Chapter

4 *Multidimensional quantitative data*

4.0 Multidimensional statistics

Basic statistics are now part of the curriculum of most ecologists. However, statistical techniques based on such simple distributions as the unidimensional normal distribution are not really appropriate for analysing complex ecological data sets. Nevertheless, researchers sometimes perform series of simple analyses on the various descriptors in the data set, expecting to obtain results that are pertinent to the problem under study. This type of approach is incorrect because it does not take into account the covariance among descriptors; see also Box 1.3 where the statistical problem created by multiple testing is explained. In addition, such an approach only extracts minimum information from data which have often been collected at great cost and it usually generates a mass of results from which it is difficult to draw much sense. Finally, in studies involving species assemblages, it is usually more interesting to describe the variability of the structure of the assemblage as a whole (i.e. *mensurative* variation observed through space or time, or *manipulative* variation resulting from experimental manipulation; Hurlbert, 1984) than to look at each species independently.

Fortunately, methods derived from *multidimensional statistics*, which are used throughout this book, are designed for analysing complex data sets. These methods take into account the co-varying nature of ecological data and can evidence the structures that underlie the data. The present chapter discusses the basic theory and characteristics of multidimensional data analysis. Mathematics are kept to a minimum, so that readers can easily reach a high level of understanding. Many approaches of practical interest are discussed, including several types of linear correlation, with their statistical tests. It must be noted that this chapter is limited to linear statistics.

A number of excellent textbooks deal with detailed aspects of multidimensional statistics. For example, formal presentations of the subject are found in Muirhead (1982) and Anderson (1984). Researchers less interested in mathematical theory may refer to Cooley & Lohnes (1971), Tatsuoka (1971), Press (1972), Graybill (1983), or

Table 4.1 Numerical example of two species observed at four sampling sites. Figure 4.1 shows that each row of the data matrix may be construed as a vector, as defined in Section 2.4.

Sampling sites (objects)	Species (descriptors) 1	2	(p = 2)
1	5	1	
2	3	2	
3	8	3	
4	6	4	
(n = 4)			

Morrison (1990). These books describe a number of useful methods, among which the multidimensional analysis of variance. However, none of these books specifically deals with ecological data.

Multidimensional Multivariate

Several authors use the term *multivariate* as an abbreviation for *multidimensional variate* (the latter term meaning *random variable*; Section 1.0). As an adjective, *multivariate* is interchangeable with *multidimensional.*

4.1 Multidimensional variables and dispersion matrix

As stated in Section 1.0, the present textbook deals with the analysis of *random variables*. Ecological data matrices have n rows and p columns (Section 2.1). Each row is a *vector* (Section 2.4) which is, statistically speaking, one realization of a p-dimensional random variable. In other words, for example, when p species are observed at n sampling sites, the species are the p dimensions of a random variable "species" and each site is one realization of this p-dimensional random variable.

To illustrate this concept, four sampling units with two species (Table 4.1) are plotted in a two-dimensional Euclidean space (Fig. 4.1). Vector "site 1" is the doublet (5,1). It is plotted in the same two-dimensional space as the three other vectors "site i". Each row of the data matrix is a two-dimensional vector, which is one realization of the (bivariate) random variable "species". The random variable "species" is said to be two-dimensional because the sampling units (objects) contain two species (descriptors), the two dimensions being species 1 and 2, respectively.

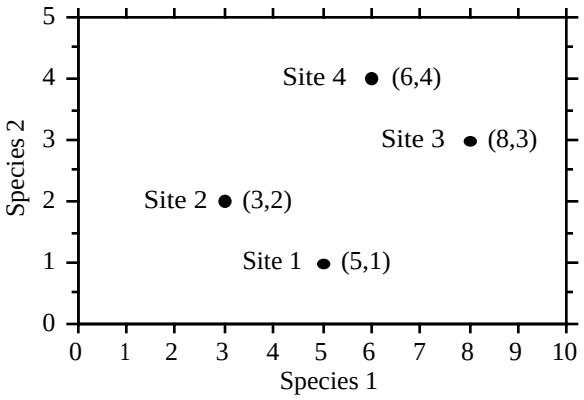

Figure 4.1 Four realizations (sampling sites from Table 4.1) of the two-dimensional random variable "species" are plotted in a two-dimensional Euclidean space.

As the number of descriptors (e.g. species) increases, the number of dimensions of the random variable "species" similarly increases, so that more axes are necessary to construct the space in which the objects are plotted. Thus, the *p* descriptors make up a *p*-dimensional random variable and the *n* vectors of observations are as many realizations of the *p*-dimensional vector "descriptors". The present chapter does not deal with *samples* of observations, which result from field or laboratory work (for a brief discussion on sampling, see Section 1.1), but it focuses instead on *populations*, which are investigated by means of the samples.

Before approaching the multidimensional normal distribution, it is necessary to define a *p*-dimensional random variable "descriptors":

$$\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, \ldots \mathbf{y}_j, \ldots \mathbf{y}_p] \qquad \textbf{(4.1)}$$

Each element $\mathbf{y}_j$ of the multidimensional table **Y** is a unidimensional random variable. Every descriptor $\mathbf{y}_j$ is observed in each of the *n* vectors "object", each sampling unit *i* providing one realization of the *p*-dimensional random variable (Fig. 4.2).

In ecology, the structure of *dependence* among descriptors is, in many instances, the matter being investigated. Researchers who study multidimensional data sets using univariate statistics assume that the *p* unidimensional $\mathbf{y}_j$ variables in **Y** are *independent* of one another (this refers to the third meaning of *independence* in Box 1.1). This is the reason why univariate statistical methods are inappropriate with most ecological data and why methods that take into account the *dependence* among descriptors must be used when analysing sets of multidimensional data. Only these methods will generate proper results when there is dependence among descriptors; it is never acceptable to replace a multidimensional analysis by a series of unidimensional treatments.

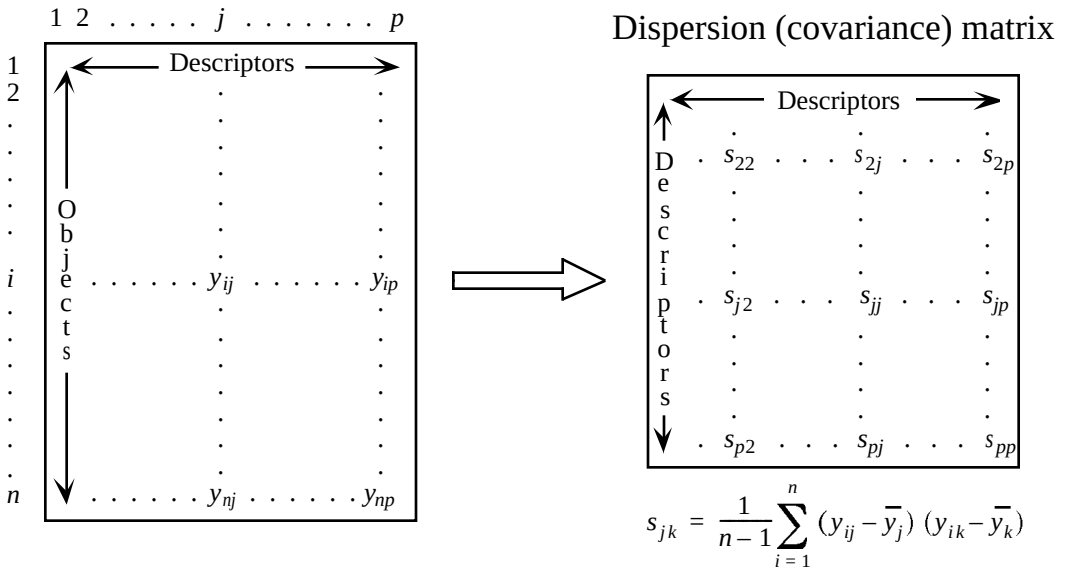

Figure 4.2 Structure of ecological data. Given their nature, ecological descriptors are *dependent* of one another. In statistics, the objects are often assumed to be *independent* observations, but this is generally not the case in ecology (Section 1.1)

The usual tests of significance require, however, "that successive sample observation vectors from the multidimensional population have been drawn in such a way that they can be construed as realizations of independent random vectors" (Morrison, 1990, p. 80). Section 1.1 has shown that this assumption of independence among observations is most often not realistic in ecology. Lack of independence among the observations (data rows) does not really matter when statistical models are used for descriptive purposes only, as it is often the case in the present book. For statistical testing, however, corrected tests of significance have to be used when the observations are autocorrelated (Section 1.1).

To sum up: (1) the p descriptors in ecological data matrices are the p dimensions of a random variable "descriptors"; (2) in general, the p descriptors are *not linearly independent* of one another; methods of multidimensional analysis are designed to bring out the structure of linear dependence among descriptors; (3) each of the n sampling units is a realization of the p-dimensional vector "descriptors"; (4) the usual tests of significance assume that the n sampling units are realizations of *independent* random vectors. The latter condition is generally not met in ecology, with consequences that were mentioned in the previous paragraph and discussed in Section 1.1. For the various meanings of the term *independence* in statistics, see Box 1.1.

The dependence among quantitative variables $\mathbf{y}_j$ brings up the concept of *covariance*. Covariance is the extension, to two descriptors, of the concept of *variance*. Variance is a measure of the *dispersion* of a random variable $\mathbf{y}_j$ around its mean; it is denoted σ_j^2. Covariance measures the *joint dispersion* of two random variables $\mathbf{y}_j$ and $\mathbf{y}_k$ around their means; it is denoted σ_{jk}. The *dispersion matrix* of $\mathbf{Y}$, called matrix $\boldsymbol{\Sigma}$ (sigma), contains the variances and covariances of the p descriptors (Fig. 4.2):

Dispersion matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} \qquad \textbf{(4.2)}$$

Matrix $\boldsymbol{\Sigma}$ is an *association matrix* [descriptors × descriptors] (Section 2.2). The elements σ_{jk} of matrix $\boldsymbol{\Sigma}$ are the covariances between all pairs of the p random variables. The matrix is symmetric because the covariance of $\mathbf{y}_j$ and $\mathbf{y}_k$ is identical to that of $\mathbf{y}_k$ and $\mathbf{y}_j$. A diagonal element of $\boldsymbol{\Sigma}$ is the covariance of a descriptor $\mathbf{y}_j$ with itself, which is the variance of $\mathbf{y}_j$, so that $\sigma_{jj} = \sigma_j^2$

Variance

The *estimate* of the *variance* of $\mathbf{y}_j$, denoted s_j^2, is computed on the *centred variable* $(y_{ij} - \bar{y}_j)$. Variable $\mathbf{y}_j$ is centred by subtracting the mean $\bar{y}_j$ from each of the n observations y_{ij}. As a result, the mean of the centred variable is zero. The unbiased estimator of the population variance s_j^2 is computed using the well-known formula:

$$s_j^2 = \frac{1}{n-1} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 \qquad \textbf{(4.3)}$$

where the sum of *squares* of the *centred data*, for descriptor *j*, is divided by the number of objects minus one ($n - 1$). The summation is over the n observations of descriptor *j*.

Covariance

In the same way, the estimate (s_{jk}) of the *covariance* (σ_{jk}) of $\mathbf{y}_j$ and $\mathbf{y}_k$ is computed on the centred variables $(y_{ij} - \bar{y}_j)$ and $(y_{ik} - \bar{y}_k)$, using the formula of a "bivariate variance". The *covariance* s_{jk} is calculated as:

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k) \qquad \textbf{(4.4)}$$

Standard deviation

When $k = j$, eq. 4.4 is identical to eq. 4.3. The positive square root of the variance is called the *standard deviation* (σ_j). Its estimate s_j is thus:

$$s_j = \sqrt{s_j^2} \qquad \textbf{(4.5)}$$

Table 4.2 Symbols used to identify (population) parameters and (sample) statistics.

	Parameter		Statistic	
	Matrix or vector	Elements	Matrix or vector	Elements
Covariance	$\boldsymbol{\Sigma}$ (sigma)	σ_{jk} (sigma)	$\mathbf{S}$	s_{jk}
Correlation	$\mathbf{P}$ (rho)	ρ_{jk} (rho)	$\mathbf{R}$	r_{jk}
Mean	$\boldsymbol{\mu}$ (mu)	μ_j (mu)	$\bar{\mathbf{y}}$	$\bar{y}_j$

The symbols for matrix $\boldsymbol{\Sigma}$ and summation $\sum$ should not be confused.

Coefficient of variation

The coefficient of variation is a dimensionless measure of variation. *CV* is used to compare the variation of variables expressed in different physical units. It is obtained by dividing the standard deviation s_j by the mean $\bar{x}_j$ of variable j:

$$CV_j = s_j / \bar{x}_j$$

Since the standard deviation and the mean of a variable have the same physical units, CV_j is dimensionless. CV_j is only defined for quantitative variables that have non-zero means and it does not makes sense for interval-scale variables (Subsection 1.4.1), for which the value of the mean is arbitrary. The coefficient of variation may be rescaled to percentages by multiplying its value by 100. For small n, an estimate with reduced bias is obtained by multiplying *CV* by $(1 + 1/(4n))$.

Contrary to the variance, which is always positive, the covariance may take positive or negative values. To understand the meaning of the covariance, imagine that the object points are plotted in a scatter diagram where the axes are descriptors $\mathbf{y}_j$ and $\mathbf{y}_k$. The data are centred by drawing new axes, whose origin is at the centroid $(\bar{y}_j, \bar{y}_k)$ of the cloud of points (centred plots of that kind with positive and negative correlations are shown in Fig. 4.7). A positive covariance means that most of the points are in quadrants I and III of the centred plot, where the centred values $(y_{ij} - \bar{y}_j)$ and $(y_{ik} - \bar{y}_k)$ have the same signs. This corresponds to a positive relationship between the two descriptors. The converse is true for a negative covariance, for which most of the points are in quadrants II and IV of the centred plot. When the covariance is null or small, the points are equally distributed among the four quadrants of the centred plot.

Parameter Statistic

Greek and roman letters are used here (Table 4.2). The properties of a *population* (called *parameters*) are denoted by *greek* letters. Their *estimates* (called *statistics*), computed from *samples*, are symbolized by the corresponding *roman* letters. These conventions are complemented by those pertaining to matrix notation (Section 2.1).

The dispersion matrix* **S** can be computed directly, by multiplying the *matrix of centred data* $[y - \bar{y}]$ with its transpose $[y - \bar{y}]'$:

$$\mathbf{S} = \frac{1}{n-1}[y - \bar{y}]'[y - \bar{y}] \qquad \textbf{(4.6)}$$

$$\mathbf{S} = \frac{1}{n-1}\begin{bmatrix} (y_{11}-\bar{y}_1) & (y_{21}-\bar{y}_1) & \cdots & (y_{n1}-\bar{y}_1) \\ (y_{12}-\bar{y}_2) & (y_{22}-\bar{y}_2) & \cdots & (y_{n2}-\bar{y}_2) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ (y_{1p}-\bar{y}_p) & (y_{2p}-\bar{y}_p) & \cdots & (y_{np}-\bar{y}_p) \end{bmatrix} \begin{bmatrix} (y_{11}-\bar{y}_1) & (y_{12}-\bar{y}_2) & \cdots & (y_{1p}-\bar{y}_p) \\ (y_{21}-\bar{y}_1) & (y_{22}-\bar{y}_2) & \cdots & (y_{2p}-\bar{y}_p) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ (y_{n1}-\bar{y}_1) & (y_{n2}-\bar{y}_2) & \cdots & (y_{np}-\bar{y}_p) \end{bmatrix}$$

$$\mathbf{S} = \frac{1}{n-1}\begin{bmatrix} \sum_{i=1}^{n}(y_{i1}-\bar{y}_1)^2 & \sum_{i=1}^{n}(y_{i1}-\bar{y}_1)(y_{i2}-\bar{y}_2) & \cdots & \sum_{i=1}^{n}(y_{i1}-\bar{y}_1)(y_{ip}-\bar{y}_p) \\ \sum_{i=1}^{n}(y_{i2}-\bar{y}_2)(y_{i1}-\bar{y}_1) & \sum_{i=1}^{n}(y_{i2}-\bar{y}_2)^2 & \cdots & \sum_{i=1}^{n}(y_{i2}-\bar{y}_2)(y_{ip}-\bar{y}_p) \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \sum_{i=1}^{n}(y_{ip}-\bar{y}_p)(y_{i1}-\bar{y}_1) & \sum_{i=1}^{n}(y_{ip}-\bar{y}_p)(y_{i2}-\bar{y}_2) & \cdots & \sum_{i=1}^{n}(y_{ip}-\bar{y}_p)^2 \end{bmatrix}$$

This elegant and rapid procedure shows once again the advantage of matrix algebra in numerical ecology, where the data sets are generally large.

Numerical example. Four species ($p = 4$) were observed at five stations ($n = 5$). The estimated population parameters, for the species, are the means ($\bar{y}_j$), the variances (s_j^2), and the covariances (s_{jk}). The original and centred data are shown in Table 4.3. Because $s_{jk} = s_{kj}$, the dispersion matrix is symmetric. The mean of each *centred variable* is zero.

In this numerical example, the covariance between species 2 and the other three species is zero. This does not necessarily mean that species 2 is independent of the other three, but simply that the joint *linear* dispersion of species 2 with any one of the other three is zero. This example will be revisited in Section 4.2.

* Some authors call $[y - \bar{y}]'[y - \bar{y}]$ a *dispersion matrix* and **S** a *covariance matrix*. For these authors, a covariance matrix is then a dispersion matrix divided by $(n - 1)$.

Table 4.3 Numerical example. Calculation of centred data and covariances.

Sites	Original data	Centred data
1 2 3 4 5	$\mathbf{Y} = \begin{bmatrix} 1 & 5 & 2 & 6 \\ 2 & 2 & 1 & 8 \\ 3 & 1 & 3 & 4 \\ 4 & 2 & 5 & 0 \\ 5 & 5 & 4 & 2 \end{bmatrix}$	$[y-\bar{y}] = \begin{bmatrix} -2 & 2 & -1 & 2 \\ -1 & -1 & -2 & 4 \\ 0 & -2 & 0 & 0 \\ 1 & -1 & 2 & -4 \\ 2 & 2 & 1 & -2 \end{bmatrix}$
Means	$\bar{\mathbf{y}}' = \begin{bmatrix} 3 & 3 & 3 & 4 \end{bmatrix}$	$[\overline{y-\bar{y}}]' = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}$
$n-1=4$	$\mathbf{S} = \frac{1}{n-1}[y-\bar{y}]'[y-\bar{y}] = \begin{bmatrix} 2.5 & 0 & 2 & -4 \\ 0 & 3.5 & 0 & 0 \\ 2 & 0 & 2.5 & -5 \\ -4 & 0 & -5 & 10 \end{bmatrix}$	

The square root of the determinant of the dispersion matrix $|\mathbf{S}|^{1/2}$ is known as the *generalized variance*. It is also equal to the square root of the product of the eigenvalues of **S**.

Any dispersion matrix **S** is *positive semidefinite* (Table 2.2). Indeed, the quadratic form of **S** ($p \times p$) with any real and non-null vector **t** (of size p) is:

$$\mathbf{t}'\mathbf{S}\mathbf{t}$$

This expression may be expanded using eq. 4.6:

$$\mathbf{t}'\mathbf{S}\mathbf{t} = \mathbf{t}'\frac{1}{n-1}[y-\bar{y}]'[y-\bar{y}]\mathbf{t}$$

$$\mathbf{t}'\mathbf{S}\mathbf{t} = \frac{1}{n-1}[[y-\bar{y}]\mathbf{t}]'[[y-\bar{y}]\mathbf{t}] = \text{a scalar}$$

This scalar is the variance of the variable resulting from the product **Yt**. Since a variance, which is a sum of squared values, can only be positive or null, it follows that:

$$\mathbf{t}'\mathbf{S}\mathbf{t} \geq 0$$

so that **S** is positive semidefinite.

An important property can be derived by computing the quadratic form of the dispersion matrix **S** using eq. 2.28 (right). In that case, $\mathbf{U}^{-1} = \mathbf{U}'$ because **S** is symmetric (property #7 of inverses, Section 2.8), and eq. 2.28 (right) becomes:

$$\mathbf{U}'\mathbf{S}\mathbf{U} = \mathbf{\Lambda}$$

As vector **t** in the quadratic form, use the successive eigenvectors $\mathbf{u}_j$ from matrix **U**. For each vector $\mathbf{u}_j$, the development above shows that

$$\mathbf{u}_j'\mathbf{S}\mathbf{u}_j \geq 0$$

Since $\mathbf{u}_j'\mathbf{S}\mathbf{u}_j = \lambda_j$, this demonstrates that *all the eigenvalues* λ_j *of* **S** *are positive or null*. This property of dispersion matrices is fundamental in numerical ecology: it allows one to partition the variance among real *principal axes* (Sections 4.4 and 9.1).

Ideally, matrix **S** (of order p) should be estimated from a number of observations n larger than the number of descriptors p. When $n \leq p$, the rank of matrix **S** is $n - 1$ and, consequently, only $n - 1$ of its rows or columns are independent, so that $p - (n - 1)$ null eigenvalues are produced. The only practical consequence of $n \leq p$ is thus the presence of null eigenvalues in the principal component solution (Section 9.1). The first few eigenvalues of **S**, which are generally those of interest, have positive eigenvalues.

4.2 Correlation matrix

The previous section has shown that the covariance provides information on the orientation of the cloud of data points in the space defined by the descriptors. That statistic, however, does not provide any information on the intensity of the relationship between variables $\mathbf{y}_j$ and $\mathbf{y}_k$. Indeed, the covariance may increase or decrease without changing the relationship between $\mathbf{y}_j$ and $\mathbf{y}_k$. For example, in Fig. 4.3, the two clouds of points correspond to different covariances (factor two in size, and thus in covariance), but the relationship between variables is identical (same shape). Since the covariance depends on the dispersion of points around the mean of each variable (i.e. their variances), determining the intensity of the relationship between variables requires to control for the variances.

The *covariance* measures the joint *dispersion* of two random variables around their means. The *correlation* is defined as a measure of the *dependence* between two random variables $\mathbf{y}_j$ and $\mathbf{y}_k$. As already explained in Section 1.5, it often happens that matrices of ecological data contain descriptors with no common scale, e.g. when some species are more abundant than others by orders of magnitude, or when the descriptors have different physical dimensions (Chapter 3). Calculating covariances on such variables obviously does not make sense, except if the descriptors are first reduced to a common scale. The procedure consists in centring all descriptors on a zero mean and reducing them to unit standard deviation (eq. 1.12). By using *standardized descriptors*,

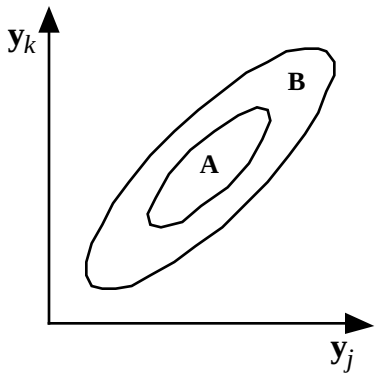

Figure 4.3 Several observations (objects), with descriptors $\mathbf{y}_j$ and $\mathbf{y}_k$, were made under two different sets of conditions (A and B). The two ellipses delineate clouds of point-objects corresponding to A and B, respectively. The covariance of $\mathbf{y}_j$ and $\mathbf{y}_k$ is twice as large for B as it is for A (larger ellipse), but the correlation between the two descriptors is the same in these two cases (i.e. the ellipses have the same shape).

it is possible to calculate meaningful covariances, because the new variables have the same scale (i.e. unit standard deviation) and are dimensionless (see Chapter 3).

Linear correlation

The covariance of two standardized descriptors is called the coefficient of linear correlation (Pearson *r*). This statistic has been proposed by the statistician Karl Pearson and is named after him. Given two standardized descriptors (eq. 1.12)

$$z_{ij} = \frac{y_{ij} - \bar{y}_j}{s_j} \qquad \text{and} \qquad z_{ik} = \frac{y_{ik} - \bar{y}_k}{s_k}$$

calculating their covariance (eq. 4.4) gives

$$s(z_j,z_k) = \frac{1}{n-1}\sum_{i=1}^{n} (z_{ij} - 0)\,(z_{ik} - 0) \qquad \text{because} \qquad \bar{z}_j = \bar{z}_k = 0$$

$$s(z_j,z_k) = \frac{1}{n-1}\sum_{i=1}^{n} \left(\frac{y_{ij} - \bar{y}_j}{s_j}\right)\left(\frac{y_{ik} - \bar{y}_k}{s_k}\right)$$

$$s(z_j,z_k) = \left(\frac{1}{s_j s_k}\right)\frac{1}{n-1}\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)\,(y_{ik} - \bar{y}_k)$$

$s(z_j,z_k) = \left(\frac{1}{s_j s_k}\right) s_{jk} = r_{jk}$, the *coefficient of linear correlation* between $\mathbf{y}_j$ and $\mathbf{y}_k$.

The developed formula is:

$$r_{jk} = \frac{s_{jk}}{s_j s_k} = \frac{\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)\,(y_{ik} - \bar{y}_k)}{\sqrt{\sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 \sum_{i=1}^{n} (y_{ik} - \bar{y}_k)^2}} \qquad \textbf{(4.7)}$$

Correlation matrix

As in the case of dispersion (Section 4.1), it is possible to construct the *correlation matrix* of $\mathbf{Y}$, i.e. the $\mathbf{P}$ (rho) matrix, whose elements are the coefficients of linear correlation ρ_{jk}:

$$\mathbf{P} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix} \qquad \textbf{(4.8)}$$

The *correlation matrix is the dispersion matrix of the standardized variables.* This concept will play a fundamental role in principal component analysis (Section 9.1). It should be noted that the diagonal elements of $\mathbf{P}$ are all equal to 1. This is because the comparison of any descriptor with itself is a case of complete dependence, which leads to a correlation $\rho_j = 1$. When $\mathbf{y}_j$ and $\mathbf{y}_k$ are independent of each other, $\rho_j = 0$. However, a correlation equal to zero does not necessarily imply that $\mathbf{y}_j$ and $\mathbf{y}_k$ are independent of each other, as shown by the following numerical example. A correlation $\rho_{jk} = -1$ is indicative of a complete, but inverse dependence of the two variables.

Numerical example. Using the values in Table 4.3, matrix **R** can easily be computed. Each element r_{jk} combines, according to eq. 4.7, the covariance s_{jk} with variances s_j and s_k:

$$\mathbf{R} = \begin{bmatrix} 1 & 0 & 0.8 & -0.8 \\ 0 & 1 & 0 & 0 \\ 0.8 & 0 & 1 & -1 \\ -0.8 & 0 & -1 & 1 \end{bmatrix}$$

Matrix **R** is symmetric, as was matrix **S**. The correlation $r = -1$ between species 3 and 4 means that these species are fully, but inversely, dependent (Fig. 4.4a). Correlations $r = 0.8$ and -0.8 are interpreted as indications of strong dependence between species 1 and 3 (direct) and species 1 and 4 (inverse), respectively. The *zero* correlation between species 2 and the other three species

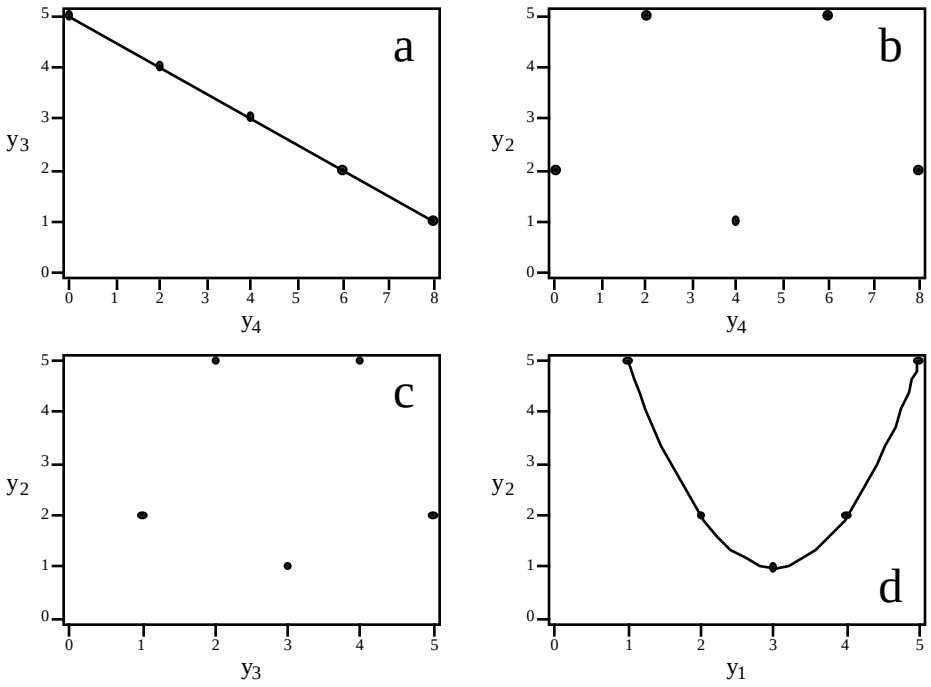

Figure 4.4 Numerical example. Relationships between species (a) 3 and 4, (b) 2 and 4, (c) 2 and 3, and (d) 2 and 1.

must be interpreted with caution. Figure 4.4d clearly shows that species 1 and 2 are completely *dependent* of each other since they are related by equation $y_2 = 1 + (3 - y_1)^2$; the zero correlation is, in this case, a consequence of the *linear* model underlying statistic *r*. Therefore, only those correlations which are *significantly* different from zero should be considered, since a null correlation has no unique interpretation.

Since the correlation matrix is the dispersion matrix of standardized variables, it is possible, as in the case of matrix **S** (eq. 4.6), to compute **R** directly by multiplying the *matrix of standardized data* with its transpose:

$$\mathbf{R} = \frac{1}{n-1}\left[(y-\bar{y})/s_y\right]'\left[(y-\bar{y})/s_y\right] = \frac{1}{n-1}\mathbf{Z}'\mathbf{Z} \qquad \textbf{(4.9)}$$

Table 4.4 shows how to calculate correlations r_{jk} of the example as in Table 4.3, using this time the *standardized data.* The mean of each *standardized variable is* zero and its standard deviation is equal to unity. The *dispersion* matrix of **Z** is identical to the *correlation* matrix of **Y**, which was calculated above using the covariances and variances.

Table 4.4 Numerical example. Calculation of standardized data and correlations.

Sites	Original data	Standardized data
1	$\mathbf{Y} = \begin{bmatrix} 1 & 5 & 2 & 6 \\ 2 & 2 & 1 & 8 \\ 3 & 1 & 3 & 4 \\ 4 & 2 & 5 & 0 \\ 5 & 5 & 4 & 2 \end{bmatrix}$	$\mathbf{Z} = \begin{bmatrix} -1.27 & 1.07 & -0.63 & 0.63 \\ -0.63 & -0.53 & -1.27 & 1.27 \\ 0 & -1.07 & 0 & 0 \\ 0.63 & -0.53 & 1.27 & -1.27 \\ 1.27 & 1.07 & 0.63 & -0.63 \end{bmatrix}$
2		
3		
4		
5		
Means	$\bar{\mathbf{y}}' = \begin{bmatrix} 3 & 3 & 3 & 4 \end{bmatrix}$	$\bar{\mathbf{z}}' = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}$
$n - 1 = 4$	$\mathbf{R}(y) = \mathbf{S}(z) = \frac{1}{n-1}\mathbf{Z}'\mathbf{Z} = \begin{bmatrix} 1 & 0 & 0.8 & -0.8 \\ 0 & 1 & 0 & 0 \\ 0.8 & 0 & 1 & -1 \\ -0.8 & 0 & -1 & 1 \end{bmatrix}$	

Matrices $\boldsymbol{\Sigma}$ and $\mathbf{P}$ are related to each other by the diagonal matrix of standard deviations of $\mathbf{Y}$. This new matrix, which is specifically designed for relating $\boldsymbol{\Sigma}$ and $\mathbf{P}$, is symbolized by $\mathbf{D}(\sigma)$ and its inverse by $\mathbf{D}(\sigma)^{-1}$:

$$\mathbf{D}(\sigma) = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & \cdot & \dots & \sigma_p \end{bmatrix} \quad \text{and} \quad \mathbf{D}(\sigma)^{-1} = \begin{bmatrix} 1/\sigma_1 & 0 & \dots & 0 \\ 0 & 1/\sigma_2 & \dots & 0 \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdots & \cdot \\ 0 & \cdot & \dots & 1/\sigma_p \end{bmatrix}$$

Using these two matrices, one can write:

$$\mathbf{P} = \mathbf{D}(\sigma^2)^{-1/2}\,\boldsymbol{\Sigma}\,\mathbf{D}(\sigma^2)^{-1/2} = \mathbf{D}(\sigma)^{-1}\,\boldsymbol{\Sigma}\,\mathbf{D}(\sigma)^{-1} \qquad \textbf{(4.10)}$$

where $\mathbf{D}(\sigma^2)$ is the matrix of the diagonal elements of $\boldsymbol{\Sigma}$. It follows from eq. 4.10 that:

$$\boldsymbol{\Sigma} = \mathbf{D}(\sigma)\,\mathbf{P}\,\mathbf{D}(\sigma) \qquad \textbf{(4.11)}$$

Significance of *r*

The theory underlying tests of significance is discussed in Section 1.2. In the case of *r*, inference about the statistical population is in most instances through the null hypothesis H_0: $\rho = 0$. H_0 may also state that ρ has some other value than zero, which would be derived from ecological hypotheses. The general formula for testing correlation coefficients is given in Section 4.5 (eq. 4.39). The Pearson correlation coefficient r_{jk} involves two descriptors (e.g. $\mathbf{y}_j$ and $\mathbf{y}_k$, hence $m = 2$ when testing a coefficient of simple linear correlation using eq. 4.39), so that $\nu_1 = 2 - 1 = 1$ and $\nu_2 = n - 2 = \nu$. The general formula then becomes:

$$F = \frac{r_{jk}^2 / 1}{(1 - r_{jk}^2) / \nu} = \nu \frac{r_{jk}^2}{1 - r_{jk}^2} \qquad \textbf{(4.12)}$$

where $\nu = n - 2$. Statistic F is tested against $F_{\alpha[1,\nu]}$.

Since the square root of a statistic $F_{[\nu_1,\nu_2]}$ is a statistic $t_{[\nu = \nu_2]}$ when $\nu_1 = 1$, *r* may also be tested using:

$$t = \frac{r_{jk}\sqrt{\nu}}{\sqrt{1 - r_{jk}^2}} \qquad \textbf{(4.13)}$$

The t statistic is tested against the value $t_{\alpha[\nu]}$. In other words, H_0 is tested by comparing the F (or t) statistic to the value found in a table of critical values of F (or t). Results of tests with eqs. 4.12 and 4.13 are identical. The number of degrees of freedom is $\nu = (n - 2)$ because calculating a correlation coefficient requires prior estimation of two parameters, i.e. the means of the two populations (eq. 4.7). H_0 is rejected when the probability corresponding to F (or t) is smaller than a predetermined *level of significance* (α for a two-tailed test, and $\alpha/2$ for a one-tailed test; the difference between the two types of tests is explained in Section 1.2). In principle, this test requires that the sample of observations be drawn from a population with a *bivariate normal distributio*n (Section 4.3). Testing for normality and multinormality is discussed in Section 4.7, and normalizing transformations in Section 1.5. When the data do not satisfy the condition of normality, t can be tested by randomization, as shown in Section 1.2.

Test of independence of variables

It is also possible to test the *independence of all variables* in a data matrix by considering the set of all correlation coefficients found in matrix **R**. The null hypothesis here is that the $p(p - 1)/2$ coefficients are all equal to zero, H_0: **R** = **I** (unit matrix). According to Bartlett (1954), the determinant of **R**, $|\mathbf{R}|$, can be transformed into a X^2 (chi-square) test statistic:

$$X^2 = -[n - (2p + 5)/6] \ln |\mathbf{R}| \qquad \textbf{(4.14)}$$

where $\ln |\mathbf{R}|$ is the natural logarithm of the determinant of **R**. This statistic is approximately distributed as χ^2 with $\nu = p(p - 1)/2$ degrees of freedom. When the probability associated with X^2 is significantly low, the null hypothesis of complete independence of the p descriptors is rejected. In principle, this test requires the

Table 4.5 Main properties of the coefficient of linear correlation. Some of these properties are discussed in later sections.

	Properties	Sections
1.	The coefficient of linear correlation measures the *intensity of the linear relationship* between two random variables.	4.2
2.	The coefficient of linear correlation between two variables can be calculated using their respective *variances* and their *covariance.*	4.2
3.	The correlation matrix is the *dispersion* matrix of *standardized variables.*	4.2
4.	The square of the coefficient of linear correlation is the *coefficient of determination*. It measures how much of the variance of each variable is explained by the other.	10.3
5.	The coefficient of linear correlation is a *parameter* of a multinormal distribution.	4.3
6.	The coefficient of linear correlation is the *geometric mean* of the *coefficients of linear regression* of each variable on the other.	10.3

observations to be drawn from a population with a *multivariate normal distributio*n (Section 4.3). If the null hypothesis of independence of all variables is rejected, the $p(p-1)/2$ correlation coefficients in matrix $\mathbf{R}$ may be tested individually; see Box 1.3 about multiple testing.

Other correlation coefficients are described in Sections 4.5 and 5.2. Wherever the coefficient of linear correlation must be distinguished from other coefficients, it is referred to as *Pearson's r*. In other instances, r is simply called the *coefficient of linear correlation* or *correlation coefficient*. Table 4.5 summarizes the main properties of this coefficient.