdots + \lambda_n (\mathbf{e}_n^T \mathbf{x}) \mathbf{e}_n \\ &= \lambda_1 \text{Proj}(\mathbf{x} \text{ on } \mathbf{e}_1) + \lambda_2 \text{Proj}(\mathbf{x} \text{ on } \mathbf{e}_2) + \cdots + \lambda_n \text{Proj}(\mathbf{x} \text{ on } \mathbf{e}_n) \end{aligned} \tag{A5.9b}$$

If we again apply Equation A5.3b, we can express this decomposition as

$$\mathbf{Ax} = \|\mathbf{x}\| \sum_{i=1}^n [\lambda_i \cdot \cos(\theta|\mathbf{x}, \mathbf{e}_i)] \mathbf{e}_i \tag{A5.9c}$$

where  $\theta|\mathbf{x}, \mathbf{e}_i$  denotes the angle between the vectors  $\mathbf{x}$  and  $\mathbf{e}_i$ . Thus, one can view a matrix as a series of vectors that form the **projection space** (the eigenvectors), so when a vector is multiplied by this matrix, the resulting vector is the weighted (by the eigenvalues) sum of projections over all of the vectors (the  $\mathbf{e}_i$ ) that span the space defined by the matrix.

**Example A5.2.** Determine the eigenstructure of the genetic covariance matrix  $\mathbf{G}$  shown in Example A5.1. Writing the characteristic equation, and recalling the expression for the determinant of a  $2 \times 2$  matrix (LW Equation 8.12a), yields

$$\begin{aligned} |\mathbf{G} - \lambda \mathbf{I}| &= \left| \begin{pmatrix} 4 - \lambda & -2 \\ -2 & 2 - \lambda \end{pmatrix} \right| \\ &= (4 - \lambda)(2 - \lambda) - (-2)^2 = \lambda^2 - 6\lambda + 4 = 0 \end{aligned}$$

Alternatively, if we use the Laplace expansion (Equation A5.7), and note that  $\text{tr}(\mathbf{G}) = 4+2 = 6$  and  $|\mathbf{G}| = 4 \cdot 2 - (-2)^2 = 4$ , we will also recover the characteristic equation, which has solutions

$$\lambda_1 = 3 + \sqrt{5} \simeq 5.236 \quad \lambda_2 = 3 - \sqrt{5} \simeq 0.764$$

The associated unit eigenvectors (which as easily obtained, along with the eigenvectors, by using the R command `eigen`) are

$$\mathbf{e}_1 \simeq \begin{pmatrix} -0.851 \\ 0.526 \end{pmatrix} \quad \mathbf{e}_2 \simeq \begin{pmatrix} 0.526 \\ 0.851 \end{pmatrix}$$

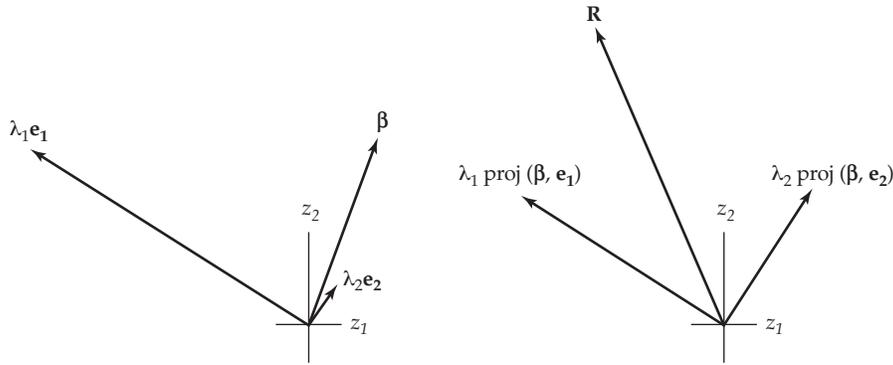
These are orthogonal as  $\mathbf{e}_1^T \mathbf{e}_2 = 0$ .

The eigenstructure of  $\mathbf{G}$  shows why the vector of responses,  $\mathbf{R}$ , is rotated away from the direction of the vector that corresponds to the direction of selection,  $\boldsymbol{\beta}$ . From Example A5.1,  $\|\boldsymbol{\beta}\| = \sqrt{10}$ , while  $\mathbf{e}_1^T \boldsymbol{\beta} \simeq 0.727$  and  $\mathbf{e}_2^T \boldsymbol{\beta} \simeq 3.079$ . Because  $\|\mathbf{e}_1\| = \|\mathbf{e}_2\| = 1$ , Equation A5.2a simplifies to

$$\cos(\theta|\mathbf{e}_1, \boldsymbol{\beta}) \simeq \frac{0.727}{\sqrt{10}} \simeq 0.230 \quad \text{and} \quad \cos(\theta|\mathbf{e}_2, \boldsymbol{\beta}) \simeq \frac{3.079}{\sqrt{10}} \simeq 0.974$$

giving the angle between  $\mathbf{e}_1$  and  $\boldsymbol{\beta}$  as  $\theta(\mathbf{e}_1, \boldsymbol{\beta}) \simeq 76.7^\circ$ , while  $\theta(\mathbf{e}_2, \boldsymbol{\beta}) \simeq 13.2^\circ$ . Applying Equation A5.3b, the corresponding scaled projections of  $\boldsymbol{\beta}$  on these eigenvectors are

$$\begin{aligned} \lambda_1 \text{Proj}(\boldsymbol{\beta} \text{ on } \mathbf{e}_1) &= \lambda_1 \cos(\theta|\mathbf{e}_1, \boldsymbol{\beta}) \|\boldsymbol{\beta}\| \mathbf{e}_1 = \left( 5.236 \cdot 0.230 \cdot \sqrt{10} \right) \mathbf{e}_1 \\ &= 3.803 \begin{pmatrix} -0.851 \\ 0.526 \end{pmatrix} = \begin{pmatrix} -3.236 \\ 2 \end{pmatrix} \\ \lambda_2 \text{Proj}(\boldsymbol{\beta} \text{ on } \mathbf{e}_2) &= \lambda_2 \cos(\theta|\mathbf{e}_2, \boldsymbol{\beta}) \|\boldsymbol{\beta}\| \mathbf{e}_2 = \left( 0.764 \cdot 0.974 \cdot \sqrt{10} \right) \mathbf{e}_2 \\ &= 2.353 \begin{pmatrix} 0.526 \\ 0.851 \end{pmatrix} = \begin{pmatrix} 1.236 \\ 2 \end{pmatrix} \end{aligned}$$



**Figure A5.3** **Left:** The scaled eigenvectors associated with the covariance matrix,  $\mathbf{G}$ , from Example A5.1, plotted along with the selection gradient,  $\beta$ . Note that  $\mathbf{e}_1$  and  $\mathbf{e}_2$  are orthogonal and hence can be thought of as describing a new coordinate system. Because  $\lambda_1 \gg \lambda_2$ , the leading eigenvector,  $\mathbf{e}_1$ , largely dominates the transformation. **Right:** This is shown by taking the projections of  $\beta$  on each of these eigenvectors (shown here on a magnified scale relative to the left figure). Even though  $\beta$  is nearly parallel to  $\mathbf{e}_2$  ( $\theta|\mathbf{e}_1, \beta = 13.2^\circ$ ), the projection of  $\beta$  on  $\mathbf{e}_1$  yields a vector of greater length than the projection of  $\beta$  on  $\mathbf{e}_2$  (3.803 versus 2.353). From Equation A5.9b, the vector of responses to selection,  $\mathbf{R}$ , is the sum of these two projections.

From Equation A5.9b, we can express the response,  $\mathbf{R}$ , as the sum of the projections of  $\beta$  onto the eigenvalues of  $\mathbf{G}$ , returning

$$\begin{aligned} \mathbf{R} &= \mathbf{G}\beta = \lambda_1 \text{Proj}(\beta \text{ on } \mathbf{e}_1) + \lambda_2 \text{Proj}(\beta \text{ on } \mathbf{e}_2) \\ &= \begin{pmatrix} -3.236 \\ 2 \end{pmatrix} + \begin{pmatrix} 1.236 \\ 2 \end{pmatrix} = \begin{pmatrix} -2 \\ 4 \end{pmatrix} \end{aligned}$$

As Figure A5.3 shows, the eigenstructure of  $\mathbf{G}$  explains the unusual behavior of the selection response seen in Figure A5.2. The eigenvector associated with the **leading eigenvalue**,  $\lambda_1$ , accounts for most of the variation inherent in  $\mathbf{G}$  (87%, as  $\lambda_1/(\lambda_1 + \lambda_2) = 0.87$ ), and this eigenvector corresponds to a strong negative correlation between the additive-genetic values of  $z_1$  and  $z_2$ . Hence, even though  $\beta$  points in very much the same direction as  $\mathbf{e}_2$ , because  $\lambda_1 \gg \lambda_2$ , the projection of  $\beta$  on  $\mathbf{e}_1$  yields a vector of greater length than the projection of  $\beta$  on  $\mathbf{e}_2$  (3.803 versus 2.353), and it is this  $\mathbf{e}_1$  projection vector that results in the decrease in  $\mu_{z_1}$ .

### PROPERTIES OF SYMMETRIC MATRICES

Many of the matrices encountered in quantitative genetics are **symmetric**, satisfying  $\mathbf{A} = \mathbf{A}^T$  (and therefore necessarily square). Examples include covariance matrices and the  $\gamma$  matrix of quadratic coefficients in the Pearson-Lande-Arnold fitness regression (Chapter 30). Symmetric matrices have a number of useful properties (proofs of which can be found in Dhrymes 1978; Horn and Johnson 1985; and Wilf 1978):

1. If  $\mathbf{A}$  is symmetric, then if  $\mathbf{A}^{-1}$  exists, it is also symmetric.
2. The eigenvalues and eigenvectors of a symmetric matrix are all real.
3. For any  $n$ -dimensional symmetric matrix, a corresponding set of  $n$  orthonormal eigenvectors can be constructed, namely, we can obtain a set of eigenvalues  $\mathbf{e}_i$  for  $1 \leq i \leq n$  that satisfies

$$\mathbf{e}_i^T \mathbf{e}_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

In particular, this guarantees that a spectral decomposition of  $\mathbf{A}$  exists (Equation A5.9a).

4. A symmetric matrix  $\mathbf{A}$  can be **diagonalized** as

$$\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T \tag{A5.10a}$$

where  $\mathbf{\Lambda}$  is a diagonal matrix and  $\mathbf{U}$  is an orthonormal matrix ( $\mathbf{U}^{-1} = \mathbf{U}^T$ ). If  $\lambda_i$  and  $\mathbf{e}_i$  are the  $i$ th eigenvalue and its associated unit eigenvector of  $\mathbf{A}$ , then

$$\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_n) = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & \lambda_n \end{pmatrix} \tag{A5.10b}$$

and

$$\mathbf{U} = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n) \tag{A5.10c}$$

Geometrically,  $\mathbf{U}$  is a unity matrix and thus describes a rigid rotation of the original coordinate system to a new coordinate system given by the eigenvectors of  $\mathbf{A}$ , while the diagonal elements of  $\mathbf{\Lambda}$  give the amount by which vectors of unit length in the original coordinate system are scaled in the transformed system. If we use the decomposition  $\mathbf{\Lambda} = \sum_{i=1}^n \mathbf{\Lambda}_i$ , where  $\mathbf{\Lambda}_i$  is a diagonal matrix whose elements are all zero, except for  $\lambda_i$ , then Equation A5.10a becomes

$$\mathbf{A} = \mathbf{U} \left( \sum_{i=1}^n \mathbf{\Lambda}_i \right) \mathbf{U}^T = \sum_{i=1}^n \mathbf{U}\mathbf{\Lambda}_i\mathbf{U}^T = \sum_{i=1}^n \lambda_i \mathbf{e}_i \mathbf{e}_i^T \tag{A5.10d}$$

recovering the spectral decomposition (Equation A5.9a). The last step in Equation A5.10d follows because  $\mathbf{e}_i^T \mathbf{e}_j = 0$  for  $i \neq j$ . Because of this feature, Equation A5.10a is also called the **spectral factorization** or **eigendecomposition** of  $\mathbf{A}$ .

Using Equation A5.10a, it is easy to show that

$$\mathbf{A}^{-1} = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^T \tag{A5.11a}$$

To see this, note that

$$\mathbf{A}^{-1}\mathbf{A} = (\mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{U}^T) (\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T) = \mathbf{U}\mathbf{\Lambda}^{-1} (\mathbf{U}^T\mathbf{U}) \mathbf{\Lambda}\mathbf{U}^T = \mathbf{U}\mathbf{\Lambda}^{-1}\mathbf{\Lambda}\mathbf{U}^T = \mathbf{U}\mathbf{U}^T = \mathbf{I}$$

Similar logic yields

$$\mathbf{A}^{1/2} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{U}^T \tag{A5.11b}$$

$$\mathbf{A}^{-1/2} = \mathbf{U}\mathbf{\Lambda}^{-1/2}\mathbf{U}^T \tag{A5.11c}$$

$$\mathbf{A}^k = \mathbf{U}\mathbf{\Lambda}^k\mathbf{U}^T \quad \text{for any integer } k \tag{A5.11d}$$

where the **square root matrix**,  $\mathbf{A}^{1/2}$ , satisfies  $\mathbf{A}^{1/2}\mathbf{A}^{1/2} = \mathbf{A}$ , and  $\mathbf{A}^{-1/2}$  satisfies  $\mathbf{A}^{-1/2}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1/2} = \mathbf{A}^{1/2}$ , as well as  $\mathbf{A}^{-1/2}\mathbf{A}^{1/2} = \mathbf{A}^{1/2}\mathbf{A}^{-1/2} = \mathbf{I}$ .

Because  $\mathbf{\Lambda}$  is diagonal, the  $i$ th diagonal elements of  $\mathbf{\Lambda}^{-1}$ ,  $\mathbf{\Lambda}^{1/2}$ ,  $\mathbf{\Lambda}^{-1/2}$ , and  $\mathbf{\Lambda}^k$  are  $\lambda_i^{-1}$ ,  $\lambda_i^{1/2}$ ,  $\lambda_i^{-1/2}$ , and  $\lambda_i^k$ , respectively, implying that if  $\lambda_i$  is an eigenvalue of  $\mathbf{A}$ , then  $\lambda_i^{-1}$ ,  $\lambda_i^{1/2}$ ,  $\lambda_i^{-1/2}$ , and  $\lambda_i^k$ , respectively, are eigenvalues of the matrices  $\mathbf{A}^{-1}$ ,  $\mathbf{A}^{1/2}$ ,  $\mathbf{A}^{-1/2}$ , and  $\mathbf{A}^k$ . Note that Equations A5.11a–A5.11d further imply that the matrices  $\mathbf{A}$ ,  $\mathbf{A}^{-1}$ ,  $\mathbf{A}^{1/2}$ ,  $\mathbf{A}^{-1/2}$ , and  $\mathbf{A}^k$  all have the same eigenvectors, namely the columns of  $\mathbf{U}$ . Finally, using Equation A5.10a, we see that premultiplying  $\mathbf{A}$  by  $\mathbf{U}^T$  and then postmultiplying by  $\mathbf{U}$  gives a diagonal matrix whose elements are the eigenvalues of  $\mathbf{A}$

$$\mathbf{U}^T\mathbf{A}\mathbf{U} = \mathbf{U}^T(\mathbf{U}\mathbf{\Lambda}\mathbf{U}^T)\mathbf{U} = (\mathbf{U}^T\mathbf{U})\mathbf{\Lambda}(\mathbf{U}^T\mathbf{U}) = \mathbf{\Lambda} \tag{A5.12}$$

5. The **Rayleigh-Ritz** theorem gives useful bounds on quadratic products associated with the symmetric matrix  $\mathbf{A}$ . It states that if the eigenvalues of  $\mathbf{A}$  are ordered as  $\lambda_{max} = \lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n = \lambda_{min}$ , then for any vector of constants  $\mathbf{c}$  (for  $\|\mathbf{c}\| > 0$ )

$$\lambda_1 \|\mathbf{c}\| \geq \mathbf{c}^T \mathbf{A} \mathbf{c} \geq \lambda_n \|\mathbf{c}\| \tag{A5.13a}$$

If  $\mathbf{c}$  is of unit length, then all quadratic products are bounded by

$$\lambda_1 \geq \mathbf{c}^T \mathbf{A} \mathbf{c} \geq \lambda_n \tag{A5.13b}$$

The maximum and minimum quadratic products occur, respectively, when  $\mathbf{c} = \mathbf{e}_1$  and  $\mathbf{c} = \mathbf{e}_n$ , the eigenvectors associated with  $\lambda_1$  and  $\lambda_n$ . This is a useful result for bounding variances. Consider a univariate random variable,  $y = \mathbf{c}^T \mathbf{x}$ , formed by a linear combination of the elements of a random vector,  $\mathbf{x}$ . Recall from LW Equation 8.19 that the variance of a sum  $y = \mathbf{c}^T \mathbf{x}$  is  $\sigma^2(y) = \mathbf{c}^T \mathbf{V}_\mathbf{x} \mathbf{c}$ , where  $\mathbf{V}_\mathbf{x}$  is the covariance matrix for  $\mathbf{x}$ . If we apply Equation A5.13a we obtain

$$\lambda_1 \|\mathbf{c}\|^2 \geq \sigma^2(y) \geq \lambda_n \|\mathbf{c}\|^2 \tag{A5.14}$$

where  $\lambda_1$  is the largest (leading or **dominant**) eigenvalue and  $\lambda_n$  is the smallest eigenvalue of the covariance matrix  $\mathbf{V}_\mathbf{x}$ .

**Example A5.3.** Consider the additive-genetic covariance matrix  $\mathbf{G}$  from Examples A5.1 and A5.2. Recalling the results from Example A5.2 and using Equation A5.10a, we can express  $\mathbf{G}$  as  $\mathbf{U}\mathbf{A}\mathbf{U}^T$ , where

$$\mathbf{A} = \begin{pmatrix} 5.241 & 0 \\ 0 & 0.765 \end{pmatrix} \quad \text{and} \quad \mathbf{U} = (\mathbf{e}_1 \quad \mathbf{e}_2) = \left( \begin{pmatrix} -0.851 \\ 0.526 \end{pmatrix} \quad \begin{pmatrix} 0.526 \\ 0.851 \end{pmatrix} \right)$$

From Equation A5.11a, the eigenvalues of  $\mathbf{A}^{-1}$  are  $(5.241)^{-1} \simeq 0.191$  and  $(0.765)^{-1} \simeq 1.307$ , while from Equation A5.11b, the eigenvalues of  $\mathbf{A}^{1/2}$  are  $\sqrt{5.241} \simeq 2.289$  and  $\sqrt{0.765} \simeq 0.875$ .

### Correlations Can Be Removed by a Matrix Transformation

A powerful use of diagonalization is that it allows one to extract a set of  $n$  uncorrelated variables for any  $n \times n$  nonsingular covariance matrix,  $\mathbf{V}_\mathbf{x}$ . Consider the transformation

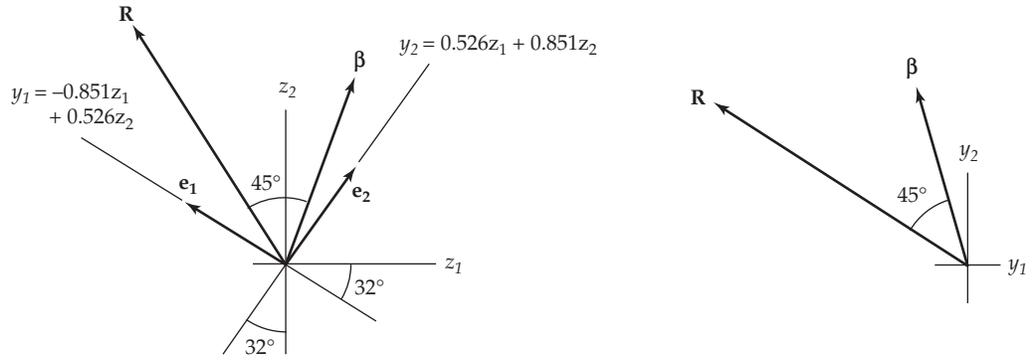
$$\mathbf{y} = \mathbf{U}^T \mathbf{x} \tag{A5.15a}$$

where  $\mathbf{U} = (\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_n)$  contains the normalized eigenvectors of  $\mathbf{V}_\mathbf{x}$ . Because  $\mathbf{U}$  is an orthonormal matrix, this transformation is a rigid rotation of the axes of the original  $(x_1, \dots, x_n)$  coordinate system to a new system given by  $(e_1, \dots, e_n)$ . Applying LW Equation 8.21b and Equation A5.12, respectively, the covariance matrix for  $\mathbf{y}$  is

$$\mathbf{V}_\mathbf{y} = \mathbf{U}^T \mathbf{V}_\mathbf{x} \mathbf{U} = \mathbf{\Lambda} \tag{A5.15b}$$

where  $\mathbf{\Lambda}$  is a diagonal matrix whose elements are the eigenvalues of  $\mathbf{V}_\mathbf{x}$ ,

$$\sigma(y_i, y_j) = \begin{cases} \lambda_i & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$



**Figure A5.4** The transformation (Equation A5.15a) generating a set of independent variables for the covariance matrix  $\mathbf{G}$  from Example A5.4 results in a rigid rotation of axes of the original traits onto the new, uncorrelated set. **Left:** The direction of the new axes are given by the eigenvectors  $\mathbf{e}_1$  and  $\mathbf{e}_2$ . The angle between the new axis,  $\mathbf{e}_1$ , and the original  $\mathbf{z}_1$  axis is given by the angle between  $\mathbf{e}_1$  and  $\mathbf{z}_1 = (1, 0)^T$ . Here,  $\|\mathbf{e}_1\| = \|\mathbf{z}_1\| = 1$  and  $\mathbf{e}_1^T \mathbf{z}_1 = 0.851$ , giving  $\theta = \cos^{-1}(0.851) \simeq 32^\circ$ . As this transformation is a rigid rotation, the angle between  $\mathbf{e}_2$  and the  $\mathbf{z}_2 = (0, 1)^T$  axis is also  $32^\circ$ . **Right:** On the  $(y_1, y_2)$  coordinates, the angle between  $\mathbf{R}$  and  $\beta$  remains unchanged. See Example A5.4 for further details.

The rigid rotation introduced by  $\mathbf{U}$  creates a set of  $n$  uncorrelated variables, the  $i$ th of which is

$$y_i = \mathbf{e}_i^T \mathbf{x} \tag{A5.15c}$$

Because the  $\mathbf{e}_i$  are of unit length, from Equation A5.3b we have that  $y_i = \mathbf{e}_i^T \mathbf{x}$  is the length of the projection of  $\mathbf{x}$  onto the  $i$ th eigenvector of  $\mathbf{V}_\mathbf{x}$ , which implies that the axes of the new coordinate system are given by the orthogonal set of eigenvectors of  $\mathbf{V}_\mathbf{x}$ .

Defining the matrix  $\mathbf{B}$  as

$$\mathbf{B} = \mathbf{U} \mathbf{\Lambda}^{-1/2} \tag{A5.15d}$$

the vector  $\mathbf{y} = \mathbf{B}^T \mathbf{x}$  has a covariance matrix of  $\mathbf{V}_\mathbf{y} = \mathbf{I}$ , which means that this transformation creates a set of uncorrelated variables, each with unit variance. To see this, note that

$$\begin{aligned} \mathbf{V}_\mathbf{y} &= \mathbf{B}^T \mathbf{V}_\mathbf{x} \mathbf{B} = (\mathbf{U} \mathbf{\Lambda}^{-1/2})^T (\mathbf{U} \mathbf{\Lambda} \mathbf{U}^T) (\mathbf{U} \mathbf{\Lambda}^{-1/2}) \\ &= \mathbf{\Lambda}^{-1/2} (\mathbf{U}^T \mathbf{U}) \mathbf{\Lambda} (\mathbf{U}^T \mathbf{U}) \mathbf{\Lambda}^{-1/2} \\ &= \mathbf{\Lambda}^{-1/2} \mathbf{\Lambda} \mathbf{\Lambda}^{-1/2} = \mathbf{I} \end{aligned} \tag{A5.15e}$$

An alternative to Equation A5.15d is the **Cholesky decomposition**,  $\mathbf{A} = \mathbf{C}^T \mathbf{C}$ , of a square, symmetric matrix  $\mathbf{A}$ , where  $\mathbf{C}$  is an **lower triangular matrix** (all elements above the diagonal are zero). If  $\mathbf{C}$  is the Cholesky decomposition for  $\mathbf{V}_\mathbf{x}$ , then  $\mathbf{y} = \mathbf{C}^{-1} \mathbf{x}$  also returns a covariance matrix of  $\mathbf{I}$ .

**Example A5.4.** If we apply the change of variables suggested by Equation A5.15a to the vector,  $\mathbf{z}$ , of characters with associated  $\mathbf{G}$  matrix used in Example A5.1 and using the eigenvalues and vectors obtained in Example A5.2 yields

$$\begin{aligned} \mathbf{y} &= \mathbf{U}^T \mathbf{z} = \begin{pmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \\ &= \begin{pmatrix} -0.851 & 0.526 \\ 0.526 & 0.851 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \\ &= \begin{pmatrix} -0.851z_1 + 0.526z_2 \\ 0.526z_1 + 0.851z_2 \end{pmatrix} \end{aligned}$$

From Equation A5.15b,  $\mathbf{V}_y = \mathbf{A}$  as given in Example A5.3, showing that  $y_1$  and  $y_2$  are uncorrelated with  $\sigma^2(y_1) = \lambda_1 = 5.241$  and  $\sigma^2(y_2) = \lambda_2 = 0.765$ . Hence, by considering the new coordinate system with

$$y_1 = \mathbf{e}_1^T \mathbf{z} = -0.851z_1 + 0.526z_2 \quad \text{and} \quad y_2 = \mathbf{e}_2^T \mathbf{z} = 0.526z_1 + 0.851z_2$$

we can transform the original coordinate system into a new system on which there are no additive-genetic correlations between these new characters. Figure A5.4 shows that this transformation is simply a rigid rotation of the axes.

Likewise, from Equation A5.15d, the transformation that yields uncorrelated variables with unit variance is

$$\begin{aligned} \mathbf{y} &= \mathbf{A}^{-1/2} \mathbf{U}^T \mathbf{z} = \begin{pmatrix} 1/\sqrt{\lambda_1} & 0 \\ 0 & 1/\sqrt{\lambda_2} \end{pmatrix} \begin{pmatrix} \mathbf{e}_1^T \\ \mathbf{e}_2^T \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \\ &= \begin{pmatrix} 1/\sqrt{5.236} & 0 \\ 0 & 1/\sqrt{0.764} \end{pmatrix} \begin{pmatrix} -0.851 & 0.526 \\ 0.526 & 0.851 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \\ &= \begin{pmatrix} -0.372 & 0.230 \\ 0.602 & 0.974 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \end{aligned}$$

Hence, the transformed variables  $y_1 = -0.372z_1 + 0.230z_2$  and  $y_2 = 0.602z_1 + 0.974z_2$  are uncorrelated, and each has unit variance.

An alternative set of uncorrelated random variables follows from the Cholesky decomposition, which can be computed in R using the `chol` command. (As an aside, `chol` returns the upper-triangular version of the decomposition, which is simply the transpose of the lower-triangular version). The resulting decomposition is

$$\mathbf{G} = \begin{pmatrix} 4 & -2 \\ -2 & 2 \end{pmatrix} = \mathbf{C} \mathbf{C}^T = \begin{pmatrix} 2 & 0 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ 0 & 1 \end{pmatrix}$$

yielding

$$\mathbf{y} = \mathbf{C}^{-1} \mathbf{z} = \begin{pmatrix} 0.5 & 0 \\ 0.5 & 1 \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}$$

or

$$y_1 = z_1/2 \quad \text{and} \quad z_2 = z_1/2 + z_2$$

as a new set of uncorrelated variables, each with unit variance. One nice feature about using a Cholesky decomposition is that we can always isolate a given variable of interest (simply by putting first in the vector). Because  $\mathbf{C}$  is lower-triangular, it always returns the first new uncorrelated variable as a scalar times the first original variable (rather than some linear combination of all the variables, as was the case for the first decomposition in this example).

### Simultaneous Diagonalization

An extension of the notion of diagonalization is the **simultaneous diagonalization** of two symmetric matrices,  $\mathbf{P}$  and  $\mathbf{G}$ , of the same dimension. There exists a matrix  $\mathbf{T}$  such that

$$\mathbf{T}^T \mathbf{P} \mathbf{T} = \mathbf{I} \quad \text{and} \quad \mathbf{T}^T \mathbf{G} \mathbf{T} = \mathbf{D} \tag{A5.16}$$

where  $\mathbf{D}$  is a diagonal matrix, whose elements are the eigenvalues of  $\mathbf{P}^{-1} \mathbf{G}$ . Hence, the same transformation simultaneously diagonalizes both  $\mathbf{P}$  and  $\mathbf{G}$ . If one has a series of traits with both genetic ( $\mathbf{G}$ ) and phenotypic ( $\mathbf{P}$ ) covariances, they can be transformed to a scale where the new traits (based on linear combinations of the original traits) are genetically and phenotypically uncorrelated, where the elements of  $\mathbf{D}$  correspond to the heritabilities of these new traits.

**Example A5.5.** To find the matrix,  $\mathbf{T}$ , that simultaneously diagonalizes both  $\mathbf{P}$  and  $\mathbf{G}$ , we first use Equation A5.10a to write

$$\mathbf{P} = \mathbf{U}\mathbf{A}\mathbf{U}^T$$

where  $\mathbf{A}$  is a diagonal matrix and  $\mathbf{U}^T\mathbf{U} = \mathbf{U}\mathbf{U}^T = \mathbf{I}$ . Defining  $\mathbf{B} = \mathbf{U}\mathbf{A}^{-1/2}$ , Equation A5.15e showed that  $\mathbf{B}^T\mathbf{P}\mathbf{B} = \mathbf{I}$ . Next, note for  $\mathbf{M} = \mathbf{B}^T\mathbf{G}\mathbf{B}$ , that  $\mathbf{M} = \mathbf{M}^T$  (i.e.,  $\mathbf{M}$  is symmetric), as

$$\mathbf{M}^T = (\mathbf{B}^T\mathbf{G}\mathbf{B})^T = \mathbf{B}^T\mathbf{G}^T\mathbf{B} = \mathbf{B}^T\mathbf{G}\mathbf{B} = \mathbf{M}$$

Hence, we can also diagonalize  $\mathbf{M}$ ,

$$\mathbf{C}^T\mathbf{M}\mathbf{C} = \mathbf{D}$$

where  $\mathbf{D}$  is a diagonal matrix and  $\mathbf{C}^T\mathbf{C} = \mathbf{C}\mathbf{C}^T = \mathbf{I}$ . Thus,

$$\mathbf{C}^T\mathbf{M}\mathbf{C} = \mathbf{C}^T(\mathbf{B}^T\mathbf{G}\mathbf{B})\mathbf{C} = (\mathbf{B}\mathbf{C})^T\mathbf{G}(\mathbf{B}\mathbf{C}) = \mathbf{D}$$

Defining

$$\mathbf{T} = \mathbf{B}\mathbf{C} = \mathbf{U}\mathbf{A}^{-1/2}\mathbf{C}$$

we have from the previous expression that

$$\mathbf{T}^T\mathbf{G}\mathbf{T} = \mathbf{D}$$

Likewise,

$$\mathbf{T}^T\mathbf{P}\mathbf{T} = (\mathbf{B}\mathbf{C})^T\mathbf{P}(\mathbf{B}\mathbf{C}) = \mathbf{C}^T(\mathbf{B}^T\mathbf{P}\mathbf{B})\mathbf{C} = \mathbf{C}^T\mathbf{C} = \mathbf{I}$$

showing that the matrix  $\mathbf{T}$  satisfies Equation A5.16.

## CANONICAL AXES OF QUADRATIC FORMS

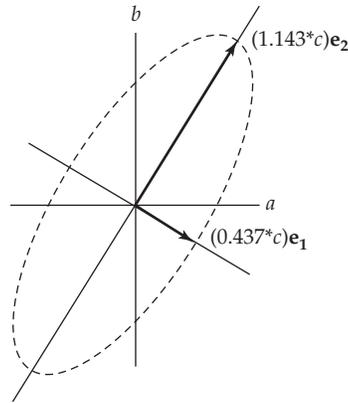
The transformation  $\mathbf{y} = \mathbf{U}^T\mathbf{x}$  given by Equation A5.15a applies to any symmetric matrix, and is referred to as its **canonical transformation**. This simplifies the interpretation of the quadratic form  $\mathbf{x}^T\mathbf{A}\mathbf{x}$ , as rotation of the original axes to align them with the eigenvectors of  $\mathbf{A}$  removes all cross-product terms ( $x_i x_j$  for  $i \neq j$ ) on this new coordinate system. Recall (Equation A5.5b) that  $\mathbf{U}$  is a unitary matrix and hence  $\mathbf{U}^T = \mathbf{U}^{-1}$ . Thus,

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

Applying Equations A5.15a and A5.12 transforms a quadratic form to one in which the square matrix is diagonal, which greatly simplifies the resulting quadratic product, as

$$\begin{aligned} \mathbf{x}^T\mathbf{A}\mathbf{x} &= (\mathbf{U}\mathbf{y})^T\mathbf{A}\mathbf{U}\mathbf{y} = \mathbf{y}^T(\mathbf{U}^T\mathbf{A}\mathbf{U})\mathbf{y} \\ &= \mathbf{y}^T\mathbf{A}\mathbf{y} \\ &= \sum_{i=1}^n \lambda_i y_i^2, \quad \text{with } y_i = \mathbf{e}_i^T\mathbf{x} \end{aligned} \tag{A5.17a}$$

where  $\lambda_i$  and  $\mathbf{e}_i$  are the eigenvalues and associated (normalized, i.e.,  $\|\mathbf{e}_i\| = 1$ ) eigenvectors of  $\mathbf{A}$ . The new axes defined by the  $\mathbf{e}_i$  vectors are the **canonical** (or **principal**)



**Figure A5.5** The general shape of surfaces of constant variance for the additive-genetic covariance matrix,  $\mathbf{G}$ , given in Example A5.1. Defining a new composite character  $y = az_1 + bz_2$ , the rotated ellipse represents the set of weights  $(a, b)$  that give  $y$  the same additive-genetic variance,  $c^2$ . The major axis of the ellipse is along  $\mathbf{e}_2$ , the eigenvector associated with the smallest eigenvalue of  $\mathbf{G}$ , where  $\lambda_2 \simeq 0.765$ , giving  $1/\sqrt{\lambda_2} \simeq 1.143$ . The minor axis of the ellipse is along  $\mathbf{e}_1$ , the eigenvector associated with the largest eigenvalue of  $\mathbf{G}$ , where  $\lambda_1 \simeq 5.241$ , giving  $1/\sqrt{\lambda_1} \simeq 0.437$ .

axes of  $\mathbf{A}$ . Because  $y_i^2 \geq 0$ , Equation A5.17a immediately shows the connection between the signs of the eigenvalues of a matrix and whether that matrix is positive definite, negative definite, or indefinite.

If all eigenvalues are positive (all  $\lambda_i > 0$ ), then any quadratic form is always positive (unless all the  $y_i$  are zero) and hence  $\mathbf{A}$  is **positive definite**. If one or more of the eigenvalues are zero, while the rest are positive, then  $\mathbf{A}$  is said to be **positive semidefinite**, implying that quadratic products are either zero (corresponding to  $\lambda_i = 0$ ) or positive. If all eigenvalues are negative (all  $\lambda_i < 0$ ), then  $\mathbf{A}$  is **negative definite** as any quadratic form is always negative, while  $\mathbf{A}$  is said to be **negative semidefinite** if the eigenvalues are either zero or negative. If  $\mathbf{A}$  has both positive and negative eigenvalues it is said to be **indefinite**, as quadratic products can be either positive or negative.

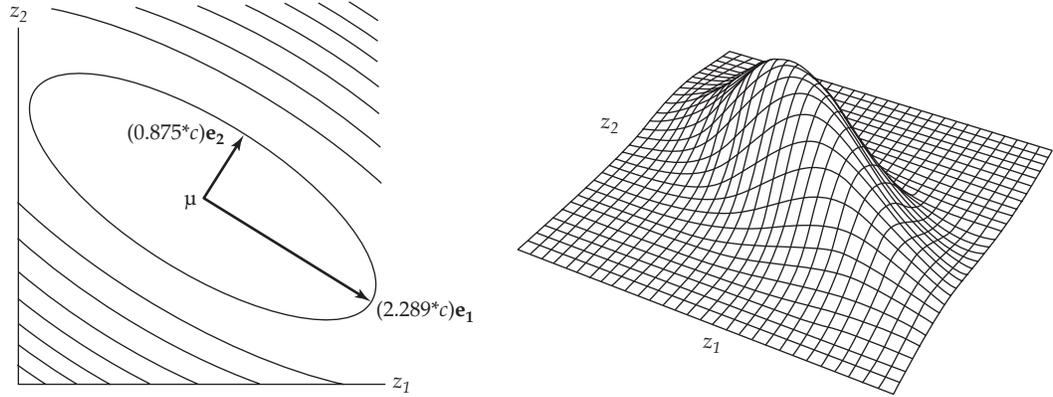
Equations of the form

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \sum_{i=1}^n \sum_{j=1}^n A_{ij} x_i x_j = c^2 \tag{A5.17b}$$

arise fairly frequently in quantitative genetics. For example, they describe surfaces of constant variance (Figure A5.5) or constant fitnesses in quadratic fitness regressions (Chapter 30). Solutions to Equation A5.17b describe **quadratic surfaces**—for two dimensions, these are the familiar conic sections (ellipses, parabolas, or hyperbolas). Equation A5.17a greatly simplifies the interpretation of these surfaces by removing all cross product terms, yielding

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \sum_{i=1}^n \lambda_i y_i^2 = c^2 \tag{A5.17c}$$

Because  $(y_i)^2$  and  $(-y_i)^2$  have the same value, the canonical axes of  $\mathbf{A}$  are also the **axes of symmetry** for the quadratic surface generated by quadratic forms involving  $\mathbf{A}$ . When all eigenvalues of  $\mathbf{A}$  are positive (as occurs with nonsingular covariance and other positive-definite matrices), Equation A5.17c describes an ellipsoid whose axes of symmetry are given by the eigenvectors of  $\mathbf{A}$ . The distance from the origin to the surface along the  $\mathbf{e}_i$  axis is  $\lambda_i y_i^2 = c^2$  or  $y_i = c\lambda_i^{-1/2}$ , as can be seen by setting all the  $y_k$  equal to zero except for  $y_i$ , which yields



**Figure A5.6** Surfaces for a multivariate normal (MVN) distribution. **Left:** Surfaces of equal probability assuming that the additive-genetic values associated with the characters  $z_1$  and  $z_2$  in Example A5.1 are  $\sim \text{MVN}(\boldsymbol{\mu}, \mathbf{G})$ . These surfaces are ellipses centered at  $\boldsymbol{\mu}$ , with the major axis of the ellipse along  $\mathbf{e}_1$  and the minor axis along  $\mathbf{e}_2$ , whose lengths (for a fixed  $c$ ) are, respectively,  $\sqrt{\lambda_1} = 2.289$  and  $\sqrt{\lambda_2} = 0.875$ . **Right:** A plot of the associated probability density. Slicing along either the major or minor axis gives a normal curve. Because the variance in the major axis is greater, the curve is much broader along this axis. The covariance between the breeding values of  $z_1$  and  $z_2$  rotates the distribution so that the principal axes ( $\mathbf{e}_1, \mathbf{e}_2$ ) do not coincide with the original ( $z_1, z_2$ ) axes.

$$\mathbf{x}^T \mathbf{A} \mathbf{x} = \lambda_i y_i^2 = c^2.$$

Consider a new variable ( $y$ ) that is a weighted combination  $y = ax_1 + bx_2 = \mathbf{b}^T \mathbf{x}$  of the original vector ( $\mathbf{x}$ ) of random variables, where  $\mathbf{b}^T = (a, b)$ . Its resulting variance is

$$\sigma^2(y) = a^2\sigma^2(x_1) + 2ab\sigma(x_1, x_2) + b^2\sigma^2(x_2) = \mathbf{b}^T \mathbf{V}_\mathbf{x} \mathbf{b}$$

As shown in Figure A5.5, the collection of  $a, b$  values that result in the same variance ( $c^2$ ) is the ellipse given by  $c^2 = \mathbf{b}^T \mathbf{V}_\mathbf{x} \mathbf{b}$ . Variables with a large amount of variance require smaller weights to achieve the constant value ( $c^2$ ) than do variables with lower variances. Thus, on a **constant-variance surface**, minor axes correspond to directions with *the most* variance, while major axes correspond to the directions with *the least* variability. This is in contrast to **surfaces of equal probability** (Figure A5.6), where major axes correspond to directions with the most variance. The reason for this reversal of roles is that constant-variance surfaces are functions of  $\lambda_i^{-1/2}$ , whereas constant-probability surfaces are functions of  $\lambda_i^{1/2}$ .

**Implications for the Multivariate Normal Distribution**

Recall the probability density function for the multivariate normal distribution (LW Chapter 8)

$$\phi(\mathbf{x}) = (2\pi)^{-n/2} |\mathbf{V}_\mathbf{x}|^{-1/2} \exp \left[ -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}_\mathbf{x}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right] \tag{A5.18a}$$

Because only the quadratic product in the exponential varies with  $\mathbf{x}$ , surfaces of equal probability for MVN distributed vectors satisfy

$$(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}_\mathbf{x}^{-1} (\mathbf{x} - \boldsymbol{\mu}) = c^2 \tag{A5.18b}$$

From the discussion following Equation A5.17c, these surfaces are  $n$ -dimensional ellipsoids centered at  $\boldsymbol{\mu}$  whose axes of symmetry are given by the principal components (the eigenvectors) of the covariance matrix,  $\mathbf{V}_\mathbf{x}$ . The length of the ellipsoid along the  $i$ th axis is  $c\sqrt{\lambda_i}$  where  $\lambda_i$  is the eigenvalue associated with the eigenvector  $\mathbf{e}_i$  (Figure A5.6).

Equation A5.18b motivates the **Mahalanobis distance**

$$D = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{V}_\mathbf{x}^{-1} (\mathbf{x} - \boldsymbol{\mu})} \tag{A5.19}$$

which measures the distance of a point from its mean  $\boldsymbol{\mu}$ , correcting for its covariance structure,  $\mathbf{V}_\mathbf{x}$  (Mahalanobis 1938). As we detail shortly,  $D$  provides one metric for tests of multivariate normality.

Applying the canonical transformation (Equation A5.15a), we can change coordinate systems by a rigid rotation to remove any correlations between the variables in  $\mathbf{x}$ . If  $\mathbf{x} \sim \text{MVN}(\boldsymbol{\mu}, \mathbf{V}_\mathbf{x})$ , then for  $\mathbf{y} = \mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu})$ , it follows that

$$\mathbf{y} \sim \text{MVN}(\mathbf{0}, \boldsymbol{\Lambda}) \tag{A5.20a}$$

where  $\boldsymbol{\Lambda}$  and  $\mathbf{U}$  are the matrices defined by Equations A5.10b and A5.10bc for the diagonalization of  $\mathbf{V}_\mathbf{x}$ . In particular,

$$y_i = \mathbf{e}_i^T(\mathbf{x} - \boldsymbol{\mu}) \quad \text{where} \quad y_i \sim \text{N}(0, \lambda_i) \tag{A5.20b}$$

Note from Equation A5.20a that because the  $y_i$  are uncorrelated, they are also independent as the joint probability density is the product of  $n$  individual univariate normal densities. We can further transform the original vector by taking

$$z_i = \frac{\mathbf{e}_i^T(\mathbf{x} - \boldsymbol{\mu})}{\sqrt{\lambda_i}} \quad \text{giving} \quad z_i \sim \text{N}(0, 1) \tag{A5.20c}$$

Applying the transformation

$$\mathbf{z} = \boldsymbol{\Lambda}^{-1/2}\mathbf{U}^T(\mathbf{x} - \boldsymbol{\mu}) \tag{A5.20d}$$

results in  $\mathbf{z} \sim \text{MVN}(\mathbf{0}, \mathbf{I})$ , namely that the  $n$  elements of the vector  $\mathbf{y}$  are each independent unit normal random variables.

### Principal Components of the Variance-Covariance Matrix

We are often interested in how the variance of a random vector 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. The procedure of **principal component analysis (PCA)** extracts these combinations by decomposing the variance of  $\mathbf{x}$  into the contributions from a series of orthogonal vectors, the first of which explains the most variation possible for any single vector, the second the next possible amount, and so on until we account for the entire variance of  $\mathbf{x}$ .

Consider Figure A5.5. Because the set of points comprising the ellipse represents the set of linear combinations (i.e., the set of weights) of the random variables of  $\mathbf{z}$  that yield **equal** variance, a little thought shows that the closer a point on this curve is to the origin, the more variance there is in that direction. The points closest to the origin are those that lie along the axis defined by  $\mathbf{e}_1$ , while those furthest away lie along the axis defined by  $\mathbf{e}_2$ . Here  $\mathbf{e}_1$  and  $\mathbf{e}_2$  are the principal components of  $\mathbf{G}$ , with the first principal component accounting for most of the variation of  $\mathbf{G}$ . In particular, the ratio of additive variances for the characters  $y_1 = \mathbf{e}_1^T \mathbf{z}$  and  $y_2 = \mathbf{e}_2^T \mathbf{z}$  is  $\sigma^2(y_1)/\sigma^2(y_2) = \sigma^2(\mathbf{e}_1^T \mathbf{z})/\sigma^2(\mathbf{e}_2^T \mathbf{z}) = \mathbf{e}_1^T \mathbf{G} \mathbf{e}_1/\mathbf{e}_2^T \mathbf{G} \mathbf{e}_2 = \lambda_1/\lambda_2 \simeq 5.241/0.765 \simeq 6.85$ , so that a character in the direction of  $\mathbf{e}_1$  has almost seven times as much additive variance as a character lying in the direction of  $\mathbf{e}_2$ .

In general, suppose we have an  $n$ -dimensional covariance matrix,  $\mathbf{V}_\mathbf{x}$ . If we order the eigenvalues of  $\mathbf{V}_\mathbf{x}$  as  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ , then Equation A5.13b gives 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  $\|\mathbf{c}_1\| = 1$ ), as

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

which occurs when  $\mathbf{c}_1 = \mathbf{e}_1$  (the normalized eigenvector associated with the leading eigenvalue  $\lambda_1$ ). This vector is the **first principal component** (often abbreviated as **PC1**), and

accounts for the fraction  $\lambda_1/\text{tr}(\mathbf{V}_\mathbf{x})$  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  $\lambda_2/\text{tr}(\mathbf{V}_\mathbf{x})$  of the total variation in  $\mathbf{x}$  (e.g., Morrison 1976; Johnson and Wichern 1988). By proceeding in this fashion, we can see that the  $i$ th PC is given by  $\mathbf{e}_i$ , and that the amount of variation it accounts for is

$$\lambda_i / \sum_{k=1}^n \lambda_k = \frac{\lambda_i}{\text{tr}(\mathbf{V}_\mathbf{x})} \quad (\text{A5.21})$$

Hence  $\sum \lambda_i = \text{tr}(\mathbf{V}_\mathbf{x})$  is the total variance of the vector  $\mathbf{x}$ , while  $\lambda_i/\text{tr}(\mathbf{V}_\mathbf{x})$  is the fraction of that total variance explained by the linear combination  $\mathbf{e}_i^T \mathbf{x}$ .

**Example A5.6.** Again let us consider the additive-genetic covariance matrix,  $\mathbf{G}$ , as shown in Examples A5.1 and A5.2. Because  $\lambda_1 \simeq 5.241$ ,  $\lambda_2 \simeq 0.765$ , and  $\text{tr}(\mathbf{G}) = 4 + 2 = 6$ , the first PC explains  $5.241/6 \simeq 0.8735$ , or 87% of the variance in  $\mathbf{G}$ . While the first PC accounts for the majority of variation over the entire space of the variables ( $\mathbf{x}$ ), the amount of variation explained by PC1 for any *particular* weighted combination,  $y = \mathbf{b}^T \mathbf{x}$ , of the original variables depends on the projection of  $\mathbf{b}$  onto PC1. For example, if  $\mathbf{b} = \mathbf{e}_2$  (the weight vector corresponds to the second eigenvector), then the projection of  $\mathbf{b}$  onto PC1 has a length of zero, because PC1 is orthogonal to  $\mathbf{e}_2$ , and hence PC1 explains none of the variation of this new variable.

Example A5.7 serves as a brief introduction to the important field of **morphometrics**, which is concerned with quantification and comparison of sizes and shapes of organisms. The reader is referred to Pimentel (1979), Reyment et al. (1984), Elewa (2004), Claude (2008) and especially Bookstein et al. (1985), Rohlf and Bookstein (1990), Reyment (1991), Bookstein (1997), Slice (2005), and Zelditch et al. (2012) for detailed treatments.

**Example A5.7.** Jolicoeur and Mosimann (1960) measured three carapace characters in 24 males of the painted turtle (*Chrysemys picta marginata*). Letting  $z_1$  be the carapace length,  $z_2$  be the maximum carapace width, and  $z_3$  be the carapace height, the resulting sample covariance matrix ( $\mathbf{S}_\mathbf{z}$ , the sample estimate of  $\mathbf{V}_\mathbf{z}$ ) for these data was found to be

$$\mathbf{S}_\mathbf{z} = \begin{pmatrix} 138.77 & 79.15 & 37.38 \\ 79.15 & 50.04 & 21.65 \\ 37.38 & 21.65 & 11.26 \end{pmatrix}$$

Hence,  $\text{tr}(\mathbf{S}_\mathbf{z}) = 138.77 + 50.04 + 11.26 = 200.07$ . Using R, the eigenvalues of  $\mathbf{S}_\mathbf{z}$  are found to be

$$\lambda_1 = 195.280, \quad \lambda_2 = 3.687, \quad \lambda_3 = 1.103$$

which (as expected) sum to the value of the trace, 200.07. The associated (normalized) eigenvectors are similarly found to be

$$\mathbf{e}_1 = \begin{pmatrix} 0.840 \\ 0.492 \\ 0.229 \end{pmatrix}, \quad \mathbf{e}_2 = \begin{pmatrix} 0.488 \\ -0.870 \\ 0.079 \end{pmatrix}, \quad \mathbf{e}_3 = \begin{pmatrix} 0.213 \\ 0.043 \\ -0.971 \end{pmatrix}$$

PC1 accounts for 97.6% of the variation ( $195.281/200.07 = 0.976$ ), while PC2 and PC3 account for 1.84% and 0.55%, respectively. Jolicoeur and Mosimann interpret PC1 as measuring overall size, as the new variable

$$y_1 = \mathbf{e}_1^T \mathbf{z} = 0.840z_1 + 0.492z_2 + 0.229z_3$$

corresponds to a simultaneous change in all three variables in the same direction, as is expected as individuals change their overall size. Likewise, PC2 and PC3 are

$$y_2 = \mathbf{e}_2^T \mathbf{z} = 0.488z_1 - 0.870z_2 + 0.079z_3$$

$$y_3 = \mathbf{e}_3^T \mathbf{z} = 0.213z_1 + 0.043z_2 - 0.971z_3$$

which Jolicoeur and Mosimann interpreted as measures of shape. Because the coefficient on  $z_3$  is small relative to the others in PC2, they interpret PC2 as measuring the tradeoff between length ( $z_1$ ) and width ( $z_2$ ). Thus, after removing the variation in size, 1.84% of the remaining variation can be accounted for by differences in the shape measured by length versus width. Likewise, because the PC3 coefficient for  $z_2$  is very small, PC3 mainly measures shape differences due to length ( $z_1$ ) versus height ( $z_3$ ).

This example points out some of the advantages, and possible pitfalls, of using principal component analysis for dimensional reduction of the data. Namely, replacing the  $n$ -component vector  $\mathbf{z}$  by an  $m < n$  component vector  $\mathbf{y}$  composed of linear combinations of the  $\mathbf{z}$ , i.e.,  $\mathbf{y}_{m \times 1} = \mathbf{M}_{m \times n} \mathbf{z}_{n \times 1}$ , where  $\mathbf{M} = (\mathbf{e}_1, \dots, \mathbf{e}_m)^T$ , with  $y_i = \mathbf{e}_i^T \mathbf{z}$ . Essentially all (over 97%) of the variance in the three measured characters is accounted for by variation in overall size, with the remaining variation accounted for by differences in shape. While the temptation is strong to simply consider overall size and ignore all shape information, it might be the case that selection is largely ignoring variation in size and instead focusing on (size-independent) shape differences. In this case, an analysis ignoring shape (as would occur if only the new character generated by PC1 were considered) would be very misleading. A further complication with principal component analysis is that it can often be difficult to give biological interpretations to the new characters resulting from the rotation of the coordinate system.

## TESTING FOR MULTIVARIATE NORMALITY

Multivariate normality is often assumed in statistical procedures, but it is less often tested. In LW Chapter 11 we briefly discussed two approaches for testing univariate normality, one graphical and the other based on deviations of observed skewness and/or kurtosis from Gaussian expectations. As we now demonstrate, both of these approaches can be extended to testing for multivariate normality. Additional methods are reviewed by Malkovich and Afifi (1973), Gnanadesikan (1977), Cox and Small (1978), Seber (1984), Looney (1995), and Henze (2002).

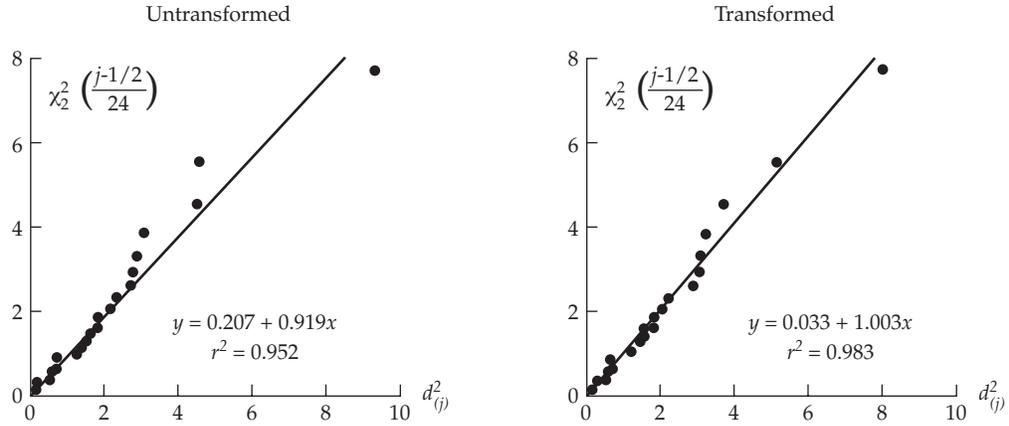
### Graphical Tests: Chi-square Plots

A fairly simple graphical test can be developed by extending the notion of the normal probability plot that is used to check univariate normality (LW Chapter 11), where observations were ranked and then plotted against their ranked expected values under normality. Departures from linearity signify departures from normality, and we can apply this same approach to check for multivariate normality. From Equation A5.20d, if  $\mathbf{z} \sim \text{MVN}(\boldsymbol{\mu}, \mathbf{V}_{\mathbf{z}})$ , then each element of the vector

$$\mathbf{y} = \mathbf{A}^{-1/2} \mathbf{U}^T (\mathbf{z} - \boldsymbol{\mu})$$

is an independent unit normal, so that  $\mathbf{y} \sim \text{MVN}(\mathbf{0}, \mathbf{I})$ . Recalling that  $\mathbf{U}^{-1} = \mathbf{U}^T$ , we can rearrange this expression to yield

$$(\mathbf{z} - \boldsymbol{\mu}) = \mathbf{U} \mathbf{A}^{1/2} \mathbf{y}$$



**Figure A5.7** Plots of ranked distance data ( $d_{(j)}^2$  being the  $j$ th smallest distance) versus the expected corresponding  $\chi^2$  value for the data of Jolicoeur and Mosimann from Example A5.8. **Left:** The untransformed data do not appear to depart significantly from linearity, although they depart slightly from the intercept (0) and slope (1) of the expected regression under multivariate normality. **Right:** Log-transforming the data gives a slightly better linear fit ( $r^2 = 0.983$  versus  $r^2 = 0.952$ ), with the best-fitting line passing through the origin as expected if the distance data follow a  $\chi^2$  distribution, and has a slope of essentially one. See Example A5.8 for more details.

Using this result and recalling Equation A5.11a, we have that

$$\begin{aligned}
 (\mathbf{z} - \boldsymbol{\mu})^T \mathbf{V}_{\mathbf{z}}^{-1} (\mathbf{z} - \boldsymbol{\mu}) &= (\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{y})^T (\mathbf{U}\boldsymbol{\Lambda}^{-1}\mathbf{U}^T) (\mathbf{U}\boldsymbol{\Lambda}^{1/2}\mathbf{y}) \\
 &= \mathbf{y}^T \boldsymbol{\Lambda}^{1/2} (\mathbf{U}^T\mathbf{U}) \boldsymbol{\Lambda}^{-1} (\mathbf{U}^T\mathbf{U}) \boldsymbol{\Lambda}^{1/2}\mathbf{y} \\
 &= \mathbf{y}^T \mathbf{y} = \sum_{i=1}^n y_i^2
 \end{aligned}
 \tag{A5.22}$$

Thus if  $\mathbf{z} \sim \text{MVN}$ , the quadratic form given by Equation A5.22 is the sum of  $n$  independent squared unit normal random variables. By definition, this sum is a  $\chi^2$  random variable with  $n$  degrees of freedom (LW Appendix 5), suggesting that one test for multivariate normality is to compare the goodness of fit of the scaled distances

$$d_i^2 = (\mathbf{z}_i - \bar{\mathbf{z}})^T \mathbf{S}_{\mathbf{z}}^{-1} (\mathbf{z}_i - \bar{\mathbf{z}}) \tag{A5.23}$$

to those generated by  $n$  (rank-ordered) draws from a  $\chi_n^2$ . Here  $\mathbf{z}_i$  is the vector of observations from the  $i$ th individual,  $\bar{\mathbf{z}}$  the vector of sample means, and  $\mathbf{S}_{\mathbf{z}}^{-1}$  the inverse of the sample covariance matrix. Note that the  $d_i$  are simply the squared Mahalanobis distances (Equation A5.19). We use the term *distance* because when  $\mathbf{z}$  is transformed to  $\mathbf{y}$ ,  $\mathbf{V}_{\mathbf{y}} = \mathbf{I}$ , giving the variance of the linear combination  $\mathbf{c}^T \mathbf{y}$  as  $\mathbf{c}^T \mathbf{V}_{\mathbf{y}} \mathbf{c} = \mathbf{c}^T \mathbf{I} \mathbf{c} = \|\mathbf{c}\|^2$ . Thus, regardless of orientation, any two  $\mathbf{y}$  vectors having the same length also have the same variance, which equals their squared Euclidean distance.

The regression test for multivariate normality is based on *ordered* distances. Hence, we first order the distances generated by Equation A5.23 from smallest to largest,

$$d_{(1)}^2 \leq d_{(2)}^2 \leq \dots \leq d_{(m)}^2$$

where  $m$  is the number of individuals sampled. Note that we use the subscription notation where  $d_{(j)}^2$  denotes the  $j$ th smallest *distance* (the  $j$ th smallest value of Equation A5.23), whereas  $d_i^2$  is the distance associated with the vector of observations for the  $i$ th *observation*.

Let  $\chi_n^2(\alpha)$  correspond to the value of a chi-square random variable,  $X$ , with  $n$  degrees of freedom that satisfies  $\text{Prob}[X \leq \chi_n^2(\alpha)] = \alpha$ . Under multivariate normality, we expect the points

$$\left( d_{(i)}^2, \chi_n^2 \left[ \frac{i - 1/2}{m} \right] \right) \quad \text{for } 1 \leq i \leq m$$

to fall along a line with a slope of one and an intercept of zero, as the  $i$ th ordered distance has  $i/m$  observations less than or equal to it (the factor of  $1/2$  is added as a correction for continuity). As with univariate normal probability plots, departures from multivariate normality are indicated by departures from linearity. More formally, one can use a standard **Kolmogorov–Smirnov test** (Conover 1999) for comparing two distributions to compare the goodness-of-fit of these ordered distances with a  $\chi_n^2$ .

**Example A5.8.** Consider again the data of Jolicoeur and Mosimann (1960) on carapace characters in 24 male turtles. Are the characters  $z_1$  (carapace length) and  $z_2$  (maximun carapace width) jointly bivariate normally distributed? Here  $n = 2$  and  $m = 24$  and

$$\bar{\mathbf{z}} = \begin{pmatrix} 113.13. \\ 88.29 \end{pmatrix}, \quad \mathbf{S}_{\mathbf{z}} = \begin{pmatrix} 138.77 & 79.15 \\ 79.15 & 50.04 \end{pmatrix}, \quad \mathbf{S}_{\mathbf{z}}^{-1} = \begin{pmatrix} 0.0737 & -0.1165 \\ -0.1165 & 0.2043 \end{pmatrix}$$

where  $\mathbf{S}_{\mathbf{z}}$  is the sample covariance matrix. A partial list of the 24 vectors of observations is

$$\mathbf{z}_1 = \begin{pmatrix} 93 \\ 74 \end{pmatrix}, \quad \dots, \quad \mathbf{z}_{11} = \begin{pmatrix} 113 \\ 88 \end{pmatrix}, \quad \dots, \quad \mathbf{z}_{24} = \begin{pmatrix} 135 \\ 106 \end{pmatrix}$$

Applying Equation A5.23, these observations translate into the distances

$$d_1^2 = 4.45, \quad \dots, \quad d_{11}^2 = 0.002, \quad \dots, \quad d_{24}^2 = 9.277$$

After rank ordering, these correspond to  $d_{(23)}^2$ ,  $d_{(1)}^2$ , and  $d_{(24)}^2$ , respectively. For  $d_{(23)}^2$ , the matching value when distances are  $\chi^2$ -distributed is

$$\chi_2^2 \left( \frac{23 - 1/2}{24} \right) = \chi_2^2(0.9375)$$

The R command `qchisq(0.9375, 2)` returns a value of  $x = 5.545$ , which satisfies  $\text{Pr}(\chi_2^2 \leq x) = 0.9375$ , and calculates the point generated from  $\mathbf{z}_1$  as  $(4.45, 5.545)$ . Likewise, the  $\chi^2$  values for  $d_{(1)}^2$  and  $d_{(24)}^2$  are 0.043 and 7.742, respectively. Proceeding similarly for the other values, we obtain the regression plotted in Figure A5.7. This departs somewhat from linearity. Further, under the assumption of multivariate normality, the best-fitting linear regression is expected to have a slope of one and to pass through the origin, while the best linear fit of these data shows slight departures from these values. Transforming the data by taking logs results in a slightly better fit (Figure A5.7).

### Mardia’s Test: Multivariate Skewness and Kurtosis

As was the case for univariate normality, we can test for multivariate normality by examining the sample skewness and kurtosis. Mardia (1970, 1974) proposed multivariate extensions of skewness and kurtosis measures and suggested a large-sample test based on the asymptotic distribution of these statistics. If there are  $m$  vectors of observations (with each vector measuring  $n$  characters), then the multivariate skewness is estimated by

$$b_{1,n} = \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m \left[ (\mathbf{z}_i - \bar{\mathbf{z}})^T \mathbf{S}_{\mathbf{z}}^{-1} (\mathbf{z}_j - \bar{\mathbf{z}}) \right]^3 \tag{A5.24a}$$

while the multivariate kurtosis is estimated by

$$b_{2,n} = \frac{1}{m} \sum_{i=1}^m \left[ (\mathbf{z}_i - \bar{\mathbf{z}})^T \mathbf{S}_{\mathbf{z}}^{-1} (\mathbf{z}_i - \bar{\mathbf{z}}) \right]^2 \quad (\text{A5.24b})$$

If  $\mathbf{z} \sim \text{MVN}$ , then  $b_{1,n}$  and  $b_{2,n}$  have expected values 0 and  $n(n+2)$ . For large values of  $m$ , Mardia showed that the (scaled) multivariate skewness is asymptotically distributed as a chi-square random variable with  $f$  degrees of freedom, with

$$\frac{m}{6} b_{1,n} \sim \chi_f^2, \quad \text{where } f = \frac{n(n+1)(n+2)}{6} \quad (\text{A5.25a})$$

Likewise for large values of  $m$ , the multivariate kurtosis (following appropriate scaling) is distributed as a unit-normal, with

$$\frac{b_{2,n} - n(n+2)}{\sqrt{8n(n+2)/m}} \sim N(0, 1) \quad (\text{A5.25b})$$

If either Equation A5.25a or A5.25b is significant, then multivariate normality is rejected.

**Example A5.9.** Do the data considered in Example A5.8 display significant skewness or kurtosis? Here  $n = 2$  and  $m = 24$ . Applying Equations A5.25a and A5.25b gives  $b_{1,2} = 0.6792$  and  $b_{2,2} = 7.6043$ . Considering skewness first, from Equation A5.25a it follows that the value

$$\frac{m}{6} b_{1,2} = \frac{24}{6} 0.6792 = 2.717$$

is (under MVN) a draw from a chi-square distribution with  $f = 2(2+1)(2+2)/6 = 4$  degrees of freedom. Because  $\text{Prob}(\chi_4^2 \geq 2.717) \simeq 0.606$ , this is not significant. Turning to kurtosis, Equation A5.25b yields

$$\frac{b_{2,n} - n(n+2)}{\sqrt{8n(n+2)/m}} = \frac{7.6043 - 8}{1.633} \simeq -0.2423$$

which is also not significant as  $\text{Prob}(|N(0, 1)| \geq 0.2423) \simeq 0.81$ . Transforming the data by taking logs gives  $b_{1,2} = 0.2767$  and  $b_{2,2} = 7.1501$ , and hence showing a slight decrease in skewness and a slight increase in kurtosis relative to the untransformed data. Reyment (1971) presented a number of other biological examples using Mardia's test.