BIPLOT GRAPHICAL DISPLAY TECHNIQUES

by

KAREN ILONI

THESIS

Submitted in fulfilment of the requirements for the degree of

MASTER OF SCIENCE

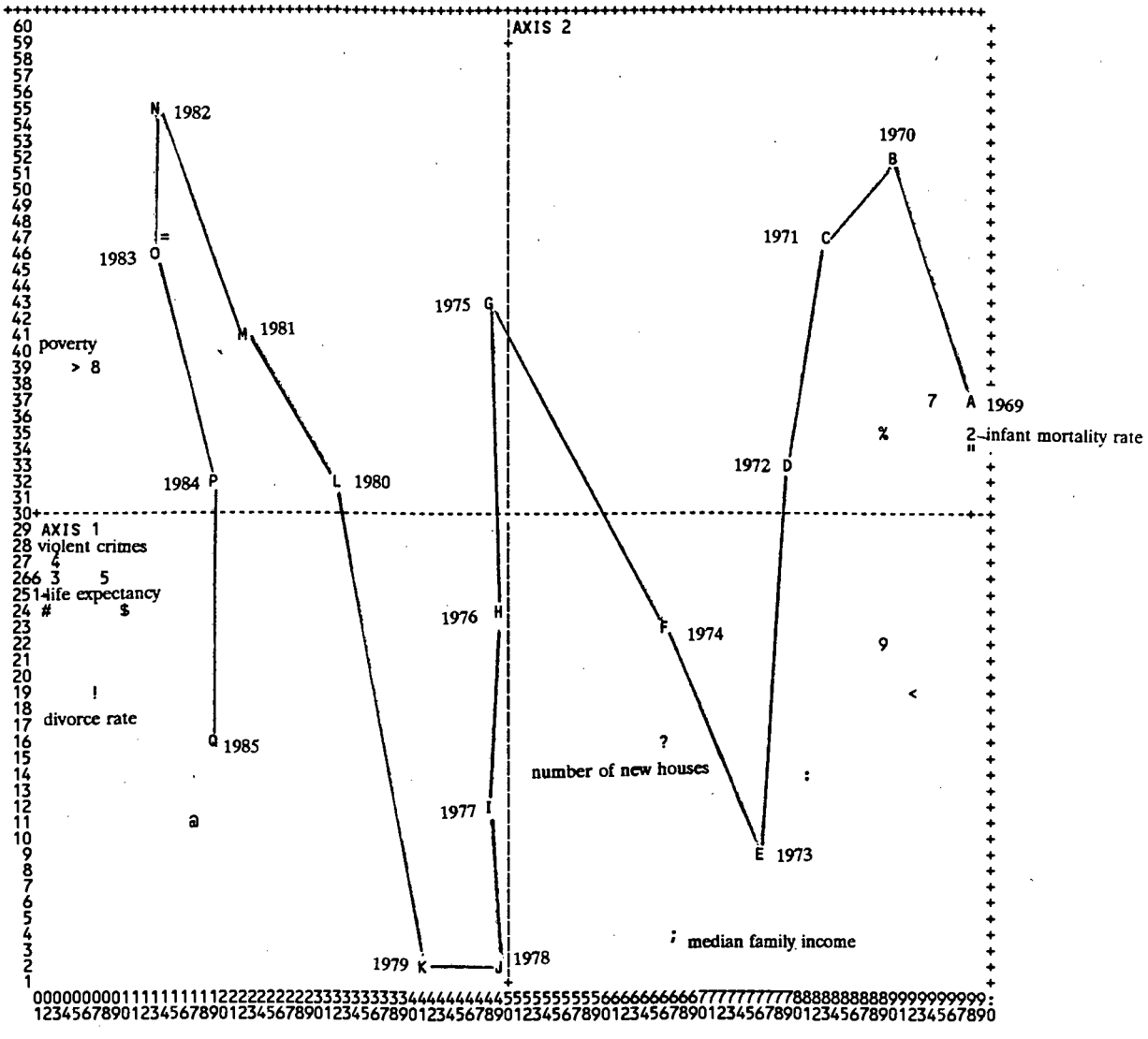
In the Department of

Statistical Sciences

University of Cape Town

SUPERVISOR: Professor L.G. Underhill

University of Cape Town, December 1991


AXIS 2
AXIS 1
1982
1983
1981
1980
1984
1985
1975
1976
1977
1978
1979
1974
1973
1970
1971
1972
1969
poverty
> 8
violent crimes
life expectancy
divorce rate
number of new houses
median family income
infant mortality rate

ACKNOWLEDGEMENTS

I wish to thank Professor Les Underhill, for supervising the thesis, and for his insights and encouragement.

I am grateful to Professor Theo Stewart, for the understanding and the flexibility allowed me during the course of writing this thesis.

The Foundation for Research Development provided financial assistance.

Final thanks go to the members of the Department of Statistical Sciences at UCT who gave advice and encouragement, and to my friends Janice, Greg and Rosanne.

CONTENTS

1

INTRODUCTION

The thesis deals with graphical display techniques based on the singular value decomposition. These techniques, known as biplots, are used to find low dimensional representations of multidimensional data matrices.

The aim of the thesis is to provide a review of biplots for a practical statistician who is not familiar with the area. It therefore focuses on the underlying theory, assuming a standard statisticians' knowledge of matrix algebra, and on the interpretation of the various plots.

The topic falls in the realm of descriptive statistics. As such, the methods are chiefly exploratory. They are a means of summarising the data. The data matrix is represented in a reduced number of dimensions, usually two, for simplicity of display. The aim is to summarise the information in the matrix and to present a visual representation of this information. The aim in using graphical display techniques is that the "gain in interpretability far exceeds the loss in information" (Greenacre, 1984).

A graphical description is often more easy to understand than a numerical one. Histograms and pie charts are familiar forms of data representation to many people with no other, or very rudimentary, statistical understanding. These are applicable to univariate data. For multivariate data sets, univariate methods do not reveal interesting relationships in the data set as a whole. In addition, a biplot can be presented in a manner which can be readily understood by non-statistically minded individuals.

Greenacre (1984) comments that only in recent years has the value of statistical graphics been recognised. Young (1989) notes that recently there has been a shift in emphasis, among statisticians towards exploratory data analysis methods. This school of thought was

given momentum by the publication of the book "Exploratory Data Analysis" (Tukey, 1977). The trend has been facilitated by advances in computer technology which have increased both the power and the accessibility of computers.

Biplot techniques include the popular correspondence analysis. The original proponents of correspondence analysis (among them Benzecri) reject probabilistic modelling. At the other extreme, some view graphical display techniques as a mere preliminary to the more traditional statistical approaches. Under the latter view, graphical display techniques are used to suggest models and hypotheses.

The emphasis in exploratory data techniques such as graphical displays is on 'getting a feel' for the data rather than on building models and testing hypotheses. These methods do not replace model building and hypothesis testing, but supplement them. The essence of the philosophy is that models are suggested by the data, rather than the frequently followed route of first fitting a model.

Some work has gone into developing inferential methods, with hypothesis tests and associated p-values for biplot-type techniques (Lebart *et al*, 1984, Greenacre, 1984). However this aspect is not important if the techniques are viewed merely as exploratory.

Chapter Two provides the mathematical concepts necessary for understanding biplots. Chapter Three explains exactly what a biplot is, and lays the theoretical framework for the biplot techniques that follow. The goal of this chapter is to provide a framewotk in which biplot techniques can be classified and described. Correlation biplots are described in Chapter Four. Chapter Five discusses the principal component biplot, and the link between these and principal component analysis is drawn. In Chapter Six, correspondence analysis is presented. In Chapter Seven practical issues such as choice of centre are discussed. Practical examples are presented in Chapter Eight. The aim is that these examples illustrate techniques commonly applicable in practice. Evaluation and choice of biplot is discussed in Chapter Nine.

2

MATHEMATICAL BACKGROUND

The mathematical concepts used in the thesis are summarised in this chapter.

Consider a data matrix A with n rows and m columns. When we want to refer to one row or one column of a matrix A, we will speak of the ith row and jth column of the matrix. The entry in the ith row and the jth column of A will be denoted a_{ij}.

The entries of a data matrix can be classified as being qualitative or quantitative. If quantitative, the entries are either ordinal, interval or ratio. Qualitative data is expressed on a nominal scale.

2.1 Generic Definition of a Data Matrix

Data matrices can be broadly categorised into two types. We will refer to these as Class I and Class II. The first type of data matrix has the characteristic that its rows and columns are classified by two different sets of categories (groupings). These are commonly referred to as two-way tables or as multivariate data matrices. The second type has its rows and columns classified by the same categories.

Class I

A multivariate data matrix typically consists of n 'individuals' on which a series of m variables have been measured. The entries of the matrix can be any, or a mixture of, the types of data previously mentioned.

An individual could for example be a person, ecological site, animal species or time period. Individuals are also referred to as cases, subjects or taxonomic units.

Examples of possible observations on a person are height (measured on a ratio scale), body temperature (interval scale), position in a queue (ordinal scale) and hair colour (nominal scale). The observations are referred to as variables.

Conventionally, individuals form the rows of the matrix, and variables the columns.

A contingency table is a special type of Class I matrix. Its entries are counts. Two methods of classification are applied to the data. These classifications are represented by the rows and the columns of the matrix; thus both the rows and the columns are variables. The entry in cell a_{ij} is a count of the number of individuals that have row classification i and column classification j.

Class II

These are square matrices whose entries are some measure of association between the respective row and column categories. The matrix consists of pairwise comparisons.

These measures of association indicate either distances or similarities between the categories, where the concepts of 'distance' and 'similarity' have a very broad interpretation. A distance measure decreases in magnitude with increasing likeness, whereas increasing similarity indicates increasing likeness.

Square matrices are either symmetric or asymmetric. In a symmetric matrix, the association between the ith and jth category is the same as that between the jth and the ith, i.e. $a_{ij}=a_{ji}$. Examples of symmetric matrices are a table of distances between towns and a correlation matrix. Association between pairs of objects need not be symmetric, as for example, when a_{ij} is the probability that the wind direction changes from direction i to direction j in an hour. (A small probability indicates that directions i and j are 'distant'.)

A special case of a symmetric matrix is a diagonal matrix. A diagonal matrix is a square matrix in which all off-diagonal entries are zero.

A skew symmetric matrix A has main diagonal elements $a_{ii}=0$ and $A=-A^T$. An example of such a matrix is emmigration/immigration data, i.e. a_{ij} is the number of people that emmigrated from country i to country j.

Class I to Class II

Class II matrices are sometimes observed directly, but can also be derived from Class I matrices by defining a distance on the rows or columns of a Class I matrix. Euclidean distance is the simplest example of a distance function.

Generalised Euclidean distance between the rows or columns is an example of a distance metric on quantitative data. An example of a Generalised Euclidean distance is chi squared distance in contingency tables, which is described in Chapter 6.

Techniques dealing with square symmetric matrices fall under the heading of multidimensional scaling (Greenacre and Underhill, 1982). We deal with Class I matrices in this thesis.

2.2 Multidimensional Vector Spaces, Basis and Dimension

The *dimension* of the space is given by the number of vectors in the basis.

Suppose that the r vectors $U=\{u_1,u_2,\ldots,u_r\}$ form a basis for the r-dimensional vector space V. The basis U has two properties:

(i) U is linearly independent. This means that none of the basis vectors can be expressed as a linear combination of the other basis vectors.

(ii) U spans V. This means that every vector v in V is expressible as a linear combination

of $u_1, u_2, \ldots, u_r$:

$$v = \sum_{i=1}^{r} \alpha_i u_i \qquad (2.1)$$

2.3 Distance, Norm and Scalar Product

We will adhere to the convention that a vector is a matrix with a single column. The transpose of a matrix is denoted by *T*. Row vectors are denoted by a transposed column vector.

Let $a^T = (a_1, a_2, \ldots, a_r)$ and $b^T = (b_1, b_2, \ldots, b_r)$ be vectors in V.

A *distance* (or *metric*) in V is a map d from VxV into R^+ with the properties

$$\begin{aligned} d(a,b) &= d(b,a) \\ d(a,b) &= 0 \Leftrightarrow a = b \\ d(a,b) &\leq d(a,c) + d(c,b) \end{aligned} \qquad (2.2)$$

for a, b, c in V.

The *norm* of a vector is the distance of the vector from the origin. The norm of a vector v in V is denoted $\|v\|$.

$$\|v\| = d(v,0) \qquad (2.3)$$

where $0=(0\ 0\ \ldots 0)^T$.

The norm also satisfies the property

$$\|cv\| = |c|\,\|v\| \qquad (2.4)$$

for any scalar c.

d(a,b) is defined to be $\|a-b\|$.

The L_p-norms are defined as

$$\|a\|_p = \left(\sum_{i=1}^{r} |a_i|^p\right)^{\frac{1}{p}} \qquad (2.5)$$

and the L_p-distances as

$$d_p(a,b) = \|a-b\|_p = \left(\sum_{i=1}^{r} |a_i - b_i|^p\right)^{\frac{1}{p}} \qquad (2.6)$$

The L_2 norm and distance are known as *Euclidean norm* and *Euclidean distance*. These are denoted $\|a\|$ and d(a,b) respectively. Squared Euclidean distance is denoted by $d^2(a,b)$.

The Euclidean metric is the usual Pythagorean distance.

For the remainder of the text, we assume Euclidean space, unless otherwise stated, so that

$$\|a\| = \|a\|_2 \qquad (2.7)$$

and

$$d(a,b) = d_2(a,b) \qquad (2.8)$$

Thus the Euclidean distance is given by

$$d(a,b) = \sqrt{\sum_{i=1}^{r} (a_i - b_i)^2} \qquad (2.9)$$

and the Euclidean norm by

$$\|a\| = \sqrt{\sum_{i=1}^{r} a_i^2} \qquad (2.10)$$

In R^2, the Euclidean norm of a vector is its length:

$$\|a\| = \sqrt{a_1^2 + a_2^2} \qquad (2.11)$$

(Pythagoras' Theorem)

The *squared norm of a matrix* is defined to be the sum of its squared entries

$$\|A\|^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} a_{ij}^2 = tr(AA^T) \qquad (2.12)$$

Thus the *norm of a matrix* is

$$\|A\| = \sqrt{\sum_{i,j} a_{ij}^2} \qquad (2.13)$$

The definition of a matrix norm enables the distance between two matrices to be defined as

$$d(A,B) = \|A-B\| = \sqrt{\sum_{i,j} (a_{ij}-b_{ij})^2} \qquad (2.14)$$

The *scalar product* (a,b) between two vectors a and b in a vector space V is a function from VxV into R with the properties:

$$\begin{aligned}(a,b)&=(b,a)\\(a,c_1b_1+c_2b_2)&=c_1(a,b_1)+c_2(a,b_2)\\(a,a)&\geq 0\\(a,a)&=0\Leftrightarrow a=0\end{aligned} \qquad (2.15)$$

where c_1 and c_2 are scalars.

Two vectors are *orthogonal* if their scalar product is zero. If, in addition, they both have norms of one then they are *orthonormal*,

i.e a basis $U=\{u_1,u_2,\ldots,u_n\}$ is orthonormal if

$$\begin{aligned}(u_i,u_j)&=1 \qquad \text{for } i=j\\&=0 \qquad \text{otherwise}\end{aligned} \qquad (2.16)$$

i.e. if $U^TU=I_r$.

If the basis is orthonormal we denote the scalar product between two vectors a and b by a^Tb.

$$\begin{aligned}a^Tb&=a_1b_1+a_2b_2+\ldots+a_rb_r\\&=\sum_{i=1}^{r}a_ib_i\end{aligned} \qquad (2.17)$$

This can equivalently be expressed as

$$a^Tb=\|a\|\,\|b\|\cos\theta \qquad (2.18)$$

where θ is the angle between a and b.

The norm of a vector, and the distance between two vectors can be expressed in terms of scalar products. In Euclidean space:

$$\|a\|^2 = a^T a, \qquad \|a\| = \sqrt{a^T a} \tag{2.19}$$

and

$$\begin{aligned} d(a,b) &= \|a-b\| \\ &= \sqrt{(a-b)^T (a-b)} \\ &= \sqrt{a^T a + b^T b - 2a^T b} \end{aligned} \tag{2.20}$$

Thus having defined the scalar product in a space, there is an associated norm and distance.

Notice that the distance between two vectors is independent of the origin of the space, but that their scalar product is defined in terms of the angle subtended at the origin, and is thus dependent on the origin and on their distances from the origin.

2.4 Generalised Euclidean Space

Consider a generalisation of the scalar product definition:

$$(a,b)_\Phi = a^T \Phi b \tag{2.21}$$

where Φ is a positive definite matrix.

The associated squared norm and squared distances are:

$$\|a\|_\Phi^2 = (a,a)_\Phi \tag{2.22}$$

and

$$d_\Phi^2(a,b) = \|a-b\|_\Phi^2 = (a-b)^T \Phi (a-b) \tag{2.23}$$

These scalar products, norms and distances are said to be *in the metric* Φ. The space with this metric is said to be a *Generalised Euclidean space.* ('Ordinary' Euclidean space is in the metric I, where I is the identity matrix.)

If Ω and Φ are positive definite symmetric matices of orders nxn and mxm respectively, where Ω defines the metric between the columns of A and Φ defines the metric between

the rows of A, we can define the generalised matrix norm of the matrix A in the metric Ω, Φ :

$$\|A\|_{\Omega,\Phi}=tr(\Omega A\Phi A^T)^{\frac{1}{2}}=tr(\Phi A^T\Omega A)^{\frac{1}{2}} \qquad (2.24)$$

If Ω is diagonal then

$$\|A\|_{\Omega,\Phi}=\left(\sum_{i=1}^{n} w_{ii}a_i^T\Phi a_i\right)^{\frac{1}{2}} \qquad (2.25)$$

If Φ is diagonal then

$$\|A\|_{\Omega,\Phi}=\left(\sum_{j=1}^{m} \phi_{jj}a_j^T\Omega a_j\right)^{\frac{1}{2}} \qquad (2.26)$$

where

a_i^T is the ith row of A

a_j is the jth column of A

w_{ii} and ϕ_{jj} are entries on the diagonal of Ω and Φ respectively.

A metric which we shall be referring to frequently is that space defined by the inverse of the covariance matrix, $\Phi=S^{-1}$. The distance associated with this norm is known as Mahalanobis distance.

2.5 Singular Value Decomposition (SVD)

The singular value decomposition (SVD) is used in graphical display techniques to find a lower rank matrix which approximates the data matrix.

The SVD of a real nxm matrix Z is

$$Z = U \Gamma V^T = \sum_{k=1}^{r} \alpha_k u_k v_k^T \qquad (2.27)$$

where

1. Γ is a diagonal matrix of positive numbers, $\Gamma = diag(\alpha_1, \ldots, \alpha_r)$, arranged so that $\alpha_1 \geq \alpha_2 \geq \ldots \geq \alpha_r > 0$.
2. u_k and v_k are the kth columns of the nxr matrix U and the pxr matrix V respectively.
3. $r \leq \min\{n,m\}$ is the rank of Z (usually, n>m and, for most applications, r=m or r=m-1).
4. U (nxr) and V (mxr) are orthonormal, i.e. $U^TU = V^TV = I_r$, the identity matrix of order r.
5. The eigenstructure of the square symmetric matrix Z^TZ is $Z^TZ = V\Gamma^2V^T$; the non-zero eigenvalues are given by α_k^2, (k=1,2,...,r), and the eigenvectors by the columns v_k of V, (k=1,2,...,r).
6. Similarly, $ZZ^T = U\Gamma^2U^T$ is the eigenstructure of the nxn matrix ZZ^T.

The scalar α_k is called the kth singular value of Z. The u_k and v_k are called the left and right singular vectors respectively.

If the α_k are all distinct then the singular value decomposition of a matrix is unique up to a simultaneous reflection of corresponding columns of U and V.

In square symmetric matrices, U=V and the SVD and eigenstructure are the same as U and V coincide. For B a symmetric nxn matrix

$$B=U\,\Gamma U^T=\sum_{k\ 1}^{r}\alpha_k u_k u_k^T \qquad (2.28)$$

The squared norm of the matrix Z can be expressed as the sum of its squared singular values:

$$\begin{aligned}\|Z\|^2&=tr(ZZ^T)\\&=tr(U\,\Gamma V^T V\,\Gamma U^T)\\&=\sum_{k=1}^{r}\alpha_k^2\end{aligned} \qquad (2.29)$$

A further application of SVD we will make use of is that the Moore-Penrose generalised inverse, Z^+, of an nxm matrix Z, can be calculated from the SVD:

$$Z^+=V\,\Gamma^{-1}U^T \qquad (2.30)$$

2.5.1 Computation of the SVD

SVD algorithms are available in many packages, including GENSTAT (Genstat 5 Committee, 1988), NAG (Numerical Algorithm Group, 1990), MATLAB (Aptech systems, 1981), GAUSS (Gauss system version 2.2). There is an SVD routine in Numerical Recipes (Press et al, 1986).

2.5.2 Geometric Interpretation of the SVD

The decomposition of the matrix Z into the product of simpler ones allows a useful geometric understanding of it. This can be understood by consideration of the following: Any matrix with real-valued entries can be expressed as the product of a rotation (possibly followed by a reflection), a stretch and then a rotation matrix.

If the α_k are all distinct, the decomposition is unique up to a simultaneous reflection of corresponding columns of U and V.

The Γ part of the decomposition, which is the stretch matrix, is unique. A matrix is

'stretched' by multiplying it by a diagonal matrix. Pre-multiplication by a stretch matrix results in rescaling of the rows, and vice versa. The corresponding rows or columns of the original matrix are differentially rescaled by a factor of α_k in dimension k. Thus the rows of V^T, i.e. the columns of V, and the columns of U are rescaled.

A rotation is the application of an orthogonal matrix. The matrix is rigidly rotated about some angle. Rotation of a matrix has no effect on its scalar products; they remain unchanged. A reflection about both axes in two dimensions is equivalent to a rotation of the points about an angle of 180^0 from their original orientation.

The α_k represent the magnitudes of the matrix Z in the r dimensions. The columns of U and V are orthonormal bases for the columns and rows of Z respectively.

A further discussion of the geometric interpretation of the SVD may be found in Mandel (1982).

2.6 Lower Rank Approximation using SVD

Having expressed Z in terms of its SVD, it is simple to find a rank p approximation to Z, denoted $Z_{[p]}$ (p < r).

$$Z_{[p]}=U_{[p]}\Gamma_{[p]}V^T_{[p]} \qquad (2.31)$$

where

$U_{[p]}$ and $V_{[p]}$ are the first p columns of U and V respectively and

$\Gamma_{[p]}$ is a pxp diagonal matrix consisting of the first p singular values of Z, i.e. $\Gamma_{[p]}=diag(\alpha_1,\ldots,\alpha_p)$.

$Z_{[p]}$ is the closest of all possible rank p approximations to Z in the sense that it minimizes the sum of the squared differences between corresponding entries of Z and $Z_{[p]}$.

Thus

$$\sum_{i,j} (z_{ij}-z_{ij[p]})^2 \quad i=1,\ldots,n \quad j=1,\ldots,m \qquad (2.32)$$

is minimized.

This is equivalent to minimizing $\| Z-Z_{[p]} \|$.

The theorem of low rank approximation was proved by Eckart and Young (1936).

We will in general be concerned with the case p=2. $Z_{[2]}$ can then be represented in two dimensions which permits its graphical appraisal.

The general idea is that if $Z_{[2]}$ is a 'good' representation of Z then much information can be gleaned about Z from the graphical display of $Z_{[2]}$. This key concept is discussed later.

2.7 Quality of the Low Rank Approximation

It is clearly necessary to measure how close the low rank approximation is to the original matrix. This can be done by comparing the squared norm of $Z_{[p]}$ to that of Z. The norm of Z in the kth dimension is given by the kth singular value, α_k. The squared norm of a matrix is thus the sum of its squared singular values (from 2.29). A measure of the quality of the low rank approximation is the ratio of the squared norms:

$$\frac{\| Z_{[p]} \|^2}{\| Z \|^2} = \frac{\sum_{k=1}^{p} \alpha_k^2}{\sum_{k=1}^{r} \alpha_k^2} \qquad (2.33)$$

This measure is usually multiplied by 100, and expressed as a percentage.

Quality of display of the individual row and column points is discussed in Section 2.11.

2.8 Generalised Singular Value Decomposition (GSVD)

The generalised SVD of a matrix Z in the norm Ω,Φ enables us to find a matrix of rank p, $Z_{[p]}$, such that $\|Z-Z_{[p]}\|_{\Omega,\Phi}$ is minimised.

$Z_{[p]}$ minimizes

$$tr(\Omega(Z-Z_{[p]})\Phi(Z-Z_{[p]}^T) \tag{2.34}$$

The GSVD theorem states that any real nxm matrix Z can be expressed as

$$Z=N\Gamma M^T \tag{2.35}$$

where

$$N^T\Omega N=M^T\Phi M=I_r \tag{2.36}$$

The rank p approximation to Z in the metric Ω,Φ is given by:

$$Z_{[p]}=N_{[p]}\Gamma M_{[p]}^T \tag{2.37}$$

Proof

Consider the ordinary SVD of

$$\Omega^{\frac{1}{2}}Z\Phi^{\frac{1}{2}}=U\Gamma V^T \tag{2.38}$$

where

$$U^TU=V^TV=I.$$

Now let

$$\begin{aligned} N&=\Omega^{-\frac{1}{2}}U \\ M&=\Phi^{-\frac{1}{2}}V \end{aligned} \tag{2.39}$$

Then

$$\Omega^{\frac{1}{2}} Z \Phi^{\frac{1}{2}} = (\Omega^{\frac{1}{2}} N) \Gamma (M^T \Phi^{\frac{1}{2}}) \qquad (2.40)$$

The generalised SVD of Z is thus

$$Z = N \Gamma M^T \qquad (2.41)$$

with

$$N^T \Omega N = I = M^T \Phi M = I \qquad (2.42)$$

2.9 Singular Value Decomposition Displays (SVDD)

Techniques that we will use for the display of a data matrix in a lower dimensional space using the GSVD can be defined in terms of three phases (Greenacre, 1984).

The type of display depends on the choices made at each of the phases. Different techniques are obtained by varying the options in these phases.

Phase I

The data matrix is transformed. The most common transformation is some centring of the data, such as subtracting the mean of each variable. We denote the output matrix from this phase by Z. The choice involved here is which transformation to use.

Phase II

The GSVD of the matrix Z is computed, in order to find a low rank approximation to Z.

Choices in this phase are the metrics Φ and Ω.

Phase III

Matrices F and G which contain the coordinates representing the *row points* and *column*

points respectively, of Z, are found. These points represent the rows and columns of Z. There are different ways in which this can be done. This can be expressed as choosing a and b in:

$$\begin{aligned} F &= N\,\Gamma^{a} \\ G &= M\,\Gamma^{b} \end{aligned} \qquad (2.43)$$

where a and b are real numbers.

To plot the rows and/or columns in a p-dimensional display we use $F_{[p]}$ and $G_{[p]}$ which are nxp and mxp respectively, and are of rank p. They are given by:

$$\begin{aligned} F_{[p]} &= N_{[p]}\,\Gamma_{[p]}^{a} \\ G_{[p]} &= M_{[p]}\,\Gamma_{[p]}^{b} \end{aligned} \qquad (2.44)$$

where $N_{[p]}$ and $M_{[p]}$ are the first p columns of M and N respectively and $\Gamma_{[p]}$ is, as before, a diagonal matrix consisting of the first p singular values of Z.

The individual row and column points are given by the rows f_i^T of $F_{[p]}$ and g_j^T of $G_{[p]}$.

Plotting the points

To facilitate interpretation of the simultaneous display of the row and the columns points, their coordinates are plotted on the same scale. If the scales differ too greatly, the coordinates of one set of points is multiplied by a suitable constant. The rescaling does not change the interpretations.

The scales of the displayed axes must be equal. This is because the interpretations utilise angles between the vectors, and the relative lengths of the vectors. The scale itself is not needed for the interpretations.

The display is rarely in anything other than two dimensions. Unless there are very few points, or there is access to sophisticated software, it is difficult to display a three dimensional plot. The additional complexity in interpreting such a plot goes against what we are aiming at - a readily interpretable representation to increase understanding of the

relationships in the data set. The detail involved in three dimensional displays is offset by loss of simplicity.

A recent development (Young, 1989) is the representation of higher dimensional biplots by means of dynamic computer graphics. Such graphics allow representation of three dimensional matrices. The viewer is able to manipulate the display by rotating or translating the cloud of points to reveal its structure. Hyper-dimensional methods also exist, which aid understanding of structures that have more than three dimensions.

The choices made in the above three phases are represented in the following expression:

SVDD (Z,Ω,Φ,a,b).

The expression completely describes the type of display.

2.10 Transition Formulae and Supplementary Points

The *transition formulae* express the row coordinates in terms of the column coordinates, and vice versa. There are two main uses of these formulae:

1. As a description of the relationship between the row and column points in the display.
2. They are used to plot the supplementary points. A supplementary point is a row or column point that is superimposed on the display after the other points have been computed and displayed. The point is not included in the calculations to determine the row and column coordinates, but is represented on the plot using the same coordinate system; thus its position relative to other points can be interpreted.

The expressions for the supplementary points are derived as follows:

Since $Z=N\Gamma M^T$ with $N^T\Omega N=M^T\Phi M=I$, we have

$$Z\Phi M\,\Gamma^{-1}=N\,\Gamma M^T\Phi M\,\Gamma^{-1}=N \qquad (2.45)$$

by postmultiplying Z by $\Phi M\Gamma^{-1}$.

Similarly, postmultiply Z^T by $\Omega N\Gamma^{-1}$ to obtain

$$Z^T \Omega N \Gamma^{-1} = M \Gamma N^T \Omega N \Gamma^{-1} = M \qquad (2.46)$$

but $F = N\Gamma^a$ and $G = M\Gamma^b$

So

$$\begin{aligned} F &= Z\Phi M \Gamma^{-1}\Gamma^a \\ &= Z\Phi G \Gamma^{-b}\Gamma^{-1}\Gamma^a \\ &= Z\Phi G \Gamma^{a-b-1} \end{aligned} \qquad (2.47)$$

and

$$G = Z^T \Omega F \Gamma^{b-a-1} \qquad (2.48)$$

Formulae (2.47) and (2.48) are called the transition formulae. The coordinates for plotting z^{*T}, a supplementary row are given by

$$f^{*T} = z^{*T} \Phi G \Gamma^{a-b-1} \qquad (2.49)$$

The coordinates for plotting z^+, a supplementary column are

$$g^{+T} = z^{+T} \Omega F \Gamma^{b-a-1} \qquad (2.50)$$

2.11 Decomposition of the Squared Norm, Absolute and Relative Contributions

If Ω and Φ are diagonal matrices, we can decompose the squared norm of Z in the metric Ω, Φ in three ways.

The first of these is (by 2.29):

$$\| Z \|^2_{\Omega,\Phi} = \sum_{k=1}^{r} \alpha_k^2 \qquad (2.51)$$

The α_k^2 represent the magnitude of Z in each of the r dimensions, and can be interpreted as the contribution of the kth axis (dimension) to the squared norm.

If a=1, then F=UΓ and

$$FF^T=U\Gamma V^T\Phi V\Gamma U^T=Z\Phi Z^T \qquad (2.52)$$

so

$$\begin{aligned}\|Z\|^2_{\Omega,\Phi}&=tr(\Omega Z\Phi Z^T)\\&=tr(\Omega FF^T)\\&=\sum_{k=1}^{r}\sum_{i=1}^{n}\omega_i f_{ik}^2\end{aligned} \qquad (2.53)$$

Since

$$\sum_{i=1}^{n}\omega_i f_{ik}^2=\alpha_k^2, \qquad (2.54)$$

$w_i f_{ik}^2$ may be interpreted as the contribution of the ith row to the squared norm due to the kth axis. This can also be expressed as the contribution of the ith row to the magnitude in the kth dimension.

Similarly, if b=1,

$$\begin{aligned}\|Z\|^2_{\Omega,\Phi}&=tr(\Phi Z^T\Omega Z)\\&=tr(\Phi GG^T)\\&=\sum_{k=1}^{r}\sum_{j=1}^{n}\phi_j g_{jk}^2\end{aligned} \qquad (2.55)$$

and since

$$\sum_{j=1}^{m}\phi_j g_{jk}^2=\alpha_k^2 \qquad (2.56)$$

$\phi_j g_{jk}^2$ may be interpreted as the contribution of the jth column to the squared norm (magnitude) due to the kth axis.

Usually, the row and column contributions to the squared norm due to the kth axis are expressed as percentages. They are called *absolute contributions*.

The absolute contributions are given by

$$a_{ik}=100\frac{w_i f_{ik}^2}{\alpha_k^2}$$
$$b_{jk}=100\frac{\phi_j g_{jk}^2}{\alpha_k^2} \qquad (2.57)$$

where

a_{ik} is the contribution of the ith row to the kth axis

b_{jk} is the contribution of the jth column to the kth axis

We can also express (2.51) as

$$\|Z\|^2_{\Omega,\Phi}=\sum_{i=1}^{n}\omega_i\sum_{k=1}^{r}f_{ik}^2 \qquad (2.58)$$

The term $w_i\sum f_{ik}^2$ can be interpreted as the contribution of the ith row to the squared norm (summed across r dimensions).

$$r_{ik}=100\frac{\omega_i f_{ik}^2}{\left(\omega_i\sum_{k=1}^{r}f_{ik}^2\right)}=100\frac{f_{ik}^2}{\sum_{k=1}^{r}f_{ik}^2} \qquad (2.59)$$

is the percentage of the squared norm contributed by the ith row which is explained by the kth axis.

Similarly,

$$s_{jk}=100\frac{g_{jk}^2}{\sum_{k=1}^{r} g_{jk}^2} \qquad (2.60)$$

is the percentage of the squared norm which is explained by the jth column.

r_{ik} is referred to as the *relative contribution* of the kth axis to the squared norm of the ith case.

s_{jk} is the relative contribution of the kth axis to the squared norm of the jth variable.

Absolute contributions are the contributions of row and column points to the axes. For a particular axis, the absolute contribution is the proportion of the squared norm for that axis that is explained by each of the points.

Relative contributions are the contributions that the axes make to the rows and columns. For a particular point, the relative contribution is the proportion of its squared norm that is explained by an axis.

It is an important to remember that the absolute and relative contributions are only defined for the rows when a=1, and for the columns when b=1.

Absolute and relative contributions are one of the ways in which the quality of the display can be assessed. This topic is discussed further in Chapter 9. Section 7.3 contains an example of assessing quality using decompositions of the norm.

3

THE BIPLOT

3.1 The Use of Scalar Products in Biplots

In this section we consider the geometric interpretation of the scalar product of two vectors. The concept of scalar products between vectors is key to the understanding of biplots. Thus before introducing biplots, the definition of scalar products in Section 2.3 is expanded.

For illustrative purposes, we consider two dimensions in Euclidean Space. In Figure 3.1, y is projected onto x. The angle subtended at the origin between the two vectors x and y is θ. The length of the projection of y onto x is p.
We saw in (2.18) that the scalar product of x and y can be expressed as

$$x^T y = \|x\| \|y\| \cos\theta \qquad (3.1)$$

and since

$$\cos\theta = \frac{p}{\|y\|} \qquad (3.2)$$

$$x^T y = p\|y\| \qquad (3.3)$$

If the two vectors form an obtuse angle, the projection has a negative sign (Figure 3.2).

Figure 3.1

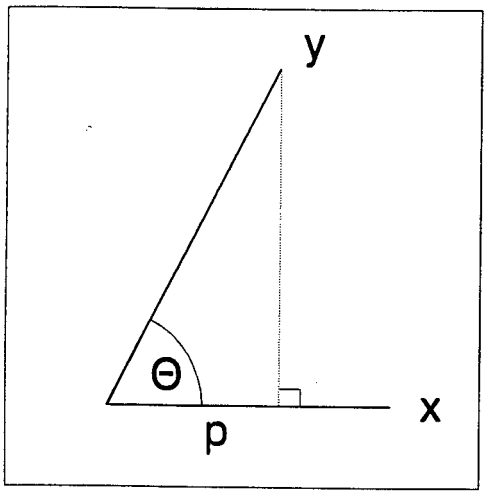


Figure 3.2

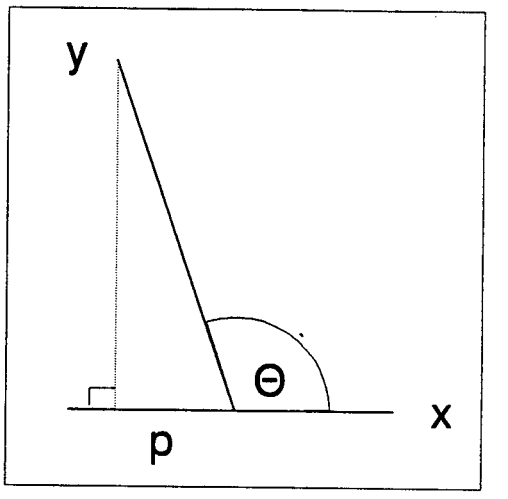


Figure 3.3

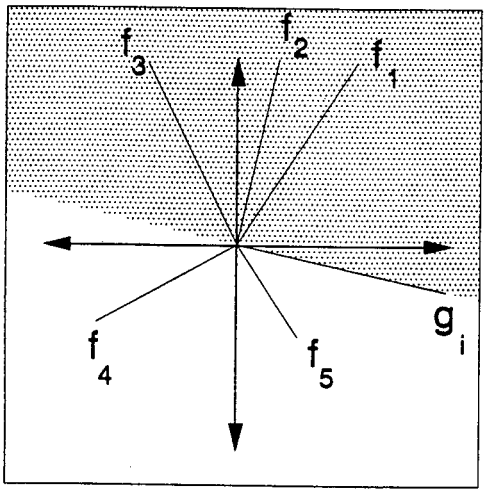
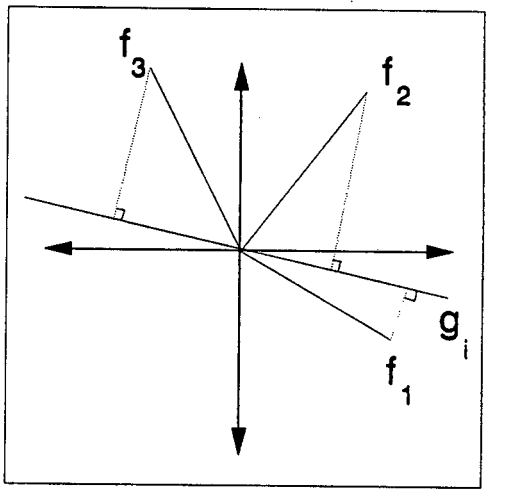


Figure 3.4



A special case is the projection of vectors onto an orthonormal basis. The norms of the vectors being projected onto are of unit length, so (3.3) reduces to

$$x^T y = p \qquad (3.4)$$

Thus the scalar product between the vectors being projected and the basis vectors gives the co-ordinates of the vectors relative to the basis. This holds true for all orthonormal bases.

If we wanted to plot the rows of $Z_{[2]}$, the rank two approximation of Z given by the singular value decomposition, the co-ordinates are two dimensional, but are expressed in terms of the old coordinate system; they are in r-space. To plot the rows relative to a two dimensional subspace the above property of scalar products is utilised.

For example, V is an orthonormal basis since $V^TV=I$.
Thus the scalar products of the rows of $Z_{[2]}$ with V, $Z_{[2]}V$ give the coordinates of the rows of $Z_{[2]}$ relative to the basis V.

We have named the matrix of the coordinates of the row points F (Section 2.9).
Note that choosing V as our basis means that

$$F=ZV=U\Gamma V^TV=U\Gamma$$

There are two key uses of scalar products in SVDD. First, expression of the coordinates of the row and column points in a form in which they can be plotted, as described above, and secondly, interpretation of the relative positions of the row and column points as described in Section 3.3 below.

3.2 Introduction to Biplots

A biplot is a particular form of graphical display of a data matrix. The 'bi' refers to the fact that in this type of plot, both the rows and the columns are displayed in such a way that the joint display of both sets of points is meaningful, and does not imply that the

display is necessarily in two dimensions. However, such planar displays are the easiest to present, and we shall be considering mostly these.

The biplot was introduced by Gabriel (1971). An extensive literature on biplots has since arisen, and biplots have been applied in diverse areas. Correspondence analysis, a biplot technique applied to contingency tables, has also aroused much interest. The literature on biplots has been summarised by Gabriel (1981), Greenacre and Underhill (1982), Gower (1984) and Greenacre (1984). Correspondence analysis is discussed in Chapter 7. Some applications of biplots have been in the fields of economics (Barr, 1990), metereology (Gabriel, 1972), medicine (Osmond, 1985), finance (van den Honert and Barr, 1988), agriculture (Underhill, 1990), market research (Shahim and Greenacre, 1988) and ecology (Ter Braak, 1983).

Originally, the definition of a biplot was limited to plots characterised by a particular interpretation of the joint display, known as the *biplot interpretation*. We will refer to this property as the *scalar product interpretation*. We will also refer to the *within set* and *between set* interpretations. A within set interpretation refers to the relationship of the row points to each other or to the relationship of the column points to each other. Between set means the relationship between a row point and a column point.

The common feature of the classical biplots is their between set scalar product interpretation. This interpretation is described below. Recently, the word biplot is being used in a broader context (Gower and Harding, 1988), and means any plot in which points representing the rows and the columns are displayed simultaneously. We use this meaning of biplot.

Various types of biplots will be discussed in subsequent chapters. These biplots are distinguished from each other in that they emphasize different features of the matrix.

3.3 General Theoretical Interpretations from the Biplot

3.3.1 Introduction to Interpretations

Suppose that Z, the matrix of interest, has n individuals (rows) each having m variables (columns).

A rank two approximation of Z, $Z_{[2]}$, can be found using the singular value decomposition. It is this approximation which is biplotted, although sometimes displays in more than two dimensions are considered. If the rank two approximation is a 'good' one, this planar display can be useful as a graphical representation of certain characteristics of Z. Therefore when interpreting the plot, the quality of the approximation (Section 7.2) is, of course, very important.

In the following discussions of interpretations of the biplot, to avoid repeatedly having to say a feature is approximated, we will assume an exact representation of a rank two matrix. Thus any row or column of the Z can be depicted exactly on a two dimensional plot. The biplot thus displays features of the data matrix exactly, as opposed to approximating the features, as is the case when Z is of rank greater than two and we use a lower rank approximation. We shall also assume unweighted Euclidean space, i.e. $\Omega = I, \Phi = I$.

3.3.2 Between Set Scalar Products Interpretation

This property allows approximation of the data matrix from the plotted row and column points.

Any nxm matrix Z of rank r can be factorised as

$$Z = FG^T \qquad (3.5)$$

where F (nxr) and G (mxr) are of rank r.

i.e

$$z_{ij}=f_i^T g_j \qquad (3.6)$$

where

f_i^T is a vector representing the ith row of F

g_j^T is a vector representing the jth row of G,

Each element z_{ij} of Z can thus be expressed as the scalar product of a vector representing the ith row and a vector representing the jth column.

Thus F and G can be considered as matrices whose rows represent the n rows and the m columns of Z respectively. Z can therefore be represented in r-space by these n+m row vectors.

For Z of rank two these vectors representing rows and columns can be displayed simultaneously and exactly on a two dimensional plot. Each row and column point on the plot represents a column or row vector of Z. The elements of the matrix Z are the scalar products of the corresponding row and column points.

Clearly, the factorisation (3.5) is not unique (Gabriel, 1971). For some Z, F and G and any nonsingular matrix R, we can factorise Z as

$$Z=(FR)(R^{-1}G^T)$$

Different factorisations of Z provide one of the ways of generating different types of biplots. The choice of factorisation is done to facilitate display of the required features of the data, as will be discussed later.

It is convenient to use the SVD of Z to express the choice of factorisations available. Using the previous notation for the decomposition of Z, $Z=U\Gamma V^T$;

for

$$F=U\Gamma^a$$

$$G=V\Gamma^b$$

where a+b=1, there are infinitely many ways of choosing a and b and hence F and G. Any choice with a+b=1, yields

$$FG^T = U\Gamma^{(a+b)}V^T = U\Gamma V^T = Z \qquad (3.7)$$

For given a and b, matrices $F_{[2]}$ and $G_{[2]}$ are matrices of rank two whose rows $f_1,...,f_n$ and $g_1,...,g_m$ are vectors representing the rows (individuals) and columns (variables) respectively. When a+b=1, each element of Z can be expressed as the scalar product of the vectors representing its row and column,

$$\begin{aligned} z_{ij} &= f_i^T g_j \\ &= \|f_i\| \|g_j\| \cos\theta_{ij} \end{aligned} \qquad (3.8)$$

where θ_{ij} is the angle between f_i and g_j, i=1,...,n j=1,...,m

The sign of z_{ij} is the same as that of $\cos\theta_{ij}$. $\cos\theta_{ij}$ is positive when the angle between f_i and g_j is between 0 and 90 degrees.

Vectors f_i that fall in the shaded area in Figure 3.3 correspond to positive entries z_{ij}. Those in the unshaded area correspond to negative entries.

Similarly, column vectors g_j that form an angle of less than 90^0 with f_i correspond to positive entries zij; the others to negative entries.

However the scalar product interpretation provides more information about the elements of X than whether they are positive or negative. It also allows the visual inspection of the relative sizes of observations on a particular variable (and *vice versa*).

In Figure 3.4, g_j is the column point for variable j, and f_1, f_2 and f_3 are row points. The projections p_i of the row points onto the column points are indicated. From (3.3) and (3.5)

$$z_{ij}=f_i^T g_j=p_i\|g_j\| \qquad (3.9)$$

Since $\|g_j\|$ is fixed, the projections indicate that $z_{1j}>z_{3j}>z_{2j}$.

The precise interpretation depends on the centring operation chosen. A commonly encountered centring is column centring. Between set interpretation for column centring is described in Section 3.4.

The between set scalar product interpretation allows the approximation of the entries of the data matrix from the plot. The product of the length of a row vector, and the value of the length of the projection of the row point onto the column point approximates the corresponding entry for that row and column. Clearly, the roles of row and column can be interchanged, with the column points being projected onto the row points.

3.3.3 Within Set Scalar Product Interpretations

This interpretation is only valid when either a=1 or b=1. When a=1, it is valid for the row points, and when b=1, for the column points.

3.3.3 (a) Column Points

When b=1 we have F=U and G=VΓ. The rows of G contain the coordinates of the column points. The scalar product between these column points are exactly the same as the scalar products between the columns of Z:

$$\begin{aligned} GG^T &= V\Gamma\Gamma V^T \\ &= V\Gamma U^T U\Gamma V^T \\ &= Z^T Z \end{aligned} \qquad (3.10)$$

i.e. considering individual column points i and j:

$$g_i^T g_j = z_i^T z_j \qquad (3.11)$$

So

$$\begin{aligned} g_i^T g_j = z_i^T z_i = \|z_i\|^2 = \|z_j\|^2 \quad & for\ i=j \\ = z_i^T z_j^T \quad & for\ i \neq j \end{aligned} \qquad (3.12)$$

The relevance of this is that by inspection of the plotted column points we can get information about the columns of Z.

The within set interpretations can be divided into two categories:

(i) *The scalar product of a column point vector with itself.*
We have seen (in 2.19) that

$$g_j^T g_j = \|g_j\|^2.$$

Therefore,

$$\begin{aligned} \|g_j\| &= \sqrt{g_j^T g_j} \\ &= \|z_j\| \end{aligned} \qquad (3.13)$$

Thus the norms of the plotted column points are the same as the norms of the columns of Z.

(ii) *The scalar product of two different column points.*
Three interpretations follow from this.

1. The angle θ subtended at the origin between the two columns k and l of Z has the following cosine:

Section 3.3.2 there are numerous ways of choosing a and b. In general, choices other than a=1 and b=1 do not lead to useful within set interpretations. We show the effect of such choices using a=0 as an example, which is of particular interest when we have a column centred matrix (Section 3.4).

If a=0 then F=U and G=ΓV.

$$FF^T=UU^T$$
$$=U\Gamma V^T V\Gamma^{-2}V^T V\Gamma U^T$$

But $Z^TZ=V\Gamma^2V^T$ and $(Z^TZ)^+=V\Gamma^{-2}V^T$, and both matrices are symmetric. Thus

$$FF^T=Z(Z^TZ)^+Z^T \qquad (3.17)$$

Similarly if b=0 then F=UΓ and G=V, so

$$GG^T=Z^T((ZZ^T)^+)^TZ \qquad (3.18)$$

In most situations these other choices are not readily interpretable because they give rise to generalised scalar products and generalised Euclidean distances in an awkward norm.

3.3.5 Interpretations in Generalised Euclidean Space

If a+b=1 then $Z=FG^T$ as before.

If b=1,

$$GG^T=M\Gamma\Gamma M$$
$$=N\Gamma M^T\Omega M\Gamma N^T$$
$$=Z\Omega Z^T \qquad (3.19)$$

If Ω is diagonal with elements w_i then the scalar products between the rows of G are the weighted scalar products between the columns of Z:

$$g_k^T g_l=z_k^T\Omega z_l=\sum_{i=1}^{n}\omega_i z_{ik}z_{il} \qquad (3.20)$$

The associated squared norms of the columns of Z and squared distances between the

columns of Z for Ω a diagonal metric are given by:

$$\|g_j\|^2=\|z_j\|_\Omega^2=\sum_{i=1}^{n}\omega_i z_{ij}^2 \qquad (3.21)$$

and

$$\|g_k-g_l\|^2=\|z_k-z_l\|_\Omega^2=\sum_{i=1}^{n}\omega_i (z_{ik}-z_{il})^2 \qquad (3.22)$$

We can also define a generalised correlation between the columns of Z in the metric Ω:

$$\frac{g_k^t g_l}{\|g_k\|\|g_l\|}=\frac{z_k^T\Omega z_l}{\|z_k\|_\Omega\|z_l\|_\Omega}=\cos\theta \qquad (3.23)$$

where θ is the angle between the column points g_k and g_l.

When $z_{ij}=x_{ij}-\bar{x}_j$ and $\Omega=I$, (3.23) reduces to Pearson's Product Moment Correlation, and (3.21) reduces to (n-1) times the sample variance. Thus (3.21) effectively defines a generalised standard deviation.

We shall see an application of the generalised correlation in Chapter 7 (Correspondence Analysis).

If Φ is positive definite, then the interpretation is in terms of a generalised norm, as discussed in Section 2.4. Such generalised norms have not yet found application.

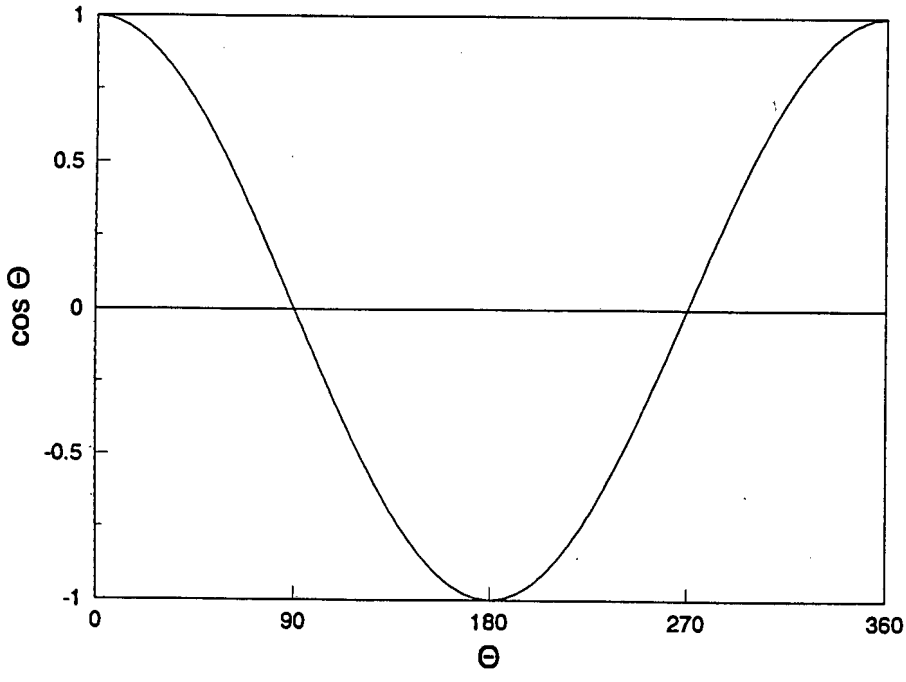
Similarly, if a=1 then

$$\begin{aligned} FF^T &= N\Gamma\Gamma N^T \\ &= N\Gamma M^T\Phi M\Gamma N^T \\ &= Z\Phi Z^T \end{aligned} \qquad (3.24)$$

The algebra could be followed through analogously for values of a and b other than one.

An example of a biplot in weighted space is given by Barr *et al* (1987).

Figure 3.5



3.4 Example - A Column Centred Matrix

Consider the matrix X with n individuals and m variables and entries x_{ij}.

$$X=\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \tag{3.25}$$

A frequently used centring is column centring. This is done by letting

$$Z=X-\overline{X} \tag{3.26}$$

where $\bar{X}$ is the matrix $\overline{X}=1\overline{x}^{T}$,

and

$$\bar{x}=\frac{1}{n}1^{T}X=(\overline{x_1},\overline{x_2},\ldots,\overline{x_m}) \tag{3.27}$$

where $\bar{x}$ is the vector of column means.

Then let the centred matrix be

$$Z=\begin{bmatrix} x_{11}-\overline{x_1} & x_{12}-\overline{x_2} & \cdots & x_{1m}-\overline{x_m} \\ x_{21}-\overline{x_1} & x_{22}-\overline{x_2} & \cdots & x_{2m}-\overline{x_m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}-\overline{x_1} & x_{n2}-\overline{x_2} & \cdots & x_{nm}-\overline{x_m} \end{bmatrix}, \tag{3.28}$$

i.e. from each observation is subtracted the mean for that column.

This transformation is known as column centring. Column centring has the effect of placing the mean row vector, or 'centre of gravity' at the origin.

We will follow through the interpretations in Sections (3.3.2)-(3.3.4) for the case of a

column centred matrix.

3.4.1 Between Set Scalar Product Interpretation

If a+b=1, the scalar product between a row and column point represents the corresponding element of Z.

$$\begin{aligned} f_i^T g_j &= z_{ij} \\ &= x_{ij} - \bar{x}_j \end{aligned} \tag{3.29}$$

Therefore:

$$\begin{aligned} \cos\theta > 0 &\Rightarrow x_{ij} - \bar{x}_j > 0 \\ &\Rightarrow x_{ij} > \bar{x}_j \end{aligned} \tag{3.30}$$

Thus the smaller the cosine of the angle between the row and column points, the closer the entry is to the mean for the applicable variable.

Similarly,

$$\begin{aligned} \cos\theta < 0 &\Rightarrow x_{ij} - \bar{x}_j < 0 \\ &\Rightarrow x_{ij} < \bar{x}_j \end{aligned} \tag{3.31}$$

If the two points are perpendicular then

$$\cos\theta = 0 \Rightarrow x_{ij} = \bar{x}_j \tag{3.32}$$

This means that for each variable of interest, examination of the plot quickly reveals which observations were above and which below the mean of the variable.

In Figure 3.3, individuals corresponding to row points that are situated in the shaded area are those that have an above average value of the variable g_j. These are entries in the Z matrix having $x_{ij} > \bar{x}_j$.

By choosing a particular column point and projecting the row points onto it, we can also order the size of observations made on that variable. For a column centred matrix, the size and magnitude of deviations from the mean can be ordered across the individuals for a particular variable.

Similarly, choosing a particular row point, we could order deviations from the means on the different variables, i.e. for row i we order $z_{ij}=x_{ij}-\overline{x_j}$ for j=1 to m.
In general, there would not be much to be gained from this unless the units in which the variables were measured were commensurate in some sense.

3.4.2 Within Set Scalar Product Interpretation

If b=1 then the within set scalar product interpretation holds for the columns. Consider two columns k and l of Z. These are denoted by z_{ik} and z_{il}, i=1,...,n.
The squared norm of variable k is

$$\begin{aligned}\|z_k\|^2 &= \sum_{i=1}^{n} z_{ik}^2 \\ &= \sum_{i=1}^{n} (x_{ik}-\overline{x_k})^2 \\ &= (n-1)*s^2(k)\end{aligned} \tag{3.33}$$

and

$$\|z_k\| = s(k)\sqrt{n-1} \tag{3.34}$$

where $s^2(k)$ and s(k) denote the sample variance and standard deviation of variable k respectively.

Therefore the lengths of the columns of Z are proportional to their standard deviations. So by considering the lengths of the plotted column points the relative sizes of the standard deviations of the corresponding variables can be compared.

The scalar product of columns k and l of Z is

$$\begin{aligned} z_k^T z_l &= \sum_{i=1}^{n} z_{ik} z_{il} \\ &= \sum_{i=1}^{n} (x_{ik} - \bar{x_k})(x_{il} - \bar{x_l}) \\ &= (n-1)s(k,l) \end{aligned} \tag{3.35}$$

where s(k,l) is the covariance between variables k and l.
Thus $ZZ^T=(n-1)S$ where S is the variance-covariance matrix of X.

The within set scalar products can also be expressed as

$$z_k^T z_l = \|z_k\| \|z_l\| \cos\theta \tag{3.36}$$

Thus, substituting (3.34) into the denominator, and (3.35) into the numerator, we have

$$\cos\theta = \frac{\sum_{i=1}^{n} (x_{ik} - \bar{x_k})(x_{il} - \bar{x_l})}{\sqrt{\sum_{i=1}^{n} (x_{ik} - \bar{x_k})^2 * \sum_{i=1}^{n} (x_{il} - \bar{x_l})^2}} \tag{3.37}$$

which is the standard Pearson Product Moment Correlation between variables k and l, r_{kl}.

Therefore, if b=1, by considering the angle between two column points, we can determine the correlation between the variables represented by those column points: Acute angles indicate positive correlation. The smaller the angle between the points, the higher the correlation. An angle of 90^0 indicates that the variables are uncorrelated, because $\cos 90^0=0$. Obtuse angles indicate increasing negative correlation, with an angle of 180^0 indicating perfect negative correlation.
Note that the relationship between angle and correlation is non-linear (Fig. 3.5). For example, an angle of 45^0 or 135^0 coincides with correlation $0.7071 \approx \sqrt{5}$.

Notice also that the projections along a column point of the other column points give an ordering of their covariances with that column point.

Euclidean distances between the column points are the Euclidean distances between the columns of Z.

3.4.3 Within Row Interpretation

Similarly, if a=1, the above algebra follows through analogously. However the interpretations do not follow directly. It is not common practice to speak of the variance or correlation of row points. However, as will be shown, such interpretations are of value.

The squared norm of row i of Z is

$$\begin{aligned} \|z_i\|^2 &= \sum_{j=1}^{m} z_{ij}^2 \\ &= \sum_{j=1}^{m} (x_{ij}-\bar{x}_j)^2 \end{aligned} \qquad (3.38)$$

This formula measures the distance of row i from the vector $\bar{x}$ of variable means defined in (3.27). A large value of the squared norm implies that row i is 'far' from the mean row vector, and *vice versa*.

The 'correlation' between two rows k and l is given by

$$\cos\theta = \frac{\sum_{j=1}^{m} (x_{kj}-\bar{x}_k)(x_{lj}-\bar{x}_l)}{\sqrt{\sum_{j=1}^{m} (x_{kj}-\bar{x}_k)^2 * \sum_{j=1}^{m} (x_{lj}-\bar{x}_l)^2}} \qquad (3.39)$$

This 'individual correlation' indicates the tendency for the relationship between two individuals to be linear. The interpretation can be understood by conceptualising it as follows. Row points with a small angle between them have a similar pattern across the variables in the sense that a high (low) value for one of the individuals on a particular

variable is associated with a high (low) value on that variable for the other individual. The two individuals have similar response patterns across the variables. This 'correlation' is made use of in Example 8.1.

If all the variables are measured in the same units, then the centred matrix Z could be obtained by subtracting the overall mean from the entries in X, i.e. $z_{ij}=x_{ij}-\overline{x}$ where $\overline{x}=\sum_{i,j} x_{ij}/(nm)$.

The squared norm of row i is then

$$\|z_i\|^2=\sum_{j=1}^{m}(x_{ij}-\overline{x})^2 \qquad (3.40)$$

and the 'correlation' between rows k and l is

$$\cos\theta=\frac{\sum_{j=1}^{m}(x_{kj}-\overline{x})(x_{lj}-\overline{x})}{\sqrt{\sum_{j=1}^{m}(x_{kj}-\overline{x})^2*\sum_{j=1}^{m}(x_{lj}-\overline{x})^2}} \qquad (3.41)$$

These are more similar in appearance (and interpretation) to the usual concepts of sample variance and correlation than are (3.40) and (3.41). This centring is applied in Example 8.1.

Note that if a=0, and the conventional column centring of (3.26) is used there is the following metric between the row points (from 3.17):

$$FF^T=\frac{1}{n-1}X^TS^{-1}X \qquad (3.42)$$

where S^{-1}, the inverse of the familiar variance-covariance matrix (see 3.3.5) is the particular form of $(Z^TZ)^+$ arising from column centring. This metric between the row points is known as the Mahalanobis metric.

3.5 Overview of the Interpretations

Many interpretations that can be made from the plots are described in the literature; references to summaries of these interpretations were mentioned in Section 3.2. However, the interpretations have not been comprehensively detailed so that they can be applied to all biplot variants. An aim of this thesis is to attempt this.

Gower and Harding (1988) gave some structure to the interpretations in noting that for biplots there are three interpretations. They express the three as:

(1) Scalar products
(2) Pythagorean distances
(3) Covariances.

"Variant forms of biplot allow the simultaneous approximation of any two of these three but not all three can be achieved optimally in one diagram." (Gower and Harding, 1988)

We have expressed the three interpretations given above in broader terms, resulting in a structure which we use to classify biplots, and describe biplot interpretations.

In any particular biplot, at most two of the three approximations can be displayed simultaneously:

(i) Scalar products between the row and the column points
(ii) Scalar products within the row points
(iii) Scalar products within the column points.

We shall refer to interpretations (i), (ii) and (iii).

We described the interpretations in terms of the above structure. Section 3.3.2 described interpretation (i), and Section 3.3.3 described interpretations (ii) and (iii). There we saw that expressing the interpretations in terms of scalar products in fact encompasses other

interpretations; the norm of a vector and the distance between two vectors can be expressed in terms of scalar products (2.16).

Which of the interpretations are displayed depends on the choices made at Phase III (Section 2.9).
For (i) we must have a+b=1
For (ii) we must have a=1
For (iii) we must have b=1

The above structure does not include between set interpretation in terms of the transition formulae of Section 2.10. This interpretation is valid for all biplots. The interpretation is invoked when a between set interpretation is not possible in terms of scalar products (i.e. when $a+b\neq 1$), and is then used to justify plotting of both sets of points on the same axes.

Strictly speaking, there are between and within set relationships for values of a and b other than those described. These are not readily interpretable. For example, the within row relationship for a=0 was included in Section 3.3.4 for the sake of completeness.

The choice of a and b determines the family to which the plot belongs. They are referred to as (a,b) plots.

Biplots in the following chapters have been classified according to the values of a and b.

Chapter 4 Correlation Biplot family ((0,1)-plots)
Chapter 5 Principal Component Biplot family ((1,0)-plots)
Chapter 6 Correspondence Analysis family ((1,1)-plots)

Another family of interest is the Symmetric Biplot family. These are (½,½) plots and hence have a between set, but no within set interpretations. Such a plot can provide useful information about the structure of a matrix (Bradu and Gabriel, 1978).

Using Table 3.1, the interpretations relevant to each biplot can be readily determined.

Within each family, common interpretations hold. For example, members of the Principal Component family have a=1, b=0 and thus interpretations (i) and (ii) in common.

Decomposition of the norm for the rows is valid when a=1, i.e. for the Principal Component and Correspondence Anâlysis families. Decomposition of the norm of the columns is valid when b=1, i.e. for the Correlation Biplot and Correspondence Analysis families. Neither of these decompositions is valid for the Symmetric Biplot family.

Plots have been classified according to the choice of a and b made in Phase III. We shall see that plots within these families are differentiated from each other by the choices made in Phases I and II.

4

CORRELATION BIPLOT FAMILY

4.1 General Interpretations for the Family

The correlation biplot family is defined by:

$$\text{SVDD } (Z,\Omega,\Phi,0,1).$$

The way that Z is obtained from X in Phase I determines which member of the family is obtained.

In the more well known biplots of this family, which we will be describing in this chapter, the cosine of the angle between two column points approximates the generalised correlation between the corresponding variables. The exception to this is Spearman's rank correlation biplot (described in Section 4.5), where a rank correlation is approximated. Hence the name of the family.

The norms of the plotted column points approximate the norm of the corresponding variables. The between set scalar product interpretation holds.

In summary, the two scalar product interpretations which are simultaneously displayed in this family are between set and within the column points (interpretations (i) and (iii)).

4.1.1 Between Set

As a=0 and b=1 is the Phase III choice in this family, the coordinates of the row and column points are given by

$$F=U,$$

$$G=V\Gamma.$$

nocol x nocol nocol nocol

So

$$FG^T=U\Gamma V^T=Z.$$

Thus the scalar products between the plotted row and column points approximate the respective elements of Z. The individual elements of z_{ij} are given by $z_{ij}=f_i^T g_j$.

4.1.2 Within Column Points

In Section 3.3.3 it was shown that when b=1, as is the case for this family, the scalar products of the column points with each other are the scalar products of the columns of Z, i.e. $GG^T=Z^TZ$.

Thus, in this family of biplot, distances between the column points in the display represent the Euclidean distances between the columns of Z. Euclidean distance between the variables is approximated indirectly in the sense that distances can be expressed as the sum of scalar products. These scalar products are directly approximated. Column norms and the angles between the column points are also approximated. The algebra of these approximations is contained in Section 3.3.3.

4.1.3 Within Row Points

As a=0,

$$FF^T=Z((Z^TZ)^+)^TZ^T \qquad (4.1)$$

from (3.13).

For two rows k and l,

$$f_k^T f_l=z_k^T(Z^TZ)^+z_l \qquad (4.2)$$

So the distances between the row points in the display represent the Mahalanobis-like distances between the rows of these matrices (Gabriel 1971, Greenacre 1984, Underhill 1990a).

The interpretations specific to each plot depend on the underlying nature of Z. They will be discussed separately for each member of the correlation biplot family.

4.2 Covariance Biplot

This is formed by centring X in Phase I as $Z=\frac{(X-\bar{X})}{\sqrt{n-1}}$ where $\bar{X}$ is as defined in equation (3.2.7).

The display is described by

$$\text{SVDD}\left(\frac{(X-\bar{X})}{\sqrt{n-1}}, \text{I}, \text{I}, 0, 1\right).$$

The covariance biplot takes its name from its property that the within set scalar products of the column points, G^TG, approximate the variance-covariance matrix of the variables of X, Z^TZ .

The entries of Z are $z_{ij}=\frac{(x_{ij}-\overline{x_j})}{\sqrt{n-1}}$, so

$$\begin{aligned}\|z_k\|^2&=\sum_{i=1}^{n} z_{ik}^2\\&=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ik}-\overline{x_k})^2\\&=s^2(k)\end{aligned} \qquad (4.3)$$

and

$$\begin{aligned}z_k^T z_l&=\sum_{i=1}^{n} z_{ik}z_{il}\\&=\frac{1}{n-1}\sum_{i=1}^{n}(x_{ik}-\overline{x_k})(x_{il}-\overline{x_l})\\&=s(k,l)\end{aligned} \qquad (4.4)$$

Thus the following features are displayed:

1. The scalar product of a column with itself, the squared norm, is the variance of the corresponding variable (from 4.3). The standard deviation of the variable is therefore given by the length of the associated column point. i.e. $s(j)=\|g_j\|$.

2. The scalar product of two column points approximates their covariance, $g_k^T g_l = s(k,l)$ (from 4.4). The cosines of the angles between the column points are correlations between those variables (as in 3.33), i.e. $r_{kl} = \cos\theta_{kl}$ where r_{kl} is the correlation between variables k and l and θ_{kl} is the angle subtended at the origin between the two column points k and l.

3. The between set scalar product interpretation holds. The scalar product of row point i and column point j approximates z_{ij} as defined above.

4.3 Correlation Biplot

This is the display

SVDD (Z,I,I,0,1)

where Z is formed by centring X in Phase I as

$$Z=\left(\frac{1}{\sqrt{n-1}}\right)(X-\bar{X})S^{-1} \qquad (4.5)$$

and S^{-1} is a diagonal matrix with elements s_j, j=1,...,m, the standard deviations of the columns of X, and $\bar{X}$ is as defined in Section 3.4.

The z_{ij} are therefore given by

$$z_{ij}=\frac{x_{ij}-\overline{x_j}}{\sqrt{n-1}\,s_j} \qquad (4.6)$$

Interpretations following from this are:

1. The cosine of the angle between two column points approximates their correlation.

2. Points which have a strong positive correlation are plotted close together, since $\|g_k-g_l\|^2=2(1-r_{kl})$, where r_{kl} is the correlation between variables k and l.

3. The norms of the columns of Z are one.

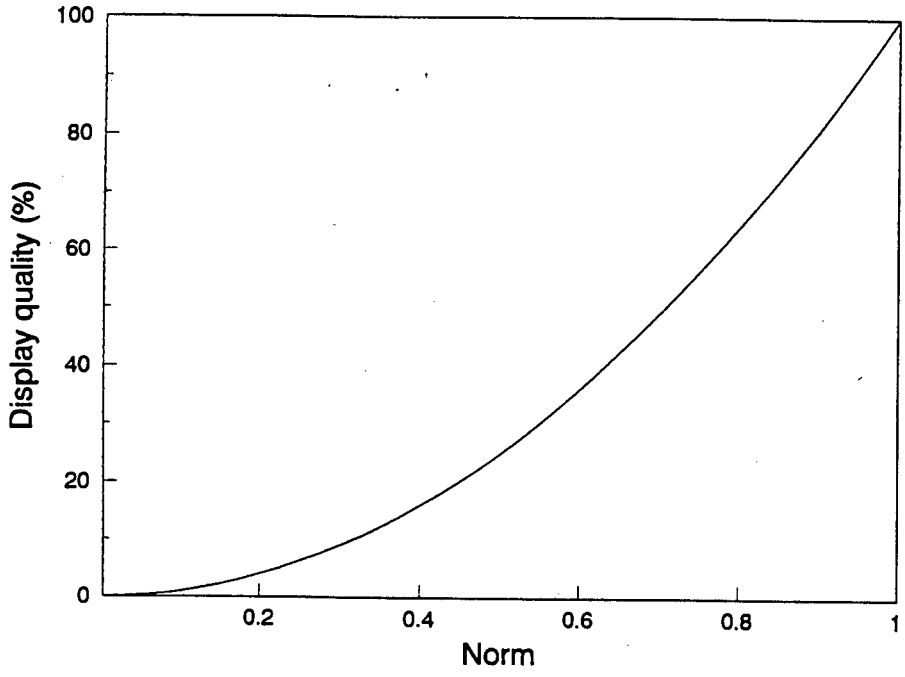
A feature of the plot that follows from the third interpretation is that the norm of the plotted column point indicates the quality of display of the corresponding variable. This is because the standard deviations of the columns of Z are one. The display quality of a column point is

$$\frac{\|g_{j[p]}\|^2}{\|g_j\|^2}=\|g_{j[p]}\|^2 \qquad (4.7)$$

Since the norms of the columns of Z are one, all displayed column points in the lower

rank approximation have norms of at most one and therefore must fall inside the unit circle. In two dimensions, the closer the column points to the unit circle, the higher their quality. Notice that this relationship is not linear, but quadratic (Fig. 4.1). For example, a column point with a norm of 0.5 and therefore plotted halfway to the unit circle, has a quality of display of only 25%.

Figure 4.1



4.4 Coefficient of Variation Biplot

The coefficient of variation biplot (Underhill 1990a) is a (0-1)-plot of the matrix

$$Z=\left(\frac{1}{1-n}\right)^{\frac{1}{2}}(X-\overline{X})\,(diag(\overline{x_1},\ldots,\overline{x_m}))^{-1} \qquad (4.8)$$

The z_{ij} are given by

$$z_{ij}=\frac{1}{\sqrt{n-1}}\frac{(x_{ij}-\overline{x_j})}{\overline{x_j}} \qquad (4.9)$$

The coefficient of variation of a variable j is the ratio of its standard deviation to its mean. It is the most commonly used measure of *relative* variability. The standard deviation is a measure of *absolute* variability. The coefficient of variation is, like the correlation coefficient, dimensionless. The coefficient of variation biplot highlights the relative variability of the columns.

In the coefficient of variation biplot, the norms of the column points give the coefficients of variation of the variables, $\|g_j\|=\frac{s_j}{\overline{x_j}}$. As for all plots in the covariance biplot family, the cosine of the angle between two column points gives the correlation between the corresponding variables.

The squared distance between two column points is given by

$$\|z_k-z_l\|^2=\frac{1}{(n-1)\,\overline{x_k}\,\overline{x_l}}\sum_{i=1}^{n}(x_{ik}-\overline{x_k})\,(x_{il}-\overline{x_l}) \qquad (4.10)$$

(from 3.15 and 4.9), which is a generalised covariance.

The between set scalar product interpretation holds. The scalar product between a row and column point approximates the corresponding element of z_{ij}. Projections of the row points onto a column point allow us to order the x_{ij} as $\overline{x_j}$ and s_j are fixed for

a particular column point. However the use of projecting column onto row points is less clear, unless the variables are measured in commensurate units. The angle between a column and a row point will, as before, indicate whether a particular observation is above or below the mean for a variable.

The coefficient of variation is defined only when the observations are all positive and measured on a ratio scale. For the interpretations from this biplot to be meaningful therefore, all the variables must be positive and measured on a ratio scale.

4.5 Spearman's Rank Correlation Biplot

The above biplot is a member of the correlation biplot family where the Phase I transformation of X to Z is a ranking of the observations within each column point. This results in the cosine of the angle between two column points being their rank correlation.

The correlation measure between two variables to which we have been referring thus far is Pearson's Product Moment correlation. Spearman's rank correlation coefficient (rho) is a frequently used non-parametric measure of correlation, based on ranks. It can in fact be derived directly from from Pearson's Product Moment correlation by replacing the original data with ranks. Like Pearson's correlation, Spearman's rho ranges from -1 to 1. These values have an analogous interpretation to Pearson's measure; the closer the value to zero, the smaller the correlation between the relevant variables. Spearman's rank correlation coefficient can be applied to data that is ordinal, or to the ranks derived from higher order data. Ties are dealt with by assigning average ranks.

If there are no ties then the columns of Z have the same norm. The squared norms are then given by

$$\|z_j\|^2 = \sum_{i=1}^{n} r_{ij}^2 \qquad (4.11)$$

where r_{ij} is the rank of the ith observation of variable j.
The implication of this is that the quality of display of each column point can be assessed by considering its distance from the circle having a radius of $\sqrt{\sum_{i=1}^{n} r_{ij}^2}$. All column points fall inside this circle. This interpretation is analogous to the display quality of the columns in the correlation biplot (Section 4.3).

In order to centre the display at the origin, ranks can be assigned so that the middle rank is zero.

For n odd, ranks from $\frac{-(n-1)}{2}$ to $\frac{(n-1)}{2}$ are assigned.

For n even, a ranking system such as
{...-1.5, -0.5, 0.5, 1.5, ...} could be used.

Two columnn points will tend to be plotted close together if their ranks are similar across the observations.

Thus data that is ordinal in nature can be displayed.

The plot is also of value as a method for dealing with matrices having certain vectors with very large numbers that would otherwise dominate the plot. Converting such points into supplementary points effectively excludes their contribution to the plot.

A disadvantage of the plot is one common to all non-parametric methods - loss of information about the magnitude of differences between observations.

The Spearman's rank correlation biplot is applied to contingency table data in Section 8.4.

4.6 Comparison of Centrings

An advantage of using the coefficient of variation biplot rather than the covariance biplot is that when the scales of measurement are different, standard deviations cannot be meaningfully compared. However the relative variabilities can be compared. The correlation biplot and Spearman's rank correlation biplot do not display variability.

The interpretation of the display for the correlation biplot is almost the same as that of the covariance biplot. However, depending on the nature of the data, and the features that are required to be displayed it can be extremely important which display is used. If the scales of the variables differ greatly, the variables on larger scales have a far better quality of representation in the covariance biplot, at the expense of the other variables.

In the correlation biplot all the variables are standardized, having means of 0 and variances of 1. This prevents the plot from being dominated by a few variables but has the disadvantage that the relative variabilities are not displayed.

By standardizing the scales of measurement in the correlation biplot, we are effectively inflating the relative weight of variables having small standard deviations (and *vice versa*). In some applications this is undesirable, for example when we do not want such variables to have a large influence on the display.

Variables with large coefficients of variation tend to be associated with the large singular values and therefore have a high quality display on the coefficient of variation biplot (and *vice versa*). This does not depend on the original scale of measurement. Variables highly correlated to those with large coefficients of variation will also be well displayed.

The coefficient of variation biplot is useful when the scales of measurement of the variables are different. Variables with large relative standard deviations would dominate a covariance biplot of such a matrix. In a correlation biplot, information about the variability would be lost.

A disadvantage of the coefficient of variation biplot is its limitation to positive variables measured on a ratio scale. This can sometimes be overcome by performing some transformation on the offending variables.

Spearman's rank correlation biplot is useful when there are large discrepancies between the magnitudes of the observations. It is robust with respect to outliers.

The effect of different standardisations on the quality of the display is illustrated in Example 8.2, where the biplots described in this chapter are applied to the same data set.

5

PRINCIPAL COMPONENTS BIPLOT FAMILY

The Principal Components Biplot (PCB) family is defined to be the singular value decomposition displays

$$SVDD(Z,\Omega,\Phi,1,0)$$

where Z is a transformation of the original data matrix X.

These are (1,0)-plots, instead of the (0,1)-plots of the correlation biplot family. Thus the two categories of interpretation that hold are between set and within row scalar products (interpretations (i) and (ii) as defined in Section 3.5).

The family takes its name from the statistical technique of principal components analysis (PCA). A brief review of PCA and how it relates to the PCB follows:

5.1 Principal Components Analysis

PCA is a relatively old mathematical technique. It was first described by Pearson (1901).

Jolliffe (1986) gives the following concise description of PCA: "The central idea ... is to reduce the dimensionality of a data set which consists of a large number of interrelated variables, while retaining as much as possible of the variation present in the data set. This is achieved by transforming to a new set of variables, the principal components, which are uncorrelated, and which are ordered so that so that the first few retain most of the variation present in all of the original variables."

Suppose that there are m variables. The relationships between these variables,

particularly their covariance or correlation structure, is of interest. For large m, it is extremely difficult to examine these relationships directly. In PCA, new variables are derived in such a way as to retain as much as possible of the information given by the old variables. This information is measured in terms of the variance. The technique is especially useful when the number of new variables required to give a good approximation is much smaller than m.

The new variables, the principal components, are linear combinations of the original variables. The principal components are mutually orthogonal i.e. there are no linear relationships among them.

The principal components are not always readily interpretable in their own right. However, when they are interpretable, the principal components can contribute valuable insights into the data. The interpretation is discussed in Section 5.2.

More formally:
Let the vector of the m random variables be denoted by x. The first principal component is given by $\beta_1 x$ where $\beta_1 = (\beta_{11} + \ldots + \beta_{1m})$ is a vector of m constants chosen such that $\beta_1 x$ has the greatest possible variance.

$$\beta_1 x = \beta_{11} x_1 + \ldots + \beta_{1m} x_m = \sum_{j=1}^{m} \beta_{ij} x_j \qquad (5.1)$$

Thus β_1 defines a linear combination of the m original variables. A new variable is defined by $\beta_1 x$.

The kth principal component $(k \neq 1)$ is constructed such that $\beta_k x$ has maximum variance subject to being uncorrelated with the previous principal components.

If the first two principal components are used as axes, and the row points are plotted in two dimensions with respect to these axes, then the points have the maximum variation of all two dimensional representations.

In general, the higher the intercorrelation between the m variables, the more the variation the first few principal components will account for. The last few principal components identify relationships in which there is not much variation.

5.2 Interpretation of the Individual Principal Components

There are no precise rules for interpreting the principal components. As Steffans (1983) notes, interpretation of the principal components in physical terms is subjective.

The first principal component often has positive coefficients for all the variables. This reflects a 'size' component of the individuals. An example of this is anatomical measurements made on species. A 'size' component is expected if all the variables are pairwise positively correlated.

After size has been accounted for, the main sources of variation are indicated on the later principal components. When the size effect itself is not of interest, these later principal components are plotted against one another.

Later principal components are interpretable as 'shape' components. They highlight contrasts between the variables. The contrasts are broadly indicated by coefficients of opposite sign. When the first principal component has positive coeffficients, later principal components have coefficients of opposite sign in order to be orthogonal to the first principal component. Hawkins and Fatti (1983) describe ways of interpreting the information in the later principal components.

When interpreting each principal component, the general pattern or shape of the coefficients across the variables for that principal component should be considered. A simple method for doing this is given in Jolliffe (1986). Small differences in the coefficients are not important. The principal components are more readily interpretable when only a few have coefficients far from zero. The sign of the coefficients is arbitrary. The variance of the principal component remains unchanged if all the signs are changed.

It is sometimes of interest to compare sources of variation between subgroups of individuals using the principal components.

5.3 The Link Between Principal Components Analysis and the Principal Components Biplot

Principal components analysis can be represented as

SVDD (Z,I,I,1,-).

The hyphen means that the column points are not displayed. In PCA, the focus of interest is usually the *columns* of G which contain the coefficients in the linear combinations that define the new variables from the original ones (equation 5.1). As described in Section 5.1, the idea is to work with fewer, uncorrelated variables. The row points (representing the individuals), given by the rows of F, are often plotted.

In the principal components biplot, which is given by

SVDD (Z,I,I,1,0),

the *rows* of G, which contain the coordinates of the points that represent the variables, are displayed. The row points are displayed on the same set of axes, which is justified because the between set scalar product interpretation holds. Note that interpretation of the individual principal components, as described for principal component analysis, holds.

Two ways of preprocessing the data matrix for principal components analysis and biplots are commonly performed. These are column centring and column standardization. These are the same centrings (Phase I) as those for the correlation and covariance biplots (Chapter 4).

5.4 Interpretations for the Principal Components Biplot Family

As noted in the previous section, column centring and column standardization are the usual preprocessing (Phase I) options applied to principal component biplots. There is no reason why other centrings are not more commonly applied. For example, Underhill (1990a) notes that a coefficient of variation centring (Section 4.4) may be useful under some conditions. We apply a Spearman's rank preprocessing (Section 4.5) in Example 8.4.

The same principles for choice of centre as discussed in Section 4.6 for the correlation biplot family apply for the principal components biplot family. Some choices of preprocessing are listed in Section 7.1. Phase I choice is discussed further in Chapter 9.

Principal components biplots are (1,0)-plots and therefore have interpretations (i) and (ii) (Section 3.5). In addition, interpretations of the individual principal components (Section 5.2) for the row points can be performed.

Since $a+b=1$, scalar products between the row and column points approximate the entries of Z (interpretation (i)) as described in Section 3.3.2.

Within row scalar products are preserved (interpretation (ii)). Thus the interpretations are those presented in Section 3.3.3(b). The distances between row points in the display represents Euclidean distances between the row points in the full rank matrix. The norms of the row points approximate the norms of the rows of Z.

The angles between the row points in the covariance, correlation and coefficient of variation type centrings represent the 'individual correlations' between the row points as described in Section 3.4.3.

The within row points interpretations for each centring are analogous to those described for the column points in Chapter 4. Thus, for example, in the correlation biplot centring, the distance of a plotted row point from the unit circle indicates its quality of display.

Rows with large norms for a column centred Z indicate individuals with large 'standard deviations'.

For the case of a column centred matrix, within row interpretations are described in Section 3.4.3. The distance between the column points approximates Mahalanobis distance between the columns.

6

CORRESPONDENCE ANALYSIS FAMILY

6.1 Introduction

Correspondence analysis has been popularised by its application to the display of two-way contingency tables. It is usually described with reference to contingency tables, although it can be applied to other matrices with non-negative entries.

A history of correspondence analysis appears in Greenacre (1984). The algebraic basis of the technique can be traced back to 1935. Greenacre (1981) notes that the technique is theoretically equivalent to other techniques that have appeared since the mid 1930s. Among these are: simultaneous linear regression, reciprocal averaging and dual scaling. Much of the development of correspondence analysis was in the psychometric literature. Correspondence analysis in its current form originated in France in the early 1960s, in the context of linguistics. Benzecri was a leading figure in these developments (see, for example, Benzecri, 1969). Correspondence analysis has become very popular in other parts of the world in recent years, and is applied to diverse fields. Greenacre (1984) presents numerous applications, including genetics, social psychology, linguistics and education and gives further references to published applications. Examples of more recent applications include ecology (Digby and Kempton, 1987), chemistry (Underhill and Peisach, 1985) and market research (Shahim and Greenacre, 1988). We apply correspondence analysis to ornithology (Section 8.4). Much of the literature on correspondence analysis is written in French. Books by Greenacre (1984), who was a student of Benzecri, and by Lebart, Morineau and Warwick (1984), helped to make this work accessible to an English readership.

Correspondence analysis is actually a type of biplot, as both the row and column points of the matrix are displayed. We describe correspondence analysis with reference to the framework for biplots presented in Chapter 3. Correspondence analysis is essentially a (1-1)-plot with a centring and metric which result in within set interpretations in terms of chi squared distances. Because a+b=2, the display does not have the scalar product interpretation. The joint display of row and column points is motivated separately - between set interpretations can be made with reference to the transition formulae.

Other (1-1)-plots are discussed in Section 6.5, and are applied in Examples 8.1 and 8.4.

6.2 Contingency Tables and the Chi Squared (X^2) Metric

A contingency table contains count data. The data is classified according to two variables - a row and a column classification. The count in a particular cell is the number of entries that fall into the corresponding row and column classification. For example, we could classify a group of people by age group and by opinion on a certain issue. Contingency tables are reported in a variety of contexts.

When studying contingency tables, dependencies between the row and column classifications, and the nature of this relationship are of interest.

Two events A and B are said to be statistically independent if

$$P(A \cap B) = P(A)\,P(B) \qquad (6.1)$$

In the context of the above contingency table example, this says that the probability that a person is of a certain age group and opinion (i.e. falls into a particular cell) is equal to the product of these two separate probabilities, if the age and opinion classifications are independent of each other. Thus under the null hypothesis that the rows and columns of the table are statistically independent, the expected frequency of each cell can be calculated. Therefore a 'large' difference between the expected and the observed

frequency for a particular cell is an indication of dependence.

We will use the following notation:

X is an nxm contingency table with elements x_{ij}

X_{ri} and X_{cj} are the row and column totals of X, respectively.

$$N=\sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}.$$

Under the null hypothesis of independence, the probability that an observation falls into cell ij (i.e. the ith row classification and the jth column classification) is

$\frac{X_{ri}}{N}\frac{X_{cj}}{N}$ (from 6.1), and the expected frequency of cell ij is

$$\frac{X_{ri}X_{cj}}{N} \qquad (6.2)$$

The Pearson X^2 statistic is a measure of the dependence between the rows and the columns of a matrix. It is frequently used in tests of association in contingency tables. This test statistic for the matrix X is given by

$$D^2=\sum_{i,j}\frac{(x_{ij}-X_{ri}X_{cj}/N)^2}{X_{ri}X_{cj}/N} \qquad (6.3)$$

where

x_{ij} is the observed frequency in cell k,

$X_{ri}X_{cj}/N$ is the expected frequency in cell k, under the null hypothesis of independence (from 6.2),

and the summation is done over all cells k in the table (k=1,...,nxm).

This statistic has approximately a X^2 distribution with (n-1)(m-1) degrees of freedom.

High values provide evidence for a departure from the null hypothesis of row-column independence.

Consider P, the matrix with entries $p_{ij}=x_{ij}/N$. It is a matrix of the relative frequencies of X.

The row and column totals of P are r_i and c_j respectively .

The vectors and diagonal matrices of the r_i and c_j are r, c and Dr, Dc respectively.

Note that $\sum_{i=1}^{n}\sum_{j=1}^{m} p_{ij}=\sum_{i=1}^{n} r_i=\sum_{j=1}^{m} c_j=1$.

P, the matrix of relative frequencies, has a grand total of 1. Thus the expected frequency of cell ij in (6.2) reduces to r_ic_j and its associated X^2 statistic is

$$I=\sum \frac{(p_{ij}-r_ic_j)^2}{r_ic_j} \qquad (6.4)$$

In correspondence analysis, I is known as the *inertia*, and is related to the Pearson X^2 statistic for the matrix X by

$$D^2=NxI.$$

Inertias, rather than the chi squared values are referred to in correspondence analysis.

The centred matrix Z used in correspondence analysis has entries:

$$z_{ij}=\frac{p_{ij}-r_ic_j}{r_ic_j} \qquad (6.5)$$

The inertia, I, is actually

$$I=\|Z\|^2_{Dr,Dc} \qquad (6.6)$$

In matrix notation, $Z=D_r^{-1}(P-rc^T)D_c^{-1}$ $(=D_r^{-1}PD_c^{-1}-\underline{1}\underline{1}^T)$ (Phase I). In order to achieve the generalization (6.6), the choice of the Phase II metrics must be $\Omega=D_r$ and $\Phi=D_c$.

Thus the centring of Z as in (6.5) allows within set interpretations in terms of chi squared distances.

6.3 Correspondence Analysis

Correspondence analysis is the display

$$\text{SVDD } (Dr^{-1}(P-rc^T)Dc^{-1},Dr,Dc,1,1)$$

Note that correspondence analysis is, unlike the (1-0) and (0-1)-plots in the previous sections, symmetric in its treatment of rows and columns. This is due to the choices made in each of the three phases. Thus the correspondence analyses of the matrices X and X^T are identical.

In the following sections we describe biplot interpretations for correspondence analysis. The key concept here is that of statistical independence:
From (6.5) we can express the elements of the transformed matrix Z as

$$z_{ij}=\frac{p_{ij}}{r_i c_j}-1 \qquad (6.7)$$

Under the null hypothesis of row-column independence, the observed and expected elements of P are equal, i.e. $p_{ij}=r_i c_j$ and so $z_{ij}=0$. Values of z_{ij} far from zero indicate dependence.

A *row profile* is the vector obtained by dividing a row vector by the sum of its elements. A profile can be conceptualised as the shape of a vector, i.e. as the graph of its relative frequencies in the case of a row of a contingency table.

The *average row profile* is the profile of the column totals of the matrix. It is the weighted average or centre of gravity of the row profiles. The weight for each row is its row sum. For the P matrix, the average row profile is $c=(c_1,...,c_m)$. The row weights are given by the r_i. The *average column profile* is $r=(r_1,...,r_n)$.

The centre of gravity of Z is located at the origin of the axes. This point represents independent row and column vectors. It is the position for the average row and column

profiles.

6.3.1 Within Set Interpretations

Because a=b=1, scalar products within the row points and within the column points of the displayed matrix approximate the corresponding scalar products of Z.

The interpretation for the row points is described first. Because correspondence analysis treats the rows and columns in the same way, the within column points interpretations are analogous.

The general expression for scalar products between the row points is given by

$$FF^T = Z\Phi Z^T = Dr^{-1}(P-rc^T)Dc^{-1}(P-rc^T)^T Dr^{-1} \qquad (6.8)$$

(from 3.3.5).

For row k, we have from (3.17) that

$$\begin{aligned} \|f_k\|^2 &= \|z_k\|^2_{Dc} \\ &= \sum_{j=1}^{m} c_j \left(\frac{p_{kj}}{r_k c_j} - 1 \right)^2 \\ &= \sum_{j=1}^{m} c_j \left(\frac{p_{kj}}{r_k} - c_j \right)^2 \end{aligned} \qquad (6.9)$$

which is the weighted distance from the vector representing perfect independence. It is also interpretable as the distance between row profile k and the average row profile. It is a type of generalised variance. Thus a row point with a large norm (far from the origin) indicates dependence. Profiles plotted near the origin do not differ much from the average profile (or alternatively, they may be badly approximated).

The scalar product between row points k and l is (from 3.16)

$$\begin{aligned} f_k^T f_l &= z_k^T D_c z_l \\ &= \sum_{j=1}^{m} c_j \left(\frac{p_{kj}}{r_k c_j} - 1 \right) \left(\frac{p_{lj}}{r_l c_j} - 1 \right) \end{aligned} \qquad (6.10)$$

which is a generalised covariance between the row points.

A generalised correlation is given by

$$\cos\theta = \frac{z_k^T D_c z_l}{\|z_k\|_{Dc} \|z_l\|_{Dc}} \qquad (6.11)$$

where θ is the angle between f_k and f_l (from 3.19).

The squared distance between two row points k and l is given by

$$\begin{aligned} \|f_k - f_l\|^2 &= \|z_k - z_l\|^2_{Dc} \\ &= \sum_{j=1}^{m} \left(c_j \left(\frac{p_{kj}}{r_k} - \frac{p_{lj}}{r_l} \right) \right)^2 \end{aligned} \qquad (6.12)$$

(from 3.18). If the distance between row profile and the average profile is small then its frequency is approximately proportional to the average frequency.

This distance between the row points is X^2 distance. It is a generalised Euclidean metric. Each term is weighted by the inverse of its frequency. Large differences, in cases where the total column frequency is large, are reduced. The smaller differences are increased.

In this metric, row vectors with similar profiles have small distances between them. This is irrespective of the column totals.

The Euclidean distances between the rows of Z, the displayed matrix, are the chi-squared distances between the rows of X, the matrix of observed frequencies.

6.3.2 Between Set Interpretations

The between set interpretation has been the subject of some controversy in recent literature (e.g. Carroll *et al* (1986)). Erroneous interpretations of the relationship between

row and column points are common. Greenacre (1989) notes that "... (in correspondence analysis) ... a debate has always existed over the legitimate interpretation of the display". As Greenacre and Hastie (1987) emphasize, there is no specifically intended between set distance concept in correspondence analysis. (This holds true for all (1-1) plots.)

Because $a+b\neq 1$, the scalar product interpretation does not hold. Between set scalar products are not approximated. This means that the distance between a row and a column point is not directly interpretable.

There is, however, an interpretable relationship between the row and column points. Row points are attracted away from the origin in the direction of columns in which they have large entries. The interpretation is due to the transition formulae (Section 2.10):

$$F=Dr^{-1}PG\Gamma^{-1}$$

$$G=Dc^{-1}P^{T}F\Gamma^{-1}$$

The coordinate for the ith row on the kth axis is given by

$$f_{ik}=\frac{1}{\alpha_k}\sum_{j=1}^{m}\frac{p_{ij}}{r_i}g_{jk} \qquad (6.13)$$

Therefore a row point is attracted towards the direction of the column points in which it is most prominent. Note however that the positions of all the column points determine the positions of the row points. Similarly, column points are attracted to row points in which they are prominent.

6.3.3 Decomposition of the Norm

As a+b=1, decomposition of the norm can be done for both the rows and the columns.

$$\|Z\|^2_{Dr,Dc} = \sum_{i=1}^{n}\sum_{j=1}^{m} r_i c_j \left(\frac{p_{ij}-r_{ic_j}}{r_i c_j}\right)^2 = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{(p_{ij}-r_i c_j)^2}{r_i c_j} \qquad (6.14)$$

This is Pearsons chi squared statistic for the P matrix, or the inertia.

Decomposition of inertia is detailed in Underhill and Peisach (1985).

6.4 Symmetric and Asymmetric Plots

Greenacre (1989) describes the between set interpretation with reference to three plots: two of these are termed asymmetric plots. The third, a symmetric plot, can be thought of as the concurrent display of the two asymmetric plots.

The two asymmetric plots are given by :

Plot 1 : SVDD $(Dr^{-1}(P-rc^T)Dc^{-1},Dr,Dc,0,1)$

$F=Dr^{-1}U$

$G=Dc^{-1}V\Gamma$

and

Plot 2 : SVDD $(Dr^{-1}(P-rc^T)Dc^{-1},Dr,Dc,1,0)$

$F=Dr^{-1}U\Gamma$

$G=Dc^{-1}V$

In the first plot distances within the column points are interpretable, while the second has interpretable distances within the row points. They are in fact (0-1) and (1-0)-plots with the same centring as correspondence analysis. The within set metrics are chi squared.

The two asymmetric plots have a+b=1, and therefore a between set interpretation in terms of scalar products. The between set interpretation is that since $z_{ij}=\frac{p_{ij}}{r_i c_j}-1$, by

considering a column point j, projections of the row points onto this point allow us to order z_{ij} for the i=1,...,n. Large projections indicate entries far from from those expected under the null hypothesis of row-column independence.

Greenacre and Underhill (1982) note that display

$$\text{SVDD}(P\text{-}rc^T, Dc^{-1}, Dr^{-1}, \tfrac{1}{2}, \tfrac{1}{2})$$

displays the deviations $\frac{(p_{ij}-r_i c_j)}{\sqrt{r_i c_j}}$, as a between set interpretation.

6.5 Other (1-1)-Plots

Plots in this family have interpretations (ii) and (iii) which are within set interpretations (Section 3.5). They do not have a between set scalar product interpretation. Justification for plotting the row and column points on the same axes is that their relative positions can be interpreted with the aid of the transition formulae (equations 2.47 and 2.48).

For example, a column centered matrix with a (1-1) Phase III choice in ordinary Euclidean space has the within row interpretation of the principal components biplot and the within column interpretation of the covariance biplot.

In all (1-1)-plots, decomposition of both the row and the column norms are applicable (Section 2.11)

For any particular centring and metrics, the relevant within set interpretations follow directly from Section 3.3.3, in the same way as was followed through for correspondence analysis in Section 6.3 above. Between set interpretations follow from the transition formulae of Section 2.10.

Correspondence analysis is not the only biplot that is symmetric in its treatment of rows and columns. An example of another such biplot is given in Example 8.1.2 where the

overall mean was subtracted from the elements of the X matrix. Further applications of (1-1)-plots appear in Examples 8.1 and 8.4.

A discussion of the merits of (1-1)-plots appears in Section 9.1.

7

INTRODUCTION TO THE PRACTICAL EXAMPLES

7.1 Preprocessing the Data Matrix

There are an infinite number of preprocessing possibilities. The choice of preprocessing is one of the issues addressed in the practical examples. Here we describe preprocessing possibilities, but do not enter into the debate of which one to apply in a given situation, as this is discussed in Chapters 3 and 9.

Preprocessing (Phase I) affects interpretations possible from the plot, such as the distance measures. Some transformations have drawbacks in that they complicate the interpretations.

The quality of the plot is determined *inter alia* by the preprocessing chosen.

Commonly used transformations are to take logarithms of the data (Example 8.4.3) and the square root transform. Log transformations lessen the differences in the relative sizes of the entries. Choosing a transformation is discussed by many authors, e.g. Dolby (1963), Tukey (1977). Families of transformations are described in Draper and Smith (1981).

Centring is one form of preprocessing. A key purpose in centring the data is to ensure that the origin is inside the cloud of points to be displayed. The origin is an element of every subspace. Thus the origin will be included in every display. As the low rank approximation is a subspace that is also closest to the original data, an origin 'far' from this data will represent the data less well. This is illustrated by the comparison of Examples 8.1.1 and 8.1.2, as discussed in Section 9.2.

Underhill (1990b) lists many centrings found to be useful. Some of these are listed below.

If a method is applied to one of the practical examples in Chapter 8, mention of this is made after the description of the centring.

Centring operations can be broadly classified into the following categories: column centring, row centring, double centring, overall centring, rank one centring and model centring.

The following notation is used:

X is an nxm data matrix with elements x_{ij}

Z is the centred matrix with elements z_{ij}

$x_{ij[1]}$ - the i,j th element of the rank one approximation to X

r_i - the sum of the elements in the ith row

c_j - the sum of the elements in the jth column

$\bar{r}_i$ - the mean of the ith row

$\bar{c}_j$ - the mean of the jth column

t - the sum of all the elements of X

$\bar{t}$ - the mean of all the elements of X

1. Column Centring

Subtract the mean of each column from each of the elements in the column:

$$z_{ij}=x_{ij}-\overline{c_j} \qquad (7.1)$$

This is the most commonly used centring method. Often the columns represent the variables. Thus the mean of a column is the mean of the observations for that column. Column centring has the effect of shifting the origin to the centre of gravity of the *rows* of the data matrix. The covariance biplot centring is a type of column centring. Applications of column centring are Examples 8.2.1 and 8.1.3.

2. Row Centring

Subtract the mean of each row from each of the elements in the row:

$$z_{ij}=x_{ij}-\overline{r_i} \qquad (7.2)$$

If the columns of the data matrix are observations on variables, then this option only makes sense if the variables that are row centered are measured in the same units. This centring is applied in Examples 8.1.4 and 8.2.2.

3. Overall Centring

Subtract from each element of the matrix the overall mean of the matrix.

$$z_{ij}=x_{ij}-\overline{t} \qquad (7.3)$$

The overall mean of a matrix only makes sense if all the variables are measured in the same units. The centring is used by Bradu and Gabriel (1978) as an aid to fitting a model to a matrix. We apply the centring in Example 8.1.2.

4. Double Centring

Subtract from each element of the matrix the means for its row and column, and add the overall mean.

$$z_{ij}=x_{ij}-\overline{r_i}-\overline{c_j}+\overline{t} \qquad (7.4)$$

The matrix analysed consists of the residuals from fitting both row and column effects. Thus interactions between rows and columns are emphasized. Example 8.1.5 makes use of this centring.

5. Rank One Centring

From each element of the matrix subtract the corresponding element of the rank one approximation matrix.

$$z_{ij}=x_{ij}-x_{[1]ij} \qquad (7.5)$$

This is effectively the centring used in Correspondence Analysis, and was used by Gabriel (1971) in his original biplot paper.

6. Model Centring

Subtract from each element of the matrix its estimated value according to a model. Suppose the model fitted to the matrix is $\hat{X}=(\hat{x}_{ij})$. Then

$$z_{ij}=x_{ij}-\hat{x}_{ij} \qquad (7.6)$$

Other transformations which are made use of are *doubling* (Example 8.3.2), ranked data (Example 8.4.6) and expressing the entries as a percentage of their row or column total (Example 8.4.1).

Different transformations can be applied to different vectors of the matrix. An example of this is given in Section 8.2 (runners) where row centring is only applied to some of the variables.

Choice of Phases is discussed further in Chapter 9.

7.2 Computing and Interpreting the Examples

Computing of the biplots applied to data sets in the thesis was performed using the SVDD program written by Underhill (1990b).

In the presentation of the examples (e.g. Table 7.1), the following abbreviations are used:

percent	- The percentage of the squared norm accounted for on each axis (equation 3.2).
cumul	- The cumulative totals of 'percent'.
mass	- The weight assigned to the column or row point.
inrt	- The percentage of the norm of Z that the point accounts for.
fact	- The coordinate of the point on the axis.
cor	- The relative contribution (equations 2.59 and 2.60). This is the quality of

display of the point on the axis.

ctr - When decomposition of the norm is applicable, this represents the absolute contribution (equation 2.57). It is interpreted as the contribution of the point to the axis. If a or b equals zero, it is the relative squared distance from the origin for the point on the axis, described by equation (7.7) below.

qual - The quality of display of the point in two dimensions. This is equal to the sum of the 'cor' on the first two axes.

Decomposition of the norm for the rows is valid when a=1, and for the columns when b=1. If b=0, the decomposition of the norm for the columns breaks down. However, useful diagnostics are generated by calculating

$$\frac{g_{jk}^2}{\sum_{j=1}^{m} g_{jk}^2} \qquad (7.7)$$

These may be interpreted as relative squared distances from the origin of each variable on the kth axis. Analogous results hold for the rows when a=0. These quantities are given in the 'cor' column.

If one or two variables have large values of g_{jk}^2 then these few variables dominate the kth axis. The worst possible scenario is when one variable dominates the first axis, and one variable the second, in which case the biplot is effectively a scatterplot of these two variables.

7.3 Quality of the Display - An Example

To illustrate the use of the quantities given above consider a covariance biplot of the data set (Figure 8.1.7) whose entries are the amount of pollution during the months

(rows) of the year at different places (columns) in Cape Town. The data set is explored in Section 8.1.

Eighty four percent of the squared norm of Z is approximated on the first axis (Table 7.1), and 11% on the second. Thus 95% of the squared norm is retained overall.

Decomposition of the norm, and hence interpretation of the absolute (ctr) and relative (cor) contributions only makes sense for the rows when a=1, and for the columns when b=1. Here we have a=0 and b=1, so only the column decomposition of the norm is applicable. The interpretation of the 'cor' column for the rows is in terms of proportions of squared distances from the origin (equation 7.7).

Consider for example the column point representing the site Salt River. It has a 90% quality of display in two dimensions (Table 7.1). Its inertia is 4,7% of the Z matrix. It contributes 4,8% to the first axis and 87,6% of its squared norm is displayed on that axis. The overall contribution of the point to the two dimensional display is 0.048(84)+0.01(11)=4,14%.

The row point 'July' has a relative squared distance from the origin of 20%. This is expressed relative to the other row points. Points 'close' to the origin are not well displayed. We could also look to the values of the coordinates on the axes for this information.

A low value of 'qual' for a point means that the point has substantial components in higher dimensions. Examples of such points are November and April.

The 'plot positions' listed below each figure give the coordinates of the points relative to the numbers on the border of the plots. Thus for example, the origin is at vertical position 30 and horizontal position 50.

TABLE 7.1

THE EIGENVALUES

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	129.18324	84.213	84.213
2	16.25179	10.594	94.807
3	3.56679	2.325	97.132
4	2.44614	1.595	98.727
5	1.22005	0.795	99.522
6	0.54782	0.357	99.879
7	0.18563	0.121	100.000

THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	FACT 2	COR
1	JULY85	A	1.000	0.299	0.098	0.4467	0.200	-0.0827	0.007
2	AUG	B	1.000	0.639	0.086	0.4933	0.243	-0.3760	0.141
3	SEP	C	1.000	0.387	0.117	-0.0815	0.007	-0.5572	0.310
4	OCT	D	1.000	0.468	0.047	-0.1689	0.029	-0.3531	0.125
5	NOV	E	1.000	0.024	0.114	-0.1383	0.019	-0.0056	0.000
6	DEC	F	1.000	0.281	0.077	-0.3736	0.140	0.1121	0.013
7	JAN86	G	1.000	0.228	0.081	-0.3507	0.123	0.0792	0.006
8	FEB	H	1.000	0.151	0.033	-0.1839	0.034	-0.0260	0.001
9	MAR	I	1.000	0.105	0.075	-0.1313	0.017	0.1938	0.038
10	APR	J	1.000	0.085	0.113	-0.0854	0.007	0.2441	0.060
11	MAY	K	1.000	0.483	0.047	0.1935	0.037	0.3502	0.123
12	JUNE	L	1.000	0.412	0.112	0.3801	0.144	0.4211	0.177

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	CITY HALL	1	1.000	0.999	0.259	0.5229	0.687	0.212	-0.3520	0.311	0.762
2	SALT RIVER	2	1.000	0.896	0.047	0.2500	0.876	0.048	0.0379	0.020	0.009
3	PAARDEN ISL	3	1.000	0.903	0.152	0.4500	0.868	0.157	0.0900	0.035	0.050
4	CITY HOSP	4	1.000	0.938	0.121	0.4063	0.886	0.128	0.0982	0.052	0.059
5	EPPING	5	1.000	0.943	0.221	0.5642	0.937	0.246	0.0459	0.006	0.013
6	TAMBOERSKLF	6	1.000	0.803	0.012	0.1164	0.726	0.010	0.0380	0.077	0.009
7	FORESHORE	7	1.000	0.949	0.187	0.5065	0.894	0.199	0.1260	0.055	0.098

8

PRACTICAL EXAMPLES

The data sets used to illustrate the biplots are from diverse areas - Ecology (more specifically, ornithology), Medicine (Sports Science), Social Science and Environmental Studies (atmospheric pollution).

These data sets were chosen with the aim of covering a broad spectrum of the types of data sets and problems likely to be encountered in practice. On the other hand we do not wish to overwhelm the reader with too many examples. Clearly, we cannot cover all possibilities. The emphasis is on giving a general flavour of the techniques and their interpretations.

The first example (Example 8.1 - Atmospheric Pollution) has a data matrix whose rows and columns are measured in the same units. A variety of the centring options described in Chapter 7 were found to be useful here. The consequences of different choices of family is demonstrated.

Example 8.2 (Marathon Runners) deals with a data set in which there are a large number of variables, which are measured in very different units. Biplots from the correlation biplot family are presented and compared.

Example 8.3 (Quality of Life in the United States) deals with a multivariate time series data set.

In the final example (Example 8.4 - Bird Conservation), the data takes the form of a contingency table of counts of birds at sections of coastline. There are large variations in the magnitudes of the entries, making this a particularly difficult data set to display.

Biplot techniques applied here include Ter Braak's diversity biplot (a member of the Principal components family), correspondence analysis and other (1-1)-plots, and the Spearmans rank correlation biplot.

In the presentation of the practical examples, the example, display and the qualities of approximation are given the same number. Thus the second biplot of Example 8.1 is described in Section 8.1.2, displayed in Figure 8.1.2 and the qualities of approximation are given in Table 8.1.2.

EXAMPLE 8.1
ATMOSPHERIC POLLUTION IN CAPE TOWN

The data (Table 8.1) are monthly averages from July 1985 to June 1986 of smoke (soiling figure per m^3) at seven different sites in Cape Town (City Engineer, Cape Town, 1987).

Plot 8.1.1.

This is a (1-1)-plot on uncentred data (Figure 8.1.1). Phase I was not performed on the matrix.

Quality of the Display

The overall quality of the display is 99%, with 98% accounted for by the first axis (Table 8.1.1). Thus this is essentially a one dimensional display. The unusually high quality of the display is due to the fact that both the row and the column points have high generalised correlations (Section 3.3.5), i.e. there are strong linear relationships between them. All points have a display quality of at least 97%. This illustrates that a point that does not contribute much to the axis can still be well displayed. January contributes 2,9% to the first axis yet has a display quality of 97,3% on that axis. This is due to the multicollinearity of the data.

Interpretations

Considering that the data are uncentred, this is a surprizingly good display. Most of the variation is accounted for. It displays interesting features of the data such as the high positive correlations and the pattern formed by linking the month points sequentially, showing the seasonal variation (Figure 8.1.1). The first axis orders the sites from that with the least to that with the most mean pollution. Pollution is high in winter (May to August) and lower in Summer. The sites associated with high levels of pollution (Epping and Foreshore) are attracted towards the winter months. Similarly, Tamboerskloof, with the lowest pollution levels, is situated near the summer months. The point City Hall is somewhat distinct from the other row points. It has a particularly large value in August, and is attracted away from the origin, in the direction of the August point.

TABLE 8.1 ATMOSPHERIC POLLUTION-CAPE TOWN

Monthly average for smoke (soiling figure per m^3)

	CITY HALL	SALT RIVER	PAARDEN ISLAND	CITY HOSPITAL	EPPING	TAMBOERS-KLOOF	FORE-SHORE
JULY 1985	6.7	4.2	6.3	4.1	8.4	2.2	6.8
AUGUST	8.2	4.0	6.8	4.5	7.3	2.2	7.0
SEPTEMBER	5.8	3.1	3.8	2.2	4.0	1.1	3.7
OCTOBER	4.6	2.4	3.6	1.7	3.9	1.3	3.7
NOVEMBER	3.5	2.7	2.9	2.4	4.7	1.9	4.5
DECEMBER	1.9	1.9	3.8	1.4	2.8	1.3	2.8
JANUARY 1986	2.0	2.1	3.2	1.6	3.7	1.1	2.6
FEBRUARY	3.4	2.5	3.9	2.3	3.8	1.6	3.8
MARCH	2.9	3.3	4.2	2.1	4.1	1.7	4.8
APRIL	3.0	3.2	4.0	2.5	4.4	1.5	5.5
MAY	4.0	4.0	5.9	3.7	6.3	2.0	6.3
JUNE	4.8	4.0	7.0	5.5	7.0	2.1	7.0

Comments

The origin is *not* contained within the cloud of plotted points. The points do not have a good spread over the plot, but are bunched up on the right hand side of the first axis. Despite the fact that the matrix is uncentred, interesting features of the data are highlighted. It is unusual to get a meaningful display of uncentred data. However the display does not lend itself to more detailed analyses using the within set scalar product interpretations. We are looking at the data from a position that is 'far' from the cloud of points, which suggests that centring is required.

TABLE 8.1.1

THE EIGENVALUES

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	1432.63940	98.163	98.163
2	16.62120	1.139	99.302
3	3.80243	0.261	99.562
4	3.22000	0.221	99.783
5	2.22933	0.153	99.935
6	0.54926	0.038	99.973
7	0.39271	0.027	100.000

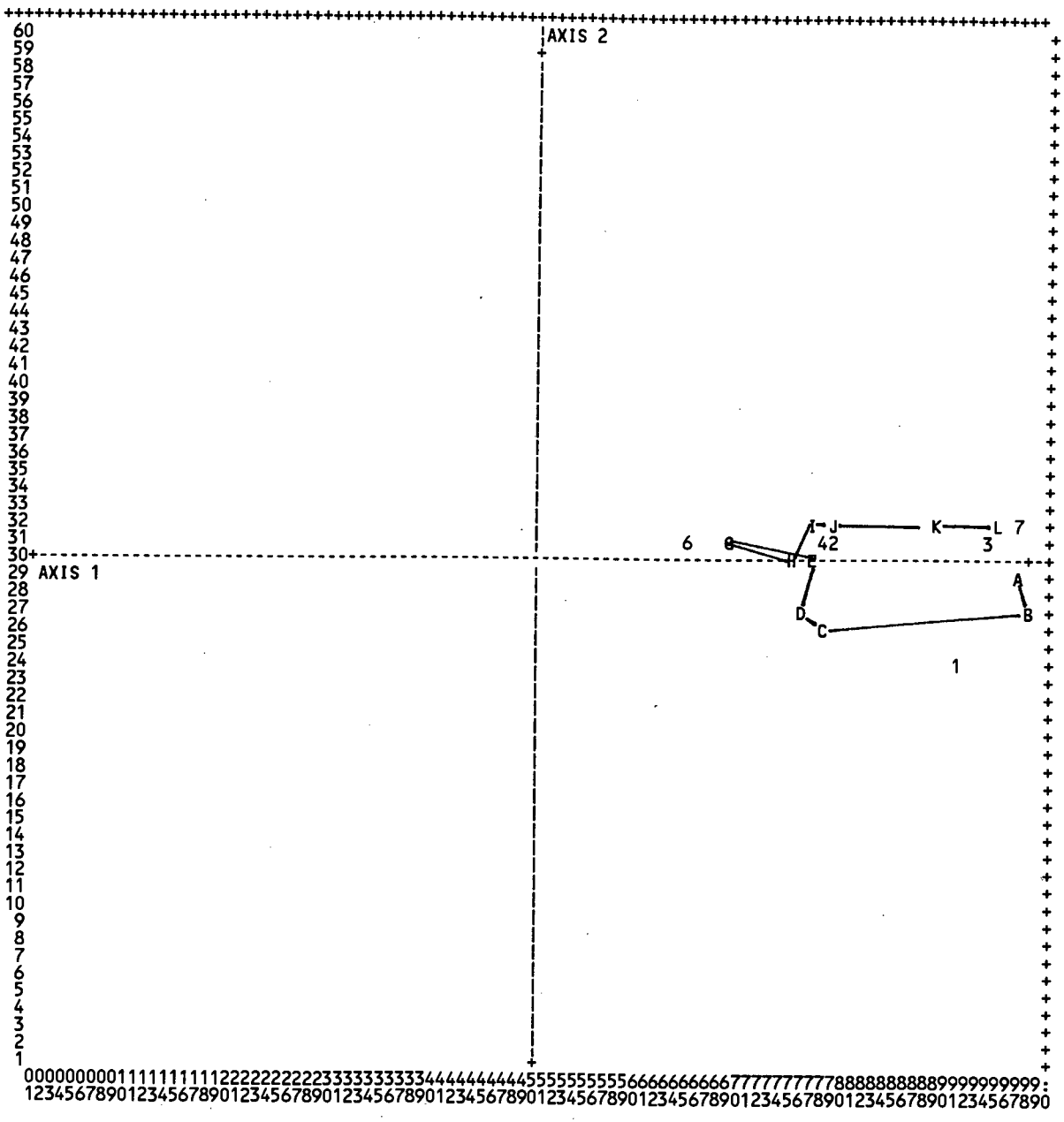
THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	JULY85	A	1.000	0.995	0.165	0.1546	0.994	0.167	-0.0057	0.001	0.019
2	AUG	B	1.000	0.998	0.176	0.1591	0.986	0.177	-0.0178	0.012	0.192
3	SEP	C	1.000	0.994	0.064	0.0940	0.946	0.062	-0.0213	0.049	0.273
4	OCT	D	1.000	0.996	0.050	0.0846	0.975	0.050	-0.0125	0.021	0.094
5	NOV	E	1.000	0.980	0.055	0.0884	0.980	0.054	0.0011	0.000	0.001
6	DEC	F	1.000	0.969	0.028	0.0626	0.955	0.027	0.0076	0.014	0.034
7	JAN86	G	1.000	0.981	0.029	0.0646	0.973	0.029	0.0059	0.008	0.021
8	FEB	H	1.000	0.997	0.048	0.0834	0.997	0.049	0.0006	0.000	0.000
9	MAR	I	1.000	0.989	0.058	0.0907	0.978	0.057	0.0096	0.011	0.055
10	APR	J	1.000	0.989	0.064	0.0954	0.976	0.064	0.0110	0.013	0.073
11	MAY	K	1.000	1.000	0.112	0.1273	0.989	0.113	0.0135	0.011	0.110
12	JUNE	L	1.000	0.992	0.151	0.1472	0.982	0.151	0.0146	0.010	0.127

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	CITY HALL	1	1.000	1.000	0.174	0.1553	0.947	0.168	-0.0366	0.053	0.806
2	SALT RIVER	2	1.000	0.990	0.085	0.1106	0.988	0.085	0.0046	0.002	0.013
3	PAARDEN ISL	3	1.000	0.992	0.191	0.1661	0.989	0.193	0.0088	0.003	0.046
4	CITY HOSP	4	1.000	0.979	0.079	0.1058	0.974	0.078	0.0073	0.005	0.032
5	EPPING	5	1.000	0.994	0.232	0.1833	0.994	0.234	0.0032	0.000	0.006
6	TAMBOERSKLF	6	1.000	0.981	0.024	0.0586	0.976	0.024	0.0045	0.006	0.012
7	FORESHORE	7	1.000	0.996	0.215	0.1764	0.991	0.217	0.0119	0.004	0.085

FIGURE 8.1.1



NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	JULY85	A	0.1546	-0.0057	29	97
2	AUG	B	0.1591	-0.0178	27	98
3	SEP	C	0.0940	-0.0213	26	78
4	OCT	D	0.0846	-0.0125	27	76
5	NOV	E	0.0884	0.0011	30	77
6	DEC	F	0.0626	0.0076	31	69
7	JAN86	G	0.0646	0.0059	31	69
8	FEB	H	0.0834	0.0006	30	75
9	MAR	I	0.0907	0.0096	32	77
10	APR	J	0.0954	0.0110	32	79
11	MAY	K	0.1273	0.0135	32	89
12	JUNE	L	0.1472	0.0146	32	95
1	CITY HALL	1	0.1553	-0.0366	24	91
2	SALT RIVER	2	0.1106	0.0046	31	79
3	PAARDEN ISL	3	0.1661	0.0088	31	94
4	CITY HOSP	4	0.1058	0.0073	31	78
5	EPPING	5	0.1833	0.0032	30	98
6	TAMBOERSKLF	6	0.0586	0.0045	31	65
7	FORESHORE	7	0.1764	0.0119	32	97

Plot 8.1.2

The data is centred by subtracting the overall mean (equation 7.3) from all the entries in the matrix. The rationale here is that the origin is arbitrary in this context, and the real focus of interest is the variation of the numbers, rather than their actual magnitudes. Thus the plot picks up relative variation of sites and months. Because all entries have the same unit of measurement, comparisons are valid within and across points. A (0-1)-plot is displayed (Figure 8.1.2). The particular form of the generarlised correlation between the sites for this centring is given in equation (3.37).

Quality of the Display

The quality of the display is 91%, of which 60% is on the first axis. The winter months have large relative squared distances from the origin on the first axis.

Interpretations

We are looking at the data from a point 'closer' to them than in the previous plot. From this perspective, the points representing the months are seen to lie approximately in the line of the axis sketched on the plot. The site points lie along a line at roughly 90^0 to this. Note the ordering of the sites along this line and along the first axis is exactly the same as the true ordering of the mean pollution of the sites. This ordering: (from smallest to biggest) is: sites 6, 4, 2, 1, 3, 7 and 5.

The winter months (June, July and August), when pollution is high, are grouped on the right hand side of the first axis, with the midsummer months (December and January), when pollution is low, furthest away from these. The rest of the months fall between the two extremes. The angle subtended at the origin between the sites with highest pollution and the months with highest pollution is close to zero, so that the scalar products between them are large.

Considering projections of month onto site points, it can be seen that some sites experience greater variation in pollution levels during the year than others. For example in Tamboerskloof, projections of the points representing the months are bunched up, revealing little change in pollution levels over the year. In Epping, there is little change

in levels during winter and quite a bit of change over the rest of the year. Similarly, there is more variation over the sites' pollution levels in winter than in midsummer.

Bradu and Gabriel (1978) proposed the use of biplot displays to diagnose the type of model to fit to a data matrix, and derived rules for doing this. If the row points and the column points are collinear, and the two lines are perpendicular to each other, an additive model is suggested. (The additive model is given by $z_{ij}=\alpha_i+\beta_j$ where α_i and β_j are the row and column effects, respectively.) The positions of the points in the biplot presented here suggest that an additive model would fit the data. (In Plot 8.1.5 below, we consider the Z matrix where the row and column means have been subtracted.)

TABLE 8.1.2

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	160.85339	60.222	60.222
2	82.57083	30.914	91.136
3	16.15146	6.047	97.183
4	3.56150	1.333	98.516
5	2.41341	0.904	99.420
6	1.00247	0.375	99.795
7	0.54776	0.205	100.000

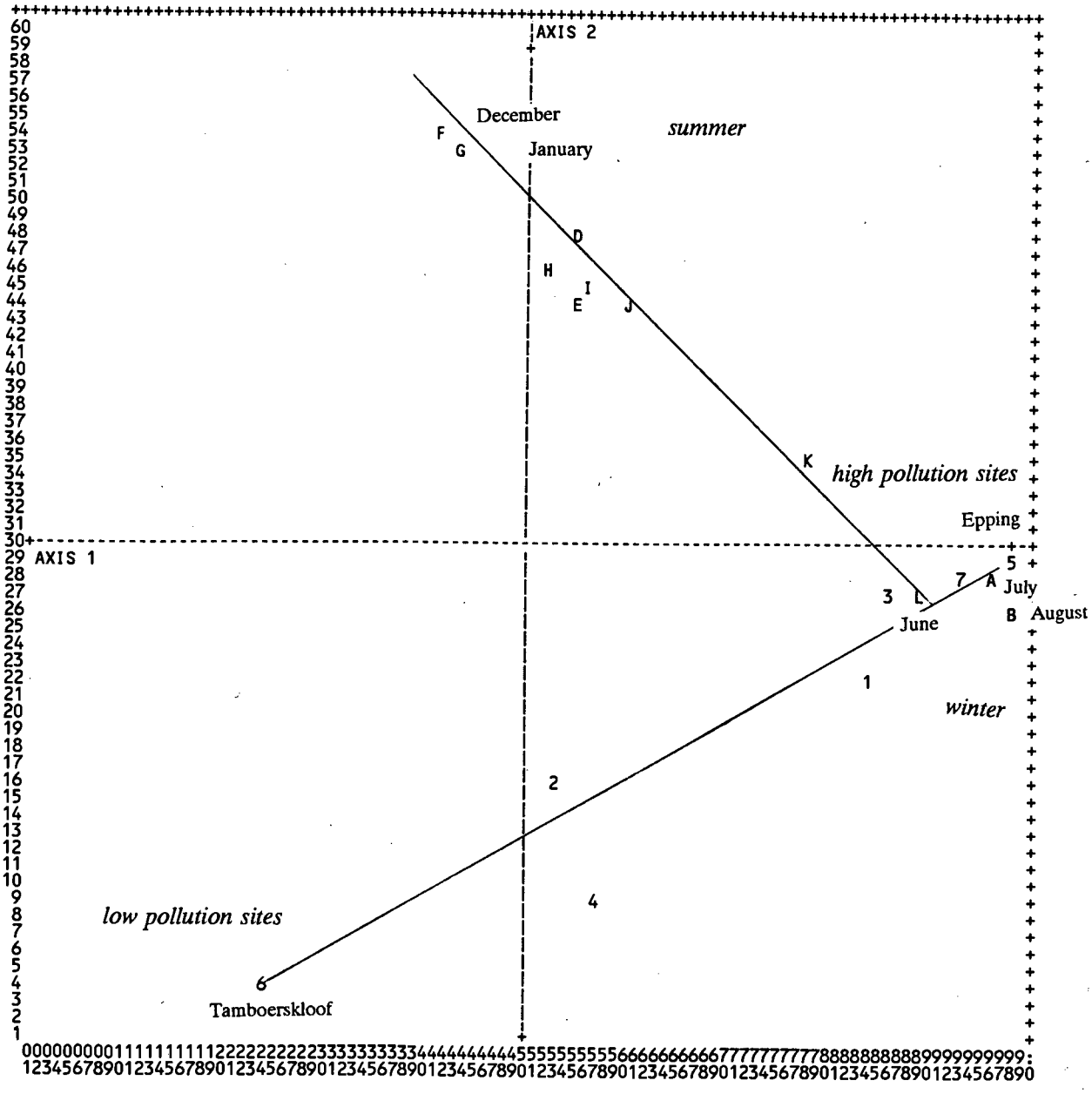
THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	FACT 2	COR
1	JULY85	A	1.000	0.390	0.108	0.5419	0.389	-0.0287	0.001
2	AUG	B	1.000	0.479	0.098	0.5670	0.469	-0.0833	0.010
3	SEP	C	1.000	0.106	0.127	0.1228	0.017	0.2814	0.089
4	OCT	D	1.000	0.435	0.043	0.0636	0.013	0.3558	0.421
5	NOV	E	1.000	0.113	0.104	0.0641	0.006	0.2803	0.107
6	DEC	F	1.000	0.361	0.089	-0.1031	0.017	0.4638	0.344
7	JAN86	G	1.000	0.361	0.086	-0.0811	0.011	0.4594	0.350
8	FEB	H	1.000	0.569	0.026	0.0323	0.006	0.3175	0.563
9	MAR	I	1.000	0.202	0.066	0.0779	0.013	0.2952	0.189
10	APR	J	1.000	0.188	0.073	0.1218	0.029	0.2857	0.159
11	MAY	K	1.000	0.301	0.056	0.3312	0.279	0.0943	0.023
12	JUNE	L	1.000	0.250	0.124	0.4614	0.246	-0.0575	0.004

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	CITY HALL	1	1.000	0.706	0.159	0.5098	0.614	0.162	-0.1974	0.092	0.047
2	SALT RIVER	2	1.000	0.928	0.046	0.0493	0.020	0.002	-0.3333	0.908	0.135
3	PAARDEN ISL	3	1.000	0.902	0.120	0.5335	0.891	0.177	-0.0615	0.012	0.005
4	CITY HOSP	4	1.000	0.950	0.109	0.1048	0.038	0.007	-0.5152	0.912	0.321
5	EPPING	5	1.000	0.961	0.199	0.7148	0.961	0.318	-0.0192	0.001	0.000
6	TAMBOERSKL	6	1.000	0.993	0.205	-0.3734	0.254	0.087	-0.6365	0.738	0.491
7	FORESHORE	7	1.000	0.924	0.163	0.6326	0.922	0.249	-0.0317	0.002	0.001

FIGURE 8.1.2



NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	JULY85	A	0.5419	-0.0287	28	96
2	AUG	B	0.5670	-0.0833	26	98
3	SEP	C	0.1228	0.2814	44	60
4	OCT	D	0.0636	0.3558	48	55
5	NOV	E	0.0641	0.2803	44	55
6	DEC	F	-0.1031	0.4638	54	41
7	JAN86	G	-0.0811	0.4594	53	43
8	FEB	H	0.0323	0.3175	46	52
9	MAR	I	0.0779	0.2952	45	56
10	APR	J	0.1218	0.2857	44	60
11	MAY	K	0.3312	0.0943	35	78
12	JUNE	L	0.4614	-0.0575	27	89
1	CITY HALL	1	0.5098	-0.1974	22	84
2	SALT RIVER	2	0.0493	-0.3333	16	53
3	PAARDEN ISL	3	0.5335	-0.0615	27	86
4	CITY HOSP	4	0.1048	-0.5152	9	57
5	EPPING	5	0.7148	-0.0192	29	98
6	TAMBOERSKL	6	-0.3734	-0.6365	4	24
7	FORESHORE	7	0.6326	-0.0317	28	93

Plot 8.1.3

The display (Figure 8.1.3) is a (1-1)-plot of a column centred matrix (equation 7.1). Thus the column point interpretation is that of the covariance biplot. The within row scalar product interpretation holds and there is no between set scalar product interpretation. The generalised correlation between the row points for this centring is discussed in Section 3.4. By centring the columns, the rows (months) are emphasized.

The City Hall site is noticably distinct from the other points; the pollution values in July and August are particularly high relative to the mean pollution level for City Hall. The other site points are almost collinear.

Quality of the Display

The quality of the display is 95%, of which 84% is on the first axis. The first axis, and hence most of the display, is mainly constituted by the winter months.

Interpretations

The information provided is largely on the first axis. The pattern formed by linking the month points sequentially is evident. Differences between the months have been emphasized by this choice of centring, while we still have no further information about the sites other than their approximate ordering according to pollution levels and that they are generally linearly related. Sites are pairwise positively correlated except City Hall, which is attracted in the direction of August, as in previous plots.

TABLE 8.1.3

T H E E I G E N V A L U E S

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	129.18324	84.213	84.213
2	16.25179	10.594	94.807
3	3.56679	2.325	97.132
4	2.44614	1.595	98.727
5	1.22005	0.795	99.522
6	0.54782	0.357	99.879
7	0.18563	0.121	100.000

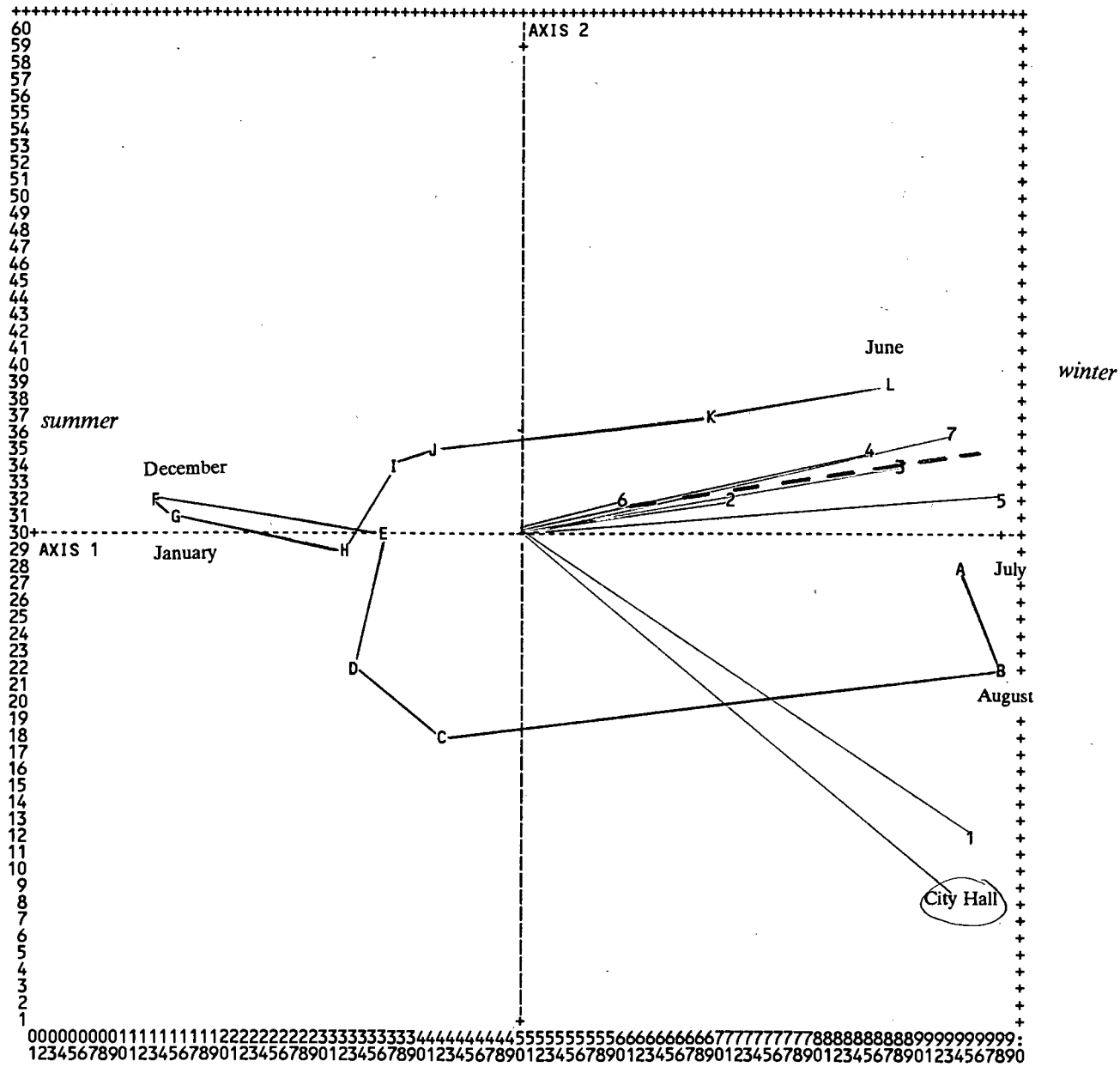
T H E R O W O B J E C T S

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	JULY85	A	1.000	0.958	0.176	0.5077	0.954	0.200	-0.0334	0.004	0.007
2	AUG	B	1.000	0.992	0.222	0.5607	0.924	0.243	-0.1516	0.068	0.141
3	SEP	C	1.000	0.937	0.041	-0.0926	0.136	0.007	-0.2246	0.801	0.310
4	OCT	D	1.000	0.990	0.038	-0.1919	0.639	0.029	-0.1423	0.351	0.125
5	NOV	E	1.000	0.595	0.027	-0.1571	0.595	0.019	-0.0023	0.000	0.000
6	DEC	F	1.000	0.957	0.124	-0.4247	0.946	0.140	0.0452	0.011	0.013
7	JAN86	G	1.000	0.951	0.110	-0.3986	0.945	0.123	0.0319	0.006	0.006
8	FEB	H	1.000	0.963	0.030	-0.2091	0.960	0.034	-0.0105	0.002	0.001
9	MAR	I	1.000	0.836	0.022	-0.1493	0.656	0.017	0.0781	0.180	0.038
10	APR	J	1.000	0.666	0.019	-0.0971	0.329	0.007	0.0984	0.337	0.060
11	MAY	K	1.000	0.978	0.046	0.2199	0.693	0.037	0.1412	0.285	0.123
12	JUNE	L	1.000	0.959	0.146	0.4320	0.831	0.144	0.1698	0.128	0.177

T H E C O L U M N O B J E C T S

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	CITY HALL	1	1.000	0.999	0.259	0.5229	0.687	0.212	-0.3520	0.311	0.762
2	SALT RIVER	2	1.000	0.896	0.047	0.2500	0.876	0.048	0.0379	0.020	0.009
3	PAARDEN ISL	3	1.000	0.903	0.152	0.4500	0.868	0.157	0.0900	0.035	0.050
4	CITY HOSP	4	1.000	0.938	0.121	0.4063	0.886	0.128	0.0982	0.052	0.059
5	EPPING	5	1.000	0.943	0.221	0.5642	0.937	0.246	0.0459	0.006	0.013
6	TAMBOERSKLF	6	1.000	0.803	0.012	0.1164	0.726	0.010	0.0380	0.077	0.009
7	FORESHORE	7	1.000	0.949	0.187	0.5065	0.894	0.199	0.1260	0.055	0.098

FIGURE 8.1.3



NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	JULY85	A	0.5077	-0.0334	28	94
2	AUG	B	0.5607	-0.1516	22	98
3	SEP	C	-0.0926	-0.2246	18	42
4	OCT	D	-0.1919	-0.1423	22	33
5	NOV	E	-0.1571	-0.0023	30	36
6	DEC	F	-0.4247	0.0452	32	13
7	JAN86	G	-0.3986	0.0319	31	15
8	FEB	H	-0.2091	-0.0105	29	32
9	MAR	I	-0.1493	0.0781	34	37
10	APR	J	-0.0971	0.0984	35	41
11	MAY	K	0.2199	0.1412	37	69
12	JUNE	L	0.4320	0.1698	39	87
1	CITY HALL	1	0.5229	-0.3520	12	95
2	SALT RIVER	2	0.2500	0.0379	32	71
3	PAARDEN ISL	3	0.4500	0.0900	34	88
4	CITY HOSP	4	0.4063	0.0982	35	85
5	EPPING	5	0.5642	0.0459	32	98
6	TAMBOERSKLF	6	0.1164	0.0380	32	60
7	FORESHORE	7	0.5065	0.1260	36	93

d_1 d_2

Plot 8.1.4

The display (Figure 8.1.4) is a (1-0)-plot of a row centred matrix (equation 7.2). The scalar cosine of the angle between the row points is Pearsons product moment correlation. The norms of the rows are proportional to their standard deviations, and the squared Euclidean distance between two row points is their covariance (Section 3.3.4). The rows have the interpretations that the columns usually have in the covariance biplot. By row centring, the relative importance of the rows is made more equitable, and the columns (sites) are emphasized in the display.

Plot 8.1.5

A double centre (equation 7.4) is performed on the data (Fig. 8.1.5), as described under Plot 8.1.2. We are effectively analysing the residuals from an additive model. Seasonal variation is very marked still. The sites are approximately collinear. The plotted values are those remaining after both row and column effects have been taken into account.

For example, Epping has higher pollution than expected in winter, after taking row and column means into account. Tamboerskloof has higher pollution than expected in summer and City Hall has higher pollution than expected in August and September.

That there is still a pattern suggests that the additive model does not describe fully the information contained in the data set. This biplot is not all 'noise'.

Plots 8.1.6 and 8.1.7

These are (1-0) and (0-1) column centred biplots of the data respectively. The aim of including these plots is to illustrate the effect of choice of biplot family on the resulting plot. Plot 8.1.3 is a (1-1)-plot with the same centring. Section 9.2 describes the effects of choice of family further. Notice that the quality of the display is unaffected by the choice of family. The row coordinates are the same in the (1-0) and (1-1)-plots. The column coordinates are the same in the (0-1) and (1-1)-plots.

TABLE 8.1.4

THE EIGENVALUES

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	127.56165	83.014	83.014
2	16.38395	10.662	93.676
3	3.79911	2.472	96.148
4	3.19633	2.080	98.228
5	2.17445	1.415	99.643
6	0.54796	0.357	100.000
7	0.00000	0.000	100.000
8	0.00000	0.000	100.000
9	0.00000	0.000	100.000
10	0.00000	0.000	100.000
11	0.00000	0.000	100.000
12	0.00000	0.000	100.000

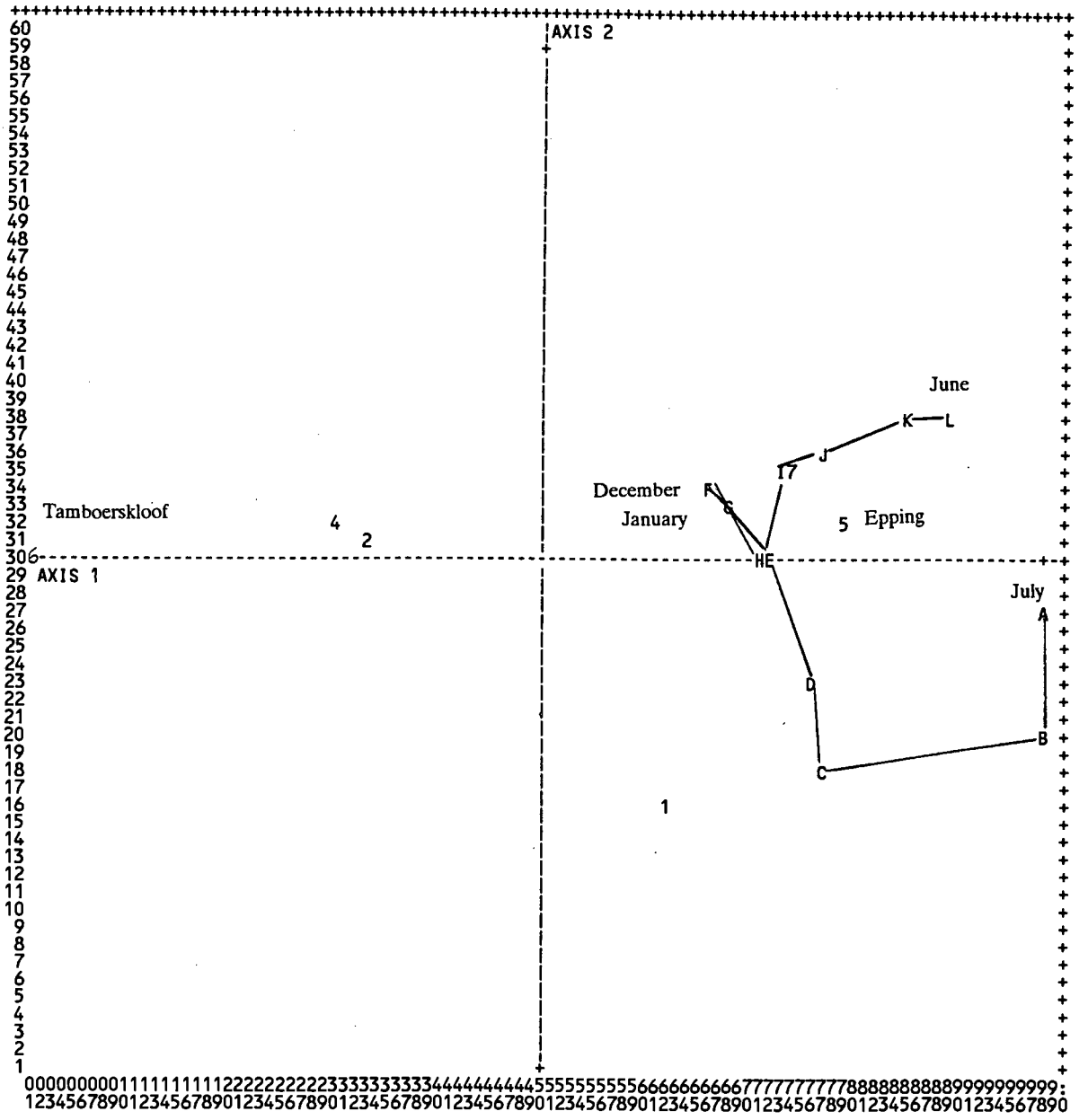
THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	JULY85	A	1.000	0.961	0.174	0.5048	0.954	0.200	-0.0437	0.007	0.012
2	AUG	B	1.000	0.983	0.184	0.4996	0.882	0.196	-0.1691	0.101	0.174
3	SEP	C	1.000	0.962	0.086	0.2879	0.629	0.065	-0.2095	0.333	0.268
4	OCT	D	1.000	0.971	0.060	0.2732	0.816	0.059	-0.1195	0.156	0.087
5	NOV	E	1.000	0.782	0.044	0.2285	0.780	0.041	0.0102	0.002	0.001
6	DEC	F	1.000	0.735	0.032	0.1729	0.613	0.023	0.0772	0.122	0.036
7	JAN86	G	1.000	0.839	0.032	0.1933	0.760	0.029	0.0624	0.079	0.024
8	FEB	H	1.000	0.973	0.032	0.2191	0.973	0.038	0.0063	0.001	0.000
9	MAR	I	1.000	0.883	0.051	0.2449	0.763	0.047	0.0971	0.120	0.058
10	APR	J	1.000	0.900	0.068	0.2832	0.772	0.063	0.1153	0.128	0.081
11	MAY	K	1.000	0.998	0.103	0.3707	0.871	0.108	0.1412	0.126	0.122
12	JUNE	L	1.000	0.915	0.136	0.4106	0.808	0.132	0.1500	0.108	0.137

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	CITY HALL	1	1.000	0.997	0.113	0.2032	0.238	0.032	-0.3626	0.759	0.803
2	SALT RIVER	2	1.000	0.858	0.053	-0.2630	0.847	0.054	0.0290	0.010	0.005
3	PAARDEN ISL	3	1.000	0.800	0.077	0.2917	0.723	0.067	0.0954	0.077	0.056
4	CITY HOSP	4	1.000	0.802	0.081	-0.3103	0.778	0.075	0.0545	0.024	0.018
5	EPPING	5	1.000	0.925	0.156	0.4687	0.916	0.172	0.0487	0.010	0.014
6	TAMBOERSKLF	6	1.000	0.993	0.400	-0.7816	0.993	0.479	0.0045	0.000	0.000
7	FORESHORE	7	1.000	0.920	0.120	0.3913	0.828	0.120	0.1306	0.092	0.104

FIGURE 8.1.4



NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	CITY HALL	1	0.2032	-0.3626	16	62
2	SALT RIVER	2	-0.2630	0.0290	31	33
3	PAARDEN ISL	3	0.2917	0.0954	33	68
4	CITY HOSP	4	-0.3103	0.0545	32	30
5	EPPING	5	0.4687	0.0487	32	79
6	TAMBOERSKLF	6	-0.7816	0.0045	30	1
7	FORESHORE	7	0.3913	0.1306	35	74
1	JULY85	A	0.5048	-0.0437	27	98
2	AUG	B	0.4996	-0.1691	20	98
3	SEP	C	0.2879	-0.2095	18	77
4	OCT	D	0.2732	-0.1195	23	76
5	NOV	E	0.2285	0.0102	30	72
6	DEC	F	0.1729	0.0772	34	66
7	JAN86	G	0.1933	0.0624	33	68
8	FEB	H	0.2191	0.0063	30	71
9	MAR	I	0.2449	0.0971	35	73
10	APR	J	0.2832	0.1153	36	77
11	MAY	K	0.3707	0.1412	38	85
12	JUNE	L	0.4106	0.1500	38	89

TABLE 8.1.5

THE EIGENVALUES

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	19.50121	48.797	48.797
2	12.81466	32.065	80.862
3	3.56493	8.920	89.783
4	2.31756	5.799	95.582
5	1.21813	3.048	98.630
6	0.54762	1.370	100.000

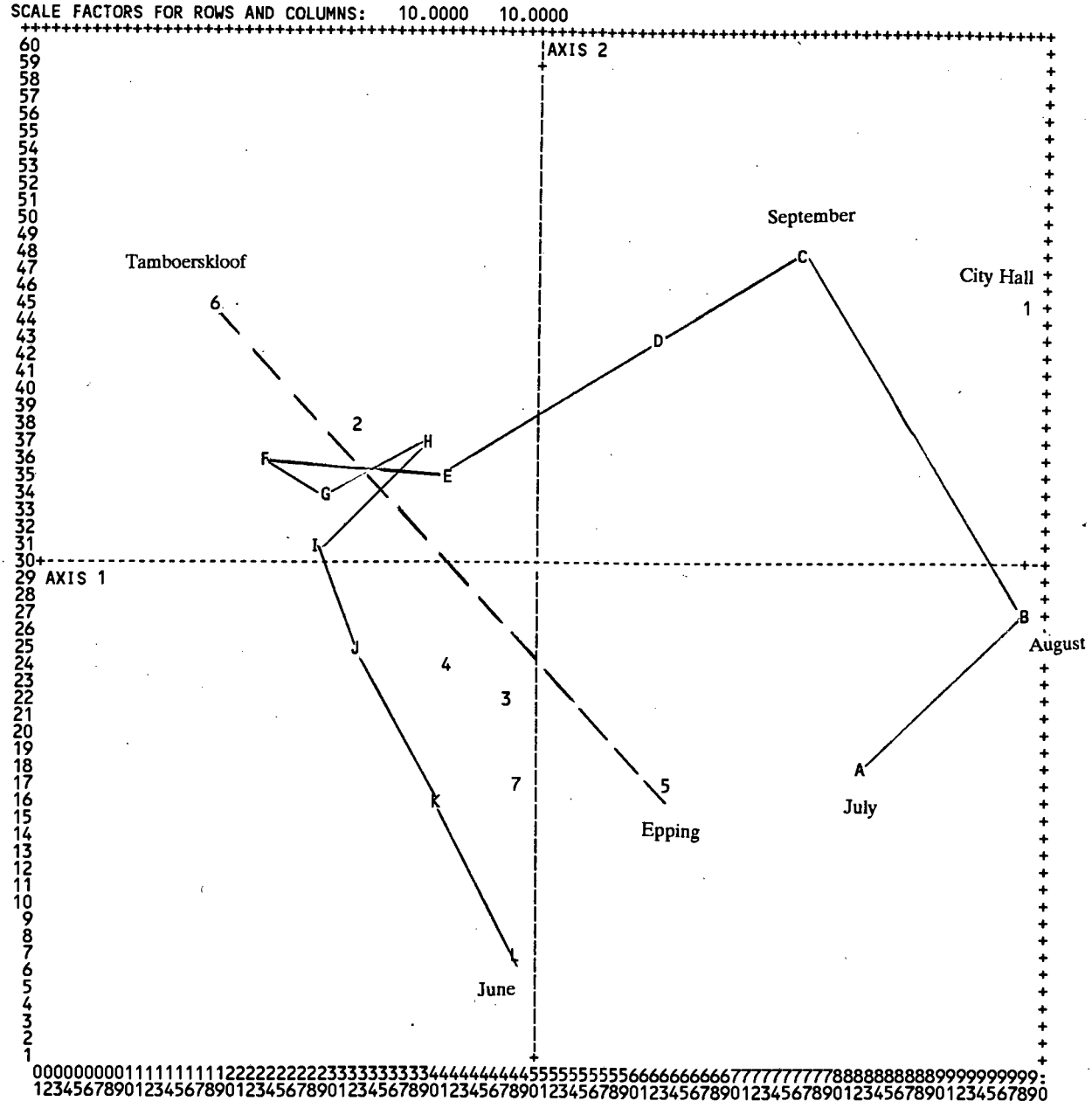
THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	JULY85	A	1.000	0.809	0.133	0.1771	0.589	0.161	-0.1082	0.220	0.091
2	AUG	B	1.000	0.960	0.187	0.2671	0.953	0.366	-0.0234	0.007	0.004
3	SEP	C	1.000	0.930	0.132	0.1428	0.386	0.105	0.1695	0.544	0.224
4	OCT	D	1.000	0.982	0.049	0.0654	0.220	0.022	0.1217	0.761	0.116
5	NOV	E	1.000	0.216	0.053	-0.0491	0.114	0.012	0.0465	0.102	0.017
6	DEC	F	1.000	0.748	0.085	-0.1476	0.642	0.112	0.0602	0.107	0.028
7	JAN86	G	1.000	0.654	0.058	-0.1155	0.575	0.068	0.0428	0.079	0.014
8	FEB	H	1.000	0.826	0.022	-0.0570	0.371	0.017	0.0630	0.454	0.031
9	MAR	I	1.000	0.758	0.047	-0.1186	0.755	0.072	0.0073	0.003	0.000
10	APR	J	1.000	0.552	0.053	-0.0987	0.459	0.050	-0.0445	0.093	0.015
11	MAY	K	1.000	0.931	0.054	-0.0541	0.136	0.015	-0.1306	0.794	0.133
12	JUNE	L	1.000	0.821	0.128	-0.0116	0.003	0.001	-0.2042	0.818	0.325

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	CITY HALL	1	1.000	0.997	0.369	0.3418	0.792	0.599	0.1736	0.204	0.235
2	SALT RIVER	2	1.000	0.809	0.077	-0.1280	0.532	0.084	0.0925	0.278	0.067
3	PAARDEN ISL	3	1.000	0.281	0.078	-0.0198	0.013	0.002	-0.0916	0.269	0.065
4	CITY HOSP	4	1.000	0.394	0.048	-0.0579	0.176	0.017	-0.0643	0.218	0.032
5	EPPING	5	1.000	0.650	0.119	0.0960	0.193	0.047	-0.1477	0.457	0.170
6	TAMBOERSKLF	6	1.000	0.962	0.214	-0.2207	0.568	0.250	0.1836	0.393	0.263
7	FORESHORE	7	1.000	0.570	0.094	-0.0114	0.003	0.001	-0.1463	0.567	0.167

FIGURE 8.1.5



NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	JULY85	A	0.1771	-0.1082	18	82
2	AUG	B	0.2671	-0.0234	27	98
3	SEP	C	0.1428	0.1695	48	76
4	OCT	D	0.0654	0.1217	43	62
5	NOV	E	-0.0491	0.0465	35	41
6	DEC	F	-0.1476	0.0602	36	23
7	JAN86	G	-0.1155	0.0428	34	29
8	FEB	H	-0.0570	0.0630	37	39
9	MAR	I	-0.1186	0.0073	31	28
10	APR	J	-0.0987	-0.0445	25	32
11	MAY	K	-0.0541	-0.1306	16	40
12	JUNE	L	-0.0116	-0.2042	7	48
1	CITY HALL	1	0.3418	0.1736	45	98
2	SALT RIVER	2	-0.1280	0.0925	38	32
3	PAARDEN ISL	3	-0.0198	-0.0916	22	47
4	CITY HOSP	4	-0.0579	-0.0643	24	41
5	EPPING	5	0.0960	-0.1477	17	63
6	TAMBOERSKLF	6	-0.2207	0.1836	45	18
7	FORESHORE	7	-0.0114	-0.1463	17	48

TABLE 8.1.6

T H E E I G E N V A L U E S

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	129.18324	84.213	84.213
2	16.25179	10.594	94.807
3	3.56679	2.325	97.132
4	2.44614	1.595	98.727
5	1.22005	0.795	99.522
6	0.54782	0.357	99.879
7	0.18563	0.121	100.000

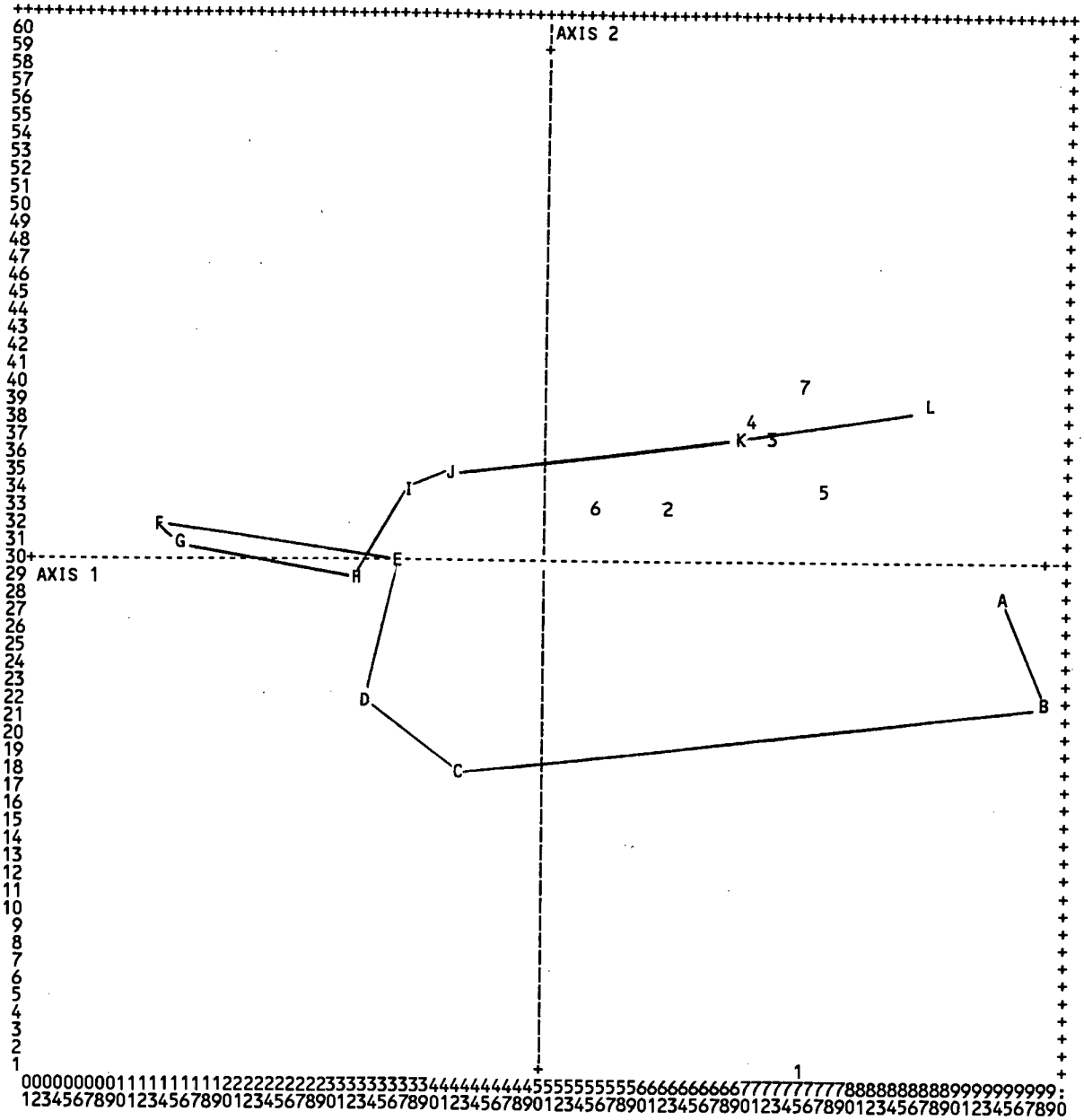
T H E R O W O B J E C T S

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	JULY85	A	1.000	0.958	0.176	0.5077	0.954	0.200	-0.0334	0.004	0.007
2	AUG	B	1.000	0.992	0.222	0.5607	0.924	0.243	-0.1516	0.068	0.141
3	SEP	C	1.000	0.937	0.041	-0.0926	0.136	0.007	-0.2246	0.801	0.310
4	OCT	D	1.000	0.990	0.038	-0.1919	0.639	0.029	-0.1423	0.351	0.125
5	NOV	E	1.000	0.595	0.027	-0.1571	0.595	0.019	-0.0023	0.000	0.000
6	DEC	F	1.000	0.957	0.124	-0.4247	0.946	0.140	0.0452	0.011	0.013
7	JAN86	G	1.000	0.951	0.110	-0.3986	0.945	0.123	0.0319	0.006	0.006
8	FEB	H	1.000	0.963	0.030	-0.2091	0.960	0.034	-0.0105	0.002	0.001
9	MAR	I	1.000	0.836	0.022	-0.1493	0.656	0.017	0.0781	0.180	0.038
10	APR	J	1.000	0.666	0.019	-0.0971	0.329	0.007	0.0984	0.337	0.060
11	MAY	K	1.000	0.978	0.046	0.2199	0.693	0.037	0.1412	0.285	0.123
12	JUNE	L	1.000	0.959	0.146	0.4320	0.831	0.144	0.1698	0.128	0.177

T H E C O L U M N O B J E C T S

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	FACT 2	COR
1	CITY HALL	1	1.000	0.974	0.143	0.4600	0.212	-0.8731	0.762
2	SALT RIVER	2	1.000	0.057	0.143	0.2199	0.048	0.0940	0.009
3	PAARDEN ISL	3	1.000	0.207	0.143	0.3959	0.157	0.2232	0.050
4	CITY HOSP	4	1.000	0.187	0.143	0.3575	0.128	0.2437	0.059
5	EPPING	5	1.000	0.259	0.143	0.4964	0.246	0.1139	0.013
6	TAMBOERSKLF	6	1.000	0.019	0.143	0.1024	0.010	0.0942	0.009
7	FORESHORE	7	1.000	0.296	0.143	0.4456	0.199	0.3126	0.098

FIGURE 8.1.6



NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	JULY85	A	0.5077	-0.0334	28	94
2	AUG	B	0.5607	-0.1516	22	98
3	SEP	C	-0.0926	-0.2246	18	42
4	OCT	D	-0.1919	-0.1423	22	33
5	NOV	E	-0.1571	-0.0023	30	36
6	DEC	F	-0.4247	0.0452	32	13
7	JAN86	G	-0.3986	0.0319	31	15
8	FEB	H	-0.2091	-0.0105	29	32
9	MAR	I	-0.1493	0.0781	34	37
10	APR	J	-0.0971	0.0984	35	41
11	MAY	K	0.2199	0.1412	37	69
12	JUNE	L	0.4320	0.1698	39	87
1	CITY HALL	1	0.4600	-0.8731	1	75
2	SALT RIVER	2	0.2199	0.0940	33	62
3	PAARDEN ISL	3	0.3959	0.2232	37	72
4	CITY HOSP	4	0.3575	0.2437	38	70
5	EPPING	5	0.4964	0.1139	34	77
6	TAMBOERSKLF	6	0.1024	0.0942	33	55
7	FORESHORE	7	0.4456	0.3126	40	75

TABLE 8.1.7

THE EIGENVALUES

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	129.18324	84.213	84.213
2	16.25179	10.594	94.807
3	3.56679	2.325	97.132
4	2.44614	1.595	98.727
5	1.22005	0.795	99.522
6	0.54782	0.357	99.879
7	0.18563	0.121	100.000

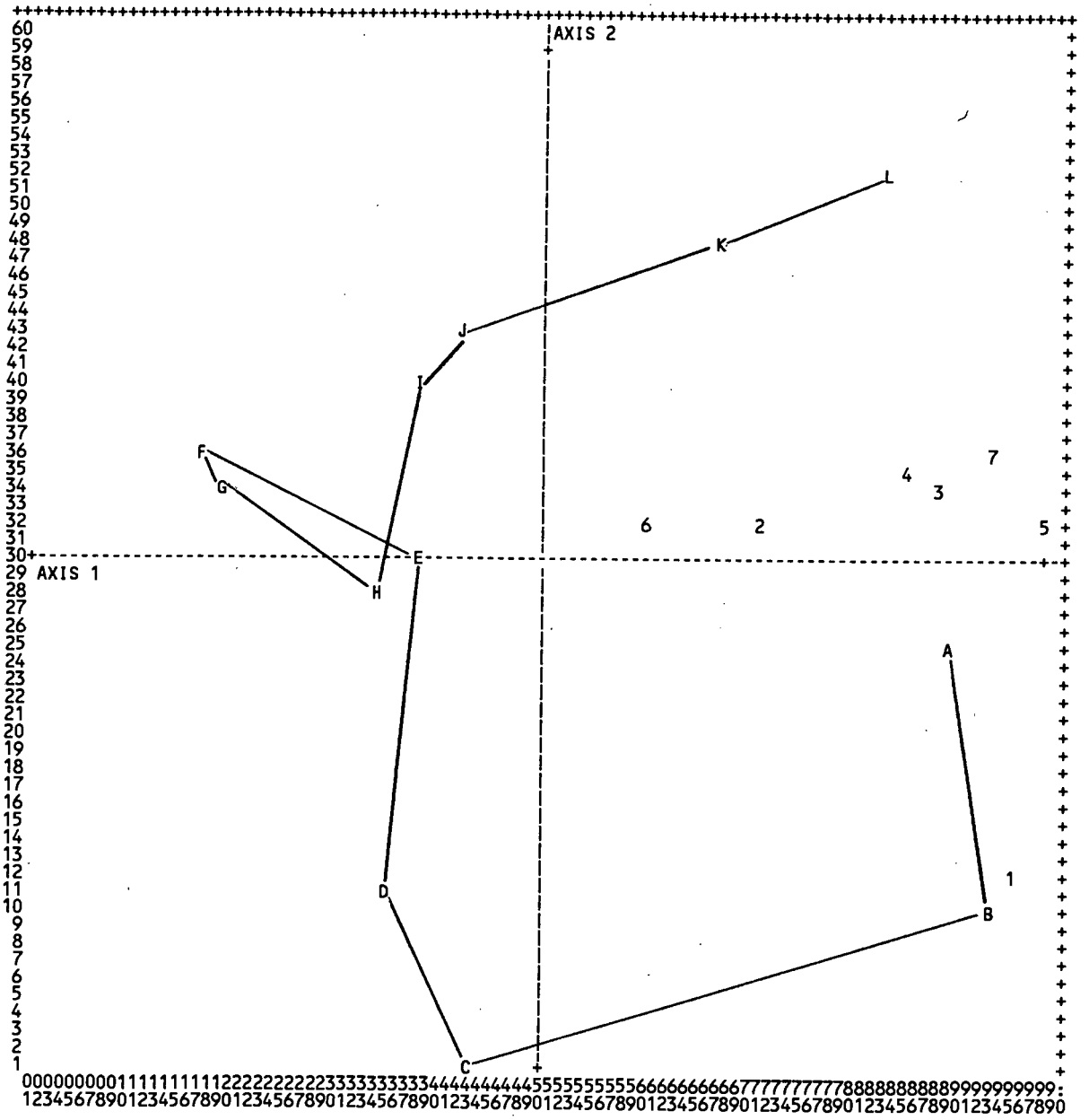
THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	FACT 2	COR
1	JULY85	A	1.000	0.299	0.098	0.4467	0.289	-0.0827	0.010
2	AUG	B	1.000	0.639	0.086	0.4933	0.404	-0.3760	0.235
3	SEP	C	1.000	0.387	0.117	-0.0815	0.008	-0.5572	0.379
4	OCT	D	1.000	0.468	0.047	-0.1689	0.087	-0.3531	0.381
5	NOV	E	1.000	0.024	0.114	-0.1383	0.024	-0.0056	0.000
6	DEC	F	1.000	0.281	0.077	-0.3736	0.257	0.1121	0.023
7	JAN86	G	1.000	0.228	0.081	-0.3507	0.217	0.0792	0.011
8	FEB	H	1.000	0.151	0.033	-0.1839	0.148	-0.0260	0.003
9	MAR	I	1.000	0.105	0.075	-0.1313	0.033	0.1938	0.072
10	APR	J	1.000	0.085	0.113	-0.0854	0.009	0.2441	0.076
11	MAY	K	1.000	0.483	0.047	0.1935	0.113	0.3502	0.370
12	JUNE	L	1.000	0.412	0.112	0.3801	0.185	0.4211	0.227

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	CITY HALL	1	1.000	0.999	0.259	0.5229	0.687	0.212	-0.3520	0.311	0.762
2	SALT RIVER	2	1.000	0.896	0.047	0.2500	0.876	0.048	0.0379	0.020	0.009
3	PAARDEN ISL	3	1.000	0.903	0.152	0.4500	0.868	0.157	0.0900	0.035	0.050
4	CITY HOSP	4	1.000	0.938	0.121	0.4063	0.886	0.128	0.0982	0.052	0.059
5	EPPING	5	1.000	0.943	0.221	0.5642	0.937	0.246	0.0459	0.006	0.013
6	TAMBOERSKLF	6	1.000	0.803	0.012	0.1164	0.726	0.010	0.0380	0.077	0.009
7	FORESHORE	7	1.000	0.949	0.187	0.5065	0.894	0.199	0.1260	0.055	0.098

FIGURE 8.1.7



NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	JULY85	A	0.4467	-0.0827	25	89
2	AUG	B	0.4933	-0.3760	10	93
3	SEP	C	-0.0815	-0.5572	1	43
4	OCT	D	-0.1689	-0.3531	11	35
5	NOV	E	-0.1383	-0.0056	30	38
6	DEC	F	-0.3736	0.1121	36	17
7	JAN86	G	-0.3507	0.0792	34	19
8	FEB	H	-0.1839	-0.0260	28	34
9	MAR	I	-0.1313	0.1938	40	38
10	APR	J	-0.0854	0.2441	43	42
11	MAY	K	0.1935	0.3502	48	67
12	JUNE	L	0.3801	0.4211	52	83
1	CITY HALL	1	0.5229	-0.3520	12	95
2	SALT RIVER	2	0.2500	0.0379	32	71
3	PAARDEN ISL	3	0.4500	0.0900	34	88
4	CITY HOSP	4	0.4063	0.0982	35	85
5	EPPING	5	0.5642	0.0459	32	98
6	TAMBOERSKLF	6	0.1164	0.0380	32	60
7	FORESHORE	7	0.5065	0.1260	36	93

EXAMPLE 8.2
MARATHON RUNNERS

In the following example, biplots displaying data taken from a study on marathon runners (Noakes et al, 1988) are presented. A series of measurements was made on 30 recreational runners who completed a standard 42.2 km marathon - the 1987 Cape Peninsula marathon. Plots from the correlation biplot family are discussed and compared.

The research was aimed at identifying factors that lead to heat injury (heatstroke) as a result of marathon running. The extent of heat injury is measured by body temperature. Dehydration and body temperature have been found to be highly correlated in many studies. The popularly held belief is that dehydration is the most important determinant of heat injury; runners are thus advised to consume adequate quantities of fluid in order to avoid becoming dehydrated. However the authors wished to investigate whether this correlation could be explained by a third variable. For example, high temperatures are also measured in faster runners, who tend to drink less and sweat more than slower runners. Many variables were measured, in order to investigate which of these influence body temperature. As such the study was primarily exploratory in nature.

We consider the data set as an example of how biplots can be used in exploratory data analysis, and do not enter into the debate as to the cause of heat injury, which is best left to sports scientists, and which was described in order to provide the reader with some background knowledge of the application area. Of importance here is the exploratory nature of the research, hence the large number of variables, which often pose problems for practical researchers.

Variables measured other than those mentioned above include observations made on the runners before, during and after the race, measurements made on treadmills, and physiological measurements such as height and percentage body fat (Table 8.2).

Row centring (equation 7.2) is applied to variables 10, 11, 15, 16, 21 and 22 (Table 8.2). Using this centring, values relative to the mean of each particular individual are obtained.

Row centring is a fairly unusual centring; it emphasises change relative to each individual rather than to the sample mean. For example, variable 21 is obtained by subtracting variable 20 (resting oxygen capacity) from variable 17 (oxygen capacity at 42 kilometers). The variable therefore represents the difference in oxygen capacity for each individual from what it usually is to what it is at the end of the marathon.

In Table 8.2, the runners have been ordered from fastest to slowest over the distance. In the plots (Figures 8.2.1 to 8.2.4), symbols for the row points go through the upper case letters of the alphabet consecutively, and then through the lower case letters so that 'A' denotes the fastest runner, 'B' the second fastest, and 'd' the slowest. Three of the runners (B, S and c) have some observations missing. They are given supplementary point status (i.e. weights of zero), with mean values for the missing variables.

In order to facilitate comparisons between the plots, the same variables were used in all the different plots. As noted in Section 4.4, variables for the coefficient of variation biplot should conform to certain requirements - non-negativity and be measured on a ratio scale - in order that the plot be meaningfully interpreted.

The measurement of body temperature in degrees centigrade (C) does not provide a meaningful zero (0^0C is the freezing point of water). It would thus not be appropriate to include it in the coefficient of variation biplot. To rectify this, normal body temperature, $37{,}4^0$C, was subtracted from each observation. This transformation was used because exercise causes an increase in body temperature, and this is conveniently measured as an increase above normal body temperature. All observations of body temperature referred to this origin, are positive.

The requirement that there are no negative observations can be relaxed in practice provided the effect of a few negative observations is small. However, the mean (and therefore the coefficient of variation) of the variable must be positive. In this data set, the dehydration variable provides an example of a variable in which the one negative observation does not have an important effect on the interpretation of the plot. Variables with many negative observations and with a negative mean should be excluded from a coefficient of variation biplot.

TABLE 8.2 MARATHON RUNNERS DATA.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
R01	31	175	71.6	2.4	13.2	1.85	15.44	15.0	13.0	2.44	3.62	3.50	2.95	0.67	78.7	76.1	64.2	67.0	11.7	9.1	1.50	1.472	4.19	2.2	5.4	2.4
R02	39	173	66.3	1.4	13.2	1.79	14.72	14.5	13.3	1.42	3.48	3.43	3.12	0.36	86.5	85.2	77.4	63.0	23.5	22.2	1.69	1.258	3.62	3.0	4.8	2.3
R03	38	179	85.5	1.6	17.9	2.01	14.22	14.5	12.5	1.72	4.29	4.37	3.82	0.47	87.2	88.8	77.7	59.6	27.6	29.2	2.08	1.503	3.51	3.1	5.9	1.3
R04	48	177	62.4	1.6	15.7	1.78	14.22	14.5	13.0	1.22	3.28	3.39	2.87	0.41	83.1	85.9	72.6	64.1	19.0	21.8	1.84	0.731	1.28	1.8	4.0	4.3
R05	38	168	68.1	2.1	16.5	1.81	14.22	14.5	13.0	1.22	3.49	3.60	3.13	0.37	74.0	76.2	66.3	69.6	4.4	6.6	1.78	0.532	0.44	3.2	5.5	1.9
R06	33	173	73.0	2.1	15.9	1.83	14.15	16.0	14.0	0.15	3.42	4.04	3.39	0.03	70.9	83.7	70.3	68.9	2.0	14.8	1.79	1.440	4.11	1.9	5.5	2.3
R07	42	181	77.7	3.1	18.3	1.95	14.07	13.5	12.0	2.07	3.99	3.83	3.46	0.53	85.7	82.3	74.4	61.9	23.8	20.4	1.80	1.207	3.35	2.9	5.1	2.2
R08	46	181	63.4	1.1	14.8	1.82	13.91	12.5	10.5	3.41	2.82	2.61	2.28	0.54	84.1	77.6	68.0	54.8	29.3	22.8	1.82	1.188	3.47	1.6	3.1	4.5
R09	36	190	90.6	0.6	17.4	1.88	13.40	13.0	11.0	2.40	4.21	4.17	3.43	0.78	88.5	87.8	72.2	54.0	34.5	33.8	1.89	1.221	2.87	3.5	5.2	2.5
R10	26	180	73.4	1.2	13.0	1.90	13.05	12.5	10.5	2.55	2.84	2.72	2.36	0.48	68.8	65.9	57.2	58.1	10.7	7.8	2.01	1.225	3.27	4.3	4.9	2.2
R11	30	179	68.0	0.8	15.9	1.76	12.79	11.5	10.0	2.79	2.99	2.64	2.25	0.74	80.8	71.4	60.7	56.9	23.9	14.5	1.98	1.364	4.41	2.2	3.8	3.9
R12	37	171	67.0	1.1	18.2	1.77	12.79	13.0	11.0	1.79	3.05	3.11	2.50	0.55	72.9	74.4	59.8	63.1	9.8	11.3	2.05	0.715	1.19	2.3	5.8	1.9
R13	29	177	75.4	1.8	11.9	1.76	12.72	12.0	10.9	1.82	3.22	3.05	2.79	0.43	69.3	65.7	60.2	63.8	5.5	1.9	2.41	1.383	3.58	2.8	4.8	1.8
R14	38	175	64.0	2.1	17.1	1.78	12.60	13.0	11.2	1.40	2.91	2.96	2.70	0.21	73.7	75.1	68.5	61.7	12.0	13.4	2.08	0.509	0.16	2.3	3.5	2.9
R15	46	172	66.2	1.6	18.0	2.05	12.35	12.0	10.0	2.35	2.76	2.68	2.30	0.46	68.3	66.4	56.9	61.5	6.8	4.9	2.08	0.652	0.91	2.4	4.0	3.0
R16	32	178	79.2	1.1	14.4	2.12	12.35	13.6	11.6	0.75	3.42	3.76	3.19	0.23	83.6	92.1	78.1	52.8	30.8	39.3	2.32	1.012	2.15	2.1	5.3	2.4
R17	38	183	84.7	2.6	18.5	2.04	12.29	13.0	11.0	1.29	3.72	3.89	3.43	0.30	84.2	87.9	77.5	54.4	29.8	33.5	2.06	1.381	3.90	3.9	4.3	2.2
R18	37	165	68.0	1.0	19.2	2.08	12.23	12.0	10.0	2.23	2.69	2.66	2.32	0.37	65.3	64.6	56.4	61.8	3.5	2.8	2.17	0.905	2.06	3.0	5.2	1.4
R19	21	159	50.8	1.0	21.7	2.16	11.83	12.0	10.0	1.83	2.27	2.29	2.08	0.19	85.3	86.2	78.4	53.6	31.7	32.6	1.43	0.627	-0.76	3.0	4.8	2.3
R20	42	176	78.8	1.6	23.1	1.95	11.72	12.0	10.0	1.72	3.08	3.15	2.71	0.37	72.7	74.3	63.9	53.3	19.4	21.0	2.02	0.577	2.36	4.0	5.0	1.7
R21	26	185	81.0	1.4	16.2	2.05	11.51	11.5	10.0	1.51	3.08	3.08	2.65	0.43	60.5	60.5	52.0	64.1	-3.6	-3.6	2.20	0.887	1.98	3.5	5.3	2.4
R22	30	179	82.2	1.4	17.7	2.01	11.41	12.0	10.0	1.41	2.64	2.78	2.31	0.33	65.2	68.6	57.0	50.5	14.7	18.1	2.22	0.985	2.31	2.7	5.9	1.4
R23	24	186	83.0	1.0	15.3	1.79	11.20	10.5	9.0	2.20	3.51	3.26	2.73	0.78	77.2	71.6	59.9	55.4	21.8	16.2	2.26	0.642	0.96	3.3	4.6	2.3
R24	38	183	85.7	1.4	14.0	2.11	11.11	12.0	10.0	1.11	2.92	3.14	2.65	0.27	59.8	64.4	54.2	59.0	0.8	5.4	2.13	1.208	3.50	4.9	3.2	2.9
R25	32	185	96.1	1.9	15.8	1.83	11.06	11.0	9.0	2.06	3.07	3.05	2.75	0.32	56.1	55.7	50.2	59.7	-3.6	-4.0	2.29	1.657	4.79	3.2	5.5	0.5
R26	49	187	84.0	0.9	22.9	1.49	11.01	9.5	7.7	3.31	2.97	2.62	2.21	0.76	64.0	56.5	47.7	56.2	7.8	0.3	2.38	0.894	1.90	3.9	4.4	2.9
R27	51	182	91.8	1.9	24.4	1.89	10.55	11.0	9.5	1.05	3.39	3.48	3.18	0.21	82.4	84.6	77.4	45.6	36.8	39.0	2.36	0.826	1.74	4.3	5.2	1.2
R28	37	178	76.4	2.1	14.8	1.93	9.89	10.0	8.0	1.89	3.07	3.10	2.63	0.44	78.8	79.7	67.7	52.9	25.9	26.8	2.56	1.107	3.66	2.7	4.3	2.5
R29	41	186	93.5	1.8	20.8	2.16	9.89	11.0	9.0	0.89	3.73	4.05	3.48	0.25	81.1	88.0	75.7	50.5	30.6	37.5	2.56	1.005	2.67	3.7	4.4	1.8
R30	41	163	62.3	0.6	24.0	1.66	9.66	9.5	7.5	2.16	2.70	2.65	2.22	0.48	85.7	84.2	70.5	51.4	34.3	32.8	1.75	0.498	1.61	3.0	4.8	2.3
Mean	37	179	77.4	1.6	17.0	1.90	12.5	12.5	10.6	1.8	3.30	3.30	2.8	0.43	74.7	75.0	64.7	58.6	16.2	16.4	2.10	1.000	2.60	3.0	4.8	2.3
S.D.	7	6.0	9.7	5.9	3.1	0.15	1.44	1.56	1.56	0.76	0.44	0.53	0.46	0.19	9.12	10.4	9.15	5.9	12.1	13.1	0.25	0.33	1.28	0.86	0.8	0.9
C. V.	19.7	3.3	12.6	36.7	18.0	7.8	11.6	12.5	14.6	42.1	13.6	16.0	16.3	44.1	12.2	13.9	14.1	10.1	75.3	79.4	12.1	31.4	48.7	28.3	16.7	39.0

TABLE 8.2 (cont)

List of Variables

NO	NAME
1	Age (years)
2	Height (cm)
3	Weight (kg)
4	Body temperature (^{0}C)
5	Percentage body fat (%)
6	Surface area (m^2)
7	Running speed over 42 kilometers ($km.h^{-1}$)
8	Running speed over 21 kilometers ($km.h^{-1}$)
9	Running speed over 6 kilometers ($km.h^{-1}$)
10	Running speed between 6 and 42 kilometers ($km.h^{-1}$)
11	Metabolic rate at 42 kilometers ($l.min^{-1}$)
12	Metabolic rate at 21 kilometers (l.min-1)
13	Metabolic rate at 6 kilometers ($l.min^{-1}$)
14	Change in metabolic rate between 42 and 6 kilometers ($l.min^{-1}$)
15	Oxygen capacity at 42 km (ml O_2 kg^{-1} min^{-1})
16	Oxygen capacity at 21 km (ml O_2 kg^{-1} min^{-1})
17	Oxygen capacity at 6 km (ml O_2 kg^{-1} min^{-1})
18	Oxygen capacity at rest (VO_2 max) (ml O_2 kg^{-1} min^{-1})
19	Difference between oxygen capacity at 42 km and when resting
20	Difference between oxygen capacity at 21 km and when resting
21	Total fluid intake (l)
22	Sweat rate ($l.hr^{-1}$)
23	Dehydration (%)
24	Endomorphy
25	Mesomorphy
26	Ectomorphy

TABLE 8.2(a)

CORRELATION MATRIX - MARATHON RUNNER DATA

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	1.00	-0.04	-0.06	0.14	0.63	-0.10	-0.04	-0.04	-0.06	0.05	0.02	0.03	0.07	-0.11	0.26	0.24	0.29	-0.21	0.30	0.29	-0.01	-0.30	-0.26	0.03	-0.32	0.23
2	-0.04	1.00	0.72	-0.16	0.09	-0.04	-0.38	-0.45	-0.44	0.17	0.32	0.19	0.18	0.29	0.11	0.01	0.01	-0.53	0.34	0.25	0.31	0.29	0.28	0.44	-0.17	0.05
3	-0.06	0.72	1.00	0.02	0.32	0.26	-0.54	-0.39	-0.41	-0.19	0.41	0.41	0.43	-0.07	-0.05	0.04	0.06	-0.53	0.21	0.27	0.49	0.33	0.30	0.65	0.28	-0.58
4	0.14	-0.16	0.02	1.00	0.00	0.14	0.18	0.33	0.35	-0.38	0.31	0.37	0.48	-0.42	0.14	0.26	0.38	0.25	-0.01	0.09	-0.19	0.16	0.18	-0.03	0.08	-0.23
5	0.63	0.09	0.32	0.00	1.00	0.00	-0.37	-0.30	-0.32	-0.05	0.07	0.09	0.11	-0.08	0.08	0.10	0.13	-0.44	0.27	0.28	0.23	-0.39	-0.32	0.36	0.13	-0.28
6	-0.10	-0.04	0.26	0.14	0.00	1.00	-0.21	0.07	0.01	-0.44	0.11	0.29	0.29	-0.46	0.04	0.27	0.27	-0.24	0.15	0.33	0.17	0.06	0.09	0.22	0.08	-0.27
7	-0.04	-0.38	-0.54	0.18	-0.37	-0.21	1.00	0.86	0.87	0.11	0.32	0.25	0.24	0.17	0.36	0.28	0.26	0.64	-0.04	-0.06	-0.88	0.21	0.09	-0.50	0.16	0.29
8	-0.04	-0.45	-0.39	0.33	-0.30	0.07	0.86	1.00	0.98	-0.38	0.39	0.52	0.49	-0.26	0.33	0.49	0.46	0.59	-0.03	0.12	-0.76	0.17	0.06	-0.43	0.28	0.09
9	-0.06	-0.44	-0.41	0.35	-0.32	0.01	0.87	0.98	1.00	-0.38	0.41	0.51	0.49	-0.23	0.34	0.48	0.46	0.63	-0.05	0.09	-0.75	0.14	0.02	-0.44	0.27	0.11
10	0.05	0.17	-0.19	-0.38	-0.05	-0.44	0.11	-0.38	-0.38	1.00	-0.22	-0.55	-0.55	0.80	-0.01	-0.44	-0.44	-0.06	0.01	-0.32	-0.13	0.10	0.13	-0.06	-0.26	0.32
11	0.02	0.32	0.41	0.31	0.07	0.11	0.32	0.39	0.41	-0.22	1.00	0.91	0.91	0.13	0.67	0.68	0.67	0.00	0.50	0.53	-0.20	0.30	0.18	0.09	0.33	-0.19
12	0.03	0.19	0.41	0.37	0.09	0.29	0.25	0.52	0.51	-0.55	0.91	1.00	0.97	-0.21	0.57	0.77	0.75	0.02	0.42	0.60	-0.15	0.26	0.15	0.07	0.38	-0.27
13	0.07	0.18	0.43	0.48	0.11	0.29	0.24	0.49	0.49	-0.55	0.91	0.97	1.00	-0.28	0.57	0.74	0.77	0.01	0.42	0.58	-0.12	0.28	0.17	0.10	0.35	-0.32
14	-0.11	0.29	-0.07	-0.42	-0.08	-0.46	0.17	-0.26	-0.23	0.80	0.13	-0.21	-0.28	1.00	0.19	-0.21	-0.28	-0.01	0.15	-0.16	-0.18	0.02	0.02	-0.03	-0.07	0.31
15	0.26	0.11	-0.05	0.14	0.08	0.04	0.36	0.33	0.34	-0.01	0.67	0.57	0.57	0.19	1.00	0.88	0.87	-0.26	0.88	0.82	-0.25	0.06	0.04	-0.26	-0.06	0.27
16	0.24	0.01	0.04	0.26	0.10	0.27	0.28	0.49	0.48	-0.44	0.68	0.77	0.74	-0.21	0.88	1.00	0.97	-0.22	0.77	0.89	-0.18	0.04	0.01	-0.22	0.07	0.10
17	0.29	0.01	0.06	0.38	0.13	0.27	0.26	0.46	0.46	-0.44	0.67	0.75	0.77	-0.28	0.87	0.97	1.00	-0.23	0.77	0.87	-0.14	0.06	0.02	-0.18	0.04	0.04
18	-0.21	-0.53	-0.53	0.25	-0.44	-0.24	0.64	0.59	0.63	-0.06	0.00	0.02	0.01	-0.01	-0.26	-0.22	-0.23	1.00	-0.69	-0.63	-0.56	0.04	-0.05	-0.36	0.14	0.11
19	0.30	0.34	0.21	-0.01	0.27	0.15	-0.04	-0.03	-0.05	0.01	0.50	0.42	0.42	0.15	0.88	0.77	0.77	-0.69	1.00	0.92	0.08	0.02	0.05	-0.02	-0.11	0.15
20	0.29	0.25	0.27	0.09	0.28	0.33	-0.06	0.12	0.09	-0.32	0.53	0.60	0.58	-0.16	0.82	0.89	0.87	-0.63	0.92	1.00	0.11	0.01	0.03	-0.01	-0.00	0.03
21	-0.01	0.31	0.49	-0.19	0.23	0.17	-0.88	-0.76	-0.75	-0.13	-0.20	-0.15	-0.12	-0.18	-0.25	-0.18	-0.14	-0.56	0.08	0.11	1.00	-0.10	-0.08	0.35	-0.04	-0.34
22	-0.30	0.29	0.33	0.16	-0.39	0.06	0.21	0.17	0.14	0.10	0.30	0.26	0.28	0.02	0.06	0.04	0.06	0.04	0.02	0.01	-0.10	1.00	0.93	-0.02	0.13	-0.13
23	-0.26	0.28	0.30	0.18	-0.32	0.09	0.09	0.06	0.02	0.13	0.18	0.15	0.17	0.02	0.04	0.01	0.02	-0.05	0.05	0.03	-0.08	0.93	1.00	0.02	0.03	-0.08
24	0.03	0.44	0.65	-0.03	0.36	0.22	-0.50	-0.43	-0.44	-0.06	0.09	0.07	0.10	-0.03	-0.26	-0.22	-0.18	-0.36	-0.02	-0.01	0.35	-0.02	0.02	1.00	0.02	-0.42
25	-0.32	-0.17	0.28	0.08	0.13	0.08	0.16	0.28	0.27	-0.26	0.33	0.38	0.35	-0.07	-0.06	0.07	0.04	0.14	-0.11	-0.00	-0.04	0.13	0.03	0.02	1.00	-0.74
26	0.23	0.05	-0.58	-0.23	-0.28	-0.27	0.29	0.09	0.11	0.32	-0.19	-0.27	-0.32	0.31	0.27	0.10	0.04	0.11	0.15	0.03	-0.34	-0.13	-0.08	-0.42	-0.74	1.00

Plot 8.2.1 Covariance Biplot

In a covariance biplot (described in Section 4.2), the data is column centred, and scaled so that the norms of the variables are their standard deviations.

Quality of the Display

The overall quality of approximation of the two dimensional representation is 83%. The first axis retains 66% of the variance of the original data matrix, and the second axis, 17% (Table 8.2.1).

The first axis is almost completely composed of variables 15, 16, 17, 19, 20 (the oxygen capacity variables). These five variables constitute 95.7% of this axis. The second axis is constituted mainly by weight (56%), and also by height and oxygen capacity.

Notice that in the covariance biplot, the variables with the largest standard deviations (given in Table 8.2) tend to dominate the axes and be well approximated. The plot does not provide much information about variables whose standard deviations are relatively small.

Interpretations

The covariance biplot (Figure 8.2.1) immediately reveals the relative sizes of the standard deviations of the variables. The change in oxygen capacity variables (variables 19 and 20) have the largest norms (furthest from the origin) i.e. the largest standard deviations. Variables 15, 16, 17 and 3 also have large norms. The variables with the smallest standard deviations are bunched up in the vicinity of the origin; the points representing variables 6, 10, 11, 13, 14, 21, 22 and 25 are virtually superimposed.

The lines connecting the points for variables 2, 3, and 5 (height, weight and fat) with the origin have small angles between them, representing positive correlations. The angles between these variables and variables 15, 16, and 17 (the oxygen capacity variables) are approximately 90 degrees. This indicates that the two groups of variables are uncorrelated. Variable 18 (oxygen capacity at rest, or VO_2 max) is displayed as being negatively correlated to variables 2, 3 and 5.

The above is borne out by the correlation matrix (Table 8.2(a)) in all cases except for variable 5. For example, the correlation matrix reveals that variables 2 and 5 are not highly correlated. In fact they are close to being uncorrelated. (r=0.09). Variable 5 (fat) is poorly represented in the first two dimensions. (Its quality of representation in the plot is 0.067 + 0.09 = 0.157, i.e less than 16%.) This illustrates the pitfalls of interpreting specific points which are poorly approximated. Interpretations should be confirmed by reference to the original matrix.

The correlation structure of the variables bunched up at the origin is not clear from the plot. These variables all have small standard deviations and with their present scaling, have little impact on the display.

The scalar product interpretation is valid. As this plot does not display most of the variables satisfactorily, this interpretation is not discussed here.

When the influence of a particular observation is considered to be 'too' great (some judgement is required as to what 'too' great is), the observation could be deleted (or made a supplementary point) and the two plots compared. The observation may well be an outlier. Another possibility is that such an observation has a relatively high value of a well represented variable.

The effect of influential columns can also be investigated. This is illustrated in plot 8.2.4 below, where the speed variables are suppressed.

Comments

The plot serves as an illustration of the consequences of using the covariance biplot when the variables are measured on widely differing scales. The net effect is that some variables are represented well, at the expense of others. A good 'all round' view of the data set is not obtained, despite the high percentage of the variation explained in two dimensions. A further criticism of the plot is that due to the different scales on which the variables are measured, Euclidean distances between the column points are not particularly meaningful.

TABLE 8.2.1

T H E E I G E N V A L U E S

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	14386.58008	66.383	66.383
2	3681.38525	16.987	83.369
3	1473.45435	6.799	90.168
4	1078.43152	4.976	95.144
5	584.46497	2.697	97.841
6	222.12778	1.025	98.866
7	93.58221	0.432	99.298
8	70.67296	0.326	99.624
9	37.63599	0.174	99.797
10	23.08973	0.107	99.904
11	10.03416	0.046	99.950
12	4.43219	0.020	99.971
13	3.12409	0.014	99.985
14	2.01060	0.009	99.994
15	0.53359	0.002	99.997
16	0.25548	0.001	99.998
17	0.19673	0.001	99.999
18	0.13864	0.001	100.000

TABLE 8.2.1 (cont)

THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	FACT 2	COR
1	R1	A	1.000	0.038	0.038	-0.0591	0.004	0.1862	0.035
2	R3	C	1.000	0.058	0.036	0.2339	0.058	0.0065	0.000
3	R4	D	1.000	0.157	0.030	0.1123	0.016	0.3329	0.141
4	R5	E	1.000	0.103	0.038	-0.1101	0.012	0.2984	0.090
5	R6	F	1.000	0.053	0.038	-0.0511	0.003	0.2223	0.050
6	R7	G	1.000	0.030	0.033	0.1378	0.022	0.0852	0.008
7	R8	H	1.000	0.048	0.038	0.1309	0.017	0.1748	0.031
8	R9	I	1.000	0.119	0.038	0.2807	0.079	-0.1993	0.040
9	R10	J	1.000	0.022	0.038	-0.1448	0.021	-0.0220	0.000
10	R11	K	1.000	0.007	0.038	0.0052	0.000	0.0826	0.007
11	R12	L	1.000	0.056	0.033	-0.0880	0.009	0.2005	0.047
12	R13	M	1.000	0.036	0.038	-0.1874	0.035	0.0365	0.001
13	R14	N	1.000	0.057	0.038	-0.0368	0.001	0.2335	0.055
14	R15	O	1.000	0.058	0.037	-0.1698	0.030	0.1643	0.028
15	R16	P	1.000	0.096	0.034	0.2911	0.095	0.0296	0.001
16	R17	Q	1.000	0.072	0.038	0.2572	0.068	-0.0691	0.005
17	R18	R	1.000	0.077	0.038	-0.2192	0.049	0.1656	0.028
18	R20	T	1.000	0.004	0.038	0.0346	0.001	-0.0556	0.003
19	R21	U	1.000	0.118	0.038	-0.3113	0.097	-0.1443	0.021
20	R22	V	1.000	0.042	0.038	-0.0642	0.004	-0.1931	0.037
21	R23	W	1.000	0.035	0.036	0.0030	0.000	-0.1792	0.035
22	R24	X	1.000	0.093	0.037	-0.2167	0.049	-0.2030	0.043
23	R25	Y	1.000	0.259	0.038	-0.3268	0.107	-0.3884	0.152
24	R26	Z	1.000	0.126	0.038	-0.2301	0.053	-0.2680	0.072
25	R27	a	1.000	0.159	0.038	0.3209	0.103	-0.2361	0.056
26	R28	b	1.000	0.017	0.038	0.1281	0.017	-0.0121	0.000
27	R29	c	1.000	0.141	0.038	0.2798	0.079	-0.2488	0.062

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	AGE	1	1.000	0.120	0.063	0.1207	0.106	0.010	0.0424	0.013	0.005
2	HEIGHT	2	1.000	0.673	0.043	0.0741	0.059	0.004	-0.2379	0.614	0.154
3	WEIGHT	3	1.000	0.877	0.114	0.1060	0.046	0.008	-0.4529	0.832	0.557
4	TEMF	4	1.000	0.058	0.000	0.0041	0.019	0.000	0.0059	0.039	0.000
5	FAT	5	1.000	0.157	0.011	0.0405	0.067	0.001	-0.0469	0.090	0.006
6	AREA	6	1.000	0.083	0.000	0.0019	0.061	0.000	-0.0011	0.021	0.000
7	SPEED42	7	1.000	0.514	0.003	0.0041	0.003	0.000	0.0527	0.511	0.008
8	SPEED21	8	1.000	0.484	0.003	0.0148	0.035	0.000	0.0533	0.449	0.008
9	SPEED6	9	1.000	0.505	0.003	0.0134	0.028	0.000	0.0547	0.476	0.008
10	SPEED42-6	;	1.000	0.060	0.001	-0.0093	0.057	0.000	-0.0021	0.003	0.000
11	MET42	<	1.000	0.394	0.000	0.0141	0.388	0.000	-0.0018	0.006	0.000
12	MET21	=	1.000	0.404	0.000	0.0170	0.403	0.000	-0.0009	0.001	0.000
13	MET6	>	1.000	0.399	0.000	0.0148	0.397	0.000	-0.0010	0.002	0.000
14	MET42-6	@	1.000	0.010	0.000	-0.0006	0.004	0.000	-0.0007	0.005	0.000
15	O2MAX42	!	1.000	0.899	0.100	0.4232	0.828	0.124	0.1244	0.072	0.042
16	O2MAX21	"	1.000	0.920	0.130	0.4895	0.853	0.167	0.1371	0.067	0.051
17	O2MAX6	#	1.000	0.901	0.100	0.4277	0.841	0.127	0.1145	0.060	0.036
18	O2MAXREST	$	1.000	0.705	0.042	-0.1655	0.300	0.019	0.1925	0.405	0.101
19	O2MAX42-R	%	1.000	0.916	0.177	0.5887	0.904	0.241	-0.0681	0.012	0.013
20	O2MAX21-R	&	1.000	0.972	0.205	0.6549	0.965	0.298	-0.0554	0.007	0.008
21	FLUID	(	1.000	0.369	0.000	0.0001	0.000	0.000	-0.0079	0.369	0.000
22	SWEAT	)	1.000	0.056	0.000	0.0007	0.002	0.000	-0.0039	0.054	0.000
23	DEHYD	*	1.000	0.067	0.002	0.0030	0.002	0.000	-0.0166	0.065	0.001
24	ENDO	+	1.000	0.473	0.001	-0.0029	0.005	0.000	-0.0298	0.468	0.002
25	MESO	,	1.000	0.012	0.001	-0.0010	0.001	0.000	-0.0043	0.011	0.000
26	ECTO	-	1.000	0.203	0.001	0.0043	0.009	0.000	0.0204	0.194	0.001

FIGURE 8.2.1

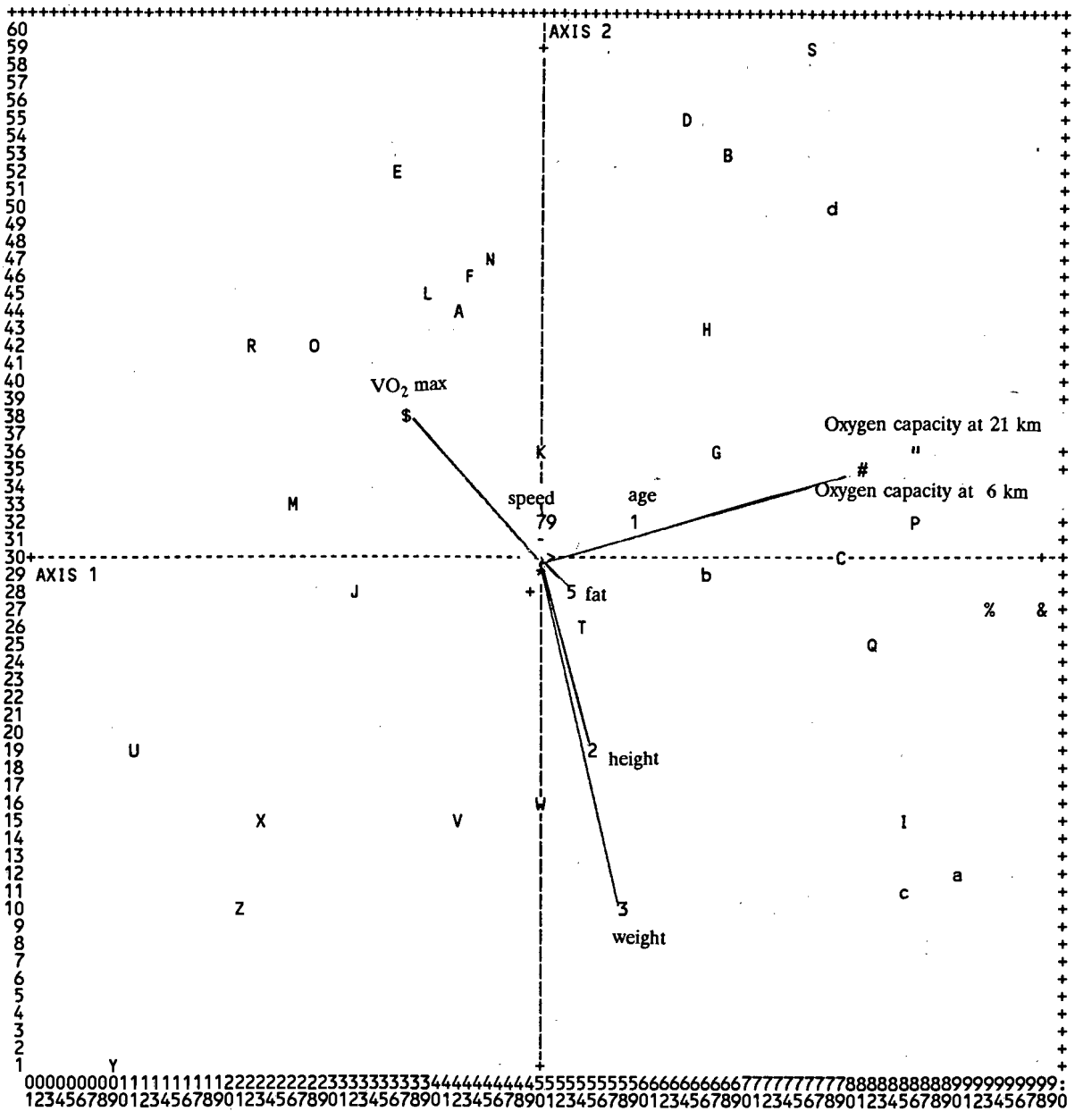

AXIS 2
AXIS 1
VO_2 max
Oxygen capacity at 21 km
Oxygen capacity at 6 km
speed
age
fat
height
weight

FIGURE 8.2.1 (cont)

NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	R1	A	-0.0591	0.1862	44	42
2	R3	C	0.2339	0.0065	30	79
3	R4	D	0.1123	0.3329	55	64
4	R5	E	-0.1101	0.2984	52	36
5	R6	F	-0.0511	0.2223	46	43
6	R7	G	0.1378	0.0852	36	67
7	R8	H	0.1309	0.1748	43	66
8	R9	I	0.2807	-0.1993	15	85
9	R10	J	-0.1448	-0.0220	28	32
10	R11	K	0.0052	0.0826	36	50
11	R12	L	-0.0880	0.2005	45	39
12	R13	M	-0.1874	0.0365	33	26
13	R14	N	-0.0368	0.2335	47	45
14	R15	O	-0.1698	0.1643	42	28
15	R16	P	0.2911	0.0296	32	86
16	R17	Q	0.2572	-0.0691	25	82
17	R18	R	-0.2192	0.1656	42	22
18	R20	T	0.0346	-0.0556	26	54
19	R21	U	-0.3113	-0.1443	19	11
20	R22	V	-0.0642	-0.1931	15	42
21	R23	W	0.0030	-0.1792	16	50
22	R24	X	-0.2167	-0.2030	15	23
23	R25	Y	-0.3268	-0.3884	1	9
24	R26	Z	-0.2301	-0.2680	10	21
25	R27	a	0.3209	-0.2361	12	90
26	R28	b	0.1281	-0.0121	29	66
27	R29	c	0.2798	-0.2488	11	85
28	R2	B	0.1483	0.3106	53	68
29	R19	S	0.2104	0.4725	59	76
30	R30	d	0.2229	0.2663	50	78
1	AGE	1	0.1207	0.0424	32	59
2	HEIGHT	2	0.0741	-0.2379	19	55
3	WEIGHT	3	0.1060	-0.4529	10	58
4	TEMP	4	0.0041	0.0059	30	50
5	FAT	5	0.0405	-0.0469	28	53
6	AREA	6	0.0019	-0.0011	30	50
7	SPEED42	7	0.0041	0.0527	32	50
8	SPEED21	8	0.0148	0.0533	32	51
9	SPEED6	9	0.0134	0.0547	32	51
10	SPEED42-6	;	-0.0093	-0.0021	30	49
11	MET42	<	0.0141	-0.0018	30	51
12	MET21	=	0.0170	-0.0009	30	51
13	MET6	>	0.0148	-0.0010	30	51
14	MET42-6	@	-0.0006	-0.0007	30	50
15	O2MAX42	!	0.4232	0.1244	35	81
16	O2MAX21	"	0.4895	0.1371	36	86
17	O2MAX6	#	0.4277	0.1145	35	81
18	O2MAXREST	$	-0.1655	0.1925	38	37
19	O2MAX42-R	%	0.5887	-0.0681	27	93
20	O2MAX21-R	&	0.6549	-0.0554	27	98
21	FLUID	(	0.0001	-0.0079	29	50
22	SWEAT	)	0.0007	-0.0039	30	50
23	DEHYD	*	0.0030	-0.0166	29	50
24	ENDO	+	-0.0029	-0.0298	28	49
25	MESO	,	-0.0010	-0.0043	30	50
26	ECTO	-	0.0043	0.0204	31	50

Plot 8.2.2 Correlation Biplot

The variables are standardised; information about the standard deviations is not displayed (Figure 8.2.2).

Quality of the display

The quality of the two dimensional display is 50% (Table 8.2.2), which although not particularly high in an absolute sense, is quite good considering that we are approximating a matrix of 26 dimensions.

The standardization forces a more equitable representation of the variables than in the previous display. In fact, the correlation biplot provides the best representation of the correlation structure (Gabriel, 1981b, Underhill, 1990a).

That the variables have a more equitable representation can be seen from the plot (Fig 8.2.2), which does not demonstrate the bunching up of variables that occured in Plot 8.2.1. This is confirmed with reference to the relative contribution (ctr) columns (Table 8.2.2). The first axis is composed mainly of the variables of oxygen capacity and metabolic rate (variables 11, 12, 13, 15, 16 and 17). The second axis is constituted by weight, speed, VO_2 max and fluid intake (variables 3, 7, 8, 9, 18 and 21).

Interpretations

The display emphasizes the relationships between the variables resting oxygen capacity, speed and ectomorphy are which are positively correlated to each other, and negatively correlated with endomorphy, fluid intake, height, weight and body fat (Table 8.2(a)). This is consistent with what sports scientists would expect: faster runners have high oxygen capacities and tend to be slim, slightly built individuals (ectomorphs). Slower individuals tend to be associated with a bigger, heavier build, and are also known to spend more time during the race consuming fluids.

A particularly interesting feature of the plot is that it groups runners of similar speed in the same quadrant. Runners A, B, D, E, F and G, the fastest runners, are attracted towards the speed variables. The slower runners (V, W, X, Y, Z, b and d) fall in the opposite quadrant to the speed variables. They are associated with large fluid intakes and

the body type endomorph.

Applying the scalar product interpretation, projection of the points representing the individual runners onto the total speed variable (variable 7) gives an approximate ordering of the speed of the runners.

TABLE 8.2.2

T H E E I G E N V A L U E S

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	195.81123	28.966	28.966
2	144.56030	21.385	50.351
3	92.40421	13.669	64.020
4	74.44394	11.012	75.032
5	45.08663	6.670	81.702
6	33.36649	4.936	86.638
7	25.53653	3.778	90.416
8	19.49965	2.885	93.300
9	15.48493	2.291	95.591
10	10.02691	1.483	97.074
11	7.59273	1.123	98.197
12	4.81166	0.712	98.909
13	2.85054	0.422	99.331
14	2.05175	0.304	99.634
15	1.05666	0.156	99.791
16	0.66532	0.098	99.889
17	0.42926	0.063	99.952
18	0.24340	0.036	99.988
19	0.04045	0.006	99.994
20	0.02146	0.003	99.998
21	0.01470	0.002	100.000

TABLE 8.2.2 (cont)

THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	FACT 2	COR
1	R1	A	1.000	0.140	0.038	0.1240	0.015	-0.3523	0.125
2	R3	C	1.000	0.139	0.037	0.3654	0.139	0.0197	0.000
3	R4	D	1.000	0.098	0.033	0.1397	0.023	-0.2538	0.075
4	R5	E	1.000	0.095	0.037	0.0849	0.008	-0.2883	0.087
5	R6	F	1.000	0.157	0.037	0.2334	0.056	-0.3125	0.101
6	R7	G	1.000	0.078	0.035	0.2575	0.073	-0.0685	0.005
7	R8	H	1.000	0.020	0.038	-0.0513	0.003	-0.1300	0.017
8	R9	I	1.000	0.080	0.038	0.2407	0.059	0.1431	0.021
9	R10	J	1.000	0.033	0.038	-0.1700	0.029	-0.0654	0.004
10	R11	K	1.000	0.026	0.037	-0.1357	0.019	-0.0799	0.007
11	R12	L	1.000	0.031	0.036	-0.0800	0.007	-0.1488	0.024
12	R13	M	1.000	0.022	0.038	-0.1087	0.012	-0.1009	0.010
13	R14	N	1.000	0.023	0.037	-0.0415	0.002	-0.1413	0.021
14	R15	O	1.000	0.053	0.037	-0.1967	0.040	-0.1101	0.013
15	R16	P	1.000	0.066	0.037	0.2414	0.061	0.0725	0.005
16	R17	Q	1.000	0.099	0.036	0.2661	0.075	0.1501	0.024
17	R18	R	1.000	0.055	0.038	-0.2142	0.047	-0.0879	0.008
18	R20	T	1.000	0.018	0.038	-0.0304	0.001	0.1291	0.017
19	R21	U	1.000	0.045	0.038	-0.2108	0.045	-0.0051	0.000
20	R22	V	1.000	0.028	0.038	-0.1369	0.019	0.0936	0.009
21	R23	W	1.000	0.041	0.031	-0.1104	0.015	0.1456	0.026
22	R24	X	1.000	0.035	0.037	-0.1693	0.030	0.0749	0.006
23	R25	Y	1.000	0.061	0.038	-0.2234	0.050	0.1021	0.010
24	R26	Z	1.000	0.177	0.038	-0.3771	0.143	0.1860	0.035
25	R27	a	1.000	0.194	0.038	0.1572	0.025	0.4088	0.169
26	R28	b	1.000	0.045	0.038	-0.0491	0.002	0.2062	0.043
27	R29	c	1.000	0.215	0.037	0.1949	0.039	0.4132	0.176

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	AGE	1	1.000	0.057	0.038	0.0798	0.025	0.003	0.0925	0.033	0.006
2	HEIGHT	2	1.000	0.412	0.038	0.0152	0.001	0.000	0.3271	0.411	0.074
3	WEIGHT	3	1.000	0.648	0.038	0.0797	0.024	0.003	0.4026	0.624	0.112
4	TEMP	4	1.000	0.210	0.038	0.2215	0.189	0.025	-0.0737	0.021	0.004
5	FAT	5	1.000	0.287	0.038	0.0308	0.004	0.000	0.2715	0.284	0.051
6	AREA	6	1.000	0.163	0.038	0.1498	0.086	0.011	0.1408	0.076	0.014
7	SPEED42	7	1.000	0.857	0.038	0.2133	0.175	0.023	-0.4211	0.682	0.123
8	SPEED21	8	1.000	0.917	0.038	0.3258	0.408	0.054	-0.3639	0.509	0.092
9	SPEED6	9	1.000	0.935	0.038	0.3208	0.396	0.053	-0.3745	0.539	0.097
10	SPEED42-6	;	1.000	0.247	0.038	-0.2510	0.242	0.032	-0.0339	0.004	0.001
11	MET42	<	1.000	0.701	0.038	0.4217	0.684	0.091	0.0674	0.017	0.003
12	MET21	=	1.000	0.834	0.038	0.4615	0.819	0.109	0.0631	0.015	0.003
13	MET6	>	1.000	0.841	0.038	0.4617	0.820	0.109	0.0746	0.021	0.004
14	MET42-6	@	1.000	0.065	0.038	-0.1282	0.063	0.008	-0.0232	0.002	0.000
15	O2MAX42	!	1.000	0.635	0.038	0.4046	0.630	0.084	0.0380	0.006	0.001
16	O2MAX21	"	1.000	0.879	0.038	0.4758	0.871	0.116	0.0474	0.009	0.002
17	O2MAX6	#	1.000	0.887	0.038	0.4762	0.872	0.116	0.0612	0.014	0.003
18	O2MAXREST	$	1.000	0.759	0.038	-0.0227	0.002	0.000	-0.4438	0.757	0.136
19	O2MAX42-R	%	1.000	0.613	0.038	0.3150	0.382	0.051	0.2452	0.231	0.042
20	O2MAX21-R	&	1.000	0.800	0.038	0.3885	0.581	0.077	0.2389	0.219	0.039
21	FLUID	(	1.000	0.668	0.038	-0.1554	0.093	0.012	0.3866	0.575	0.103
22	SWEAT	)	1.000	0.051	0.038	0.1145	0.050	0.007	-0.0169	0.001	0.000
23	DEHYD	*	1.000	0.024	0.038	0.0759	0.022	0.003	0.0191	0.001	0.000
24	ENDO	+	1.000	0.406	0.038	-0.0756	0.022	0.003	0.3160	0.384	0.069
25	MESO	,	1.000	0.066	0.038	0.1294	0.064	0.009	-0.0197	0.001	0.000
26	ECTO	-	1.000	0.127	0.038	-0.0408	0.006	0.001	-0.1770	0.121	0.022

FIGURE 8.2.2

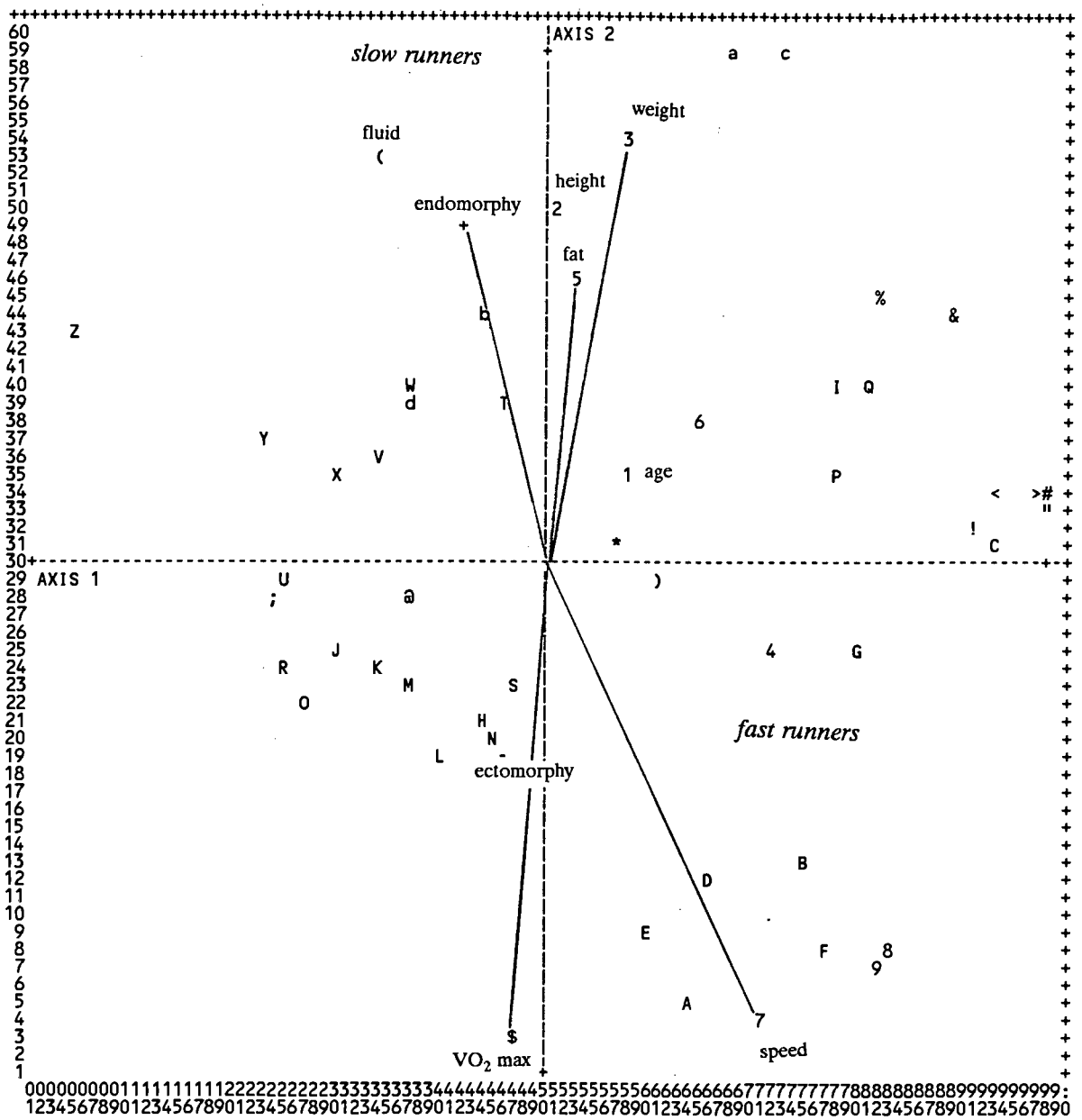

AXIS 2
AXIS 1
slow runners
fast runners
fluid
weight
height
endomorphy
fat
age
ectomorphy
speed
VO_2 max

FIGURE 8.2.2 (cont)

NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	R1	A	0.1240	-0.3523	5	64
2	R3	C	0.3654	0.0197	31	93
3	R4	D	0.1397	-0.2538	12	66
4	R5	E	0.0849	-0.2883	9	60
5	R6	F	0.2334	-0.3125	8	77
6	R7	G	0.2575	-0.0685	25	80
7	R8	H	-0.0513	-0.1300	21	44
8	R9	I	0.2407	0.1431	40	78
9	R10	J	-0.1700	-0.0654	25	30
10	R11	K	-0.1357	-0.0799	24	34
11	R12	L	-0.0800	-0.1488	19	40
12	R13	M	-0.1087	-0.1009	23	37
13	R14	N	-0.0415	-0.1413	20	45
14	R15	O	-0.1967	-0.1101	22	27
15	R16	P	0.2414	0.0725	35	78
16	R17	Q	0.2661	0.1501	40	81
17	R18	R	-0.2142	-0.0879	24	25
18	R20	T	-0.0304	0.1291	39	46
19	R21	U	-0.2108	-0.0051	29	25
20	R22	V	-0.1369	0.0936	36	34
21	R23	W	-0.1104	0.1456	40	37
22	R24	X	-0.1693	0.0749	35	30
23	R25	Y	-0.2234	0.1021	37	23
24	R26	Z	-0.3771	0.1860	43	5
25	R27	a	0.1572	0.4088	59	68
26	R28	b	-0.0491	0.2062	44	44
27	R29	c	0.1949	0.4132	59	73
28	R2	B	0.2133	-0.2321	13	75
29	R19	S	-0.0224	-0.1019	23	47
30	R30	d	-0.1044	0.1278	39	37
1	AGE	1	0.0798	0.0925	35	58
2	HEIGHT	2	0.0152	0.3271	50	51
3	WEIGHT	3	0.0797	0.4026	54	58
4	TEMP	4	0.2215	-0.0737	25	72
5	FAT	5	0.0308	0.2715	46	53
6	AREA	6	0.1498	0.1408	38	65
7	SPEED42	7	0.2133	-0.4211	4	71
8	SPEED21	8	0.3258	-0.3639	8	83
9	SPEED6	9	0.3208	-0.3745	7	82
10	SPEED42-6	:	-0.2510	-0.0339	28	24
11	MET42	<	0.4217	0.0674	34	93
12	MET21	=	0.4615	0.0631	34	97
13	MET6	>	0.4617	0.0746	34	97
14	MET42-6	@	-0.1282	-0.0232	28	37
15	O2MAX42	!	0.4046	0.0380	32	91
16	O2MAX21	"	0.4758	0.0474	33	98
17	O2MAX6	#	0.4762	0.0612	34	98
18	O2MAXREST	$	-0.0227	-0.4438	3	47
19	O2MAX42-R	%	0.3150	0.2452	45	82
20	O2MAX21-R	&	0.3885	0.2389	44	89
21	FLUID	(	-0.1554	0.3866	53	34
22	SWEAT	)	0.1145	-0.0169	29	61
23	DEHYD	*	0.0759	0.0191	31	57
24	ENDO	+	-0.0756	0.3160	49	42
25	MESO	,	0.1294	-0.0197	29	63
26	ECTO	-	-0.0408	-0.1770	19	46

Plot 8.2.3 Coefficient of Variation (CV) Biplot

In this plot (Figure 8.2.3), the norms of the points representing the variables approximate their coefficients of variation.

The quality of representation here is 65%. This plot did not result in a useful display of the variables' correlation structure or of the between set scalar products because it is dominated by the four variables (10, 14, 19 and 20) that have much larger coefficients of variation (Table 8.2) than the others. These variables are circled in Figure 8.2.3. The difference in the magnitudes of coefficients of variation results in a clustering of the remaining column points at the origin.

TABLE 8.2.3 (cont)

THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	FACT 2	COR
1	R1	A	1.000	0.017	0.038	-0.0905	0.008	0.0925	0.009
2	R3	C	1.000	0.047	0.037	0.2046	0.044	-0.0501	0.003
3	R4	D	1.000	0.008	0.031	0.0755	0.007	0.0260	0.001
4	R5	E	1.000	0.057	0.037	-0.1549	0.025	-0.1763	0.032
5	R6	F	1.000	0.195	0.038	-0.0838	0.007	-0.4292	0.188
6	R7	G	1.000	0.013	0.035	0.1045	0.012	-0.0203	0.000
7	R8	H	1.000	0.169	0.038	0.1290	0.017	0.3903	0.152
8	R9	I	1.000	0.154	0.038	0.2721	0.075	0.2793	0.079
9	R10	J	1.000	0.027	0.038	-0.1292	0.017	0.1020	0.010
10	R11	K	1.000	0.158	0.037	0.0223	0.001	0.3896	0.157
11	R12	L	1.000	0.013	0.034	-0.0974	0.011	0.0423	0.002
12	R13	M	1.000	0.041	0.038	-0.1965	0.039	-0.0489	0.002
13	R14	N	1.000	0.019	0.037	-0.0527	0.003	-0.1252	0.016
14	R15	O	1.000	0.039	0.038	-0.1728	0.030	0.0924	0.009
15	R16	P	1.000	0.118	0.036	0.3009	0.098	-0.1362	0.020
16	R17	Q	1.000	0.093	0.037	0.2573	0.069	-0.1533	0.024
17	R18	R	1.000	0.049	0.037	-0.2164	0.049	-0.0088	0.000
18	R20	T	1.000	0.006	0.038	0.0587	0.003	-0.0519	0.003
19	R21	U	1.000	0.100	0.038	-0.3096	0.096	-0.0667	0.004
20	R22	V	1.000	0.011	0.038	-0.0096	0.000	-0.1024	0.010
21	R23	W	1.000	0.068	0.033	0.0206	0.000	0.2411	0.068
22	R24	X	1.000	0.060	0.037	-0.1962	0.040	-0.1402	0.020
23	R25	Y	1.000	0.124	0.038	-0.3115	0.097	-0.1629	0.027
24	R26	Z	1.000	0.168	0.038	-0.2149	0.046	0.3477	0.121
25	R27	a	1.000	0.152	0.038	0.3451	0.120	-0.1804	0.033
26	R28	b	1.000	0.024	0.038	0.1513	0.023	0.0317	0.001
27	R29	c	1.000	0.122	0.038	0.2944	0.088	-0.1820	0.034

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	AGE	1	1.000	0.098	0.015	0.0003	0.096	0.003	0.0000	0.002	0.000
2	HEIGHT	2	1.000	0.155	0.000	0.0000	0.084	0.000	0.0000	0.071	0.000
3	WEIGHT	3	1.000	0.125	0.006	0.0002	0.071	0.001	-0.0001	0.054	0.002
4	TEMP	4	1.000	0.332	0.052	0.0002	0.008	0.001	-0.0011	0.324	0.096
5	FAT	5	1.000	0.088	0.012	0.0003	0.082	0.002	-0.0001	0.006	0.000
6	AREA	6	1.000	0.272	0.002	0.0001	0.074	0.000	-0.0002	0.197	0.003
7	SPEED42	7	1.000	0.012	0.005	0.0000	0.002	0.000	0.0001	0.010	0.000
8	SPEED21	8	1.000	0.136	0.006	0.0001	0.009	0.000	-0.0002	0.127	0.004
9	SPEED6	9	1.000	0.123	0.008	0.0001	0.005	0.000	-0.0003	0.118	0.006
10	SPEED42-6	;	1.000	0.853	0.068	-0.0005	0.050	0.007	0.0019	0.803	0.313
11	MET42	<	1.000	0.333	0.007	0.0004	0.318	0.005	-0.0001	0.016	0.001
12	MET21	=	1.000	0.535	0.010	0.0005	0.336	0.007	-0.0004	0.199	0.011
13	MET6	>	1.000	0.574	0.010	0.0005	0.329	0.007	-0.0004	0.246	0.014
14	MET42-6	@	1.000	0.808	0.075	-0.0002	0.005	0.001	0.0020	0.803	0.344
15	O2MAX42	!	1.000	0.797	0.006	0.0005	0.755	0.009	0.0001	0.042	0.001
16	O2MAX21	"	1.000	0.826	0.007	0.0006	0.778	0.012	-0.0002	0.048	0.002
17	O2MAX6	#	1.000	0.840	0.008	0.0006	0.763	0.012	-0.0002	0.077	0.003
18	O2MAXREST	$	1.000	0.441	0.004	-0.0003	0.410	0.003	-0.0001	0.031	0.001
19	O2MAX42-R	%	1.000	0.989	0.218	0.0037	0.932	0.427	0.0009	0.057	0.071
20	O2MAX21-R	&	1.000	0.992	0.243	0.0040	0.983	0.501	-0.0004	0.009	0.013
21	FLUID	(	1.000	0.023	0.006	0.0001	0.008	0.000	-0.0001	0.015	0.000
22	SWEAT	)	1.000	0.002	0.038	0.0001	0.001	0.000	-0.0001	0.001	0.000
23	DEHYD	*	1.000	0.003	0.091	0.0001	0.003	0.001	0.0000	0.000	0.000
24	ENDO	+	1.000	0.019	0.031	0.0000	0.000	0.000	-0.0002	0.019	0.003
25	MESO	,	1.000	0.136	0.011	0.0000	0.001	0.000	-0.0003	0.134	0.008
26	ECTO	-	1.000	0.309	0.059	0.0001	0.005	0.001	0.0011	0.304	0.102

FIGURE 8.2.3

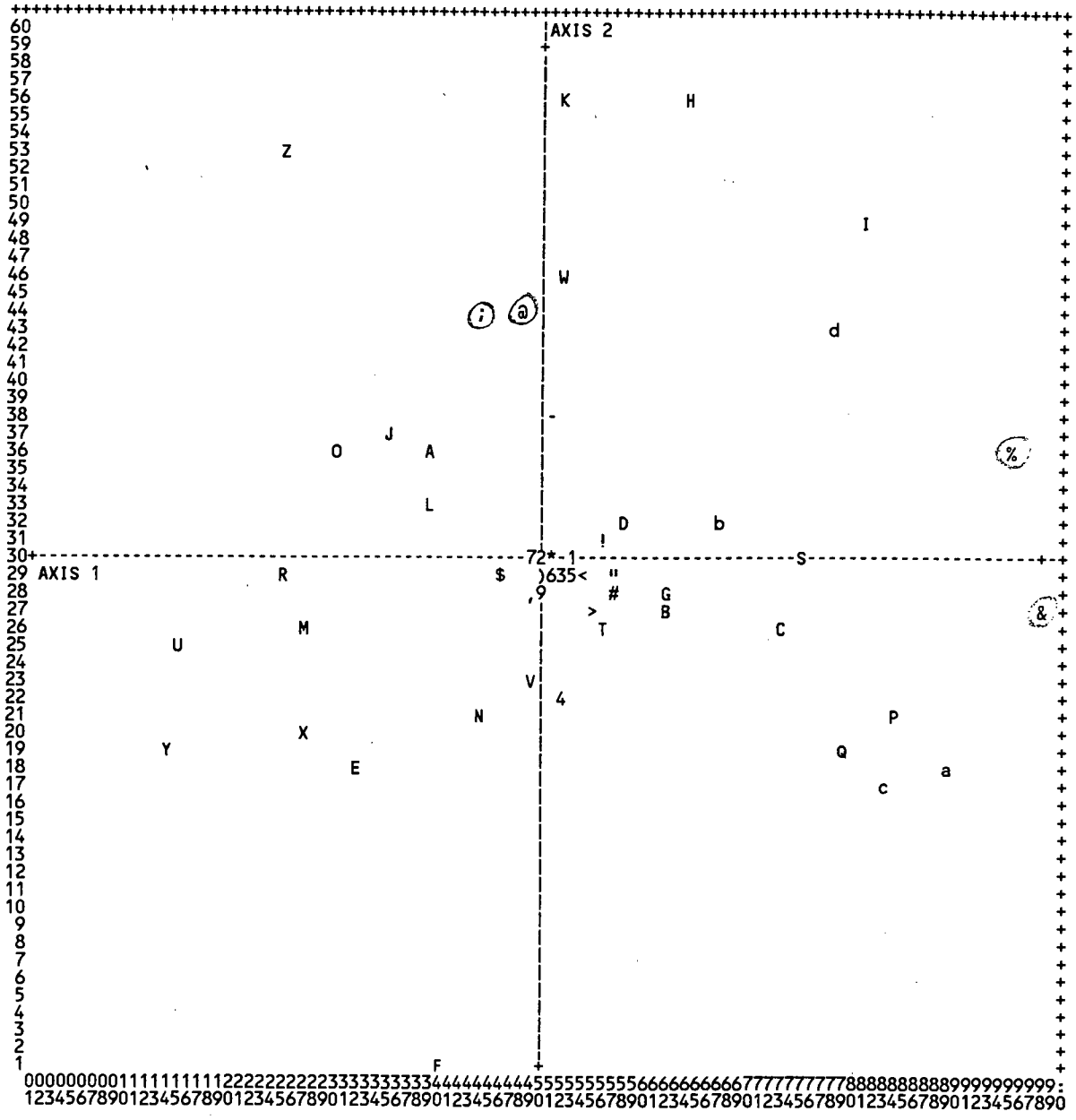


FIGURE 8.2.3 (cont)

NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	R1	A	-0.0905	0.0925	36	39
2	R3	C	0.2046	-0.0501	26	73
3	R4	D	0.0755	0.0260	32	58
4	R5	E	-0.1549	-0.1763	18	32
5	R6	F	-0.0838	-0.4292	1	40
6	R7	G	0.1045	-0.0203	28	62
7	R8	H	0.1290	0.3903	56	64
8	R9	I	0.2721	0.2793	49	81
9	R10	J	-0.1292	0.1020	37	35
10	R11	K	0.0223	0.3896	56	52
11	R12	L	-0.0974	0.0423	33	39
12	R13	M	-0.1965	-0.0489	26	27
13	R14	N	-0.0527	-0.1252	21	44
14	R15	O	-0.1728	0.0924	36	30
15	R16	P	0.3009	-0.1362	21	84
16	R17	Q	0.2573	-0.1533	19	79
17	R18	R	-0.2164	-0.0088	29	25
18	R20	T	0.0587	-0.0519	26	56
19	R21	U	-0.3096	-0.0667	25	15
20	R22	V	-0.0096	-0.1024	23	49
21	R23	W	0.0206	0.2411	46	52
22	R24	X	-0.1962	-0.1402	20	27
23	R25	Y	-0.3115	-0.1629	19	14
24	R26	Z	-0.2149	0.3477	53	25
25	R27	a	0.3451	-0.1804	18	89
26	R28	b	0.1513	0.0317	32	67
27	R29	c	0.2944	-0.1820	17	83
28	R2	B	0.1106	-0.0415	27	62
29	R19	S	0.2214	0.0031	30	75
30	R30	d	0.2455	0.1897	43	78
1	AGE	1	0.0003	0.0000	30	53
2	HEIGHT	2	0.0000	0.0000	30	50
3	WEIGHT	3	0.0002	-0.0001	29	52
4	TEMP	4	0.0002	-0.0011	22	52
5	FAT	5	0.0003	-0.0001	29	53
6	AREA	6	0.0001	-0.0002	29	51
7	SPEED42	7	0.0000	0.0001	30	49
8	SPEED21	8	0.0001	-0.0002	28	50
9	SPEED6	9	0.0001	-0.0003	28	50
10	SPEED42-6	;	-0.0005	0.0019	44	44
11	MET42	<	0.0004	-0.0001	29	54
12	MET21	=	0.0005	-0.0004	27	55
13	MET6	>	0.0005	-0.0004	27	55
14	MET42-6	@	-0.0002	0.0020	44	48
15	O2MAX42	!	0.0005	0.0001	31	56
16	O2MAX21	"	0.0006	-0.0002	29	57
17	O2MAX6	#	0.0006	-0.0002	28	57
18	O2MAXREST	$	-0.0003	-0.0001	29	46
19	O2MAX42-R	%	0.0037	0.0009	36	95
20	O2MAX21-R	&	0.0040	-0.0004	27	98
21	FLUID	(	0.0001	-0.0001	29	50
22	SWEAT	)	0.0001	-0.0001	29	50
23	DEHYD	*	0.0001	0.0000	30	51
24	ENDO	+	0.0000	-0.0002	28	49
25	MESO	,	0.0000	-0.0003	28	49
26	ECTO	-	0.0001	0.0011	38	51

Plot 8.2.4 Correlation Biplot (with the Speed Variables Suppressed)

In the previous correlation biplot (Figure 8.2.2), the speed variables account for 16,2% of the first axis and 31,3% of the second axis. It is therefore not altogether surprising that an ordering of the runners from fastest to slowest was noted. The motivation for this display is to investigate the effect of removing the speed variables. Suppression of dominating vectors allows other relationships to be displayed. In this plot (Figure 8.2.4), suppression of the speed variables enabled the 'body type' variables to be more prominently displayed.

Quality of the Display

The quality of the two dimensional display is 47,5% (Table 8.2.4), which is quite good considering that the rank of the original data matrix has high rank and that the variables have equitable representation.

The first axis is not unduly dominated by any variables. The second axis is mainly constituted by the variables measuring body type: endomorph, mesomorph, ectomorph and weight, which explain 53,2% of the axis. The lines that can be drawn in by connecting the points representing ectomorph, mesomorph and endomorph are roughly parallel to the second axis.

Interpretations

In spite of the absence of the speed variables, runners are grouped according to their marathon times, with the faster runners tending to have positive coefficients on the second axis. The slower runners have negative coefficients on that axis.

Comments

As noted above the second axis mainly represents the 'body type' variables. Thus the grouping of runners according to their speeds supports hypotheses of the importance of body type in determining running performance.

The importance of body type and weight on running performance is well known to sports scientists. Individuals with a large ectomorph measure have a slight build, and perform

better than solidly built runners. Heavy runners are also at a disadvantage relative to lighter ones.

TABLE 8.2.4

T H E E I G E N V A L U E S

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	178.57196	31.219	31.219
2	93.37119	16.324	47.543
3	83.49240	14.597	62.139
4	64.33086	11.247	73.386
5	41.50911	7.257	80.643
6	33.18497	5.802	86.444
7	23.65969	4.136	90.580
8	16.75667	2.929	93.510
9	12.56950	2.197	95.707
10	8.51601	1.489	97.196
11	7.57726	1.325	98.521
12	3.70948	0.649	99.169
13	2.25711	0.395	99.564
14	1.50760	0.264	99.828
15	0.53882	0.094	99.922
16	0.31867	0.056	99.978
17	0.10647	0.019	99.996
18	0.01961	0.003	100.000

T H E R O W O B J E C T S

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	FACT 2	COR
1	R1	A	1.000	0.041	0.039	0.0016	0.000	0.1875	0.041
2	R3	C	1.000	0.155	0.037	0.3494	0.150	-0.0628	0.005
3	R4	D	1.000	0.254	0.033	0.0359	0.002	0.4260	0.253
4	R5	E	1.000	0.044	0.040	-0.0468	0.002	0.1914	0.041
5	R6	F	1.000	0.024	0.040	0.0617	0.004	0.1316	0.020
6	R7	G	1.000	0.103	0.028	0.2316	0.088	0.0946	0.015
7	R8	H	1.000	0.133	0.044	-0.0458	0.002	0.3543	0.130
8	R9	I	1.000	0.110	0.039	0.3026	0.106	-0.0547	0.003
9	R10	J	1.000	0.059	0.034	-0.1807	0.044	-0.1060	0.015
10	R11	K	1.000	0.049	0.043	-0.1199	0.015	0.1766	0.033
11	R12	L	1.000	0.047	0.033	-0.1331	0.024	0.1272	0.022
12	R13	M	1.000	0.039	0.029	-0.1430	0.032	-0.0632	0.006
13	R14	N	1.000	0.084	0.039	-0.0992	0.011	0.2501	0.073
14	R15	O	1.000	0.084	0.040	-0.2173	0.054	0.1604	0.029
15	R16	P	1.000	0.084	0.032	0.2358	0.080	0.0515	0.004
16	R17	Q	1.000	0.107	0.040	0.3018	0.103	-0.0646	0.005
17	R18	R	1.000	0.077	0.034	-0.2384	0.077	-0.0200	0.001
18	R20	T	1.000	0.008	0.043	0.0106	0.000	-0.0884	0.008
19	R21	U	1.000	0.093	0.042	-0.2193	0.052	-0.1932	0.041
20	R22	V	1.000	0.065	0.035	-0.1165	0.018	-0.1913	0.047
21	R23	W	1.000	0.009	0.030	-0.0326	0.002	-0.0710	0.008
22	R24	X	1.000	0.080	0.040	-0.1598	0.029	-0.2105	0.051
23	R25	Y	1.000	0.231	0.042	-0.1832	0.037	-0.4218	0.194
24	R26	Z	1.000	0.104	0.042	-0.2568	0.072	-0.1720	0.032
25	R27	a	1.000	0.130	0.043	0.2854	0.086	-0.2050	0.044
26	R28	b	1.000	0.004	0.034	0.0510	0.004	-0.0047	0.000
27	R29	c	1.000	0.255	0.028	0.3250	0.174	-0.2220	0.081

T H E C O L U M N O B J E C T S

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	AGE	1	1.000	0.109	0.045	0.1185	0.054	0.008	0.1200	0.055	0.015
2	HEIGHT	2	1.000	0.351	0.045	0.1491	0.085	0.012	-0.2629	0.266	0.074
3	WEIGHT	3	1.000	0.929	0.045	0.2118	0.173	0.025	-0.4436	0.757	0.211
4	TEMP	4	1.000	0.113	0.045	0.1662	0.106	0.015	0.0424	0.007	0.002
5	FAT	5	1.000	0.135	0.045	0.1266	0.062	0.009	-0.1385	0.074	0.021
6	AREA	6	1.000	0.182	0.045	0.1701	0.111	0.016	-0.1358	0.071	0.020
7	MET42	<	1.000	0.687	0.045	0.4225	0.687	0.100	-0.0100	0.000	0.000
8	MET21	=	1.000	0.748	0.045	0.4404	0.746	0.109	-0.0221	0.002	0.001
9	MET6	>	1.000	0.765	0.045	0.4445	0.760	0.111	-0.0356	0.005	0.001
10	MET42-6	@	1.000	0.043	0.045	-0.0853	0.028	0.004	0.0631	0.015	0.004
11	O2MAX42	!	1.000	0.879	0.045	0.4160	0.666	0.097	0.2353	0.213	0.059
12	O2MAX21	"	1.000	0.953	0.045	0.4628	0.824	0.120	0.1830	0.129	0.036
13	O2MAX6	#	1.000	0.947	0.045	0.4678	0.842	0.123	0.1659	0.106	0.029
14	O2MAXREST	$	1.000	0.340	0.045	-0.1966	0.149	0.022	0.2229	0.191	0.053
15	O2MAX42-R	%	1.000	0.659	0.045	0.4084	0.642	0.093	0.0679	0.018	0.005
16	O2MAX21-R	&	1.000	0.811	0.045	0.4570	0.803	0.117	0.0444	0.008	0.002
17	FLUID	(	1.000	0.393	0.045	-0.0101	0.000	0.000	-0.3197	0.393	0.109
18	SWEAT	)	1.000	0.106	0.045	0.1067	0.044	0.006	-0.1269	0.062	0.017
19	DEHYD	*	1.000	0.094	0.045	0.0865	0.029	0.004	-0.1307	0.066	0.018
20	ENDO	+	1.000	0.558	0.045	0.0334	0.004	0.001	-0.3795	0.554	0.154
21	MESO	,	1.000	0.138	0.045	0.0923	0.033	0.005	-0.1652	0.105	0.029
22	ECTO	-	1.000	0.518	0.045	-0.0756	0.022	0.003	0.3590	0.496	0.138

FIGURE 8.2.4

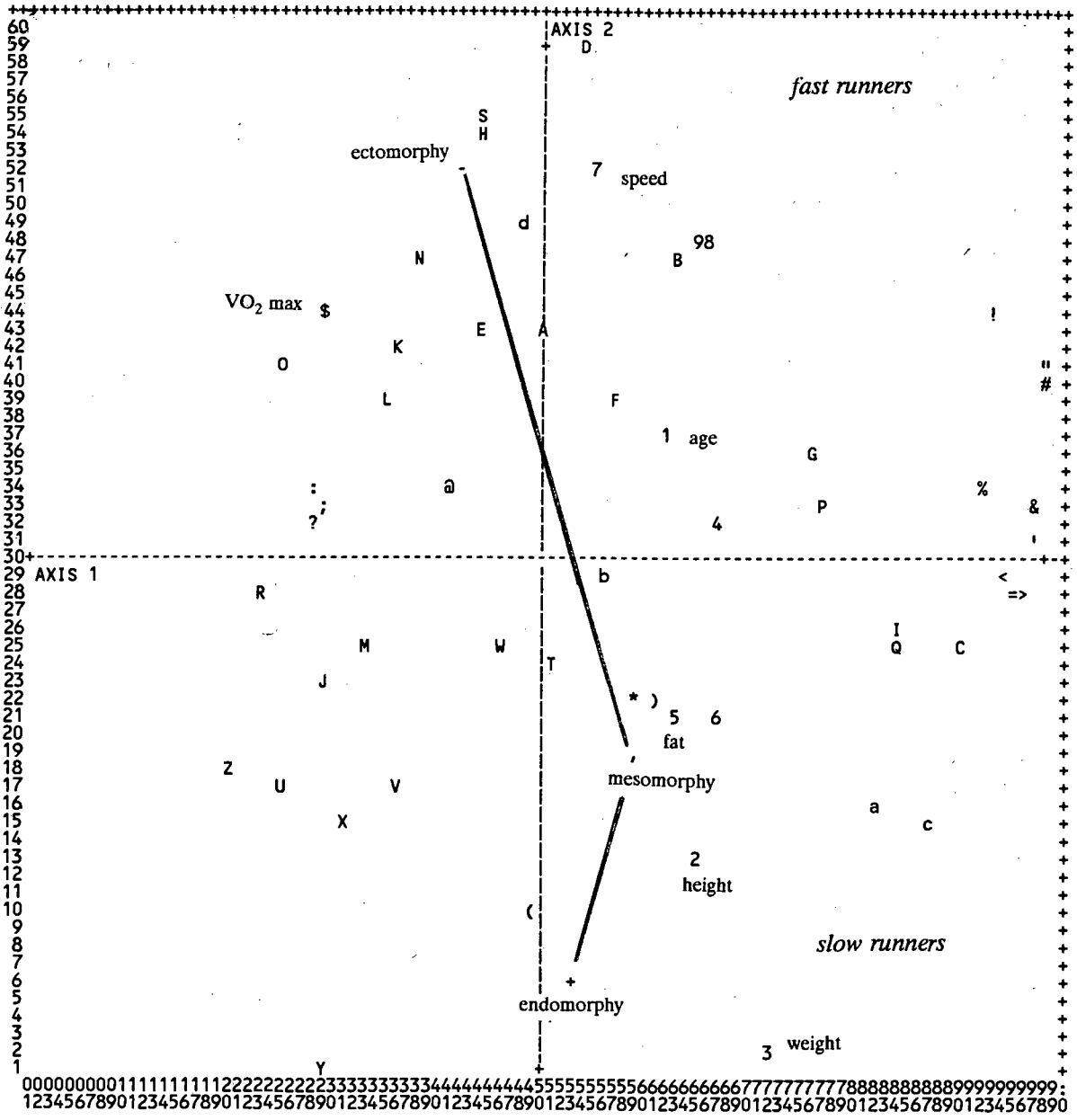

AXIS 2
AXIS 1
fast runners
slow runners
ectomorphy
speed
VO_2 max
age
fat
mesomorphy
height
endomorphy
weight

FIGURE 8.2.4 (cont)

NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	R1	A	0.0016	0.1875	43	50
2	R3	C	0.3494	-0.0628	25	90
3	R4	D	0.0359	0.4260	59	54
4	R5	E	-0.0468	0.1914	43	44
5	R6	F	0.0617	0.1316	39	57
6	R7	G	0.2316	0.0946	36	76
7	R8	H	-0.0458	0.3543	54	44
8	R9	I	0.3026	-0.0547	26	84
9	R10	J	-0.1807	-0.1060	23	29
10	R11	K	-0.1199	0.1766	42	36
11	R12	L	-0.1331	0.1272	39	35
12	R13	M	-0.1430	-0.0632	25	33
13	R14	N	-0.0992	0.2501	47	38
14	R15	O	-0.2173	0.1604	41	25
15	R16	P	0.2358	0.0515	33	77
16	R17	Q	0.3018	-0.0646	25	84
17	R18	R	-0.2384	-0.0200	28	23
18	R20	T	0.0106	-0.0884	24	51
19	R21	U	-0.2193	-0.1932	17	25
20	R22	V	-0.1165	-0.1913	17	36
21	R23	W	-0.0326	-0.0710	25	46
22	R24	X	-0.1598	-0.2105	15	31
23	R25	Y	-0.1832	-0.4218	1	29
24	R26	Z	-0.2568	-0.1720	18	20
25	R27	a	0.2854	-0.2050	16	82
26	R28	b	0.0510	-0.0047	29	56
27	R29	c	0.3250	-0.2220	15	87
28	R2	B	0.1138	0.2440	47	63
29	R19	S	-0.0460	0.3650	55	44
30	R30	d	-0.0122	0.2773	49	48
1	AGE	1	0.1185	0.1200	37	62
2	HEIGHT	2	0.1491	-0.2629	13	65
3	WEIGHT	3	0.2118	-0.4436	2	72
4	TEMP	4	0.1662	0.0424	32	67
5	FAT	5	0.1266	-0.1385	21	63
6	AREA	6	0.1701	-0.1358	21	67
7	MET42	<	0.4225	-0.0100	29	94
8	MET21	=	0.4404	-0.0221	28	95
9	MET6	>	0.4445	-0.0356	28	96
10	MET42-6	@	-0.0853	0.0631	34	41
11	O2MAX42	!	0.4160	0.2353	44	93
12	O2MAX21	"	0.4628	0.1830	41	98
13	O2MAX6	#	0.4678	0.1659	40	98
14	O2MAXREST	$	-0.1966	0.2229	44	29
15	O2MAX42-R	%	0.4084	0.0679	34	92
16	O2MAX21-R	&	0.4570	0.0444	33	97
17	FLUID	(	-0.0101	-0.3197	10	49
18	SWEAT	)	0.1067	-0.1269	22	61
19	DEHYD	*	0.0865	-0.1307	22	59
20	ENDO	+	0.0334	-0.3795	6	53
21	MESO	,	0.0923	-0.1652	19	59
22	ECTO	-	-0.0756	0.3590	52	42
23	SPEED42	7	0.0522	0.3502	52	55
24	SPEED21	8	0.1548	0.2877	48	66
25	SPEED6	9	0.1466	0.2992	48	65
26	SPEED42-21	:	-0.2072	0.0715	34	28

8.2.5 Comparison of the Plots

The choice of which member of the correlation biplot to use is discussed in Section 4.6. Here we have a practical illustration of the choice of preprocessing.

The covariance biplot does not provide useful information about the variables that are measured on a small scale. All that can be really seen from the plot (Figure 8.2.1) is that they have small variances. Because they are all bunched up together, their positions relative to other variable and row points is not clear. The variables with small variances are in general poorly represented in two dimensions. As the approximation aims at maximising the total variance, the columns with greater variance are approximated at the expense of those with smaller variances, i.e the variables are given weight proportional to their variances. Variables with small variances are, however, displayed well when they are highly correlated (positively or negatively) with well approximated variables.

In the covariance biplot, domination of the plot by a few variables occurs when the variables are measured on the same scale, but there are large differences in the variances. The problem is exacerbated when the scales are different. The covariance biplot is often unsuitable when the scales of measurement vary a lot, unless some sort of scaling is done prior to input into the SVD. Variables measured on larger scales have larger variances. Outlying observations can also inflate the variance, resulting in similar problems of dominance.

The coefficient of variation biplot of the data (Figure 8.2.3) has disadvantages analogous to those of the covariance biplot. The variables that are displayed the best tend to be those with the largest coefficients of variation or those correlated to variables with large coefficients of variation. The other variables are bunched up at the origin.

The correlation biplot (Figure 8.2.2) is a display of the variables after standardisation. The correlation biplot attempts to display all the variables equitably. All the columns are given equal weights. As discussed in Section 4.6, relative variabilities are not displayed here. The overall quality of display is reduced but the axes are composed of more variables.

The same problem occurs in the coefficient of variation biplot if only a few variables have large coefficients of variation.

As noted in Section 8.2.2, the correlation biplot provides the best display of the correlation structure.

EXAMPLE 8.3
QUALITY OF LIFE IN THE UNITED STATES

Barr, Underhill & Kahn (1990) describe the application of the covariance biplot to multivariate time series data. Besides the usual features approximated by biplots, biplots on such data display changes in the variables over time and the changing relationships between the variables over time.

A similar set of multivariate time series data is taken from an article entitled "Toward A Comprehensive Quality of Life Index" (Johnson, 1988). The overall quality of life (QOL) of a community, population group or larger society is composed of many factors. Socioeconomic indicators considered to be fairly representative were used. The variables were observed annually in the United States over the time period 1969 to 1986 (Table 8.3). The indicators were taken from nine different 'areas of social concern'; each of these areas is represented by at least two variables. The article deals with the problem of constructing a QOL index from these measures.

The differences in the scales and magnitudes of measurement of the variables leads to the recommendation that a biplot with column centring only, such as the covariance biplot, is not suitable. This is illustrated in Example 8.3. The median income variable (variable 11) has much larger values than the other variables and for reasons discussed earlier (in Section 4.6) would dominate such a plot to the near exclusion of the other variables. A suitable scaling of the variables, so that their magnitudes do not differ too much, is required. The correlation biplot, in which the variables are first standardized, was found to provide a useful display of the data.

TABLE 8.3
Selected U.S. socioeconomic indicators: 1969 to 1985

Area of concern and indicator	Year 1969 1978	1970 1979	1971 1980	1972 1981	1973 1982	1974 1983	1975 1984	1976 1985	1977 1986
A. *Health*									
(1) Life expectancy at birth (HEALTH1)	70.4 73.5	70.9 73.9	71.1 73.7	71.2 74.1	71.4 74.4	72.0 74.5	72.6 74.5	72.9 74.8	73.3
(2) Infant mortality rate (HEALTH2)	20.9 13.8	20.0 13.1	19.1 12.6	18.5 11.9	17.7 11.5	16.7 11.2	16.1 10.6	15.2 10.0	14.1
(3) Days of disability (HEALTH3)	8.5[e] 9.8	8.5 9.5	8.6[e] 9.8	9.3 9.9	9.2 10.1[e]	9.3 10.2[e]	9.7 10.3[e]	9.4 10.4[e]	9.4
B. *Public safety*									
(4) Rate of violent crimes (PS4)	329 490	364 540	396 587	401 594	417 571	461 538	482 539	460 556	468
(5) Rate of property crimes (PS5)	335 463	362 501	377 534	356 526	374 503	439 464	480 449	481 465	459
C. *Education*									
(6) % of pop. 25+ with coll. 4+ (ED6)	10.7 15.7	11.1 16.4	11.4 17.0	12.0 17.1	12.6 17.7	13.3 18.8	13.9 19.1	14.8 19.4	15.4
(7) Average SAT scores (ED7)	474 448	471 447	468 445	463 445	462 446	453 446	451 448	452 453	450
D. *Employment*									
(8) Unemployment rate (EMPLOY8)	3.4 6.0	4.8 5.8	5.8 7.0	5.5 7.5	4.8 9.5	5.5 9.5	8.3 7.4	7.6 7.1	6.9

Area of concern and indicator	Year 1969 1978	1970 1979	1971 1980	1972 1981	1973 1982	1974 1983	1975 1984	1976 1985	1977 1986
(9) % unempl. less than 15 weeks (EMPLOY9)	86.7 77.2	83.8 79.8	76.3 75.4	76.1 72.4	81.2 67.4	81.5 60.7	68.3 68.2	67.9 72.3	72.1
(10) % unempl. not job losers (EMPLOY10)	64.1 58.3	55.8 57.1	53.7 48.3	43.2 49.4	61.2 41.3	56.5 41.6	44.7 48.4	50.3 50.2	54.7
E. *Earnings & income*									
(11) Median family income (INCOME11)	25 632 26 938	25 317 26 885	25 301 25 418	26 473 24 525	27 017 24 187	26 066 24 580	25 395 25 072	26 179 26 780	26 320
(12) Average weekly earnings (INCOME12)	189.45 189.24	187.05 184.06	190.33 173.27	198.50 169.96	198.46 167.84	190.35 171.26	184.30 172.78	186.85 170.42	189.00
F. *Poverty*									
(13) % of pop. in poverty (POV13)	12.1 11.4	12.6 11.7	12.5 13.0	11.9 14.0	11.1 15.0	11.2 15.2	12.3 14.4	11.8 14.0	11.6
(14) % of children in poverty (POV14)	13.8 15.7	14.9 16.0	15.1 17.9	14.9 19.5	14.2 21.3	15.1 22.2	16.8 21.3	15.8 20.5	16.0
G. *Housing*									
(15) New "POHUs" (HOUSE15)	1467 2020	1434 1745	2052 1292	2357 1084	2045 1062	1338 1703	1160 1750	1538 1742	1987
(16) Average sales price (1977 $) (HOUSE16)	51.9 54.6	48.1 54.9	48.5 52.6	49.1 52.7	52.6 52.0	52.7 54.3	52.1 54.2	54.1 55.0[e]	54.2

(*Table 8.3 cont*)

Area of concern and indicator	Year 1969 1978	1970 1979	1971 1980	1972 1981	1973 1982	1974 1983	1975 1984	1976 1985	1977 1986
H. *Family stability*									
(17) Rate of divorce (FAM17)	3.2 5.1	3.5 5.3	3.7 5.2	4.0 5.3	4.3 5.0	4.6 4.9	4.8 4.9	5.0 5.0	5.0
(18) % of families "intact" (FAM18)	86.8 82.8	86.8 82.5	86.1 82.5	85.8 81.7	85.2 81.3	85.0 81.3	84.3 80.8	84.1 80.3	83.8
I. *Equality*									
(19) Bl:Wh. ratio, life expectancy (EQUAL19)	0.886 0.919	0.894 0.918	0.897 0.915	0.899 0.920	0.900 0.923	0.906 0.925	0.910 0.926	0.913 0.927[e]	0.915
(20) Bl:Wh. ratio, Coll. (EQUAL20)	0.41 0.44	0.39 0.46	0.38 0.44	0.40 0.46	0.46 0.48	0.39 0.49	0.44 0.52	0.43 0.56	0.45
(21) Bl:Wh. ratio, med. family income (EQUAL21)	0.619 0.592	0.614 0.566	0.603 0.579	0.594 0.564	0.577 0.553	0.597 0.563	0.615 0.570	0.595 0.576	0.571

Sources and Definitions

"e" = estimated by the author.

(1) *Life expectancy at birth* (*both sexes combined*) — National Center for Health Statistics, *Monthly Vital Statistics Reports* and annual issues of the *Statistical Abstract of the United States.*

(2) *Infant mortality Rate — Ibid.*

(3) *Days of disability* — (per person per year) — National Center for Health Statistics, *Health United States 1983*, Table 29. Data are annual only, covering the years 1970 and 1972 through 1981, age-adjusted. Data for other years are estimated by the author.

(4) *Rate of violent crime* — Expressed per 100 000 population. Federal Bureau of Investigation (FBI), *Crime in the United States. 1983*, summary table (for years 1974 through 1983). Data for other years from U.S. Department of Commerce, *Social Indicators III*, Table 5/6; data for 1985 from release dated 27Jul86.

(5) *Rate of property crime* — Expressed per 10 000 population. *Ibid.*

(6) *% of pop. 25+ with Coll. 4+* — The percentage of the population aged 25 years and over who have completed 4 or more years of college education. Bureau of the Census, *Current Population reports*, Series P—20.

(7) *Average SAT scores* — An unweighted average of the "verbal" and "mathematical" components of the standardized Scholastic Aptitude Test given to high school graduates who wish to enter a college or university. Data are annual, covering the academic year from September to June, and provided by the College Entrance Examination Board, New York.

(8) *Unemployment rate* — Defined as the percentage of the civilian labor force (both sexes combined, aged 16 and over) classified as unemployed. Bureau of Labor Statistics, *Employment and Earnings* (monthly). Data are annual averages of twelve monthly estimates.

(9) *% unemployed less than 15 weeks* — Data also from issues of *Employment and Earnings*, Table A-32, and are also annual averages of monthly estimates. This measure relates to the percentage of the unemployed who have been without work for less than 15 weeks.

(10) *% unemployed not job losers* — Data also from *Employment and Earnings*, Table A-40 or from the Bureau of Labor Statistics, *Handbook of Labor Statistics*, Bulletin 2175 (Dec. 1983). This measure is the percentage of the unemployed who did not lose their last job involuntarily.

(11) *Median family income* — Expressed in constant 1984 dollars. Bureau of the Census, *Current Population Reports*, Series P—60 and annual issues of the *Statistical Abstract of the United States.*

(12) *Average weekly earnings* — Data shown have been converted to constant 1977 dollars; they relate to average weekly earnings of production or nonsupervisory workers on private nonagricultural payrolls. BLS, *Handbook of Labor Statistics* (*op. cit.*), Table 89.

(13) *% of pop. in poverty* — Bureau of the Census, *Current Population Reports*, Series P—60. for annual estimates.

(14) *% of children in poverty* — Data relate to related children under 18 years old living in families classified as below the poverty threshold. Bureau of the Census. *Ibid.*

(15) *New "POHUs" started* — Data relate to the number of new "privately-owned housing units" whose construction has begun. *Statistical Abstract of the United States* and Bureau of the Census, Construction Reports, *Housing Starts.*

(16) *Average sales price* — Data relate to average sales price of new one-family houses sold, expressed in constant 1977 dollars. *Statistical Abstract of the United States* and Bureau of the Census/Department of Housing and Urban Development, *Construction Reports*, Series C—25 and C—27.

(17) *Rate of divorce* — Based on 1000 population. *Statistical Abstract of the United States* and National Center for Health Statistics, *Monthly Vital Statistics Report.*

(18) *% of families "intact"* — Data relate to husband-wife families as a percentage of all family units. Bureau of the Census, *Current Population Reports*, Series P—20.

(19) *Bl:Wh. ratio, life expectancy* — The ratio of average life expectancy at birth for the Black population (both sexes) to that of the white population. Source is the same as for Indicator # 1.

(20) *Bl:Wh. ratio, Coll. 4+* — The ratio of the percentage of Blacks 25 and over (both sexes) who have completed 4 or more years of college education to that of whites 25 and over. Source is the same as for Indicator # 6.

(21) *Bl:Wh. ratio, med. family income* — The ratio of the median income of all Black families to that of all white families. Source is the same as for Indicator # 11.

8.3.1. Correlation Biplot

Although the original matrix has a rank of 16, the quality of the display in the first dimension is 69% (Table 8.3.1). The quality of the biplot in two dimensions is 82%. Thus the quality of the two dimensional display is good.

The 'ctr' columns show that many of the variables are well represented in two dimensions. This is indicated on the plot by the proximity of the variable points to the unit circle. There are no variables that dominate the plot to the exclusion of the others. Variable 2, infant mortality rate, is furthest from the origin, with a display quality of 99%. Variable 15, number of new houses, at a 31% quality, is displayed the least well.

Many of the variable points are positioned on the far left hand side of the plot. The angles through the origin for these vectors are very small, showing that they are highly correlated. Because the row points are in fact points in time, it means that these variables moved closely together over the years 1969 to 1985. The variables include life expectancy at birth, rate of violent crimes, rate of property crimes and black/white life expectancy ratio. Similarly, the obtuse angle between the variables life expectancy at birth and infant mortality rate indicate that they are negatively correlated. The angle through the origin between the vectors for median family income and divorce rate indicates that these variables are uncorrelated i.e. they did not move together over the time period.

A line connecting the time points shows their change with respect to the socioeconomic indicators over the period. The time points are almost sequentially ordered by the first axis.

Because the rows are points in time and the columns are variables, the biplot interpretation allows the display of changes in the variables over time and of changes in the relationships between the variables over time.

The earlier years are depicted as having a strong association with high: infant mortality rates, SAT scores, % unemployed less than 15 weeks, % of families intact, ratio of median income of black to white families and average weekly earnings.

The row points representing the latter time period are plotted in positions corresponding to relatively high values of many of the variables, including: life expectancy, average sales price of houses, rates of violent crimes and % of population with college education.

To look at the change in a variable over time, consider the projections of the time points onto the line joining a variable point with the origin. For example, consider the variable 'life expectancy at birth'. The projections display that the values for this variable were above average in 1985, about average in 1975 and below average in 1969. In fact, the projections display an approximate ordering of the variable over time. Table 8.3 confirms that this variable tends to increase over time. Variable 11, median family income, is displayed as not having a general trend over time. This is confirmed by Table 8.3.

The above can be generalised as follows: variables whose vectors lie in the same general direction as the ordering of the time points (the first axis) have an increasing or decreasing trend over time; variables perpendicular to this do not have this property. (Note that this is not generalisable to plots in which the ordering of the time points is not close to linear.)

Relative changes over time are also displayed. For example, over the years 1975 to 1978, the standardized values for median family income increased a lot, whereas the standardized infant mortality rates did not change as much.

TABLE 8.3.1

THE EIGENVALUES

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	232.04367	69.061	69.061
2	44.93157	13.372	82.433
3	21.78064	6.482	88.915
4	17.52040	5.214	94.130
5	8.06428	2.400	96.530
6	4.90331	1.459	97.989
7	2.75169	0.819	98.808
8	1.71953	0.512	99.320
9	0.89121	0.265	99.585
10	0.67724	0.202	99.787
11	0.27937	0.083	99.870
12	0.16064	0.048	99.918
13	0.13382	0.040	99.958
14	0.07607	0.023	99.980
15	0.04937	0.015	99.995
16	0.01715	0.005	100.000

THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	FACT 2	COR
1	1969	A	1.000	0.199	0.059	0.4314	0.186	0.1137	0.013
2	1970	B	1.000	0.232	0.059	0.3569	0.127	0.3230	0.104
3	1971	C	1.000	0.149	0.059	0.2968	0.088	0.2471	0.061
4	1972	D	1.000	0.068	0.059	0.2557	0.065	0.0528	0.003
5	1973	E	1.000	0.149	0.059	0.2350	0.055	-0.3067	0.094
6	1974	F	1.000	0.031	0.059	0.1467	0.022	-0.0982	0.010
7	1975	G	1.000	0.037	0.059	-0.0146	0.000	0.1913	0.037
8	1976	H	1.000	0.007	0.059	-0.0057	0.000	-0.0831	0.007
9	1977	I	1.000	0.067	0.059	-0.0133	0.000	-0.2585	0.067
10	1978	J	1.000	0.169	0.059	-0.0066	0.000	-0.4108	0.169
11	1979	K	1.000	0.174	0.059	-0.0731	0.005	-0.4103	0.168
12	1980	L	1.000	0.027	0.059	-0.1609	0.026	0.0295	0.001
13	1981	M	1.000	0.086	0.059	-0.2443	0.060	0.1617	0.026
14	1982	N	1.000	0.243	0.059	-0.3226	0.104	0.3728	0.139
15	1983	O	1.000	0.168	0.059	-0.3296	0.109	0.2433	0.059
16	1984	P	1.000	0.077	0.059	-0.2761	0.076	0.0354	0.001
17	1985	Q	1.000	0.117	0.059	-0.2756	0.076	-0.2032	0.041

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	HEALTH1	1	1.000	0.984	0.048	-0.3919	0.960	0.066	-0.0614	0.024	0.008
2	HEALTH2	2	1.000	0.986	0.048	0.3918	0.960	0.066	0.0648	0.026	0.009
3	HEALTH3	3	1.000	0.907	0.048	-0.3777	0.892	0.061	-0.0497	0.015	0.005
4	PS4	4	1.000	0.902	0.048	-0.3784	0.895	0.062	-0.0319	0.006	0.002
5	PS5	5	1.000	0.723	0.048	-0.3364	0.707	0.049	-0.0495	0.015	0.005
6	ED6	6	1.000	0.965	0.048	-0.3897	0.949	0.065	-0.0508	0.016	0.006
7	ED7	7	1.000	0.853	0.048	0.3582	0.802	0.055	0.0902	0.051	0.018
8	EMPLOY8	8	1.000	0.829	0.048	-0.3418	0.730	0.050	0.1256	0.099	0.035
9	EMPLOY9	9	1.000	0.683	0.048	0.3134	0.614	0.042	-0.1055	0.070	0.025
10	EMPLOY10	:	1.000	0.664	0.048	0.2500	0.391	0.027	-0.2091	0.273	0.097
11	INCOME11	;	1.000	0.893	0.048	0.1381	0.119	0.008	-0.3518	0.773	0.275
12	INCOME12	<	1.000	0.857	0.048	0.3429	0.735	0.051	-0.1397	0.122	0.043
13	POV13	=	1.000	0.854	0.048	-0.2846	0.506	0.035	0.2358	0.347	0.124
14	POV14	>	1.000	0.925	0.048	-0.3642	0.829	0.057	0.1239	0.096	0.034
15	HOUSE15	?	1.000	0.338	0.048	0.1314	0.108	0.007	-0.1919	0.230	0.082
16	HOUSE16	@	1.000	0.810	0.048	-0.2622	0.430	0.030	-0.2466	0.380	0.135
17	FAM17	!	1.000	0.874	0.048	-0.3423	0.732	0.050	-0.1505	0.141	0.050
18	FAM18	"	1.000	0.964	0.048	0.3891	0.946	0.065	0.0533	0.018	0.006
19	EQUAL19	#	1.000	0.972	0.048	-0.3870	0.936	0.065	-0.0758	0.036	0.013
20	EQUAL20	$	1.000	0.677	0.048	-0.3189	0.636	0.044	-0.0814	0.041	0.015
21	EQUAL21	%	1.000	0.654	0.048	0.3167	0.627	0.043	0.0660	0.027	0.010

FIGURE 8.3.1

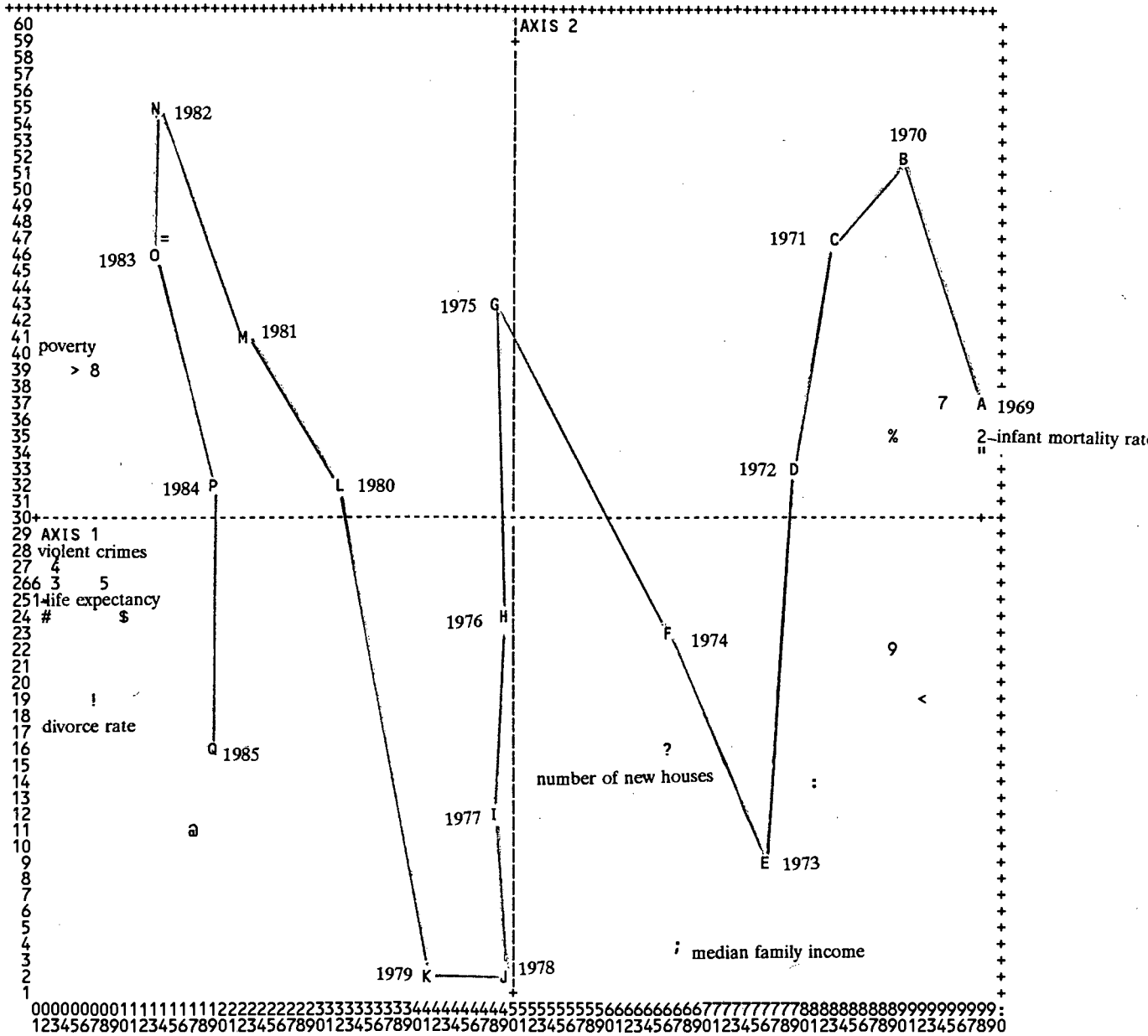

AXIS 2
AXIS 1
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
poverty
> 8
violent crimes
life expectancy
divorce rate
number of new houses
median family income
infant mortality rate

FIGURE 8.3.1 (cont)

NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	1969	A	0.4314	0.1137	37	98
2	1970	B	0.3569	0.3230	52	90
3	1971	C	0.2968	0.2471	47	83
4	1972	D	0.2557	0.0528	33	79
5	1973	E	0.2350	-0.3067	9	76
6	1974	F	0.1467	-0.0982	23	66
7	1975	G	-0.0146	0.1913	43	48
8	1976	H	-0.0057	-0.0831	24	49
9	1977	I	-0.0133	-0.2585	12	48
10	1978	J	-0.0066	-0.4108	2	49
11	1979	K	-0.0731	-0.4103	2	41
12	1980	L	-0.1609	0.0295	32	32
13	1981	M	-0.2443	0.1617	41	22
14	1982	N	-0.3226	0.3728	55	13
15	1983	O	-0.3296	0.2433	46	13
16	1984	P	-0.2761	0.0354	32	19
17	1985	Q	-0.2756	-0.2032	16	19
1	HEALTH1	1	-0.3919	-0.0614	25	1
2	HEALTH2	2	0.3918	0.0648	35	98
3	HEALTH3	3	-0.3777	-0.0497	26	3
4	PS4	4	-0.3784	-0.0319	27	3
5	PS5	5	-0.3364	-0.0495	26	8
6	ED6	6	-0.3897	-0.0508	26	1
7	ED7	7	0.3582	0.0902	37	94
8	EMPLOY8	8	-0.3418	0.1256	39	7
9	EMPLOY9	9	0.3134	-0.1055	22	89
10	EMPLOY10	:	0.2500	-0.2091	14	81
11	INCOME11	;	0.1381	-0.3518	4	67
12	INCOME12	<	0.3429	-0.1397	19	92
13	POV13	=	-0.2846	0.2358	47	14
14	POV14	>	-0.3642	0.1239	39	5
15	HOUSE15	?	0.1314	-0.1919	16	66
16	HOUSE16	@	-0.2622	-0.2466	11	17
17	FAM17	!	-0.3423	-0.1505	19	7
18	FAM18	"	0.3891	0.0533	34	98
19	EQUAL19	#	-0.3870	-0.0758	24	2
20	EQUAL20	$	-0.3189	-0.0814	24	10
21	EQUAL21	%	0.3167	0.0660	35	89

Plot 8.3.2. Doubling

A high QOL is indicated sometimes with a high value of an indicator and sometimes by a low one. For example, a high quality of life is indicated by a high life expectancy, but a low rate of infant mortality. The position of the point representing life expectancy does allow, by means of the biplot interpretation, the interpretation that the earlier years are associated with less than average life expectancies. However this relationship can be emphasized by means of the display technique known as 'doubling' (Greenacre, 1984). Doubling is usually associated with correspondence analysis, but can equally well be used with other biplot displays.

In this technique both the positive and negative 'aspects' of the variable are displayed. The number of variables in the display is doubled. This is done by forming a 'new' variable from each old variable by subtracting from the maximum score for that variable.

The quality of display of the plot does not change when a correlation biplot is 'doubled'. The quality of display of each of the variables and its associated double is the same as the quality for that variable in the ordinary correlation biplot (Table 8.3.1). The doubling operation does not change the rank of the matrix that is biplotted. Each new variable can be thought of as lying in the same dimension as its double. This is because the new variable is perfectly (negatively) correlated to, and a linear combination of, the old one. On the plot this is shown by the 180^0 angle between a vector and its double. Because the norm of the variable points in a correlation biplot represent their quality of approximation, the points for a variable point and its double are equidistant from the origin. Therefore, if the doubled variables are plotted as supplementary points instead, the same plot is obtained.

Figure 8.3.2 shows the correlation biplot of the doubled data matrix. Variables for which high values are considered to be a positive contribution to the high quality of life are marked with a '+', their 'double' with a '-', and vice versa.

The plot reveals the change in composition of factors associated with quality of life over time. Consider the rate of violent crimes, variable 4. The point representing a high rate of violent crimes, denoted '4-', is situated near the points for the later years. A low rate

of violent crimes, denoted '4+' is associated with the earlier time periods. By projecting the time points onto these vectors, the biplot interpretation indicates that the rate of violent crimes is increasing over time. Conversely, '1+', representing a high life expectancy at birth, is associated with a good quality of life and is also shown to be increasing over the time period.

FIGURE 8.3.2

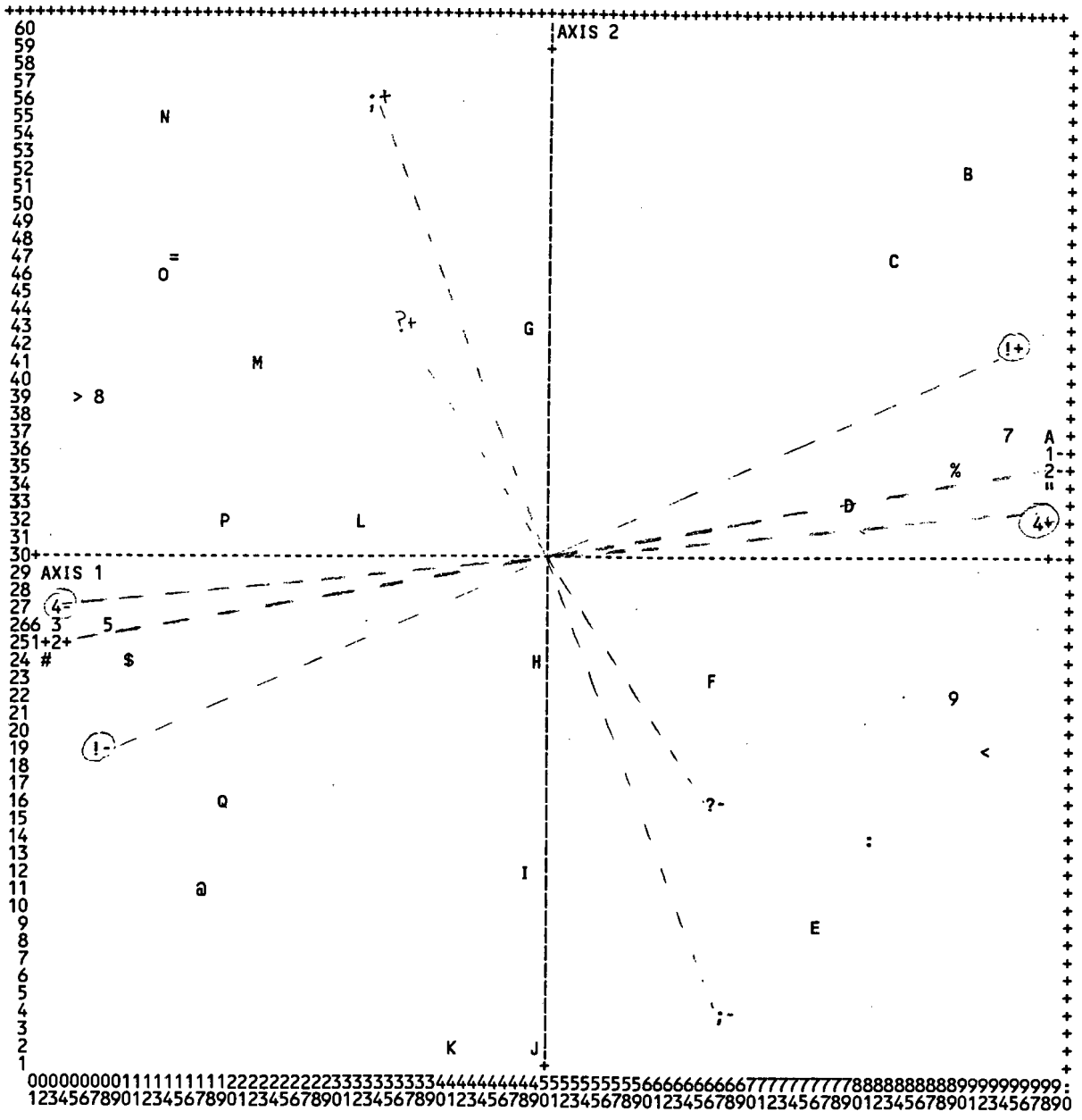

AXIS 2
AXIS 1
N
O
M
P
L
G
H
K
J
I
Q
F
E
B
C
A
7
9
D
%
<
:
;+
?+
?-
;-
!+
!-
4+
4-
> 8
#
$
@
3
5
2+
1-
2-
0000000001111111111222222222233333333334444444444555555555566666666667777777777888888888899999999990:
1234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567890

FIGURE 8.3.2 (cont)

NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	1969	A	0.4314	0.1137	37	98
2	1970	B	0.3569	0.3230	52	90
3	1971	C	0.2968	0.2471	47	83
4	1972	D	0.2557	0.0528	33	79
5	1973	E	0.2350	-0.3067	9	76
6	1974	F	0.1467	-0.0982	23	66
7	1975	G	-0.0146	0.1913	43	48
8	1976	H	-0.0057	-0.0831	24	49
9	1977	I	-0.0133	-0.2585	12	48
10	1978	J	-0.0066	-0.4108	2	49
11	1979	K	-0.0731	-0.4103	2	41
12	1980	L	-0.1609	0.0295	32	32
13	1981	M	-0.2443	0.1617	41	22
14	1982	N	-0.3226	0.3728	55	13
15	1983	O	-0.3296	0.2433	46	13
16	1984	P	-0.2761	0.0354	32	19
17	1985	Q	-0.2756	-0.2032	16	19
1	HEALTH1 +	1+	-0.3919	-0.0614	25	1
2	HEALTH2 -	2-	0.3918	0.0648	35	98
3	HEALTH3 -	3-	-0.3777	-0.0497	26	3
4	PS4 -	4-	-0.3784	-0.0319	27	3
5	PS5 -	5-	-0.3364	-0.0495	26	8
6	ED6 +	6+	-0.3897	-0.0508	26	1
7	ED7 +	7+	0.3582	0.0902	37	94
8	EMPLOY8 -	8-	-0.3418	0.1256	39	7
9	EMPLOY9 -	9-	0.3134	-0.1055	22	89
10	EMPLOY10-	:-	0.2500	-0.2091	14	81
11	INCOME11+	;+	0.1381	-0.3518	4	67
12	INCOME12+	<+	0.3429	-0.1397	19	92
13	POV13 -	=-	-0.2846	0.2358	47	14
14	POV14 -	>-	-0.3642	0.1239	39	5
15	HOUSE15 -	?-	0.1314	-0.1919	16	66
16	HOUSE16 -	@-	-0.2622	-0.2466	11	17
17	FAM17 -	!-	-0.3423	-0.1505	19	7
18	FAM18 +	"+	0.3891	0.0533	34	98
19	EQUAL19 +	#+	-0.3870	-0.0758	24	2
20	EQUAL20 +	$+	-0.3189	-0.0814	24	10
21	EQUAL21 +	%+	0.3167	0.0660	35	89
22	HEALTH1 -	1-	0.3919	0.0614	35	99
23	HEALTH2 +	2+	-0.3918	-0.0648	25	2
24	HEALTH3 +	3+	0.3777	0.0497	24	97
25	PS4 +	4+	0.3784	0.0319	33	97
26	PS5 +	5+	0.3364	0.0495	34	92
27	ED6 -	6-	0.3897	0.0508	34	99
28	ED7 -	7-	-0.3582	-0.0902	23	6
29	EMPLOY8 +	8+	0.3418	-0.1256	21	93
30	EMPLOY9 +	9+	-0.3134	0.1055	38	11
31	EMPLOY10+	:+	-0.2500	0.2091	46	19
32	INCOME11-	;-	-0.1381	0.3518	56	33
33	INCOME12-	<-	-0.3429	0.1397	41	8
34	POV13 +	=+	0.2846	-0.2358	13	86
35	POV14 +	>+	0.3642	-0.1239	21	95
36	HOUSE15 +	?+	-0.1314	0.1919	44	34
37	HOUSE16 +	@+	0.2622	0.2466	49	83
38	FAM17 +	!+	0.3423	0.1505	41	93
39	FAM18 -	"-	-0.3891	-0.0533	26	2
40	EQUAL19 -	#-	0.3870	0.0758	36	98
41	EQUAL20 -	$-	0.3189	0.0814	36	90
42	EQUAL21 -	%-	-0.3167	-0.0660	25	11

EXAMPLE 8.4
BIRD CONSERVATION

The article from which the data was taken is part of a collection of papers of the Wader Study Group workshop. The group is an international association of researchers on waders (shorebirds). An objective of the workshop was to examine approaches to conserving wetlands and the waders that depend on them.

The data (Table 8.4) consists of a matrix of the number of waders, counted in summer on the coasts and coastal wetlands of southern Africa (Summers et al, 1987). This data set is particularly awkward because of the large variations in numbers, both between species and between areas. However, this kind of data frequently arises in ecological contexts.

This data are first used to illustrate a plotting technique known as the Ter Braaks diversity biplot (Section A, Plots 8.4.1 to 8.4.3). Correspondence analysis is performed in Plots 8.4.4 and 8.4.5 (Section B). Spearman's rank correlation biplots (described in Section 4.5) directly address problems arising due to magnitude differences in the counts. These are Plots 8.4.6 and 8.4.7 in Section C.

A. TER BRAAK'S DIVERSITY BIPLOTS

These biplots are members of the principal components family. They were developed for application to sites by species matrices. Such matrices are commonly found in ecology. In these matrices, the entries are typically some measure of abundance such as counts, yields or biomass.

A key issue in ecology is that of the diversity of an ecological site. In order to conserve many species, it is important to conserve sites that are species diverse. Diversity measures are thus important in conservation.

Two well known measures of site diversity are called alpha diversity and beta diversity.

TABLE 8.4

Numbers of Waders counted in summer on the coasts and coastal wetlands of Southern Africa.
C=open coast, W=coastal wetlands.

	OYSTER CATCHER	WHITE FRONTED PLOVER	KITT-LITZ'S PLOVER	THREE BANDED PLOVER	GREY PLOVER	RINGED PLOVER	BAR TAILED GODWIT	WHIM-BREL	MARSH SAND-PIPER	GREEN-SHANK	COMMON SAND-PIPER	TURN-STONE	KNOT	SANDER-LING	LITTLE STINT	CURLEW SAND-PIPER	RUFF	AVO-CET	BLACK-WINGED STILT
NAMIBIA North-C	12	2027	0	0	2070	39	219	153	0	15	51	8336	2031	14941	19	3566	0	5	0
NAMIBIA North-W	99	2112	9	87	3481	470	2063	28	17	145	31	1515	1917	17321	3378	20164	177	1759	53
NAMIBIA South-C	197	160	0	4	126	17	1	32	0	2	9	477	1	548	13	273	0	0	0
NAMIBIA South-W	0	17	0	3	50	6	4	7	0	1	2	16	0	0	3	69	1	0	0
CAPE North-C	77	1948	0	19	310	1	1	64	0	22	81	2792	221	7422	10	4519	12	0	0
CAPE North-W	19	203	48	45	20	433	0	0	11	167	12	1	0	26	1790	2916	473	658	55
CAPE West-C	1023	2655	0	18	320	49	8	121	9	82	48	3411	14	9101	43	3230	587	10	5
CAPE West-W	87	745	1447	125	4330	789	228	529	289	904	34	1710	7869	2247	4558	40880	7166	1632	498
CAPE South-C	788	2174	0	19	224	178	1	423	0	195	162	2161	25	1784	3	1254	0	0	0
CAPE South-W	82	350	760	197	858	962	10	511	251	987	191	34	87	417	4496	15835	5327	1312	1020
CAPE East-C	474	930	0	10	316	161	0	90	0	39	48	1183	166	4626	65	127	4	0	0
CAPE East-W	77	249	160	136	999	645	15	851	101	723	266	495	83	1253	1864	4107	1939	623	527
TRANSKEI-C	22	144	0	4	1	1	0	10	0	2	9	125	5	411	0	3	0	0	0
NATAL-C	0	791	0	0	4	38	1	56	1	30	54	95	0	1726	0	0	0	0	0
NATAL-W	0	360	128	43	364	1628	63	287	328	641	850	83	67	48	6499	9094	5647	1333	582

Both these measures can be represented on a type of PCB known as the Ter Braak's diversity biplot (Ter Braak, 1983). The biplot has been called '... among the most powerful tools for species compositional data' (Ter Braak, 1983).

The alpha diversity (α) is also known as the Simpson Index (Simpson, 1949). It is a measure of the diversity of a particular site. Beta diversity (ß) is a measure of the dissimilarity between two sites. It highlights differences in species composition. In order to define the diversities, we consider the matrix obtained from Phase I:
Let P be a matrix with entries p_{ij} where p_{ij} is the proportion of individuals of species j at site i with respect to all the individuals at site i, i.e. each entry in the original matrix is divided by the total number of individuals for that site (the row total). The rows of P sum to 1.

The diversities are defined by:

$$\alpha = \| p_i \|^2 = \sum_{j=1}^{m} p_{ij}^2 \qquad (8.1)$$

which is the squared norm of site vector i,
and

$$ß = d^2(p_k, p_l) = \sum_{j=1}^{m} (p_{kj} - p_{lj})^2 \qquad (8.2)$$

which is the squared Euclidean distance between sites k and l.

A site is considered to have a high diversity if many species are represented there. The minimum value for α occurs when the site has equal proportions of each species. Therefore a low α diversity measure corresponds to a site with high diversity. The minimum value α is 1/m and the maximum value is 1. The maximum value occurs when there is only one species at the site.

Two sites with similar species compositions have a low ß diversity. Values for ß range from 0 to m.

Ter Braaks's diversity biplots are essentially principal components biplots with a preprocessing that converts the rows into proportions. They are (1-0)-plots. Thus the two interpretations valid here are:

1. Scalar products within the sites, and
2. Scalar products between the site and the species vectors.

As a=1, decomposition of the norm is valid for the sites, but not for the species.

Ter Braak (1983) described a noncentred and a species centred principal components biplot and their interpretations. These are applied to the wader data in Plots 8.4.1 and 8.4.2 respectively. In Plot 8.4.3, the species centred plot is applied to logged data.

Plot 8.4.1 Noncentred PCB of Proportion Data

This biplot is a (1-0)-plot of P, the matrix of proportions whose rows add to 1.

Scalar products within the rows are approximated. The row elements in this case are the proportions of species on the sites. Thus the squared norm of a row point approximates the α diversity of the corresponding site (from 8.1).

The plotted origin is the 'true' origin in that it can be thought of as representing a site that has no species. The α diversity is in fact the squared distance from the site point to the origin. Sites with a high α diversity are plotted close to the origin, and sites with a low α diversity far from it.

The Euclidean distance between two site vectors represents the ß diversity between those sites (from 8.2). Sites that have similar species compositions have a low ß diversity and therefore small distances between them.

Thus both the α and the ß diversities are approximated by this biplot.

Note that if we plot the site vectors in two dimensions, they fall in the area bounded by two circles centred at the origin with radii $1/\sqrt{m}$ and 1 respectively.

From the 'individual correlation' interpretation (Section 3.4.3) we have that sites with similar profiles subtend small angles at the origin.

There is also the between set (site-species) biplot interpretation. The scalar product between a row and column point approximates the proportion of the species at the site. The points representing the species are attracted in the direction of sites having high proportions of those species. The plot therefore displays which species are contributing to the diversity measure of a site, i.e. it displays the species composition of the sites.

The display is not useful when there are large differences in species abundance. Common

species are represented well, at the expense of less common ones. Some data transformation that eliminates these differences to some extent is necessary. This is discussed in more detail later in the chapter.

The problem that occurs when species abundances differ is illustrated by this particular data set rather well. The quality of the two dimensional display is 91% (Table 8.4.1), but two species dominate the plot. These two species - Sanderlings and Curlew Sandpipers - are by far the most numerous and therefore constitute large proportions of the site compositions. These species have a combined relative distance of 82% of all relative distances from the origin for the species. The remaining species are poorly represented.

Looking at the 'cor' column for the sites may lead to the wrong conclusion that many of the sites are well represented. The danger of interpreting these figures without regard to the quality of species display is apparent. The high quality display of some of the sites is due to their high proportion of the dominant species. For example, the points representing the Transkei and Natal coasts have high proportions of Sanderlings and most of their norms are retained; they are drawn away from the origin in the direction of the Sanderling point.

In effect the two dominant birds species constitute most of the plot. The position of the site points is determined by their proportions of Sanderlings and Curlew Sandpipers. Northern Namibia wetlands is positioned between these two bird points. The data set confirms the reason for this positioning; the site has substantial and approximately equal proportions of these two species. The points for the Natal and Transkei coasts are near the point for Sanderlings. Western Cape Wetlands, positioned closest to Curlew Sandpiper, has the highest proportion (54%) of these birds. Birds at the origin constitute small proportions of the site compositions.

Thus, although the sites appear to be well represented in that a large proportion of their norms are approximated, what is being displayed is almost completely due to two species. The other species are poorly displayed.

As this plot is not an adequate representation of the data set, it has not been interpreted further. Instead, it serves as an illustration of the need to look at indicators of 'goodness of display' other than usual one of the percentage of the squared norm approximated in two dimensions. Looking at decomposition of the squared norms for the columns does not reveal the flaws of the plot - the quality of representation of the rows and the columns must be considered in conjunction.

Figure 3.1

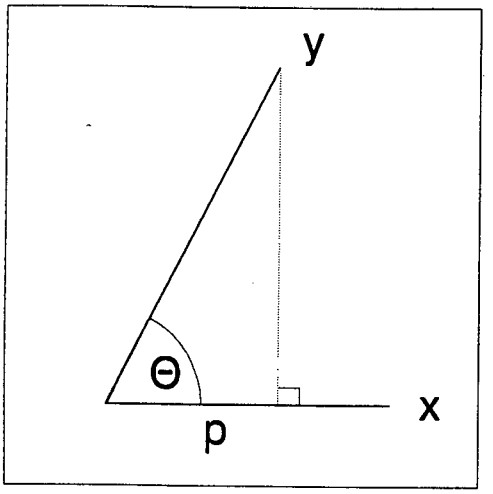


Figure 3.2

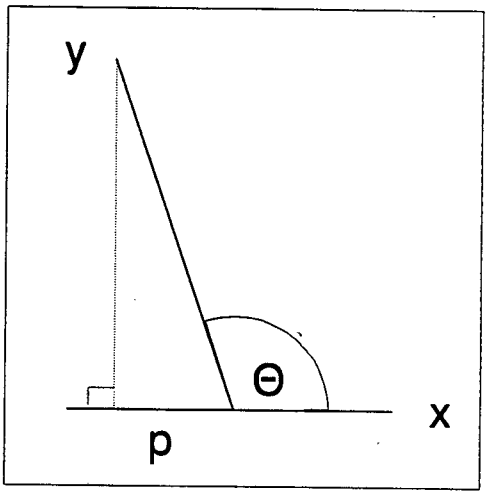


Figure 3.3

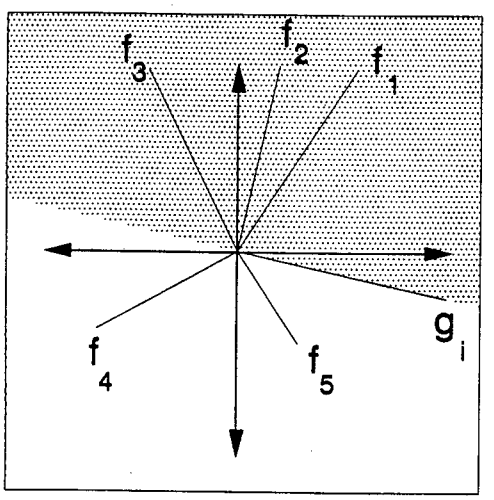


Figure 3.4

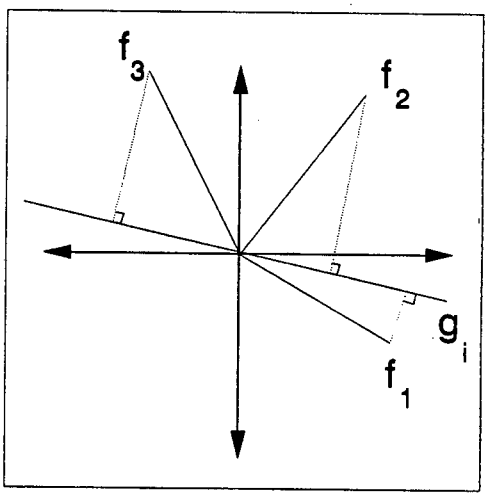


TABLE 8.4.1

THE EIGENVALUES

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	2.52220	61.721	61.721
2	1.20899	29.585	91.306
3	0.14793	3.620	94.926
4	0.07648	1.871	96.798
5	0.05948	1.456	98.253
6	0.03414	0.835	99.089
7	0.02054	0.503	99.592
8	0.00771	0.189	99.780
9	0.00478	0.117	99.897
10	0.00288	0.071	99.968
11	0.00085	0.021	99.988
12	0.00031	0.007	99.996
13	0.00009	0.002	99.998
14	0.00006	0.001	99.999
15	0.00002	0.001	100.000

THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	NAMIB-N-C	A	1.000	0.927	0.069	-0.4918	0.853	0.096	-0.1451	0.074	0.017
2	NAMIB-N-W	B	1.000	0.921	0.061	-0.4492	0.811	0.080	0.1656	0.110	0.023
3	NAMIB-S-C	C	1.000	0.848	0.048	-0.4066	0.836	0.066	-0.0491	0.012	0.002
4	NAMIB-S-W	D	1.000	0.696	0.061	-0.2603	0.274	0.027	0.3232	0.422	0.086
5	CAPE-N-C	E	1.000	0.974	0.070	-0.5265	0.973	0.110	-0.0192	0.001	0.000
6	CAPE-N-W	F	1.000	0.898	0.065	-0.2411	0.218	0.023	0.4262	0.680	0.150
7	CAPE-W-C	G	1.000	0.992	0.065	-0.5009	0.950	0.099	-0.1048	0.042	0.009
8	CAPE-W-W	H	1.000	0.940	0.078	-0.3034	0.290	0.036	0.4547	0.650	0.171
9	CAPE-S-C	I	1.000	0.700	0.042	-0.3450	0.695	0.047	-0.0303	0.005	0.001
10	CAPE-S-W	J	1.000	0.974	0.066	-0.2594	0.249	0.027	0.4420	0.724	0.162
11	CAPE-E-C	K	1.000	0.979	0.087	-0.5248	0.776	0.109	-0.2678	0.202	0.059
12	CAPE-E-W	L	1.000	0.913	0.032	-0.2373	0.437	0.022	0.2478	0.476	0.051
13	TRANSKEI-C	M	1.000	0.993	0.093	-0.5383	0.764	0.115	-0.2946	0.229	0.072
14	NATAL-C	N	1.000	0.914	0.113	-0.5663	0.692	0.127	-0.3205	0.222	0.085
15	NATAL-W	O	1.000	0.837	0.051	-0.1968	0.187	0.015	0.3675	0.651	0.112

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	FACT 2	COR
1	OYSTERCATCH	1	1.000	0.004	0.066	-0.0591	0.004	-0.0273	0.001
2	WF-PLOVER	2	1.000	0.082	0.066	-0.2590	0.067	-0.1212	0.015
3	K-PLOVER	3	1.000	0.001	0.042	-0.0067	0.000	0.0215	0.001
4	TB-PLOVER	4	1.000	0.000	0.059	-0.0068	0.000	0.0102	0.000
5	G-PLOVER	5	1.000	0.022	0.066	-0.0987	0.010	0.1110	0.012
6	R-PLOVER	6	1.000	0.006	0.057	-0.0360	0.002	0.0636	0.005
7	BT-GODWIT	7	1.000	0.000	0.066	-0.0112	0.000	0.0123	0.000
8	WHIMBREL	8	1.000	0.002	0.046	-0.0340	0.002	0.0194	0.001
9	M-SANDPIPER	9	1.000	0.004	0.002	-0.0031	0.000	0.0096	0.004
10	GREENSHANK	:	1.000	0.004	0.029	-0.0222	0.001	0.0368	0.003
11	C-SANDPIPER	;	1.000	0.001	0.039	-0.0191	0.001	0.0083	0.000
12	TURNSTONE	<	1.000	0.089	0.067	-0.2814	0.079	-0.1008	0.010
13	KNOT	=	1.000	0.003	0.066	-0.0403	0.002	0.0327	0.001
14	SANDERLING	>	1.000	0.821	0.067	-0.7800	0.609	-0.4612	0.213
15	LITTLE-STINT	?	1.000	0.091	0.059	-0.0917	0.009	0.2694	0.081
16	CURSANDPIPER	@	1.000	0.832	0.067	-0.4613	0.213	0.7870	0.619
17	RUFF	!	1.000	0.047	0.066	-0.0690	0.005	0.2044	0.042
18	AVOCET	"	1.000	0.014	0.037	-0.0292	0.002	0.0833	0.012
19	BW-STILT	#	1.000	0.002	0.033	-0.0097	0.000	0.0300	0.002

FIGURE 8.4.1

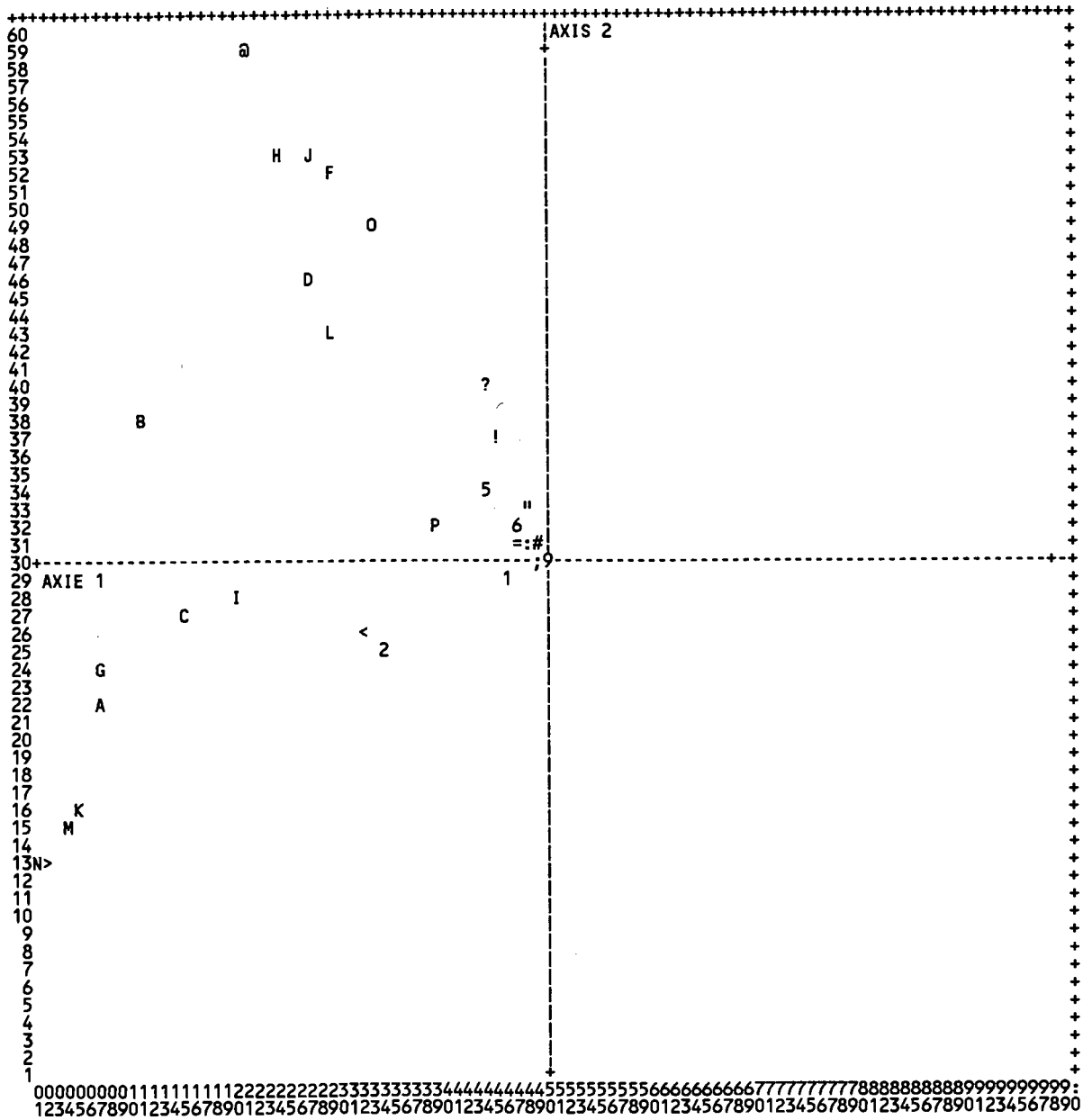


FIGURE 8.4.1 (cont)

NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	NAMIB-N-C	A	-0.4918	-0.1451	22	7
2	NAMIB-N-W	B	-0.4492	0.1656	38	11
3	NAMIB-S-C	C	-0.4066	-0.0491	27	15
4	NAMIB-S-W	D	-0.2603	0.3232	46	27
5	CAPE-N-C	E	-0.5265	-0.0192	29	5
6	CAPE-N-W	F	-0.2411	0.4262	52	29
7	CAPE-W-C	G	-0.5009	-0.1048	24	7
8	CAPE-W-W	H	-0.3034	0.4547	53	24
9	CAPE-S-C	I	-0.3450	-0.0303	28	20
10	CAPE-S-W	J	-0.2594	0.4420	53	27
11	CAPE-E-C	K	-0.5248	-0.2678	16	5
12	CAPE-E-W	L	-0.2373	0.2478	43	29
13	TRANSKEI-C	M	-0.5383	-0.2946	15	4
14	NATAL-C	N	-0.5663	-0.3205	13	1
15	NATAL-W	O	-0.1968	0.3675	49	33
16		P	-0.1220	0.0520	32	39
1	OYSTERCATCH	1	-0.0591	-0.0273	29	46
2	WF-PLOVER	2	-0.2590	-0.1212	25	34
3	K-PLOVER	3	-0.0067	0.0215	31	49
4	TB-PLOVER	4	-0.0068	0.0102	30	49
5	G-PLOVER	5	-0.0987	0.1110	34	44
6	R-PLOVER	6	-0.0360	0.0636	32	47
7	BT-GODWIT	7	-0.0112	0.0123	30	49
8	WHIMBREL	8	-0.0340	0.0194	31	48
9	M-SANDPIPER	9	-0.0031	0.0096	30	50
10	GREENSHANK	:	-0.0222	0.0368	31	48
11	C-SANDPIPER	;	-0.0191	0.0083	30	49
12	TURNSTONE	<	-0.2814	-0.1008	26	32
13	KNOT	=	-0.0403	0.0327	31	47
14	SANDERLING	>	-0.7800	-0.4612	13	2
15	LITTLE-STINT	?	-0.0917	0.2694	40	44
16	CURSANDPIPER	@	-0.4613	0.7870	59	21
17	RUFF	!	-0.0690	0.2044	37	45
18	AVOCET	"	-0.0292	0.0833	33	48
19	BW-STILT	#	-0.0097	0.0300	31	49

Plot 8.4.2. Species Centred PCB

This biplot (Figure 8.4.2) is the usual column centred PCB, but performed on proportion data, i.e. the matrix with entries $(p_{ij}-p_{.j})$ is biplotted where $p_{.j}$ is the mean proportion for the jth species.

The origin has been shifted. The true origin, i.e. the point representing a site with no species, is not at the centroid. The centroid of the plot (plotted origin) is the vector of species means.

The α diversity interpretation using site norms lengths is affected by this shift in origin. The squared norm of the ith row is $\sum_{j=1}^{m} (p_{ij}-p_{.k})^2$ which is not the same as the α diversity. In order to interpret α diversity, the 'true' origin is projected onto the plot as a supplementary site vector. The α diversity of a site is then represented by the distance between the supplementary site point and the true origin. The site diversities can therefore be compared with reference to the true origin.

Scalar products are not affected by matrix translation. The distances between sites still represent ß diversities. A better approximation for the ß diversities is obtained than in the noncentred case because the species are have been 'more evenly' weighted.

The scalar product interpretation here is in terms of deviations from the average species proportion.

Although the overall quality of display of the plot is, at 86%, good, the plot has the same problems as the previous one in that it is dominated by a few species (Table 8.4.2). The centring does not help much because it is done on the species proportions for each site, not on the raw counts. Suppression of the two most dominant species, Sanderling and Curlew Sandpiper would merely result in dominance by other species, because of the disparities in the overall frequencies of the species.

A transformation that adjusts for this disparity is called for. An example of how to interpret this type of display is left to Plot 8.1.3, where a far better representation of the data set is obtained.

TABLE 8.4.2

THE EIGENVALUES

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	1.31609	76.946	76.946
2	0.14889	8.705	85.651
3	0.10784	6.305	91.956
4	0.06072	3.550	95.506
5	0.03743	2.188	97.694
6	0.02118	1.238	98.933
7	0.00778	0.455	99.387
8	0.00598	0.350	99.737
9	0.00312	0.183	99.920
10	0.00088	0.052	99.971
11	0.00031	0.018	99.990
12	0.00009	0.005	99.995
13	0.00006	0.003	99.999
14	0.00002	0.001	100.000

THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	NAMIB-N-OC	A	1.000	0.788	0.048	-0.2470	0.746	0.046	-0.0592	0.043	0.024
2	NAMIB-N-W	B	1.000	0.171	0.019	0.0539	0.088	0.002	0.0522	0.083	0.018
3	NAMIB-S-C	C	1.000	0.739	0.027	-0.1242	0.338	0.012	-0.1352	0.401	0.123
4	NAMIB-S-W	D	1.000	0.837	0.088	0.2752	0.503	0.058	-0.2244	0.334	0.338
5	CAPE-N-C	E	1.000	0.575	0.022	-0.1458	0.575	0.016	-0.0009	0.000	0.000
6	CAPE-N-W	F	1.000	0.908	0.099	0.3772	0.842	0.108	0.1057	0.066	0.075
7	CAPE-W-C	G	1.000	0.936	0.028	-0.2135	0.936	0.035	0.0035	0.000	0.000
8	CAPE-W-W	H	1.000	0.845	0.099	0.3775	0.843	0.108	-0.0174	0.002	0.002
9	CAPE-S-C	I	1.000	0.468	0.037	-0.0817	0.106	0.005	-0.1508	0.362	0.153
10	CAPE-S-W	J	1.000	0.971	0.091	0.3841	0.948	0.112	0.0601	0.023	0.024
11	CAPE-E-C	K	1.000	0.959	0.086	-0.3726	0.946	0.105	0.0439	0.013	0.013
12	CAPE-E-W	L	1.000	0.768	0.036	0.2153	0.749	0.035	0.0341	0.019	0.008
13	TRANSKEI-C	M	1.000	0.992	0.096	-0.4023	0.983	0.123	0.0369	0.008	0.009
14	NATAL-C	N	1.000	0.902	0.135	-0.4376	0.827	0.145	0.1318	0.075	0.117
15	NATAL-W	O	1.000	0.860	0.089	0.3415	0.765	0.089	0.1198	0.094	0.096

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	FACT 2	COR
1	OYSTERCATCH	1	1.000	0.025	0.066	-0.0464	0.002	-0.1512	0.023
2	WF-PLOVER	2	1.000	0.054	0.066	-0.2109	0.045	-0.0939	0.009
3	K-PLOVER	3	1.000	0.001	0.043	0.0170	0.000	0.0180	0.001
4	TB-PLOVER	4	1.000	0.000	0.060	0.0069	0.000	-0.0169	0.000
5	G-PLOVER	5	1.000	0.220	0.067	0.0641	0.004	-0.4646	0.216
6	R-PLOVER	6	1.000	0.006	0.057	0.0456	0.002	0.0542	0.003
7	BT-GODWIT	7	1.000	0.001	0.066	0.0064	0.000	-0.0216	0.000
8	WHIMBREL	8	1.000	0.008	0.043	0.0060	0.000	-0.0710	0.008
9	M-SANDPIPER	9	1.000	0.013	0.001	0.0077	0.003	0.0151	0.010
10	GREENSHANK	:	1.000	0.005	0.031	0.0260	0.001	0.0390	0.003
11	C-SANDPIPER	;	1.000	0.000	0.041	0.0009	0.000	0.0146	0.000
12	TURNSTONE	<	1.000	0.360	0.067	-0.2006	0.040	-0.5650	0.320
13	KNOT	=	1.000	0.000	0.066	0.0125	0.000	-0.0154	0.000
14	SANDERLING	>	1.000	0.669	0.067	-0.7361	0.543	0.3555	0.127
15	LITTLE-STILT	?	1.000	0.271	0.059	0.2137	0.052	0.4383	0.219
16	CURSANDPIPER	@	1.000	0.287	0.067	0.5355	0.287	-0.0041	0.000
17	RUFF	!	1.000	0.110	0.066	0.1626	0.027	0.2867	0.083
18	AVOCET	"	1.000	0.042	0.038	0.0652	0.007	0.1400	0.035
19	BW-STILT	#	1.000	0.005	0.032	0.0241	0.001	0.0422	0.004

FIGURE 8.4.2

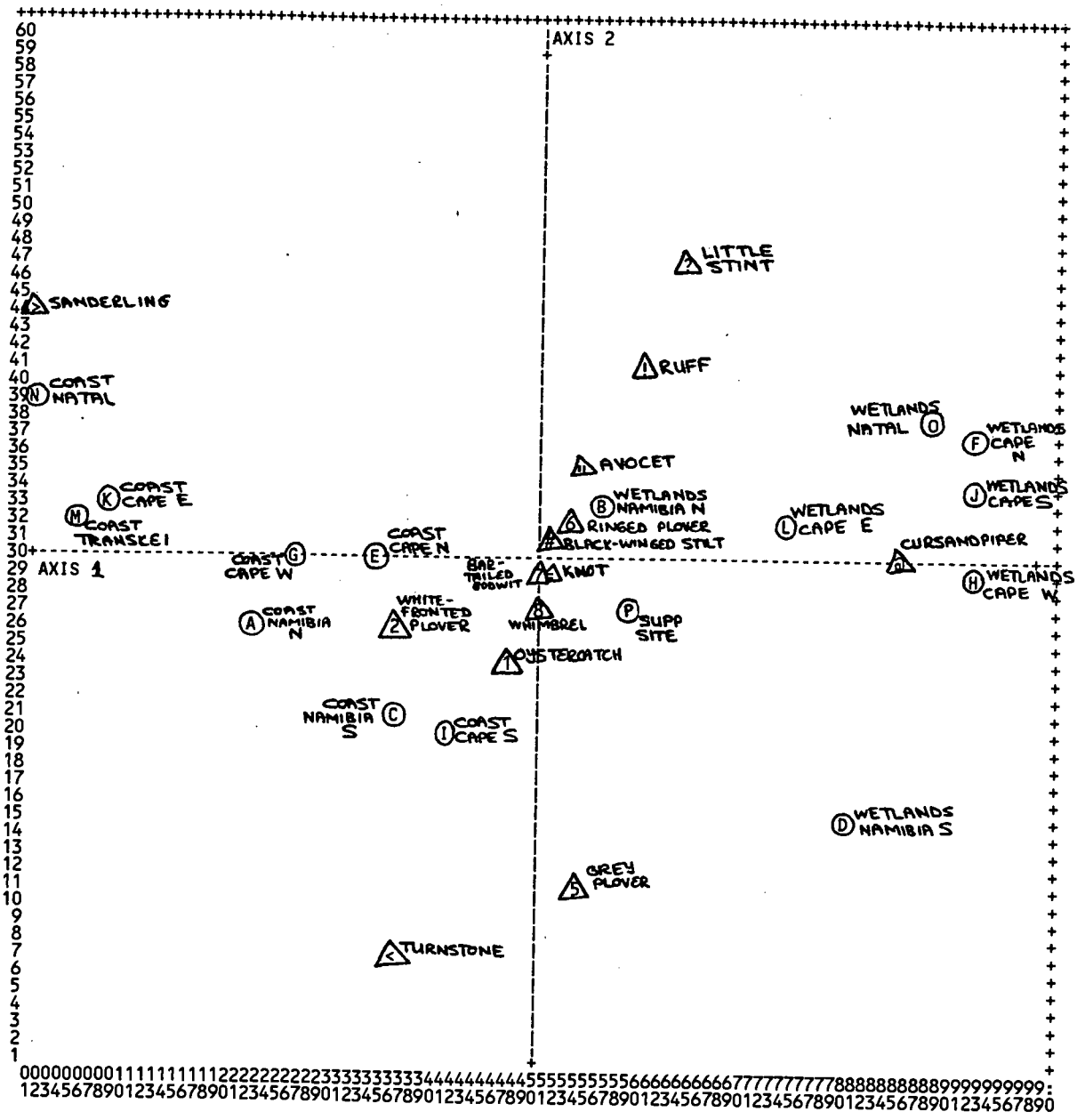

AXIS 2
AXIS 1
LITTLE STINT
SANDERLING
RUFF
COAST NATAL
WETLANDS NATAL
WETLANDS CAPE N
AVOCET
COAST CAPE E
WETLANDS NAMIBIA N
WETLANDS CAPE S
COAST TRANSKEI
RINGED PLOVER
BLACK-WINGED STILT
WETLANDS CAPE E
CURSANDPIPER
COAST CAPE N
COAST CAPE W
BAR-TAILED GODWIT
KNOT
WETLANDS CAPE W
COAST NAMIBIA N
WHITE-FRONTED PLOVER
WHIMBREL
SLUPP SITE
OYSTERCATCH
COAST NAMIBIA S
COAST CAPE S
WETLANDS NAMIBIA S
GREY PLOVER
TURNSTONE

FIGURE 8.4.2 (cont)

NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	NAMIB-N-OC	A	-0.2470	-0.0592	26	22
2	NAMIB-N-W	B	0.0539	0.0522	33	56
3	NAMIB-S-C	C	-0.1242	-0.1352	21	36
4	NAMIB-S-W	D	0.2752	-0.2244	15	80
5	CAPE-N-C	E	-0.1458	-0.0009	30	34
6	CAPE-N-W	F	0.3772	0.1057	37	92
7	CAPE-W-C	G	-0.2135	0.0035	30	26
8	CAPE-W-W	H	0.3775	-0.0174	29	92
9	CAPE-S-C	I	-0.0817	-0.1508	20	41
10	CAPE-S-W	J	0.3841	0.0601	34	92
11	CAPE-E-C	K	-0.3726	0.0439	33	8
12	CAPE-E-W	L	0.2153	0.0341	32	74
13	TRANSKEI-C	M	-0.4023	0.0369	32	5
14	NATAL-C	N	-0.4376	0.1318	39	1
15	NATAL-W	O	0.3415	0.1198	38	88
16		P	0.0822	-0.0412	27	59
1	OYSTERCATCH	1	-0.0464	-0.1512	24	47
2	WF-PLOVER	2	-0.2109	-0.0939	26	36
3	K-PLOVER	3	0.0170	0.0180	31	51
4	TB-PLOVER	4	0.0069	-0.0169	29	50
5	G-PLOVER	5	0.0641	-0.4646	11	54
6	R-PLOVER	6	0.0456	0.0542	32	53
7	BT-GODWIT	7	0.0064	-0.0216	29	50
8	WHIMBREL	8	0.0060	-0.0710	27	50
9	M-SANDPIPER	9	0.0077	0.0151	30	50
10	GREENSHANK	:	0.0260	0.0390	31	51
11	C-SANDPIPER	;	0.0009	0.0146	30	50
12	TURNSTONE	<	-0.2006	-0.5650	7	36
13	KNOT	=	0.0125	-0.0154	29	51
14	SANDERLING	>	-0.7361	0.3555	44	1
15	LITTLE-STILT	?	0.2137	0.4383	47	64
16	CURSANDPIPER	@	0.5355	-0.0041	30	85
17	RUFF	!	0.1626	0.2867	41	60
18	AVOCET	"	0.0652	0.1400	35	54
19	BW-STILT	#	0.0241	0.0422	31	51

Plot 8.4.3 Logged Ter Braak's PCB

To compensate for the large differences in the numbers of each species, the Ter Braak's diversity plot was applied to logged counts. An interesting plot (Figure 8.4.3) emerged.

Let the count data be represented by x_{ij}. The transformation made on the data is to $z_{ij}=\ln(x_{ij}+1)$. This retains zero counts in the original matrix. For each site, proportions of the z_{ij} are computed.

Quality of the Display

The overall quality of display is 73%, with the first axis accounting for 54% (Table 8.4.3).

Compared to the Ter Braaks diversity plot of the count data (Fig. 8.4.2), the birds have a far more equitable representation. The bird species Sanderling still has the largest relative distance from the origin to the first axis, but this distance has dropped from 54% to 25%. Sanderling, Curlew Sandpiper and Grey Plover have the greatest distances on the second axis.

Considering the sites, the second axis is mainly constituted by Namibia South Wetlands (52%). Because the second axis accounts for 19% of the original variation, this means that this point on the second axis contributes 10% of the overall quality. It comprises 11/73=14% of the display. The combined contribution of Namibia South Wetlands and Natal Coast (18.8%) to the second axis is 71%.

Thus although the overall quality of the display has dropped, the plot is more meaningful than previous displays.

Interpretations

Here, distances between site points represent their (logged) ß diversities and the distances between the supplementary point and the site are the (logged) α diversities. The plot groups the sites and the species associated with them into wetlands and coast. This represents two broad groupings of the ß diversities, or species compositions. The wetland sites, with the sole exception of Namibia south, are plotted close together. In the

Namibia South wetlands, the Grey Plover and Curlew Sandpiper predominate. The coastal sites are further grouped into those with sandy shores (Natal and the Transkei) and those with rocky shores. This displays the difference in species composition between the two shore types. Sandy shores are associated with White Fronted Plovers and Sanderlings. Turnstones, Whimbrels and Oystercatchers predominate on rocky shores. The Common Sandpiper is found in all the environments, and is not strongly associated with either coastal or wetland sites. Within the wetland areas, there is an approximate ordering of the sites from fresh water to salty water and lagoons.

The point representing the site with maximum diversity is given supplementary status. It coincides with the point representing the notional site with no species. The small distances between this point and the wetland areas represents the high level of diversity of these sites. The points for the coastal sites have greater distances from the site of maximum diversity. Coastal sites are not as diverse as wetland sites, having few species associated with them. The species composition of each site is indicated by the species plotted near it. For example, the Natal coast has a population predominantly of Sanderlings.

The analysis picks out six species - White-Fronted Plover, Grey Plover, Turnstone, Sanderling, Ruff and Avocet which either co-occur or avoid each other. The remaining species show relatively less patterning, and their formal 'explanations' (in terms of quality displayed) are poor. In spite of this, the positions of most of these species are in accord with what ornithologists would anticipate; certain species are key in classifying the areas.

TABLE 8.4.3

THE EIGENVALUES

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	0.14142	53.760	53.760
2	0.05050	19.197	72.957
3	0.02919	11.095	84.052
4	0.01742	6.620	90.672
5	0.00651	2.475	93.148
6	0.00613	2.331	95.479
7	0.00388	1.474	96.953
8	0.00305	1.160	98.113
9	0.00191	0.725	98.839
10	0.00170	0.644	99.483
11	0.00071	0.271	99.754
12	0.00047	0.178	99.932
13	0.00011	0.043	99.976
14	0.00006	0.024	100.000

THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	NAMIB-N-OC	A	1.000	0.383	0.046	-0.0515	0.218	0.019	0.0448	0.165	0.040
2	NAMIB-N-W	B	1.000	0.479	0.027	0.0587	0.477	0.024	-0.0040	0.002	0.000
3	NAMIB-S-C	C	1.000	0.620	0.050	-0.0731	0.408	0.038	0.0526	0.211	0.055
4	NAMIB-S-W	D	1.000	0.781	0.132	0.0315	0.029	0.007	0.1619	0.752	0.519
5	CAPE-N-C	E	1.000	0.486	0.036	-0.0596	0.371	0.025	0.0333	0.116	0.022
6	CAPE-N-W	F	1.000	0.784	0.114	0.1471	0.719	0.153	-0.0445	0.066	0.039
7	CAPE-W-C	G	1.000	0.022	0.014	-0.0089	0.022	0.001	0.0010	0.000	0.000
8	CAPE-W-W	H	1.000	0.800	0.041	0.0909	0.762	0.058	-0.0204	0.038	0.008
9	CAPE-S-C	I	1.000	0.582	0.033	-0.0679	0.527	0.033	0.0219	0.055	0.010
10	CAPE-S-W	J	1.000	0.949	0.054	0.1105	0.852	0.086	-0.0372	0.097	0.027
11	CAPE-E-C	K	1.000	0.420	0.027	-0.0541	0.419	0.021	0.0032	0.001	0.000
12	CAPE-E-W	L	1.000	0.896	0.034	0.0835	0.773	0.049	-0.0333	0.123	0.022
13	TRANSKEI-C	M	1.000	0.871	0.135	-0.1683	0.800	0.200	-0.0501	0.071	0.050
14	NATAL-C	N	1.000	0.715	0.187	-0.1602	0.522	0.181	-0.0975	0.193	0.188
15	NATAL-W	O	1.000	0.880	0.068	0.1215	0.825	0.104	-0.0316	0.056	0.020

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT1	COR	FACT2	COR
1	OYSTERCATCH	1	1.000	0.032	0.058	-0.1643	0.031	0.0235	0.001
2	WF-PLOVER	2	1.000	0.147	0.056	-0.3494	0.147	-0.0082	0.000
3	K-PLOVER	3	1.000	0.109	0.045	0.2094	0.066	-0.1710	0.044
4	TB-PLOVER	4	1.000	0.010	0.055	0.0661	0.005	0.0627	0.005
5	G-PLOVER	5	1.000	0.342	0.049	0.0073	0.000	0.5005	0.342
6	R-PLOVER	6	1.000	0.005	0.046	0.0415	0.003	-0.0385	0.002
7	BT-GODWIT	7	1.000	0.047	0.053	0.0673	0.006	0.1822	0.042
8	WHIMBREL	8	1.000	0.043	0.061	-0.1885	0.039	0.0608	0.004
9	M-SANDPIPER	9	1.000	0.104	0.037	0.1578	0.045	-0.1807	0.059
10	GREENSHANK	:	1.000	0.052	0.057	0.0129	0.000	-0.2108	0.052
11	C-SANDPIPER	;	1.000	0.053	0.044	-0.1454	0.032	-0.1199	0.022
12	TURNSTONE	<	1.000	0.207	0.064	-0.3891	0.158	0.2162	0.049
13	KNOT	=	1.000	0.002	0.064	-0.0393	0.002	0.0096	0.000
14	SANDERLING	>	1.000	0.433	0.060	-0.5009	0.279	-0.3729	0.154
15	LITTLE-STILT	?	1.000	0.087	0.055	0.2647	0.085	0.0347	0.001
16	CURSANDPIPER	@	1.000	0.342	0.062	0.1390	0.021	0.5478	0.321
17	RUFF	!	1.000	0.122	0.058	0.2934	0.100	-0.1373	0.022
18	AVOCET	"	1.000	0.166	0.051	0.2854	0.107	-0.2107	0.059
19	BW-STILT	#	1.000	0.223	0.027	0.2322	0.135	-0.1879	0.088

FIGURE 8.4.3

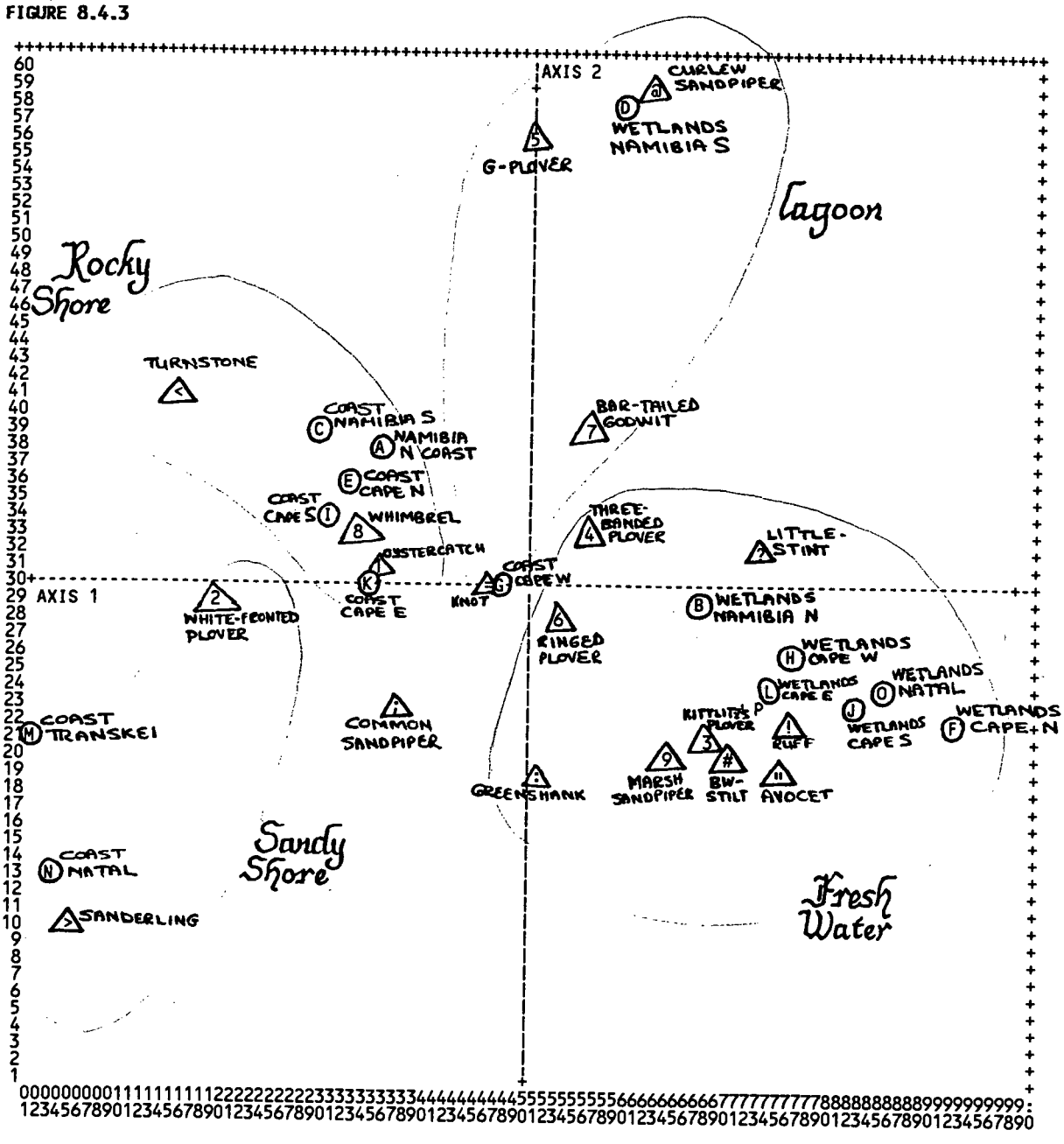

AXIS 2
AXIS 1
CURLEW SANDPIPER
WETLANDS NAMIBIA S
G-PLOVER
Lagoon
Rocky Shore
TURNSTONE
COAST NAMIBIA S
NAMIBIA N COAST
COAST CAPE N
COAST CAPE S
WHIMBREL
BAR-TAILED GODWIT
THREE-BANDED PLOVER
LITTLE-STINT
OYSTERCATCH
COAST CAPE E
KNOT
COAST CAPE W
WHITE-FRONTED PLOVER
RINGED PLOVER
WETLANDS NAMIBIA N
WETLANDS CAPE W
WETLANDS CAPE E
WETLANDS NATAL
WETLANDS CAPE S
WETLANDS CAPE N
KITTLITZ'S PLOVER
RUFF
COAST TRANSKEI
COMMON SANDPIPER
MARSH SANDPIPER
BW-STILT
AVOCET
GREENSHANK
Sandy Shore
COAST NATAL
SANDERLING
Fresh Water

FIGURE 8.4.3 (cont)

NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	NAMIB-N-OC	A	-0.0515	0.0448	38	35
2	NAMIB-N-W	B	0.0587	-0.0040	29	67
3	NAMIB-S-C	C	-0.0731	0.0526	39	29
4	NAMIB-S-W	D	0.0315	0.1619	58	59
5	CAPE-N-C	E	-0.0596	0.0333	36	32
6	CAPE-N-W	F	0.1471	-0.0445	22	92
7	CAPE-W-C	G	-0.0089	0.0010	30	47
8	CAPE-W-W	H	0.0909	-0.0204	26	76
9	CAPE-S-C	I	-0.0679	0.0219	34	30
10	CAPE-S-W	J	0.1105	-0.0372	23	82
11	CAPE-E-C	K	-0.0541	0.0032	30	34
12	CAPE-E-W	L	0.0835	-0.0333	24	74
13	TRANSKEI-C	M	-0.1683	-0.0501	21	1
14	NATAL-C	N	-0.1602	-0.0975	13	3
15	NATAL-W	O	0.1215	-0.0316	24	85
16	SUPP SITE	P	0.0793	-0.0394	23	73
1	OYSTERCATCH	1	-0.1643	0.0235	31	35
2	WF-PLOVER	2	-0.3494	-0.0082	29	19
3	K-PLOVER	3	0.2094	-0.1710	21	68
4	TB-PLOVER	4	0.0661	0.0627	33	56
5	G-PLOVER	5	0.0073	0.5005	56	50
6	R-PLOVER	6	0.0415	-0.0385	28	53
7	BT-GODWIT	7	0.0673	0.1822	39	56
8	WHIMBREL	8	-0.1885	0.0608	33	33
9	M-SANDPIPER	9	0.1578	-0.1807	20	64
10	GREENSHANK	:	0.0129	-0.2108	19	51
11	C-SANDPIPER	;	-0.1454	-0.1199	23	37
12	TURNSTONE	<	-0.3891	0.2162	41	15
13	KNOT	=	-0.0393	0.0096	30	46
14	SANDERLING	>	-0.5009	-0.3729	10	5
15	LITTLE-STINT	?	0.2647	0.0347	32	73
16	CURSANDPIPER	@	0.1390	0.5478	59	62
17	RUFF	!	0.2934	-0.1373	22	76
18	AVOCET	"	0.2854	-0.2107	19	75
19	BW-STILT	#	0.2322	-0.1879	20	70

B. CORRESPONDENCE ANALYSIS TYPE CENTRES

Plot 8.4.4 Correspondence Analysis

Quality of the display

Seventy eight percent of the inertia is retained in the two dimensional display, with the first axis retaining 61% (Table 8.4.4). The 'qual' column indicates that the points have a reasonable to good display in two dimensions.

Curlew Sandpiper, the most numerous species, has the highest mass, and makes a substantial contribution to the display. The other numerous species, Sanderling, constitutes 34% of the first axis. Three birds constitute 55% of the first axis. Similarly the well populated sites constitute the axes.

Interpretations

The plot (Figure 8.4.4) shows distinct groupings of the different site types. All the coastal sites are situated on the right hand side of the plot; the small distances between them indicate their similar profiles across the bird species. The wetland sites are divided into the Namibian wetlands, and the wetlands in Natal and in the north, west and eastern Cape. Birds are attracted away from the origin in the direction of the habitats in which they are found. Western Cape wetlands has over half its population consisting of Curlew Sandpipers, and is drawn away from the origin in the direction of this point. Although western Cape wetlands does not have a particularly high proportion of Kittlitz's Plovers, these birds are most commonly found in the western Cape wetlands, hence these points are attracted towards each other.

The origin represents the position of the average row and column profiles. Vector points far from the origin represent those furthest from what we would expect under the null hypothesis of site-species independence.

The inertia for a particular point is given by the product of its mass and its squared distance. For example, Common Sandpiper has few observations, i.e, a low mass, yet a large distance from the origin. It is well represented on the display (73%). Large norms

do not represent large masses. This illustrates a feature of correspondence analysis - sparse vectors are emphasized. Correspondence analysis may be criticised for giving undue weight to the sparse species and sites. In this particular example, domination of the plot by sparse vectors is not a problem.

Plot 8.4.5 Correspondence Analysis on Logged Data

The counts were logged before correspondence analysis was performed. That there is hardly any difference between this and the previous plot shows that correspondence analysis is a fairly robust technique.

TABLE 8.4.4

THE EIGENVALUES

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	0.49301	61.439	61.439
2	0.13163	16.404	77.843
3	0.06915	8.618	86.461
4	0.04108	5.119	91.580
5	0.02051	2.557	94.137
6	0.01702	2.122	96.258
7	0.01190	1.483	97.741
8	0.01011	1.260	99.002
9	0.00581	0.724	99.726
10	0.00124	0.154	99.880
11	0.00065	0.081	99.961
12	0.00023	0.029	99.990
13	0.00005	0.006	99.996
14	0.00003	0.004	100.000

THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	NAMIB-N-OC	A	0.108	0.817	0.155	0.0962	0.806	0.203	-0.0109	0.010	0.010
2	NAMIB-N-W	B	0.177	0.393	0.072	0.0195	0.116	0.014	-0.0302	0.278	0.122
3	NAMIB-S-C	C	0.006	0.561	0.014	0.0943	0.473	0.011	0.0406	0.088	0.008
4	NAMIB-S-W	D	0.001	0.018	0.001	-0.0030	0.000	0.000	-0.0182	0.017	0.000
5	CAPE-N-C	E	0.057	0.866	0.051	0.0791	0.866	0.072	-0.0001	0.000	0.000
6	CAPE-N-W	F	0.022	0.582	0.034	-0.0766	0.478	0.026	0.0356	0.103	0.021
7	CAPE-W-C	G	0.067	0.913	0.086	0.0929	0.840	0.117	0.0274	0.073	0.038
8	CAPE-W-W	H	0.246	0.850	0.144	-0.0494	0.517	0.121	-0.0396	0.332	0.292
9	CAPE-S-C	I	0.030	0.600	0.081	0.0888	0.369	0.048	0.0702	0.231	0.113
10	CAPE-S-W	J	0.109	0.905	0.091	-0.0752	0.841	0.125	0.0208	0.064	0.036
11	CAPE-E-C	K	0.027	0.889	0.052	0.1134	0.824	0.069	0.0320	0.066	0.021
12	CAPE-E-W	L	0.049	0.619	0.045	-0.0486	0.320	0.023	0.0470	0.299	0.082
13	TRANSKEI-C	M	0.002	0.923	0.006	0.1257	0.833	0.008	0.0413	0.090	0.003
14	NATAL-C	N	0.009	0.589	0.031	0.1180	0.513	0.025	0.0454	0.076	0.014
15	NATAL-W	O	0.091	0.892	0.138	-0.0862	0.608	0.136	0.0590	0.285	0.239

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	OYSTERCATCH	1	0.010	0.461	0.058	0.1125	0.258	0.025	0.0996	0.203	0.072
2	WF-PLOVER	2	0.048	0.754	0.079	0.0901	0.611	0.079	0.0436	0.143	0.069
3	K-PLOVER	3	0.008	0.626	0.013	-0.0842	0.568	0.012	-0.0270	0.058	0.005
4	TB-PLOVER	4	0.002	0.413	0.003	-0.0525	0.277	0.001	0.0367	0.136	0.002
5	G-PLOVER	5	0.044	0.486	0.014	0.0031	0.004	0.000	-0.0354	0.483	0.041
6	R-PLOVER	6	0.018	0.884	0.023	-0.0679	0.430	0.016	0.0698	0.454	0.065
7	BT-GODWIT	7	0.008	0.222	0.028	0.0241	0.022	0.001	-0.0725	0.200	0.034
8	WHIMBREL	8	0.010	0.343	0.021	-0.0182	0.020	0.001	0.0733	0.323	0.042
9	M-SANDPIPER	9	0.003	0.861	0.005	-0.0931	0.672	0.006	0.0494	0.189	0.006
10	GREENSHANK	:	0.013	0.759	0.015	-0.0658	0.458	0.011	0.0534	0.302	0.028
11	C-SANDPIPER	;	0.006	0.731	0.017	-0.0451	0.088	0.002	0.1216	0.642	0.067
12	TURNSTONE	<	0.072	0.771	0.132	0.1047	0.752	0.161	0.0167	0.019	0.015
13	KNOT	=	0.040	0.698	0.051	-0.0150	0.022	0.002	-0.0825	0.676	0.209
14	SANDERLING	>	0.200	0.908	0.231	0.0918	0.907	0.342	-0.0028	0.001	0.001
15	LITTLE-STI T	?	0.073	0.840	0.088	-0.0795	0.655	0.094	0.0423	0.186	0.100
16	CURSANDPIPER	@	0.343	0.919	0.095	-0.0385	0.667	0.103	-0.0237	0.252	0.146
17	RUFF	!	0.069	0.889	0.087	-0.0876	0.760	0.107	0.0360	0.129	0.068
18	AVOCET	"	0.024	0.658	0.021	-0.0659	0.623	0.021	0.0155	0.034	0.004
19	BW-STILT	#	0.009	0.742	0.019	-0.0938	0.520	0.016	0.0614	0.223	0.025

FIGURE 8.4.4

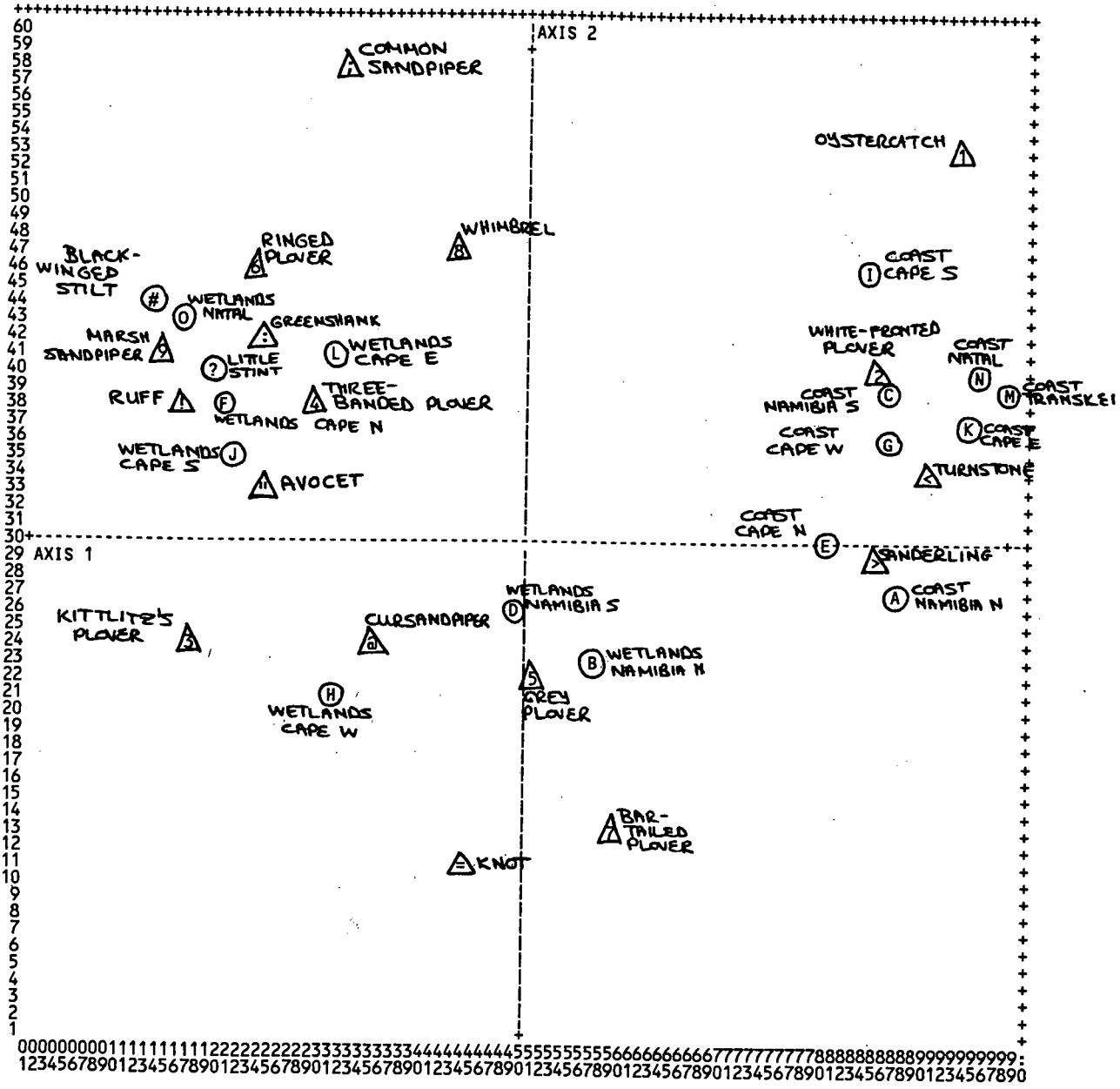

AXIS 2
AXIS 1
COMMON SANDPIPER
OYSTERCATCH
WHIMBREL
RINGED PLOVER
BLACK-WINGED STILT
COAST CAPE S
WETLANDS NATAL
GREENSHANK
MARSH SANDPIPER
LITTLE STINT
WETLANDS CAPE E
WHITE-FRONTED PLOVER
COAST NATAL
COAST NAMIBIA S
COAST TRANSKEI
RUFF
THREE-BANDED PLOVER
WETLANDS CAPE N
COAST CAPE W
COAST CAPE E
WETLANDS CAPE S
AVOCET
TURNSTONE
COAST CAPE N
SANDERLING
WETLANDS NAMIBIA S
COAST NAMIBIA N
KITTLITZ'S PLOVER
CURSANDPIPER
WETLANDS NAMIBIA N
GREY PLOVER
WETLANDS CAPE W
BAR-TAILED PLOVER
KNOT

FIGURE 8.4.4 (cont)

NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	NAMIB-N-C	A	0.0962	-0.0109	27	87
2	NAMIB-N-W	B	0.0195	-0.0302	23	57
3	NAMIB-S-C	C	0.0943	0.0406	39	86
4	NAMIB-S-W	D	-0.0030	-0.0182	26	49
5	CAPE-N-C	E	0.0791	-0.0001	30	80
6	CAPE-N-W	F	-0.0766	0.0356	38	20
7	CAPE-W-C	G	0.0929	0.0274	36	86
8	CAPE-W-W	H	-0.0494	-0.0396	21	31
9	CAPE-S-C	I	0.0888	0.0702	46	84
10	CAPE-S-W	J	-0.0752	0.0208	35	21
11	CAPE-E-C	K	0.1134	0.0320	37	94
12	CAPE-E-W	L	-0.0486	0.0470	41	31
13	TRANSKEI-C	M	0.1257	0.0413	39	98
14	NATAL-C	N	0.1180	0.0454	40	95
15	NATAL-W	O	-0.0842	0.0590	43	16
1	OYSTERCATCH	1	0.1125	0.0996	53	93
2	WF-PLOVER	2	0.0901	0.0436	40	85
3	K-PLOVER	3	-0.0842	-0.0270	24	17
4	TB-PLOVER	4	-0.0525	0.0367	38	29
5	G-PLOVER	5	0.0031	-0.0354	22	51
6	R-PLOVER	6	-0.0679	0.0698	46	23
7	BT-GODWIT	7	0.0241	-0.0725	13	59
8	WHIMBREL	8	-0.0182	0.0733	47	43
9	M-SANDPIPER	9	-0.0931	0.0494	41	14
10	GREENSHANK	:	-0.0658	0.0534	42	24
11	C-SANDPIPER	;	-0.0451	0.1216	58	32
12	TURNSTONE	<	0.1047	0.0167	34	90
13	KNOT	=	-0.0150	-0.0825	11	44
14	SANDERLING	>	0.0918	-0.0028	29	85
15	LITTLE-STINT	?	-0.0795	0.0423	40	19
16	CURSANDPIPER	@	-0.0385	-0.0237	24	35
17	RUFF	!	-0.0876	0.0360	38	16
18	AVOCET	"	-0.0659	0.0155	33	24
19	BW-STILT	#	-0.0938	0.0614	44	13

TABLE 8.4.5

THE EIGENVALUES

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	0.16070	61.157	61.157
2	0.03201	12.181	73.338
3	0.02359	8.979	82.317
4	0.01620	6.164	88.482
5	0.00891	3.392	91.874
6	0.00654	2.490	94.364
7	0.00470	1.790	96.154
8	0.00382	1.454	97.608
9	0.00227	0.862	98.471
10	0.00163	0.620	99.090
11	0.00148	0.563	99.653
12	0.00061	0.232	99.885
13	0.00019	0.072	99.957
14	0.00011	0.043	100.000

THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	NAMIB-N-C	A	0.066	0.922	0.090	-0.4119	0.476	0.070	0.3985	0.446	0.329
2	NAMIB-N-W	B	0.098	0.674	0.038	0.1173	0.134	0.008	0.2358	0.540	0.170
3	NAMIB-S-C	C	0.042	0.716	0.061	-0.5195	0.698	0.070	-0.0828	0.018	0.009
4	NAMIB-S-W	D	0.021	0.213	0.051	-0.2216	0.078	0.006	0.2927	0.136	0.057
5	CAPE-N-C	E	0.059	0.714	0.063	-0.4462	0.710	0.073	0.0327	0.004	0.002
6	CAPE-N-W	F	0.061	0.789	0.105	0.5368	0.634	0.109	-0.2658	0.155	0.134
7	CAPE-W-C	G	0.073	0.374	0.033	-0.2000	0.333	0.018	-0.0700	0.041	0.011
8	CAPE-W-W	H	0.112	0.819	0.055	0.3072	0.732	0.066	0.1055	0.086	0.039
9	CAPE-S-C	I	0.061	0.844	0.073	-0.4938	0.781	0.093	-0.1402	0.063	0.038
10	CAPE-S-W	J	0.100	0.946	0.069	0.4037	0.900	0.101	-0.0909	0.046	0.026
11	CAPE-E-C	K	0.059	0.755	0.057	-0.4260	0.721	0.067	-0.0934	0.035	0.016
12	CAPE-E-W	L	0.097	0.875	0.036	0.2854	0.834	0.049	-0.0634	0.041	0.012
13	TRANSKEI-C	M	0.026	0.802	0.073	-0.7238	0.717	0.086	-0.2488	0.085	0.051
14	NATAL-C	N	0.031	0.490	0.112	-0.5927	0.375	0.069	-0.3274	0.115	0.105
15	NATAL-W	O	0.093	0.837	0.084	0.4441	0.836	0.114	0.0150	0.001	0.001

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	OYSTERCATCH	1	0.048	0.463	0.075	-0.3910	0.370	0.045	-0.1963	0.093	0.057
2	WF-PLOVER	2	0.080	0.816	0.050	-0.3526	0.755	0.062	-0.1004	0.061	0.025
3	K-PLOVER	3	0.026	0.895	0.090	0.9033	0.883	0.130	-0.1072	0.012	0.009
4	TB-PLOVER	4	0.037	0.306	0.019	0.1477	0.163	0.005	-0.1385	0.143	0.022
5	G-PLOVER	5	0.069	0.532	0.025	-0.1347	0.189	0.008	0.1814	0.343	0.071
6	R-PLOVER	6	0.057	0.186	0.022	0.0736	0.053	0.002	-0.1168	0.133	0.024
7	BT-GODWIT	7	0.029	0.831	0.063	0.1256	0.028	0.003	0.6749	0.803	0.417
8	WHIMBREL	8	0.055	0.486	0.032	-0.2667	0.472	0.024	-0.0446	0.013	0.003
9	M-SANDPIPER	9	0.026	0.862	0.060	0.7210	0.849	0.083	-0.0891	0.013	0.006
10	GREENSHANK	:	0.053	0.478	0.014	0.0720	0.073	0.002	-0.1689	0.405	0.047
11	C-SANDPIPER	;	0.050	0.506	0.022	-0.1905	0.306	0.011	-0.1545	0.201	0.037
12	TURNSTONE	<	0.075	0.957	0.070	-0.4835	0.946	0.109	0.0540	0.012	0.007
13	KNOT	=	0.048	0.505	0.055	-0.1307	0.057	0.005	0.3666	0.448	0.201
14	SANDERLING	>	0.086	0.773	0.083	-0.4344	0.746	0.101	-0.0839	0.028	0.019
15	LITTLE-STINT	?	0.058	0.745	0.039	0.3618	0.742	0.047	0.0258	0.004	0.001
16	CURSANDPIPER	@	0.088	0.142	0.028	-0.0038	0.000	0.000	0.1089	0.142	0.033
17	RUFF	!	0.048	0.808	0.075	0.5695	0.787	0.097	-0.0914	0.020	0.013
18	AVOCET	"	0.039	0.906	0.086	0.7181	0.902	0.127	0.0495	0.004	0.003
19	BW-STILT	#	0.029	0.937	0.092	0.8822	0.931	0.139	-0.0671	0.005	0.004

FIGURE 8.4.5

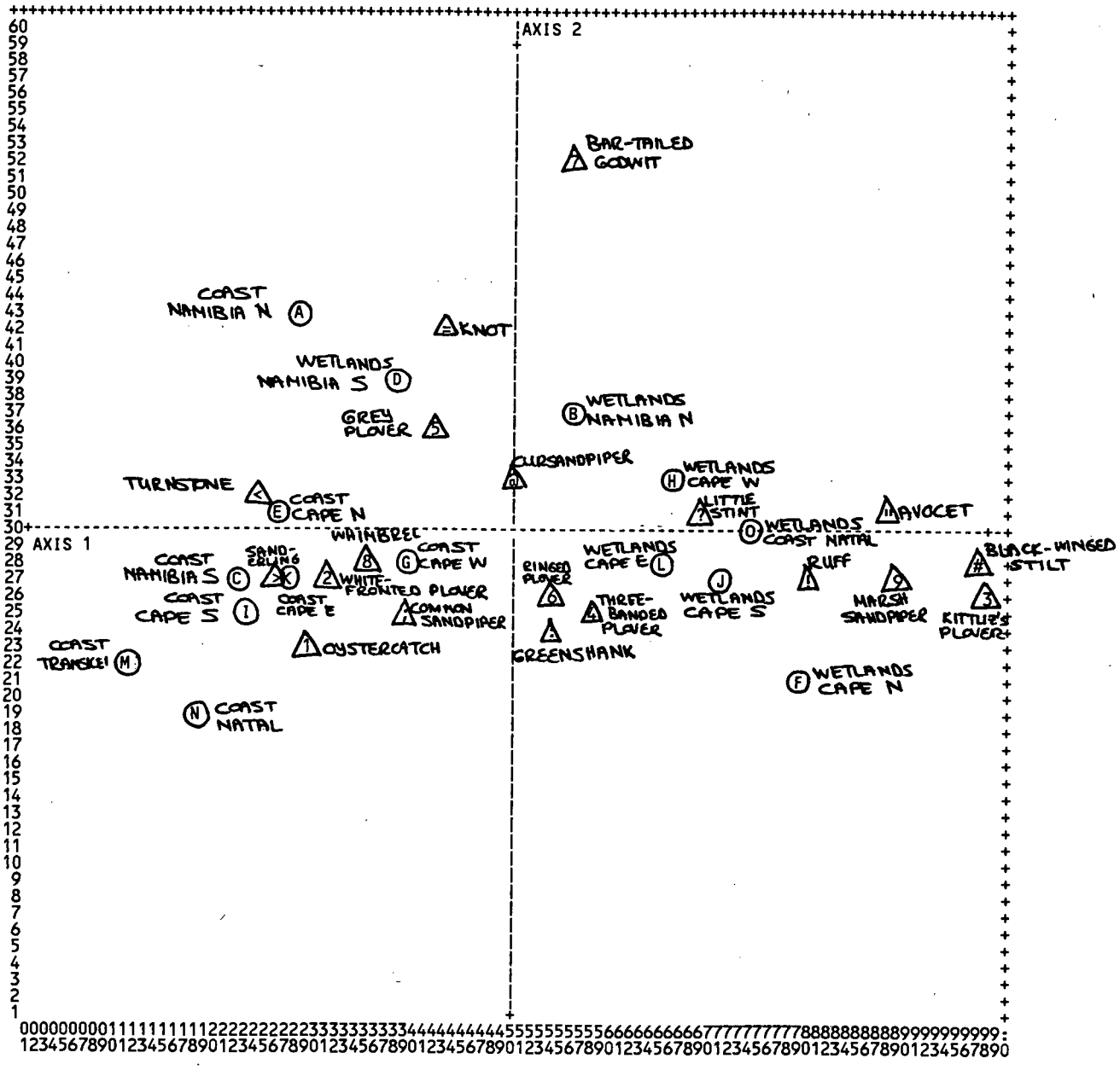

AXIS 2
AXIS 1
BAR-TAILED GODWIT
COAST NAMIBIA N
KNOT
WETLANDS NAMIBIA S
GREY PLOVER
WETLANDS NAMIBIA N
CURSANDPIPER
WETLANDS CAPE W
TURNSTONE
COAST CAPE N
LITTLE STINT
AVOCET
WETLANDS COAST NATAL
WHIMBREL
COAST NAMIBIA S
SAND-ERLING
COAST CAPE W
WETLANDS CAPE E
RUFF
BLACK-WINGED STILT
RINGED PLOVER
WHITE-FRONTED PLOVER
COAST CAPE S
COAST CAPE E
COMMON SANDPIPER
THREE-BANDED PLOVER
WETLANDS CAPE S
MARSH SANDPIPER
KITTLIE'S PLOVER
OYSTERCATCH
GREENSHANK
COAST TRANSKEI
WETLANDS CAPE N
COAST NATAL

FIGURE 8.4.5 (cont)

NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	NAMIB-N-C	A	-0.4119	0.3985	43	28
2	NAMIB-N-W	B	0.1173	0.2358	37	56
3	NAMIB-S-C	C	-0.5195	-0.0828	27	22
4	NAMIB-S-W	D	-0.2216	0.2927	39	38
5	CAPE-N-C	E	-0.4462	0.0327	31	26
6	CAPE-N-W	F	0.5368	-0.2658	21	79
7	CAPE-W-C	G	-0.2000	-0.0700	28	39
8	CAPE-W-W	H	0.3072	0.1055	33	66
9	CAPE-S-C	I	-0.4938	-0.1402	25	23
10	CAPE-S-W	J	0.4037	-0.0909	27	71
11	CAPE-E-C	K	-0.4260	-0.0934	27	27
12	CAPE-E-W	L	0.2854	-0.0634	28	65
13	TRANSKEI-C	M	-0.7238	-0.2488	22	11
14	NATAL-C	N	-0.5927	-0.3274	19	18
15	NATAL-W	O	0.4441	0.0150	30	74
1	OYSTERCATCH	1	-0.3910	-0.1963	23	29
2	WF-PLOVER	2	-0.3526	-0.1004	27	31
3	K-PLOVER	3	0.9033	-0.1072	26	98
4	TB-PLOVER	4	0.1477	-0.1385	25	58
5	G-PLOVER	5	-0.1347	0.1814	36	42
6	R-PLOVER	6	0.0736	-0.1168	26	54
7	BT-GODWIT	7	0.1256	0.6749	52	56
8	WHIMBREL	8	-0.2667	-0.0446	28	35
9	M-SANDPIPER	9	0.7210	-0.0891	27	89
10	GREENSHANK	:	0.0720	-0.1689	24	54
11	C-SANDPIPER	;	-0.1905	-0.1545	25	39
12	TURNSTONE	<	-0.4835	0.0540	32	24
13	KNOT	=	-0.1307	0.3666	42	43
14	SANDERLING	>	-0.4344	-0.0839	27	26
15	LITTLE-STINT	?	0.3618	0.0258	31	69
16	CURSANDPIPER	@	-0.0038	0.1089	33	50
17	RUFF	!	0.5695	-0.0914	27	80
18	AVOCET	"	0.7181	0.0495	31	88
19	BW-STILT	#	0.8822	-0.0671	28	97

C. SPEARMAN'S RANK CORRELATION BIPLOTS

This choice of Phase I is described in Section 4.4. The transformation of the data matrix attempts to addresses the problems of the previous plots: domination of the plot by the numerous species.

Two types of rank correlation biplots were applied to the data. In the first (Plot 8.4.6), the number of birds in each site were ranked. In the second (Plot 8.4.7) the sites were ranked within each bird species.

With the first type of ranking method, the angle between two sites in (1-0) and (1-1)-plots represents their Spearman's rank correlation. This correlation is a measure of agreement between the ranks of the sites. In order to centre the plot at the origin, the 19 bird species were given ranks from -9 to 9.

In the second plot, proportions of the species at each site were computed. For each bird species the proportions occurring at the sites was ranked from -7 (the site with the lowest proportion of that species) to 7. Again the reason for this choice of ranks is to centre the column points at the origin.

By ranking the sites within the bird species, Spearman's correlation interpretation is valid for the bird species. Taking proportions addresses the problem of well populated sites dominating the plot.

For the second method, Spearman's correlations are represented within the species in (0-1) and (1-1)-plots.

Plot 8.4.6 Ranking the Bird Species Across Each Site.

Quality of the Display

The quality of the display in two dimensions is 75%, of which 47% is accounted for on the first axis (Figure 8.4.6). As a=b=1, the quality of both the row and the column points can be considered. The quality in two dimensions of both these sets of points is good. As can be seen from the 'ctr' column and the plot, the first axis is constituted chiefly by the coastal sites.

Interpretations

A prominent feature of the plot is the clustering of the sites according to the type of habitat. The coastal sites are on the right hand side of the first axis, while those with sandy shores - the Transkei and Natal coasts separated from the other coastal points, whose shores are chiefly rocky. The wetland points form three clusters. One of these clusters is the points for Namibia, another has the Eastern and Western Cape, and the third has Natal and the Northern and Southern Cape.

Because the angles between the site vectors represent their correlations, the display highlights the similarity in the bird rankings in the coastal sites, i.e. the ordering of these birds in terms of their frequencies is similar. The wetland sites - other than those in Namibia, are uncorrelated with the shore sites.

The bird species Sanderling, White Fronted Plover and Turnstone (shorebirds) and Black Winged Stilt, Kittlitz's Plover and Marsh Sandpiper (freshwater species) are on opposite ends of the first axis and are its main contributors. Sanderling has very high rankings on the coastal sites (as well as on many of the Wetland ones.) Birds such as Black Winged Stilt and Avocet on the left hand side of the plot have low ranks on the coastal sites.

Curlew Sandpipers rank highly in almost all sites; hence the position. It is a major constituent of the second axis, accounting for 22% of the norm.

The supplementary site has the same rank (zero) for all species and is thus positioned

at the origin. In this case, it is exactly the same as the site with maximum diversity - they both consist of a single tied rank.

While the first axis shows which species are and are not common at the coast, the wetland areas come into their own on the second axis. The Little Stint and Ruff, for example, are commonly found on the wetlands. Oystercatchers are not numerous at wetland sites. It is interesting to note that the coastal areas are poorly displayed on the second axis and the wetlands (with the exception of Namibia) are well displayed. This difference is very marked. The reverse is evident on the first axis, although the wetland sites are not too badly displayed there.

In general, birds that are not generally found at the coast are situated on the left of the plot, those that prefer estuaries are in the middle, and shorebirds are on the right. Wetlands areas are grouped by the second axis.

TABLE 8.4.6

THE EIGENVALUES

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	3974.59351	47.410	47.410
2	2317.91943	27.649	75.058
3	609.75012	7.273	82.332
4	480.57077	5.732	88.064
5	328.28424	3.916	91.980
6	210.76143	2.514	94.494
7	129.13861	1.540	96.034
8	95.44963	1.139	97.173
9	85.40030	1.019	98.191
10	72.93977	0.870	99.061
11	34.58432	0.413	99.474
12	24.08654	0.287	99.761
13	13.82017	0.165	99.926
14	4.15534	0.050	99.976
15	2.04571	0.024	100.000

THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	NAMIB-N- C	A	1.000	0.774	0.067	0.2082	0.774	0.109	-0.0043	0.000	0.000
2	NAMIB-N-W	B	1.000	0.574	0.068	0.1427	0.357	0.051	-0.1111	0.217	0.053
3	NAMIB-S-C	C	1.000	0.887	0.067	0.2228	0.887	0.125	0.0040	0.000	0.000
4	NAMIB-S-W	D	1.000	0.391	0.065	0.1318	0.321	0.044	-0.0613	0.070	0.016
5	CAPE-N-C	E	1.000	0.884	0.067	0.2234	0.884	0.126	0.0011	0.000	0.000
6	CAPE-N-W	F	1.000	0.729	0.068	-0.0296	0.015	0.002	-0.2014	0.714	0.175
7	CAPE-W-C	G	1.000	0.820	0.068	0.2110	0.782	0.112	-0.0466	0.038	0.009
8	CAPE-W-W	H	1.000	0.571	0.068	0.0642	0.072	0.010	-0.1686	0.499	0.123
9	CAPE-S-C	I	1.000	0.928	0.067	0.2269	0.919	0.130	0.0225	0.009	0.002
10	CAPE-S-W	J	1.000	0.915	0.068	-0.0518	0.047	0.007	-0.2224	0.868	0.213
11	CAPE-E-C	K	1.000	0.871	0.067	0.2204	0.868	0.122	0.0133	0.003	0.001
12	CAPE-E-W	L	1.000	0.812	0.068	0.0678	0.081	0.012	-0.2042	0.731	0.180
13	TRANSKEI-C	M	1.000	0.844	0.065	0.1953	0.704	0.096	0.0869	0.140	0.033
14	NATAL-C	N	1.000	0.468	0.061	0.1432	0.402	0.052	0.0580	0.066	0.015
15	NATAL-W	O	1.000	0.755	0.068	-0.0363	0.023	0.003	-0.2042	0.732	0.180

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	OYSTERCATCH	1	1.000	0.597	0.061	0.0576	0.065	0.008	0.1644	0.532	0.117
2	WF-PLOVER	2	1.000	0.781	0.057	0.1925	0.775	0.093	0.0164	0.006	0.001
3	K-PLOVER	3	1.000	0.917	0.063	-0.2183	0.898	0.120	0.0319	0.019	0.004
4	TB-PLOVER	4	1.000	0.597	0.033	-0.0736	0.198	0.014	0.1047	0.400	0.047
5	G-PLOVER	5	1.000	0.749	0.036	0.1316	0.581	0.044	-0.0707	0.168	0.022
6	R-PLOVER	6	1.000	0.319	0.017	0.0109	0.008	0.000	-0.0668	0.311	0.019
7	BT-GODWIT	7	1.000	0.511	0.061	-0.0840	0.137	0.018	0.1387	0.374	0.083
8	WHIMBREL	8	1.000	0.360	0.031	0.0752	0.215	0.014	0.0618	0.145	0.016
9	M-SANDPIPER	9	1.000	0.942	0.062	-0.1984	0.757	0.099	0.0983	0.186	0.042
10	GREENSHANK	:	1.000	0.319	0.009	-0.0142	0.027	0.001	-0.0469	0.292	0.009
11	C-SANDPIPER	;	1.000	0.360	0.028	0.0096	0.004	0.000	0.0906	0.356	0.035
12	TURNSTONE	<	1.000	0.940	0.075	0.2291	0.833	0.132	0.0822	0.107	0.029
13	KNOT	=	1.000	0.202	0.048	0.0197	0.010	0.001	0.0880	0.192	0.033
14	SANDERLING	>	1.000	0.765	0.098	0.2499	0.763	0.157	0.0125	0.002	0.001
15	LITTLESTINT	?	1.000	0.909	0.046	-0.0416	0.045	0.004	-0.1824	0.864	0.144
16	CURSANDPIPE	@	1.000	0.918	0.101	0.1612	0.308	0.065	-0.2268	0.610	0.222
17	RUFF	!	1.000	0.829	0.061	-0.1138	0.254	0.033	-0.1711	0.575	0.126
18	AVOCET	"	1.000	0.915	0.055	-0.1769	0.680	0.079	-0.1040	0.235	0.047
19	BW-STILT	#	1.000	0.956	0.059	-0.2166	0.947	0.118	-0.0207	0.009	0.002

FIGURE 8.4.6

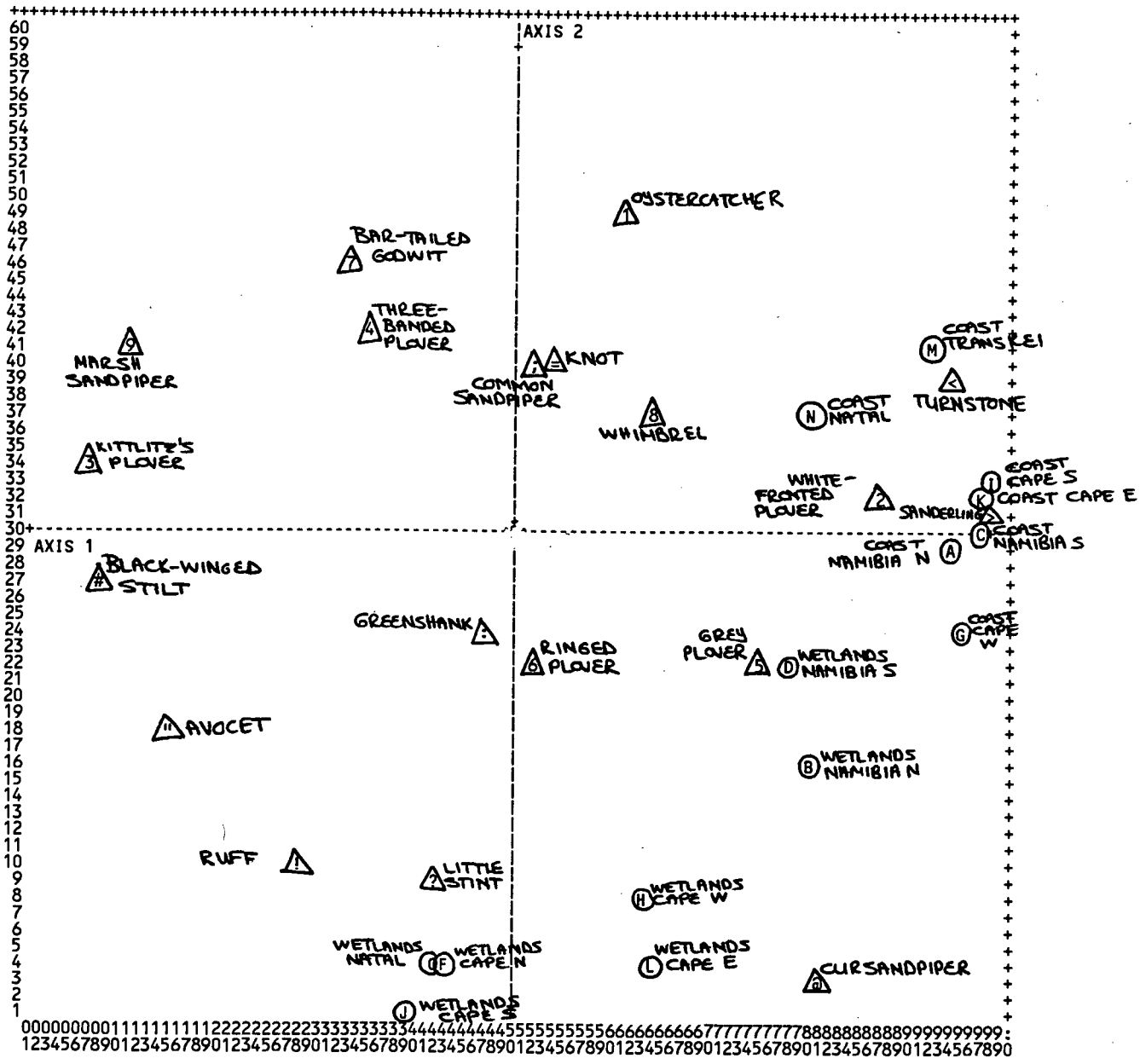

AXIS 2
AXIS 1
OYSTERCATCHER
BAR-TAILED GODWIT
THREE-BANDED PLOVER
MARSH SANDPIPER
KNOT
COMMON SANDPIPER
WHIMBREL
COAST TRANSKEI
COAST NATAL
TURNSTONE
KITTLITZ'S PLOVER
WHITE-FRONTED PLOVER
SANDERLING
COAST CAPE S
COAST CAPE E
COAST NAMIBIA S
COAST NAMIBIA N
BLACK-WINGED STILT
GREENSHANK
RINGED PLOVER
GREY PLOVER
WETLANDS NAMIBIA S
COAST CAPE W
AVOCET
WETLANDS NAMIBIA N
RUFF
LITTLE STINT
WETLANDS CAPE W
WETLANDS NATAL
WETLANDS CAPE N
WETLANDS CAPE E
CURSANDPIPER
WETLANDS CAPE S

FIGURE 8.4.6 (cont)

NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	NAMIB-N-C	A	0.2082	-0.0043	29	94
2	NAMIB-N-W	B	0.1427	-0.1111	16	80
3	NAMIB-S-C	C	0.2228	0.0040	30	97
4	NAMIB-S-W	D	0.1318	-0.0613	22	78
5	CAPE-N-C	E	0.2234	0.0011	30	98
6	CAPE-N-W	F	-0.0296	-0.2014	4	43
7	CAPE-W-C	G	0.2110	-0.0466	24	95
8	CAPE-W-W	H	0.0642	-0.1686	8	63
9	CAPE-S-C	I	0.2269	0.0225	33	98
10	CAPE-S-W	J	-0.0518	-0.2224	1	39
11	CAPE-E-C	K	0.2204	0.0133	32	97
12	CAPE-E-W	L	0.0678	-0.2042	4	64
13	TRANSKEI-C	M	0.1953	0.0869	41	92
14	NATAL-C	N	0.1432	0.0580	37	80
15	NATAL-W	O	-0.0363	-0.2042	4	42
1	OYSTERCATCH	1	0.0576	0.1644	49	61
2	WF-PLOVER	2	0.1925	0.0164	32	87
3	K-PLOVER	3	-0.2183	0.0319	34	7
4	TB-PLOVER	4	-0.0736	0.1047	42	35
5	G-PLOVER	5	0.1316	-0.0707	22	75
6	R-PLOVER	6	0.0109	-0.0668	22	52
7	BT-GODWIT	7	-0.0840	0.1387	46	33
8	WHIMBREL	8	0.0752	0.0618	37	64
9	M-SANDPIPER	9	-0.1984	0.0983	41	11
10	GREENSHANK	:	-0.0142	-0.0469	24	47
11	C-SANDPIPER	;	0.0096	0.0906	40	52
12	TURNSTONE	<	0.2291	0.0822	39	94
13	KNOT	=	0.0197	0.0880	40	54
14	SANDERLING	>	0.2499	0.0125	31	98
15	LITTLESTINT	?	-0.0416	-0.1824	9	42
16	CURSANDPIPE	@	0.1612	-0.2268	3	81
17	RUFF	!	-0.1138	-0.1711	10	28
18	AVOCET	"	-0.1769	-0.1040	18	15
19	BW-STILT	#	-0.2166	-0.0207	27	8

TABLE 8.4.7

THE EIGENVALUES

NUMBER	EIGENVALUE	PERCENT	CUMUL
1	2458.56519	47.981	47.981
2	881.54938	17.204	65.186
3	587.67645	11.469	76.655
4	399.87177	7.804	84.459
5	281.10397	5.486	89.945
6	137.74025	2.688	92.633
7	102.21026	1.995	94.628
8	80.37292	1.569	96.196
9	77.49905	1.512	97.709
10	47.35870	0.924	98.633
11	39.95340	0.780	99.413
12	19.20115	0.375	99.787
13	9.89691	0.193	99.980
14	1.00090	0.020	100.000

THE ROW OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	NAMIB-N- C	A	1.000	0.856	0.077	0.1238	0.389	0.062	-0.1358	0.467	0.209
2	NAMIB-N-W	B	1.000	0.873	0.047	-0.0412	0.071	0.007	-0.1382	0.801	0.217
3	NAMIB-S-C	C	1.000	0.383	0.056	0.1045	0.383	0.044	0.0019	0.000	0.000
4	NAMIB-S-W	D	1.000	0.097	0.072	-0.0343	0.032	0.005	0.0490	0.065	0.027
5	CAPE-N-C	E	1.000	0.749	0.043	0.1110	0.558	0.050	-0.0649	0.191	0.048
6	CAPE-N-W	F	1.000	0.592	0.101	-0.1733	0.581	0.122	0.0238	0.011	0.006
7	CAPE-W-C	G	1.000	0.334	0.032	0.0694	0.298	0.020	-0.0241	0.036	0.007
8	CAPE-W-W	H	1.000	0.857	0.067	-0.1285	0.479	0.067	-0.1140	0.378	0.147
9	CAPE-S-C	I	1.000	0.781	0.055	0.1090	0.419	0.048	0.1012	0.361	0.116
10	CAPE-S-W	J	1.000	0.925	0.077	-0.1895	0.911	0.146	0.0230	0.013	0.006
11	CAPE-E-C	K	1.000	0.504	0.043	0.1027	0.484	0.043	0.0210	0.020	0.005
12	CAPE-E-W	L	1.000	0.687	0.080	-0.1525	0.570	0.095	0.0690	0.117	0.054
13	TRANSKEI-C	M	1.000	0.752	0.072	0.1593	0.686	0.103	0.0496	0.066	0.028
14	NATAL-C	N	1.000	0.526	0.089	0.1186	0.309	0.057	0.0994	0.217	0.112
15	NATAL-W	O	1.000	0.722	0.091	-0.1789	0.688	0.130	0.0393	0.033	0.017

THE COLUMN OBJECTS

NO	NAME	SYMBOL	MASS	QUAL	INRT	FACT 1	COR	CTR	FACT 2	COR	CTR
1	OYSTERCATCH	1	1.000	0.181	0.054	0.0658	0.156	0.018	0.0264	0.025	0.008
2	WF-PLOVER	2	1.000	0.883	0.055	0.1427	0.727	0.083	0.0662	0.156	0.050
3	K-PLOVER	3	1.000	0.861	0.043	-0.1364	0.846	0.076	-0.0180	0.015	0.004
4	TB-PLOVER	4	1.000	0.394	0.054	-0.0850	0.259	0.029	0.0614	0.135	0.043
5	G-PLOVER	5	1.000	0.120	0.055	-0.0223	0.018	0.002	-0.0534	0.102	0.032
6	R-PLOVER	6	1.000	0.839	0.055	-0.1208	0.521	0.059	0.0942	0.317	0.101
7	BT-GODWIT	7	1.000	0.289	0.054	-0.0432	0.067	0.008	-0.0784	0.222	0.070
8	WHIMBREL	8	1.000	0.538	0.055	0.0125	0.006	0.001	0.1221	0.533	0.169
9	M-SANDPIPER	9	1.000	0.713	0.049	-0.1333	0.706	0.072	0.0134	0.007	0.002
10	GREENSHANK	:	1.000	0.869	0.055	-0.1251	0.559	0.064	0.0932	0.310	0.099
11	C-SANDPIPER	;	1.000	0.751	0.055	0.0067	0.002	0.000	0.1448	0.749	0.238
12	TURNSTONE	<	1.000	0.766	0.055	0.1464	0.765	0.087	-0.0049	0.001	0.000
13	KNOT	=	1.000	0.464	0.054	0.0202	0.015	0.002	-0.1117	0.449	0.142
14	SANDERLING	>	1.000	0.652	0.055	0.1340	0.642	0.073	-0.0169	0.010	0.003
15	LITTLE-STINT	?	1.000	0.870	0.054	-0.1548	0.859	0.097	-0.0177	0.011	0.004
16	CURSANDPIPER	@	1.000	0.812	0.055	-0.1427	0.727	0.083	-0.0488	0.085	0.027
17	RUFF	!	1.000	0.843	0.053	-0.1507	0.841	0.092	-0.0062	0.001	0.000
18	AVOCET	"	1.000	0.780	0.049	-0.1372	0.747	0.077	-0.0288	0.033	0.009
19	BW-STILT	#	1.000	0.868	0.043	-0.1382	0.868	0.078	0.0001	0.000	0.000

FIGURE 8.4.7

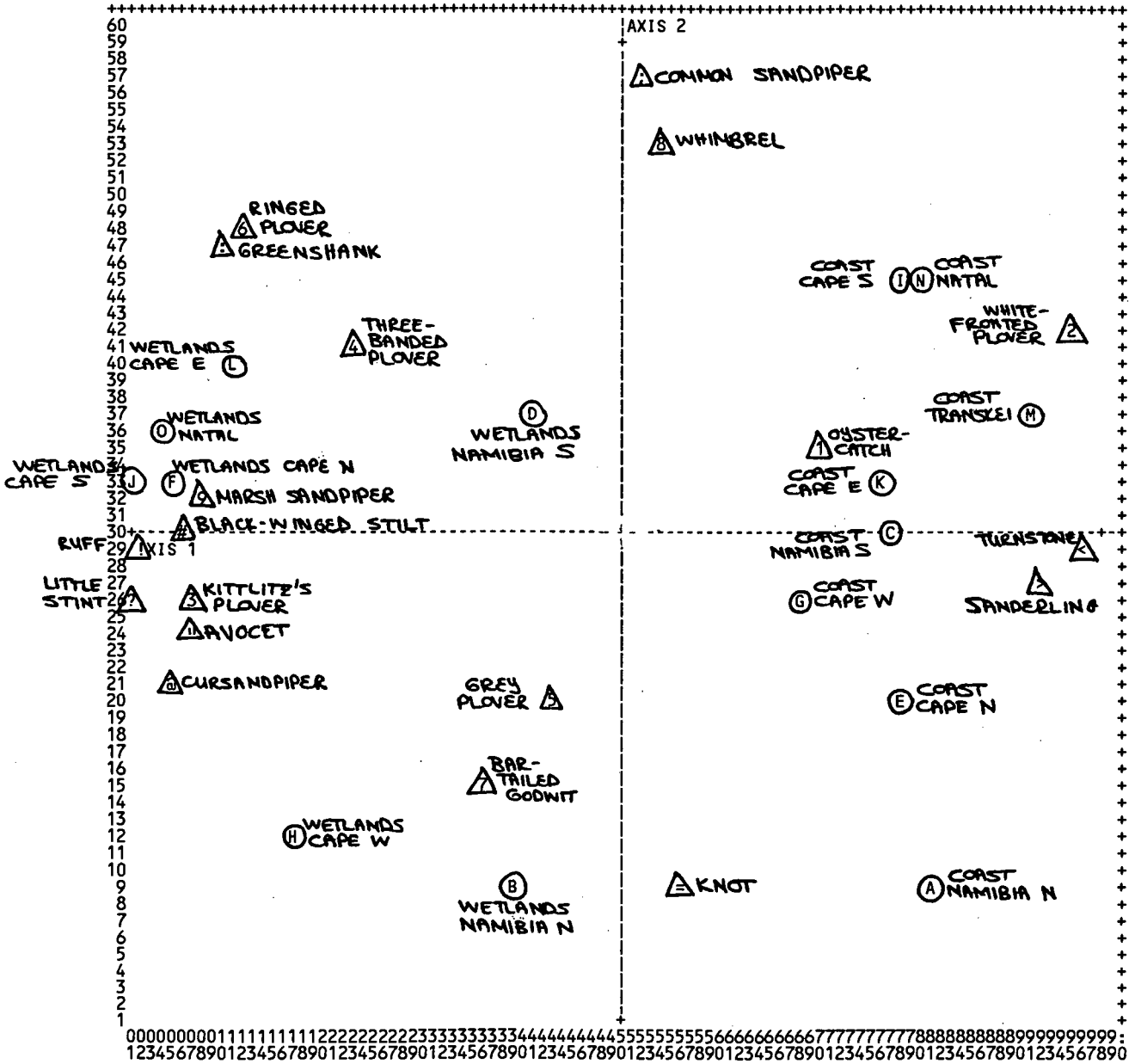

AXIS 2
AXIS 1
COMMON SANDPIPER
WHIMBREL
RINGED PLOVER
GREENSHANK
COAST CAPE S
COAST NATAL
WHITE-FRONTED PLOVER
THREE-BANDED PLOVER
WETLANDS CAPE E
WETLANDS NATAL
WETLANDS NAMIBIA S
COAST TRANSKEI
OYSTER-CATCH
WETLANDS CAPE S
WETLANDS CAPE N
MARSH SANDPIPER
COAST CAPE E
BLACK-WINGED STILT
RUFF
COAST NAMIBIA S
TURNSTONE
LITTLE STINT
KITTLITZ'S PLOVER
COAST CAPE W
SANDERLING
AVOCET
CURSANDPIPER
GREY PLOVER
COAST CAPE N
BAR-TAILED GODWIT
WETLANDS CAPE W
KNOT
COAST NAMIBIA N
WETLANDS NAMIBIA N

FIGURE 8.4.7 (cont)

NUMBER	NAME	SYMBOL	CO-ORDINATES:(X1,X2)		PLOT POSITIONS	
1	NAMIB-N-C	A	0.1238	-0.1358	9	81
2	NAMIB-N-W	B	-0.0412	-0.1382	9	39
3	NAMIB-S-C	C	0.1045	0.0019	30	77
4	NAMIB-S-W	D	-0.0343	0.0490	37	41
5	CAPE-N-C	E	0.1110	-0.0649	20	78
6	CAPE-N-W	F	-0.1733	0.0238	33	5
7	CAPE-W-C	G	0.0694	-0.0241	26	68
8	CAPE-W-W	H	-0.1285	-0.1140	12	17
9	CAPE-S-C	I	0.1090	0.1012	45	78
10	CAPE-S-W	J	-0.1895	0.0230	33	1
11	CAPE-E-C	K	0.1027	0.0210	33	76
12	CAPE-E-W	L	-0.1525	0.0690	40	11
13	TRANSKEI-C	M	0.1593	0.0496	37	91
14	NATAL-C	N	0.1186	0.0994	45	80
15	NATAL-W	O	-0.1789	0.0393	36	4
1	OYSTERCATCH	1	0.0658	0.0264	35	70
2	WF-PLOVER	2	0.1427	0.0662	42	95
3	K-PLOVER	3	-0.1364	-0.0180	26	7
4	TB-PLOVER	4	-0.0850	0.0614	41	23
5	G-PLOVER	5	-0.0223	-0.0534	20	43
6	R-PLOVER	6	-0.1208	0.0942	48	12
7	BT-GODWIT	7	-0.0432	-0.0784	15	36
8	WHIMBREL	8	0.0125	0.1221	53	54
9	M-SANDPIPER	9	-0.1333	0.0134	32	8
10	GREENSHANK	:	-0.1251	0.0932	47	10
11	C-SANDPIPER	;	0.0067	0.1448	57	52
12	TURNSTONE	<	0.1464	-0.0049	29	96
13	KNOT	=	0.0202	-0.1117	9	56
14	SANDERLING	>	0.1340	-0.0169	27	92
15	LITTLE-STINT	?	-0.1548	-0.0177	26	1
16	CURSANDPIPER	@	-0.1427	-0.0488	21	5
17	RUFF	!	-0.1507	-0.0062	29	2
18	AVOCET	"	-0.1372	-0.0288	24	7
19	BW-STILT	#	-0.1382	0.0001	30	6

9

CONCLUSIONS

This chapter addresses some issues in the selection and evaluation of biplot techniques as exploratory analyses of data matrices.

If we consider the possible combinations of choices involved in Phases I, II and III (as described in Section 2.9), there is a great variety of possible biplots. We have seen in the examples that these choices can have major impacts on the visual impression, and that therefore the selection of methods is an important consideration. Greenacre (1984) details the choice of Phases resulting in the more widely used biplots, and outlines the interpretations from these.

The problem is to decide on which plots, or series of plots to use in a particular situation. We consider the three aspects of the problem: choice of family (Phase III) (Section 9.1), choice of transformation and weightings (Phases I and II) (Section 9.2), and the overall quality, or suitability of the display (Section 9.3). These aspects are interrelated. The question of the choice and evaluation of a display has not enjoyed much attention in the literature.

We will be drawing both on the theory from the earlier chapters and on the practical applications in Chapter 8.

9.1 Choice of Family (Phase III)

The family to which a biplot belongs can be fully specified by the choice of a and b in Phase III (see Section 3.5). The family determines which two of approximations (i), (ii)

and (iii) (Section 3.5) can be made from the plot as well as which decompositions of the norm are applicable (Section 3.5).

Consider the preprocessed matrix Z with given row and column weights, having the singular value decomposition $Z=U\Gamma V^T$.

Note that the choice of family does not affect the quality of approximation of the display. The $Z_{[p]}$ matrix is obtained in Phase II and does not depend on the choices of a and b made in Phase III. It depends only on the choices made in Phases I and II. The percentage of the squared norm accounted for in each dimension and hence the quality of approximation is independent of the choices of a and b.

A principal components (1-0) and covariance (0-1) biplot of the preprocessed matrix Z are in fact not as different as they appear to be. Firstly,the (1-0)-plot of Z is exactly the same as a (0-1)-plot of its transpose. Also, the row coordinates in a (1-0)-plot of Z are given by $F=U\Gamma$. For a correlation biplot the coordinates are $F=U$. Post multiplication of a matrix by a diagonal matrix results in a 'stretch' of each of its columns by the factors in the corresponding diagonal entry, the singular value α_i in this case. Thus the coordinates for the row points on the first axis, for example are greater by a factor of α_1 in the (1-0) than in the (0-1)-plot. This means that in each dimension, the difference between the coordinates for a row point of the (1-0) and (0-1) biplots with the same choice of Phases I and II is that one point is a multiple of the other. The row points in the (1-0)-plot are 'stretched' by the α_i in each dimension.

Similarly, the column points in a correlation biplot of X are equivalent to the column points of a principal component biplot stretched by a factor of α_i in dimension i.

This differential stretching of points was illustrated in the Pollution example (Example 8.1), where biplots that differed only in their Phase III choice were applied to a data set (Figures 8.1.6 and 8.1.7).

Plots in the (1-1) family (correspondence analysis family) have the row characteristics of

(1-0)-plots and the column characteristics of (0-1)-plots. (Examples where (1-1)-plots are applied are 8.1.3, 8.1.5 and 8.4.4.)

Choice of family depends on the interpretations required. The (1-1)-plots have both within set interpretations ((ii) and (iii) of Section 3.5), but loss of the scalar product between set interpretation. However, there is a between set interpretation in terms of the transition formulae (see Section 6.3.2). Both row and column norm decompositions are valid. Despite the advantages of (1-1)-plots, they are rarely made use of other than in correspondence analysis.

We leave the summary of the above discussion to Gower (1984) who notes that (1-0)-plots deal adequately with the points representing the rows but poorly with the points representing the columns, while the opposite is true for (0,1)-plots and that the best of both methods is retained in (1-1)-plots.

9.2 Choice of Centre and Weights (Phases I and II)

Preprocessing of the data matrix is discussed in Chapter 7.1. Comparison of members of the correlation biplot family in Section 4.6, is a comparison of choice of centre and the same arguments are applicable to the other families. The effects of these choices is illustrated in Example 8.2.

Choice of centre depends on the desired emphasis of the plot. Correspondence analysis, for example, emphasizes profiles. Some general conclusions were drawn in the practical examples in Chapter 8. For example, when there are widely differing scales of measure (Example 8.2) some standardization is needed. Correspondence Analysis type centrings are sensitive to sparse vectors (Example 8.4).

Weightings increase or decrease the relative importance of rows and columns in the analysis.

An important reason for centring is to include the origin in the cloud of points, as illustrated by the different results obtained in Examples 8.1.1 and 8.1.2.

However, the effect of the preprocessing on subsequent interpretations should also be borne in mind. Some transformations have a disadvantage in that they complicate interpretations.

A further reason for preprocessing is that approximated distances in the plot should be meaningful in some sense. For this to hold, the rows or columns should be comparable. Comparability can be achieved through transformations, for example by standardizing the variables, or by weighting.

The choices of centre and weights ultimately depend on the aspects of the data set that need to be highlighted in a given situation. However, the choice is inextricably bound up with the issue of the quality of the display, as discussed in the next section.

9.3 Quality of the Plot

The measure that is usually used to assess the quality of the low rank approximation is the percentage of the squared norm (variation) of the full rank matrix that is approximated in p dimensions as discussed in Section 2.7. This is referred to as the 'quality of the approximation'. The terminology 'quality of the approximation' is perhaps misleading as the word 'quality' brings to mind other desirable criteria of the display, such as whether it is a good representation of the features of the data set and the amount of information that it conveys.

For example, Lebart *et al* (1984) point out that the percentage of variance explained is a conservative measure of the amount of information represented. If some of the vectors of the original matrix are 'random noise' the variation attributable to them is still considered as part of the overall variation.

A display with a good quality of approximation (percentage variation explained) is not necessarily a good display in a broader sense. This point is illustrated in the practical examples. For example, as shown in Example 8.3.3, where a high percentage of the squared norm can be retained but only one variable has a high display quality.

This serves to illustrate that the proportion of variance explained should be considered in conjunction with other factors, such as those that follow.

Firstly, if the original matrix has many rows or columns, in general a poorer percentage representation is expected in few dimensions than for a matrix that has few rows or columns.

The structure of the matrix is important. A better quality of approximation is obtained if the vectors of Z have a close linear relationship, exhibiting a high degree of multicollinearity. Clearly, such variables are easier to represent in a low number of dimensions.

Lastly, the number of variables contributing to the variation on each axis is important. For example, a two dimensional approximation that preserves most of the overall variation is not of much use if this variation in almost entirely contributed by one variable. This variable is effectively dominating the plot. The essence of the data is not being represented. A large percentage of the total variation accounted for is not remarkable if only a few variables contribute. The p variables comprising the approximation could be perfectly represented in at most p dimensions anyway.

A factor which determines the relative contributions that the variables and individuals make to the approximation is the relative magnitude of their scales of measurement. The effect of the choice of centring and weighting of the data is crucial here (as illustrated in Example 8.2).

Even if there is a high quality of display, there could be only a few points that contribute to this. Then the plot is effectively a plot of those points. Therefore the quality of

REFERENCES

AFFLECK-GRAVES, J.F., TROSKIE, C.G. and MONEY, A.H. (1979). A principal component index subject to constraints. Investment Analysts Journal, 14, 45-50.

ANDERSON, T.W. (1958). An Introduction to Multivariate Statistical Analysis, Wiley, New York.

BARNETT, V. (1981). Interpreting Multivariate Data. Wiley, Chichester.

BARR, G.D.I. and AFFLECK-GRAVES, J.F. (1987). The covariance biplot and stock market data: an alternative relative strength chart. South African Journal of Business Management, 18, 46-50.

BARR, G.D.I., KANTOR, B.S. and UNDERHILL, L.G. (1987). The weighted covariance biplot - an application. South African Statistical Journal, 21, 155-171.

BARR, G.D.I., UNDERHILL, L.G. and KAHN, B.S. (1990). The covariance biplot as a graphical display technique for multivariate time series data. American Journal of Mathematical and Management Sciences, 10, 1-15.

BARR, G.D.I. (1990). Macroeconomic identification of the pricing factors on the Johannesburg Stock Exchange. South African Journal of Business Management, 21, 17-26.

BRADU, D. and GABRIEL, K.R. (1978). The biplot as a diagnostic tool for models of two-way tables. Technometrics, 20, 47-68.

BREBNOR, L., BRADU, D. and SCHNEIDER, J. (1977). A model for data from a biological (cancer) research project. National Research Institute for Mathematical Sciences, CSIR Special Report, Pretoria.

representation of the individual rows and columns needs to be taken into account in the interpretation. Even when the overall approximation is good there may be some points that are not adequately represented. Confirmation of all interpretations should be made by referring back to the original data matrix.

The above situation can occur when there are substantial differences in magnitude amongst the matrix entries, for example when the scales of measurement are very different or when outliers are present. In fact, biplots are useful in the detection of outliers (Example 8.4).

Thus the percentage variation explained can give a misleading measure of the overall quality of the display. Measures of the quality of display of individual features of the plot could be developed, so that diagnostics other than percentage variation explained could be used to assess the display quality. The difference between actual and approximated features of the matrix can be directly compared. For example, true versus approximated Euclidean distances, scalar products and correlations can be compared.

In conclusion, evaluation of display quality is not as clear cut as merely looking at percentage variation explained (as is often done in practice). Each data set should be subjected to a series of plots, all of which should be carefully examined in the light of Sections 9.1, 9.2 and 9.3.

Bearing in mind that the ultimate goal of biplot techniques is to allow visual exploration of the data, there is an important subjective component in evaluating the display quality. The personal preference of the researcher will determine the displays that are actually presented.

GOWER, J.C. (1968). Adding a point to vector diagrams in multivariate analysis. Biometrika, 55, 582-585.

GOWER, J.C. (1984). Multivariate analysis: ordination, multidimensional scaling and allied topics, In Handbook of Applicable Mathematics, (W. Lederman, chief ed.) Vol. VI, Statistics, Part B, (E. Lloyd, ed). 727-781. Wiley, Chichester.

GOWER, J.C. and DIGBY, P.G.N. (1981). Expressing complex relationships in two dimensions. *In* Interpreting Multivariate Data (Barnett, V. ed), 119-146. Wiley, Chichester.

GOWER, J.C. and HARDING, S.A. (1988). Nonlinear biplots. Biometrika, 75, 3, 445-455.

GREEN, P.E. and CARROL, J.D. (1976). Mathematical Tools for Applied Multivariate Analysis. Academic Press, London.

GREENACRE, M.J. (1980). Basic Structure Displays of a Data Matrix. University of South Africa Research Report.

GREENACRE, M.J. (1981). Practical Correspondence Analysis. *In* Interpreting Multivariate Data (Barnett, V. ed), 119-146, Wiley, Chichester.

GREENACRE, M.J. (1984). Theory and Application of Correspondence Analysis. Academic Press Inc., London.

GREENACRE, M.J. (1987). The Carroll-Green-Schaffer scaling in correspondence analysis: A theoretical and empirical appraisal. University of South Africa Research Report.

GREENACRE, M.J. (1988). Correspondence analysis of multivariate categorical data by weighted least-squares. Biometrika, 75, 457-67.

GREENACRE, M.J. (1988). Some limitations of multiple correspondence analysis. Paper presented at MULTIWAY '88, International Meeting on the Analysis of Multiway Matrices, Rome, Italy.

GREENACRE, M.J. (1988). Clustering the Rows and Columns of a Contingency Table. Journal of Classification, 5, 39-51.

GREENACRE, M.J. and HASTIE, T. (1987). The geometric interpretation of correspondence analysis. Journal of the American Statistical Association, 82, 437-447.

GREENACRE, M.J. and UNDERHILL, L.G. (1982). Scaling a data matrix in a low-dimensional euclidean space. *In* Topics in Applied Multivariate Analysis (Hawkins, D.M. ed), 183-287, Cambridge University Press, Cambridge.

HAWKINS, D.M. and FATTI, L.P. (1983). The information in the minor principal components. Proceedings of the Seminar on Principal Components Analysis in the Atmospheric and Earth Sciences. CSIR. Pretoria, South Africa.

JOHNSTON, D.F. (1988). Toward a comprehensive 'quality-of-life' index. Social Indicators Research, 20, 473-496.

JOLLIFFE, I.T. (1986). Principal Component Analysis. Springer Verlag, New York.

LEBART, L., MORINEAU, A. and WARWICK, K.M. (1984). Multivariate Descriptive Statistical Analysis. Wiley, New York.

MANDEL, J. (1982). Use of the singular value decomposition in regression analysis. American Statistician, 36, 15-24.

NISHISATO, S. (1980). Analysis of Categorical Data: Dual Scaling and its Applications. University of Toronto Press, Canada.

NOAKES, T.D., MYBURGH, K.H., DU PLESSIS, J., LANG, L., LAMBERT, M., VAN DER RIET, C. and SCHALL, R. (1988). Metabolic rate not percent dehydration predicts rectal temperature in marathon runners. University of Cape Town Medical School, Cape Town.

ORLOCI, L. (1978). Multivariate Analysis in Vegetation Research Dr W. Junk, The Hague.

OSMOND, C. (1985). Biplot models applied to cancer mortality rates. Applied Statistics, 34, 63-70.

PRESS, W.H., FLANNERY, B.P., TEUKOLSKY, S.A. and VETTERLING, W.T. (1986). Numerical Recipes. Cambridge University Press, Cambridge.

SCHOMER, H.H. and DUNNE, T.T. (1986). Affective involvement in the transition to a physically active lifestyle. South African Journal for Research in Sport, Physical Education and Recreation, 9, 61-71.

SHAHIM, V. and GREENACRE, M.J. (1988). Brand maps - ideal points and market gaps. Procedings of the South African Marketing Research Association 10th convention. Johannesburg.

STEFFENS, F.E. (1983). What is principal components analysis? Proceedings of the Seminar on Principal Components Analysis in the Atmospheric and Earth Sciences. CSIR. Pretoria, South Africa.

SUMMERS, R.W., UNDERHILL, L.G., PEARSON, D.J. and SCOTT, D.A. (1987). Wader migration systems in southern and eastern Africa and western Asia. Wader Study Group Bulletin, 49, Supplement, 15-34.

TER BRAAK, C.J.F. (1983). Principal components biplots and alpha and beta diversity. Ecology, 64, 454-462.

TUKEY, J.W. (1977). Exploratory Data Analysis. Addison-Wesley, Reading, Massachusetts.

UNDERHILL, L.G. and PEISACH, M. (1985). Correspondence analysis and its application in multielemental trace analysis. Journal of Trace and Microprobe Techniques, 3, 41-65.

UNDERHILL, L.G. (1990a). The coefficient of variation biplot. Journal of Classification, 7, 241-256.

UNDERHILL, L.G. (1990b). Guide to the use of SVDD, the Singular Value Decomposition Display Program. Department of Mathematical Statistics, University of Cape Town.

UNDERHILL, L.G. (in press). The biplot as a graphical method for evaluating multivariate ecological monitoring data. Journal of Applied Ecology.

VAN DEN HONERT, R.C. and BARR, G.D.I. (1988). Simultaneous representations of explanatory characteristics of mergers. South African Journal of Business Management, 19, 161-169.

WANDT, M.A.E. and UNDERHILL, L.G. (1988). Covariance Biplot Analysis of Trace Element Concentrations in Urinary Stones. British Journal of Urology, 61, 474-481.

YOUNG, F.W. (1989). Visualizing Six-Dimensional Structure with Dynamic Statistical Graphics. Chance, 2, 22-30.

YOUNG, G. and HOUSEHOLDER, A.S. (1938). Discussion of a set of points in terms of their mutual distances. Psychometrika, 3, 19-22.