Appl. Statist. (2002)
51, *Part* 4, *pp.* 375–392

Biplots of compositional data

John Aitchison

University of Glasgow, UK

and Michael Greenacre

Universitat Pompeu Fabra, Barcelona, Spain

[Received June 2001. Final revision May 2002]

Summary. The singular value decomposition and its interpretation as a linear biplot have proved to be a powerful tool for analysing many forms of multivariate data. Here we adapt biplot methodology to the specific case of compositional data consisting of positive vectors each of which is constrained to have unit sum. These relative variation biplots have properties relating to the special features of compositional data: the study of ratios, subcompositions and models of compositional relationships. The methodology is applied to a data set consisting of six-part colour compositions in 22 abstract paintings, showing how the singular value decomposition can achieve an accurate biplot of the colour ratios and how possible models interrelating the colours can be diagnosed.

Keywords: Log-ratio transformation; Principal component analysis; Relative variation biplot; Singular value decomposition; Subcomposition

1. Introduction

Compositional data (Aitchison, 1986) consist of vectors of positive values summing to a unit, or in general to some fixed constant for all vectors. Such data arise in many disciplines, e.g. in geology as major oxide compositions of rocks, in sociology and psychology as time budgets, i.e. parts of a time period allocated to various activities, in politics as proportions of the electorate voting for different political parties and in genetics as frequencies of genetic groups within populations. The biplot (Gabriel, 1971) is a method which has been regularly applied to visualize the rows and columns of many different kinds of data matrices. In almost all cases, the original data values require transforming to depict correctly the structures that are appropriate to the particular nature of the data. Compositional data are also special in this respect and a careful consideration of the relationships between parts of a composition is required before we embark on applying biplot methodology to such data.

We consider the data of Table 1, showing six-part colour compositions in 22 paintings created for teaching purposes. Each painting was divided into a number of rectangles, in the style of a Mondrian abstract painting, and the rectangles were each coloured in one of six colours: black, white, blue, red, yellow and one further colour, labelled 'other', which varies from painting to painting. The data are the proportions of surface area occupied by the six colours. For example, the first painting has 12.5% of the area in black, 24.3% in white, and so on. One of the questions that was posed to the students was to orientate the pictures in

Address for correspondence: Michael Greenacre, Department of Economics and Business, Universitat Pompeu Fabra, Ramon Trias Fargas 25–27, 08005 Barcelona, Spain.
E-mail: michael@upf.es

 0035–9254/02/51375

Table 1. Colour composition data for 22 abstract paintings

Painting	*Proportions of area occupied by the following colours:*					
	Black	*White*	*Blue*	*Red*	*Yellow*	*Other*
1	0.125	0.243	0.153	0.031	0.181	0.266
2	0.143	0.224	0.111	0.051	0.159	0.313
3	0.147	0.231	0.058	0.129	0.133	0.303
4	0.164	0.209	0.120	0.047	0.178	0.282
5	0.197	0.151	0.132	0.033	0.188	0.299
6	0.157	0.256	0.072	0.116	0.153	0.246
7	0.153	0.232	0.101	0.062	0.170	0.282
8	0.115	0.249	0.176	0.025	0.176	0.259
9	0.178	0.167	0.048	0.143	0.118	0.347
10	0.164	0.183	0.158	0.027	0.186	0.281
11	0.175	0.211	0.070	0.104	0.157	0.283
12	0.168	0.192	0.120	0.044	0.171	0.305
13	0.155	0.251	0.091	0.085	0.161	0.257
14	0.126	0.273	0.045	0.156	0.131	0.269
15	0.199	0.170	0.080	0.076	0.158	0.318
16	0.163	0.196	0.107	0.054	0.144	0.335
17	0.136	0.185	0.162	0.020	0.193	0.304
18	0.184	0.152	0.110	0.039	0.165	0.350
19	0.169	0.207	0.111	0.057	0.156	0.300
20	0.146	0.240	0.141	0.038	0.184	0.250
21	0.200	0.172	0.059	0.120	0.136	0.313
22	0.135	0.225	0.217	0.019	0.187	0.217

the same way as the artist. The results of this experiment showed that successful orientation followed a binomial distribution with success probability $\frac{1}{4}$, a result that was in fact replicated with real Mondrian paintings. Another question was to have the students estimate the proportions of each colour, both to illustrate variability of estimation of proportions and the nature of compositional variability. Our present interest in these data, however, is to see whether any pattern is discernible in the construction of these paintings. There is considerable variation from painting to painting in their colour compositions and the challenge is to describe the patterns of variability appropriately in simple terms while maintaining the unit sum constraint that is inherent in the data. An important aspect is how to treat so-called *subcompositions*; for example if the analysis is restricted to the three primary colours then the results should be consistent with those obtained for these three colours when analysing the full composition.

In Section 2 we define the linear biplot and briefly summarize some known results which will be relevant to its application to compositional data. In Section 3 we discuss what makes compositional data different from interval- or ratio-scaled measurements and how to transform such data to perform what we call a relative variation biplot. In Section 4 we apply the relative variation biplot to the colour composition data and discuss issues of interpretation and modelling. Section 5 concludes with a discussion and comparison with methods such as regular principal component analysis and correspondence analysis. The data that are analysed in the paper can be obtained from

`http://www.blackwellpublishers.co.uk/rss/`

2. Biplots

A biplot is a graphical display of the rows and columns of a rectangular $n \times p$ data matrix $\mathbf{X}$, where the rows are often individuals or other sample units and the columns are variables. In almost all applications, biplot analysis starts with performing some transformation on $\mathbf{X}$, depending on the nature of the data, to obtain a transformed matrix $\mathbf{Z}$ which is the matrix that is actually displayed. Examples of transformations are centring with respect to the overall mean, centring with respect to variable means, normalization of variables, square root and logarithmic transforms.

Suppose that the transformed data matrix $\mathbf{Z}$ has rank r. Then $\mathbf{Z}$ can be factorized as the product

$$\mathbf{Z} = \mathbf{F}\mathbf{G}^{\mathrm{T}}, \tag{1}$$

where $\mathbf{F}$ is $n \times r$ and $\mathbf{G}$ is $p \times r$. The rows of $\mathbf{F}$ and the rows of $\mathbf{G}$ provide the co-ordinates of n points for the rows and p points for the columns in an r-dimensional Euclidean space, called the *full space* since it has as many dimensions as the rank of $\mathbf{Z}$. This joint plot of the two sets of points can be referred to as the exact biplot in the full space. There are an infinite number of ways to choose $\mathbf{F}$ and $\mathbf{G}$, and certain choices favour the display of the rows; others the display of the columns. For any particular choice, however, the biplot in r dimensions has the property that the scalar product between the ith row point and jth column point with respect to the origin is equal to the (i, j)th element z_{ij} of $\mathbf{Z}$.

We are mainly interested in low dimensional biplots of $\mathbf{Z}$, especially in two dimensions, and these can be conveniently achieved by using the singular value decomposition (SVD) of $\mathbf{Z}$:

$$\mathbf{Z} = \mathbf{U}\boldsymbol{\Gamma}\mathbf{V}^{\mathrm{T}}, \tag{2}$$

where $\mathbf{U}$ and $\mathbf{V}$ are the matrices of left and right singular vectors, each with r orthonormal columns, and $\boldsymbol{\Gamma}$ is the diagonal matrix of positive singular values in decreasing order of magnitude: $\gamma_1 \geqslant \ldots \geqslant \gamma_r > 0$. The Eckart–Young theorem (Eckart and Young, 1936) states that if one calculates the $n \times p$ matrix $\hat{\mathbf{Z}}$ using the first r^* singular values and corresponding singular vectors, e.g. for $r^* = 2$,

$$\hat{\mathbf{Z}} = (\,\mathbf{u}_1 \quad \mathbf{u}_2\,) \begin{pmatrix} \gamma_1 & 0 \\ 0 & \gamma_2 \end{pmatrix} (\,\mathbf{v}_1 \quad \mathbf{v}_2\,)^{\mathrm{T}} \tag{3}$$

then $\hat{\mathbf{Z}}$ is the least squares rank r^* matrix approximation of $\mathbf{Z}$, i.e. $\hat{\mathbf{Z}}$ minimizes the fit criterion

$$\|\mathbf{Z} - \mathbf{Y}\|^2 = \sum_i \sum_j (z_{ij} - y_{ij})^2$$

over all possible matrices $\mathbf{Y}$ of rank r^*, where $\|\ldots\|$ denotes the Frobenius matrix norm. It is this approximate matrix $\hat{\mathbf{Z}}$ which is biplotted in the lower r^*-dimensional space, called the *reduced space*. This biplot will be as accurate as is the approximation of $\hat{\mathbf{Z}}$ to $\mathbf{Z}$. The sum of squares of $\mathbf{Z}$ decomposes into two parts: $\|\mathbf{Z}\|^2 = \|\hat{\mathbf{Z}}\|^2 + \|\mathbf{Z} - \hat{\mathbf{Z}}\|^2$, where $\|\hat{\mathbf{Z}}\|^2 = \gamma_1^2 + \ldots + \gamma_{r^*}^2$, and $\|\mathbf{Z} - \hat{\mathbf{Z}}\|^2 = \gamma_{r^*+1}^2 + \ldots + \gamma_r^2$ and the goodness of fit is measured by the proportion of explained sum of squares $(\gamma_1^2 + \ldots + \gamma_{r^*}^2)/(\gamma_1^2 + \ldots + \gamma_r^2)$, usually expressed as a percentage.

The SVD also provides a decomposition which is a natural choice for the biplot. For example, from equation (3) in two dimensions. $\hat{\mathbf{Z}} = \mathbf{F}\mathbf{G}^{\mathrm{T}}$ with

$$\begin{aligned} \mathbf{F} &= (\,\gamma_1^{\alpha}\mathbf{u}_1 \quad \gamma_2^{\alpha}\mathbf{u}_2\,), \\ \mathbf{G} &= (\,\gamma_1^{1-\alpha}\mathbf{v}_1 \quad \gamma_2^{1-\alpha}\mathbf{v}_2\,) \end{aligned} \tag{4}$$

for some constant α. The most common choices of α are the values 1 or 0, when the singular values are assigned entirely either to the left singular vectors of $\mathbf{U}$ or to the right singular vectors of $\mathbf{V}$ respectively, or 0.5 when the square roots of the singular values are split equally between left and right singular vectors. Each choice, while giving exactly the same matrix approximation, will highlight a different aspect of the data matrix. The term *principal co-ordinates* refers to the singular vectors scaled by the singular values (e.g. $\mathbf{F}$ with $\alpha = 1$ or $\mathbf{G}$ with $\alpha = 0$), whereas *standard co-ordinates* are the unscaled singular vectors (Greenacre, 1984).

The most common biplot is of an individuals-by-variables data matrix $\mathbf{X}$ that has been transformed by centring with respect to column means $\bar{x}_j$:

$$z_{ij} = x_{ij} - \bar{x}_j. \tag{5}$$

Optionally, if normalization of the variables is required, there can be a further division of each column of the matrix by s_j, the estimated standard deviation of the jth variable: $z_{ij} = (x_{ij} - \bar{x}_j)/s_j$.

After calculating the SVD of $\mathbf{Z}$, the co-ordinate matrices $\mathbf{F}$ and $\mathbf{G}$ are calculated as in equations (4) by using either

(a) $\alpha = 1$, i.e. rows in principal co-ordinates and columns in standard co-ordinates, called the *form biplot*, which favours the display of the individuals (see below), or
(b) $\alpha = 0$, i.e. rows in standard co-ordinates and columns in principal co-ordinates, called the *covariance biplot*, which favours the display of the variables (Greenacre and Underhill, 1982).

The alternative solutions differ only by scale changes along the horizontal and vertical axes of the biplot (see Figs 2 and 3 in Section 4). In either biplot we conventionally depict the variables by *rays* emanating from the origin, since both their lengths and directions are important to the interpretation.

The covariance biplot is characterized by the least squares approximation of the covariance matrix $\mathbf{S} = \mathbf{Z}^{\mathrm{T}}\mathbf{Z}/(n-1)$ by $\mathbf{G}\mathbf{G}^{\mathrm{T}}/(n-1)$, the matrix of scalar products between the row vectors of $\mathbf{G}/\sqrt{(n-1)}$. Thus, apart from the constant $\sqrt{(n-1)}$, the lengths of the rays will approximate the standard deviations of the respective variables and angles between rays will have cosines which estimate the intervariable correlations. Distances between row points in the full space are measured in the Mahalanobis metric, using the inverse covariance matrix $\mathbf{S}^{-1}$. Geometrically this means that row points have been 'sphered' to have the same variance in all directions.

In the form biplot, it is the *form matrix* $\mathbf{Z}\mathbf{Z}^{\mathrm{T}}$, or matrix of scalar products between the rows of $\mathbf{Z}$, that is approximated optimally by the corresponding form matrix $\mathbf{F}\mathbf{F}^{\mathrm{T}}$ of $\mathbf{F}$. Thus the scalar products and squared norms (lengths) of the row vectors in the full space are approximated optimally in the reduced space biplot, whereas now the rays corresponding to the variables have been sphered.

Apart from the rules of interpretation of biplots, discussed further by Gabriel (1971, 1981), Greenacre and Underhill (1982) and Gower and Hand (1996), there are also the lesser known issues of calibration, approximation of differences and modelling that are particularly relevant to our study of compositional biplots.

2.1. Calibration of biplots

The oblique axis through a ray is called the *biplot axis* of the corresponding variable. Each z_{ij} is approximated by the scalar product between a row point and a column point in the biplot, and this scalar product is equal to the projection of the row point onto the biplot axis, multiplied by the length of the ray. It follows that the inverse of the length of the ray gives the length of a unit

along the biplot axis. For example, if the length of ray A is equal to 5, according to the scale of the display, then $1/5 = 0.2$ will be the length of 1 unit along this axis, so that two individuals projecting at a distance of 0.2 apart on this axis are predicted to be $0.2 \times 5 = 1$ unit apart on variable A. Knowledge of

(a) this unit length,
(b) the positive direction of the scale as indicated by the ray and
(c) the fact that the mean is at the centre of the display

allows us to calibrate the biplot axis in units of the original variable. For examples of calibration, see Gabriel and Odoroff (1990), Greenacre (1993) and Gower and Hand (1996).

2.2. Difference axes

Any linear combination of rays in the biplot defines a vector which represents the corresponding linear combination of the variables (Gabriel, 1978). In particular, the difference between two variables can be indicated by the vector connecting the end points, or *apexes*, of the two corresponding rays (Fig. 1). These difference vectors are called *links*. Thus, the difference between variables A and B is shown by the broken link in Fig. 1. Because the link points towards variable A, the difference represented is variable A minus variable B.

Like the biplot axes through rays, axes can be defined through links defining difference axes. Row points can then be projected onto a difference axis to obtain approximations of those differences for the individuals. The point of average difference on a difference axis is given by the projection of the origin onto this axis. In the covariance biplot the rays are optimal least squares representations of the corresponding full space columns, but in general the links are not necessarily optimal approximations of the true difference vectors. Differences will be accurately represented and predicted when the fit is high, of course, but when it is low differences are often represented much better with respect to other dimensions of the variable space which are optimal for displaying these differences themselves. For a discussion of this topic and an explicit analysis of differences, see Greenacre (2001).

Contrary to the general situation described above, however, the relative variation biplot

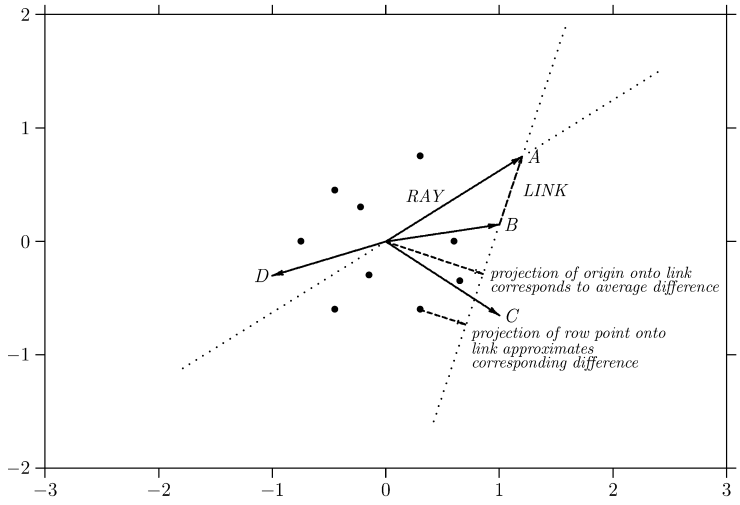

Fig. 1. Biplot axes through rays and links: •, rows (individuals); ⟶, columns (variables)

which we shall define in Section 3 for compositional data will be shown to fit both the variables and their differences optimally, and the percentage of variance explained will be the same in each case.

2.3. Diagnosis of simple models

Bradu and Gabriel (1978) gave guidelines for diagnosing simple models from straight line patterns formed by subsets of row and/or column points in a two-dimensional biplot, assuming that the biplot gives an excellent fit to the data (see also Gabriel (1981)). For example, if in a biplot we observe that a subset I of row points lies approximately in a straight line, and a subset J of column points also lies in a straight line which is perpendicular to the line of row points, then the submatrix formed by the rows I and columns J can be diagnosed to follow closely the simple additive model $z_{ij} = \mu + \alpha_i + \beta_j$. When these straight lines are not perpendicular, a slightly more general model is indicated, and even more general still when just one set of points, say the column points, falls on a straight line. The beauty of such diagnostics is that it is easier to notice groups of points lining up in a biplot than to undertake a study of all submatrices of the data.

3. Compositional data

By the very nature of the initial centring transformation (5), the biplots described above apply to interval–scale variables, since the results are invariant with respect to additive changes in the variables. If the data were ratio–scale measurements, i.e. if multiplicative differences were important in the comparison of individuals, then the data should be logarithmically transformed before centring. We now consider compositional data and the transformations which can be considered suitable to bring them onto an interval scale for biplotting.

A compositional data matrix $\mathbf{X}$ has columns corresponding to the parts, or components, of a p-part composition. A typical row vector of this matrix is $(x_1 \ \ldots \ x_p)$ with positive components subject to the unit sum constraint $x_1 + \ldots + x_p = 1$. Although standard statistical methodology, such as the calculation of covariances and correlations, is commonly applied to compositional data, there is an extensive literature on the pitfalls of such a practice (see, for example, Aitchison (1986), chapter 3). Of particular importance in the study of compositional data is the concept of a subcomposition, and the requirement that any form of analysis should have what is called *subcompositional coherence*. This is best considered in terms of two scientists A and B, with A able to record all the p parts of a composition and so to arrive at the full composition $(x_1 \ \ldots \ x_p)$, whereas B is aware of, or can record, only some parts, say $1, \ldots, p^*$, hence arriving at the subcomposition

$$(s_1 \ \ldots \ s_{p^*}) = (x_1 \ \ldots \ x_{p^*})/(x_1 + \ldots + x_{p^*}). \tag{6}$$

Subcompositional coherence requires that any inference which scientist A makes about the subcompositional parts $1, \ldots, p^*$ from knowledge of the full composition should coincide with the corresponding inference made about these parts by scientist B from the subcomposition. Regular product–moment correlations and principal component analysis, based on covariances calculated on the raw compositional data, do not have subcompositional coherence (Aitchison (1986), section 3.3).

Recognition that the study of compositions is concerned with relative and not absolute magnitudes of the components has led to considering ratios of the components. From equation (6) ratios are clearly invariant under the formation of subcompositions: $s_j/s_{j'} = x_j/x_{j'}$. Note that these are ratios *within* the compositional data vector, i.e. across the columns of the data matrix.

When it comes to calculating scalar products and covariances for the biplot it is necessary to consider on what scale these ratios themselves are, when compared across individuals. Here we maintain that the ratios themselves are on a ratio scale. Hence it is appropriate to take logarithms of the ratios, called *log-ratios*, and to consider differences between these log-ratios from individual to individual. Several justifications for the log-ratio transformation may be found in Aitchison (1986, 2001). At first this might seem unduly complicated but differences in log-ratios are already commonplace in the calculation of the log-odds in the log-linear model of categorical data and in logistic regression.

Aitchison (1986) showed that there are three equivalent ways of considering ratios within a compositional vector:

(a) the $\frac{1}{2}p(p-1)$ ratios $x_j/x_{j'}$ between pairs of components (we assume that $j < j'$ when selecting the pair),
(b) the $p-1$ ratios x_j/x_p between the first $p-1$ components and the last and
(c) the p ratios $x_j/g(\mathbf{x})$ between the components and their geometric average $g(\mathbf{x}) = (x_1 x_2 \ldots x_p)^{1/p}$.

On the logarithmic scale these are the respective differences

(a) $\log(x_j) - \log(x_{j'})$,
(b) $\log(x_j) - \log(x_p)$ and
(c) the deviations from the mean $\log(x_j) - (1/p)\,\Sigma_j \log(x_j)$.

The second option is the least interesting in the present context, because it is not symmetric with respect to all the components, and we do not discuss it further. We shall be primarily interested in the study of *pairwise log-ratios* $\log(x_j/x_{j'}) = \log(x_j) - \log(x_{j'})$, but we shall need to refer to the *centred log-ratios* $\log\{x_j/g(\mathbf{x})\}$ as well.

Suppose that we denote the logarithms $\log(x_{ij})$ of the compositional data matrix by l_{ij} and collect them in a matrix $\mathbf{L}$ $(n \times p)$. Suppose that the dot subscripts in $l_{i\cdot}$, $l_{\cdot j}$ and $l_{\cdot\cdot}$ denote the averages over the corresponding indices, so that the pairwise log-ratios are $l_{ij} - l_{ij'}$ and the centred log-ratios are $l_{ij} - l_{i\cdot}$. Let $\mathbf{T}$ be the $n \times \frac{1}{2}p(p-1)$ matrix of pairwise log-ratios with general element $\tau_{i,jj'} = l_{ij} - l_{ij'}$, $j < j'$. Although our interest is chiefly in the matrix $\mathbf{T}$ of pairwise log-ratios, we shall now show that it is possible to obtain all the results about $\mathbf{T}$ by using the smaller matrix of the centred log-ratios which has only p columns.

If we were to make a biplot of the larger matrix $\mathbf{T}$, we would centre $\mathbf{T}$ with respect to column means $\tau_{\cdot jj'} = l_{\cdot j} - l_{\cdot j'}$, as in equation (5), to obtain a matrix $\mathbf{Y}$: $y_{i,jj'} = \tau_{i,jj'} - \tau_{\cdot jj'} = l_{ij} - l_{\cdot j} - (l_{ij'} - l_{\cdot j'})$. Suppose that $\mathbf{Y}$ has SVD $\mathbf{Y} = \mathbf{A}\mathbf{\Psi}\mathbf{B}^{\mathrm{T}}$, where $\mathbf{B}$ has $\frac{1}{2}p(p-1)$ rows representing each log-ratio (jj') as a ray emanating from the origin. Notice that the corresponding 'inverse' log-ratio $(j'j)$ would be the ray of the same length emanating from the origin and pointing in the opposite direction. $\mathbf{T}$ has $\frac{1}{2}p(p-1)$ columns, but we show below that its rank is equal to $p-1$; hence it has $\frac{1}{2}(p-1)(p-2)$ columns that are effectively redundant.

Now let $\mathbf{Z}$ be the $n \times p$ matrix of (row-)centred log-ratios $l_{ij} - l_{i\cdot}$ which have been centred with respect to column means $z_{\cdot j} = l_{\cdot j} - l_{\cdot\cdot}$, i.e. $\mathbf{Z}$ is the matrix of elements of $\mathbf{L}$ which are double centred: $z_{ij} = l_{ij} - l_{i\cdot} - l_{\cdot j} + l_{\cdot\cdot}$. Let $\mathbf{Z}$ have SVD $\mathbf{Z} = \mathbf{U}\mathbf{\Gamma}\mathbf{V}^{\mathrm{T}}$. Since $\mathbf{Z}$ is double centred, its singular vectors in $\mathbf{U}$ and $\mathbf{V}$ are all centred, and the rank of $\mathbf{Z}$ is equal to $p-1$.

The SVDs of $\mathbf{Y}$ and of $\mathbf{Z}$ are directly related in the following way (see Appendix A for a proof of these results).

(a) The singular values of the two SVDs are related by a constant scaling factor: $\mathbf{\Psi} = \mathbf{\Gamma}\sqrt{p}$.
(b) The left singular vectors are identical: $\mathbf{A} = \mathbf{U}$.

(c) The right singular vectors **B** of **Y** are proportional to the corresponding differences in the row vectors of **V**; specifically $b_{jj',k} = (v_{jk} - v_{j'k})/\sqrt{p}$.

This result means that we need only to perform the analysis of the smaller matrix **Z**, from which all the results for the larger matrix **Y** can be obtained. We call the biplot of **Z** the *relative variation biplot* because it represents variation in all the component ratios. Important geometric consequences come from the equivalence of the SVDs of **Y** and **Z**. Most importantly, in the biplot of **Z** all the links from rays j to j' representing pairwise log-ratios ($j < j'$) can be transferred to the origin to obtain the solution which would have been obtained from the biplot of **Y**. This means that looking for straight line patterns in the biplot can be widened to include links which are parallel. Furthermore, the pairwise log-ratios are optimally displayed and with the same percentage explained variance as the display of the centred log-ratios.

4. Results

Figs 2 and 3 show the relative variation biplots of the data in Table 1: first the form biplot and second the covariance biplot. In the covariance biplot of Fig. 3 the column points have been rescaled by the constant $1/\sqrt{(n-1)} = 1/\sqrt{21}$ to bring the column solution onto the scale of log-ratio variance and covariance.

We collect below the properties of these relative variation biplots.

4.1. Property 1

The row points and column points are both centred at the origin of the display. This is a direct consequence of the double-centring transformation of the matrix. Thus the average row point in the display is at the origin and the average column point as well.

4.2. Property 2

In the form biplot, where rows are displayed in principal co-ordinates, distances between row points are approximations of the distances between the individuals, calculated either from the matrix of pairwise log-ratios or equivalently from the matrix of centred log-ratios:

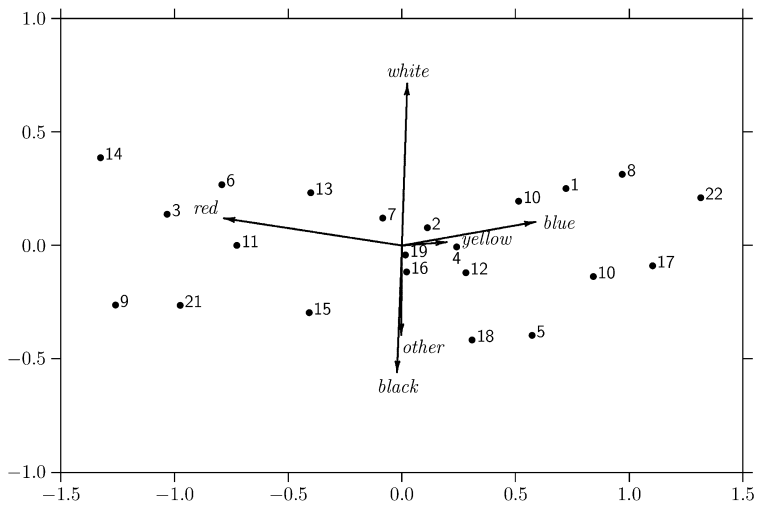

Fig. 2. Relative variation biplot of the colour composition data, preserving distances between rows (form biplot)

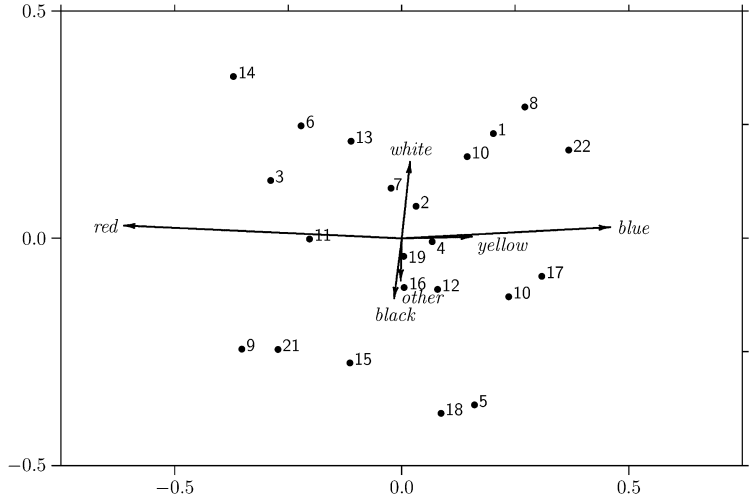

Fig. 3. Relative variation biplot of the colour composition data, preserving covariance structure between log-ratios (covariance biplot)

$$d_{ii'}^2 = \frac{1}{p}\sum_{j<j'}\sum\left\{\log\left(\frac{x_{ij}}{x_{ij'}}\right) - \log\left(\frac{x_{i'j}}{x_{i'j'}}\right)\right\}^2$$
$$= \sum_j \left[\log\left\{\frac{x_{ij}}{g(\mathbf{x}_i)}\right\} - \log\left\{\frac{x_{i'j}}{g(\mathbf{x}_{i'})}\right\}^2\right]. \quad (7)$$

The dispersion of the points along the horizontal and vertical principal axes is quantified by the corresponding eigenvalues and percentages of the sum of squares explained: 90.0% and 8.2% respectively in this application, giving an excellent overall fit of 98.2%.

4.3. Property 3

In the covariance biplot, distances between column points are approximations of the standard deviation of the corresponding log-ratio. For example, the largest link between red and blue indicates that there is the most relative variation in these two colours across the paintings. The exact standard deviations of all log-ratios are given in the upper triangle of the matrix in Table 2, whereas those estimated from the link lengths in the biplot are in the lower triangle.

For example, the exact standard deviation of the log-ratio involving black and white is equal to 0.308 whereas the displayed link has length 0.302. The estimated values are always less than the exact values, since the approximation is 'from below': the link lengths in the full five-dimensional space are exactly the standard deviations but are shorter when projected onto the reduced space of the biplot. The accuracy of recovery of the standard deviations is again reflected in the percentage of variance explained, namely 98.2%.

4.4. Property 4

Angle cosines between links in the covariance biplot estimate correlations between log-ratios. Thus the fact that the links between blue, yellow and red lie perpendicularly to the links between white, other and black indicates that log-ratios among the first set have near zero correlations with those among the second set. To support this claim, we show in Table 3 the relevant

Table 2. Standard deviations of log-ratios

Colour	*Standard deviations for the following colours:*					
	black	*white*	*blue*	*red*	*yellow*	*other*
black	0	0.308	0.504	0.616	0.225	0.130
white	0.302	0	0.466	0.645	0.221	0.270
blue	0.501	0.463	0	1.071	0.315	0.488
red	0.616	0.646	1.071	0	0.767	0.628
yellow	0.218	0.214	0.305	0.767	0	0.213
other	0.041	0.262	0.476	0.621	0.184	0

Table 3. Subset of the correlation matrix between log-ratios

Colour ratio	*Correlation for the following colour ratios:*					
	red/yellow	*red/blue*	*yellow/blue*	*white/other*	*other/black*	*white/black*
red/yellow	1.000	0.996	0.949	−0.048	−0.095	−0.082
red/blue	0.996	1.000	0.974	−0.074	−0.108	−0.110
yellow/blue	0.949	0.974	1.000	−0.133	−0.138	−0.175
white/other	−0.048	−0.074	−0.133	1.000	0.069	0.907
other/black	−0.095	−0.108	−0.138	0.069	1.000	0.482
white/black	−0.082	−0.110	−0.175	0.907	0.482	1.000

subset of the correlation matrix between log-ratios, showing that the two sets can be considered independent of each other:

4.5. Property 5

In either biplot column points lying in a straight line reveal log-ratios of high correlation, and a model summarizing this interdependence can be deduced from the relative lengths of their links. By inspection in Figs 2 and 3, the distance from red to yellow is roughly 2.5 times the distance from yellow to blue. Since all links can be transferred to the origin, it follows that

$$\log(\text{red}/\text{yellow}) - \text{ave}\{\log(\text{red}/\text{yellow})\} = 2.5[\log(\text{yellow}/\text{blue}) - \text{ave}\{\log(\text{yellow}/\text{blue})\}]$$

where ave$(\cdot)$ indicates the mean of the corresponding log-ratio across individuals. This reduces to the constant log-contrast

$$2.5\ \log(\text{blue}) + \log(\text{red}) - 3.5\ \log(\text{yellow}) = \text{constant}$$

where the constant is estimated by averaging the log-contrast over individuals. This diagnoses a proportionality relationship between the colours as

$$\text{red}/\text{yellow} \propto (\text{yellow}/\text{blue})^{2.5}.$$

Fig. 4(a) demonstrates this proportionality relationship whereas Fig. 4(b) shows the relationship in triangular co-ordinates between the three primary colours for the three-part subcomposition, showing an excellent fit to the data. Interestingly, this representation of primary colours as

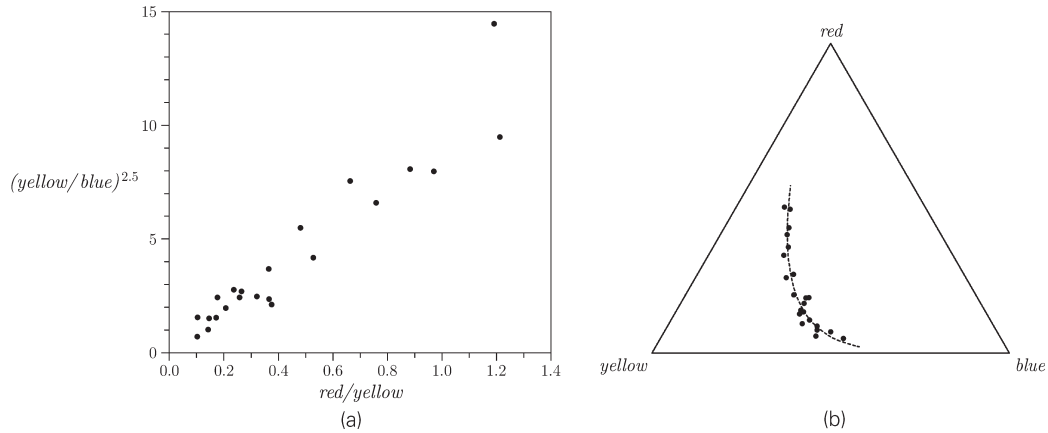

Fig. 4. (a) Relationship between colour ratios red/yellow and (yellow/blue)$^{2.5}$, showing the proportionality relationship, and (b) Goethe's colour triangle, showing mixtures of primary colours in 22 paintings, and the model diagnosed by the relative variation biplot red/yellow $\propto$ (yellow/blue)$^{2.5}$

vertices of a triangle is due to Goethe (1810) and is the earliest reference, to our knowledge, to the triangular co-ordinate system. The same system was used independently 50 years later by Maxwell (1860) to explain his own colour theory in terms of red, green and blue.

In general, if three components A, B and C lie in an approximate straight line with distances AB and BC equal to λ and μ respectively, then the constant log-contrast is of the form $\mu \log(A) + \lambda \log(C) - (\lambda + \mu) \log(B) = \text{constant}$, i.e. $(A/B)^{\mu} \propto (B/C)^{\lambda}$.

4.6. Property 6

In either biplot four column points A, B, C and D forming a parallelogram reveal a simple constant log-contrast of the form

$$\log(A) - \log(B) + \log(C) - \log(D) = \text{constant}.$$

In Figs 2 and 3 the colours black, red, white and blue lie approximately on a parallelogram. We can transfer the links black–red and blue–white to the origin and thus obtain the relationship

$$\log(\text{black/red}) - \text{ave}\{\log(\text{black/red})\} = \log(\text{blue/white}) - \text{ave}\{\log(\text{blue/white})\}$$

leading to the constant log-contrast

$$\log(\text{black}) - \log(\text{red}) + \log(\text{white}) - \log(\text{blue}) = \text{constant}$$

and thus the proportionality relationship black/red $\propto$ blue/white or equivalently black/blue $\propto$ red/ white. This relationship can be demonstrated by plotting the ratio of any two adjacent colours in the parallelogram against the ratio of the other two. Fig. 5 shows black/red against blue/white and the relationship is strongly linear through the origin, as diagnosed successfully by the parallelogram in the biplot.

4.7. Property 7

If a subset I of the individuals (rows) and a subset J of the components (columns) lie approximately on respective straight lines that are orthogonal, then the compositional submatrix formed by the rows I and columns J has approximately constant log-ratios among the components, i.e.

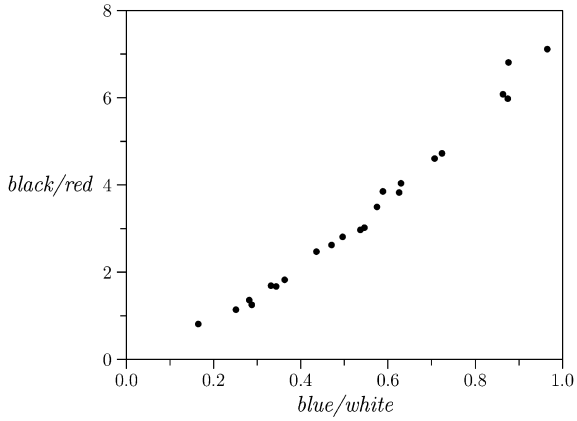

Fig. 5. Relationship between the colour ratios black/red and blue/white, showing the proportionality relationship

the double-centred submatrix of log(compositions) has near-zero entries. For example in both biplots it is possible to see a group of three row points in the lower left quadrant (rows 9, 21 and 15) which are in a straight line that is orthogonal to the line defined by the three column points white–other–black. The relevant data from Table 1 are shown in Table 4 along with the corresponding ratios.

This property of log-ratio constancy in submatrices of the data can be deduced directly from the additive model mentioned in Section 2.3 or from the concept of biplot calibration, described by the next property.

4.8. Property 8

Either biplot can be calibrated in log-ratio units or in ratio units. Calibration of the rays will apply to centred log-ratios, whereas calibration of the links will apply to pairwise log-ratios. For example, in the form biplot of Fig. 2, the length of the blue–red link is calculated as 1.372. Thus 1 unit on the biplot axis through this link has a length of $1/1.372 = 0.7290$. The mean value of log(blue/red) is calculated from the data to be 0.6153 which is the value corresponding to the origin of the display projected onto this link axis. So to calibrate this axis in tenths (0.1) of a log-ratio unit, for example, we must put tick marks on the axis at a distance of 0.0729 apart, so that the scale increases towards the right (since we are calibrating the blue − red difference) and has the value 0.6153 at the point where the origin projects onto the axis. Equivalently, we can transfer the link to the origin, in which case the origin will correspond to the average log-ratio. The calculations needed to position the tick marks for a biplot axis through a link or a ray are given in Appendix B.

Table 4. Colour composition data and corresponding ratios

Painting	*black*	*white*	*other*	*black/white*	*other/white*	*other/black*
9	0.178	0.167	0.347	1.07	2.08	1.95
15	0.199	0.170	0.318	1.17	1.88	1.60
21	0.200	0.172	0.313	1.16	1.82	1.65

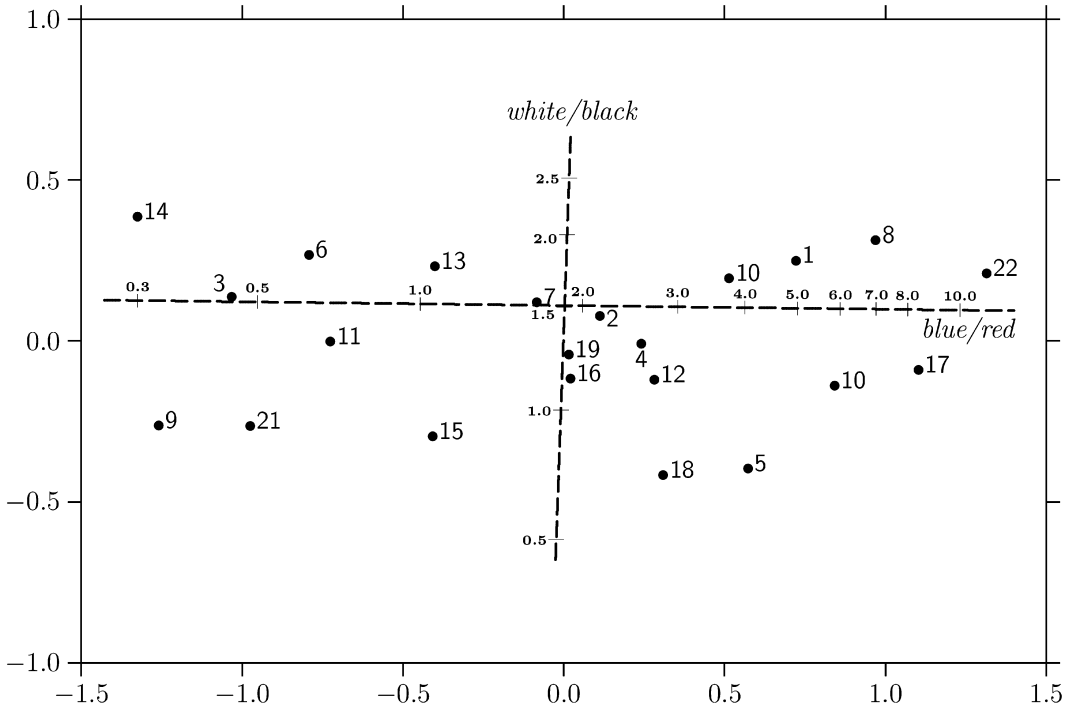

Fig. 6. Log-ratio form biplot of the colour composition data, showing calibration of the ratios blue/red and white/black on a logarithmic scale

It is also possible to calibrate either biplot directly in units of ratios. Using the log-ratio scale previously established, tick marks for appropriately rounded values on a ratio scale are determined (see the example in Appendix B). The tick marks will be on a logarithmic scale for each calibrated axis. Fig. 6 shows the form biplot calibrated in ratio units, for the colour ratios blue/red and white/black as examples. As an example, in Fig. 6 painting 8 is estimated to have twice as much white as black, and seven times as much blue as red, which is confirmed by the original data.

Calibration gives the biplot a concrete interpretation in terms of the original data and provides a new meaning to some of the properties already stated. For example, property 7 is now obvious since any points lying on a line perpendicular to a link project onto the same value on its biplot axis and thus have constant estimated values of the corresponding ratios.

4.9. Property 9

The whole compositional data matrix can be reconstructed approximately from either biplot, but we need to know the means of the centred log-ratios as well as the geometric means of the rows to be able to 'uncentre' the estimates obtained from the biplot. For this estimation we calibrate each ray representing the centred log-ratio of a column (a colour in our example). For this calibration we need to know the average centred log-ratio to be able to anchor the scale at the origin. Then the projection of each row i (a painting) onto each colour axis j gives an estimate of the centred log-ratio $\log\{x_{ij}/g(\mathbf{x}_i)\}$, and with knowledge of the geometric mean $g(\mathbf{x}_i)$ of the row we can uncentre the estimate to arrive at the estimate of x_{ij} itself. The reconstructed data from either of the two-dimensional biplots are given in Table 5 and are very close to the original data, thanks to the 98.2% explained variance in the biplot.

Table 5. Reconstructed compositional data from the two-dimensional calibrated biplot (compare with the original data in Table 1)

Painting	*Proportions of area occupied by the following colours:*					
	Black	*White*	*Blue*	*Red*	*Yellow*	*Other*
1	0.131	0.245	0.156	0.031	0.182	0.254
2	0.154	0.225	0.113	0.052	0.170	0.287
3	0.155	0.232	0.059	0.131	0.137	0.285
4	0.160	0.210	0.119	0.046	0.172	0.293
5	0.187	0.152	0.132	0.032	0.173	0.324
6	0.144	0.257	0.069	0.111	0.145	0.272
7	0.153	0.233	0.102	0.062	0.165	0.285
8	0.122	0.250	0.176	0.025	0.186	0.240
9	0.189	0.168	0.048	0.145	0.127	0.324
10	0.160	0.183	0.158	0.027	0.182	0.290
11	0.167	0.212	0.070	0.102	0.146	0.302
12	0.169	0.192	0.120	0.044	0.172	0.304
13	0.146	0.253	0.087	0.081	0.157	0.276
14	0.133	0.269	0.050	0.167	0.128	0.254
15	0.192	0.170	0.080	0.075	0.152	0.332
16	0.172	0.195	0.104	0.055	0.166	0.309
17	0.150	0.185	0.180	0.021	0.187	0.277
18	0.193	0.152	0.115	0.040	0.167	0.332
19	0.166	0.206	0.105	0.056	0.166	0.301
20	0.139	0.239	0.140	0.038	0.179	0.265
21	0.190	0.171	0.057	0.117	0.136	0.328
22	0.123	0.224	0.205	0.018	0.190	0.239

5. Discussion

The present approach is based on a certain choice of prerequisites which a method of compositional data analysis should reasonably be expected to satisfy. Most importantly, the unit sum constraint—or equivalently the fact that all compositional data vectors occupy a simplex space—should be respected throughout the analysis, and all results should have subcompositional coherence. It is clear from the above aspects of interpretation that the fundamental elements of a relative variation biplot are the links, rather than the rays as in the usual case of biplots. The complete set of links, specifying the relative variances, determines the compositional covariance structure and provides direct information about subcompositional variability and independence.

The relative variation biplot implies a certain metric, or distance function, between sample points i and i'. As we have seen in Sections 3 and 4, the squared distance can be defined either in terms of all $\frac{1}{2}p(p-1)$ (pairwise) log-ratios, or—more parsimoniously—in terms of the p centred log-ratios; see equation (7). This metric satisfies the property that the distance between any two compositions must be at least as great as the distance between any corresponding subcompositions of the compositions. For an account of how to determine an appropriate metric for compositional vectors, see Aitchison (1992). A study of the drawbacks of other metrics in the simplex space was reported by Martín-Fernández *et al.* (1998).

Attempts have been made, e.g. by Miesch *et al.* (1966), David *et al.* (1974) and Teil and Cheminée (1975), to explore compositional variability through the use of SVDs based on the raw or standardized compositional data. These approaches do not recognize specifically the

compositional nature of the data and do not have the property of subcompositional coherence. A reconstruction of compositional vectors by using biplots based on correspondence analysis (see Benzécri (1973) and Greenacre (1984, 1993)) can sometimes lead to estimated components that are negative and hence outside the simplex.

As far as identifying relationships between the components x_j of a composition is concerned, straight or parallel line patterns in the relative variation biplot indicate a particular class of models that can be written as a constant log-contrast:

$$\sum_j a_j \log(x_j) = \text{constant},$$

where $\Sigma_j a_j = 0$. Constant log-contrast relationships are important in many disciplines; for example the Hardy–Weinberg equilibrium in population genetics (Hardy, 1908) is a constant log-contrast in gene frequencies, and various equilibrium equations in geochemistry also reduce to constant log-contrasts (Krauskopf, 1979); see also Aitchison (1999) for further discussion of log-contrast laws. It can be argued that constant log-contrasts do not cover all compositional relationships of possible interest, but this is no different from the situation with the regular biplot in which only a certain class of models can be diagnosed from straight line patterns in the display.

The biplot is a natural consequence of the SVD of a matrix. To use standard SVD technology, defined on conventional multidimensional vector spaces, the compositional data are log-transformed and then double centred to ensure that component ratios are analysed on a ratio scale. Even though the initial log-transformation takes the data out of the simplex into unconstrained real vector space, the compositional nature of the data vectors is respected throughout the analysis. Aitchison (2001) showed that the same methodology can be described equivalently by an SVD which is defined directly in terms of compositions in the constrained simplex space. The simplex is established as a vector space in its own right by using compositional group operators of addition and scalar multiplication. The addition operation in this 'stay in the simplex', or *simplicial*, approach is called *perturbation*, denoted by $\oplus$, and scalar multiplication is called *powering*, denoted by $\otimes$. Without going into details about these operations, we can use them to reconstruct the ith row $\mathbf{x}_i$ of the compositional data matrix in the following way, analogous to principal component analysis:

$$\mathbf{x}_i = \boldsymbol{\xi} \oplus (s_1 u_{i1} \otimes \boldsymbol{\beta}_1) \oplus \ldots \oplus (s_r u_{ir} \otimes \boldsymbol{\beta}_r)$$

where $\boldsymbol{\xi}$ is the compositional centre of the data set, the s_k are positive 'singular values', the $\boldsymbol{\beta}_k$s are the 'right singular vectors' which form a compositional basis in the simplex, thus providing the 'principal axes' of the data compositions, and $s_k u_{ik}$ are the 'principal co-ordinates' with respect to the simplicial basis. For our colour data, the first two simplicial basis vectors turn out to be

$$\boldsymbol{\beta}_1 = (0.156 \;\; 0.149 \;\; 0.085 \;\; 0.333 \;\; 0.125 \;\; 0.153)^{\mathrm{T}},$$
$$\boldsymbol{\beta}_2 = (0.088 \;\; 0.312 \;\; 0.170 \;\; 0.173 \;\; 0.155 \;\; 0.103)^{\mathrm{T}}.$$

The way to interpret these compositional basis vectors is—as before— to look at ratios between their components. Thus the constancy of the black, white and other values (first, second and sixth) in $\boldsymbol{\beta}_1$ shows that this subcomposition is stable in the first simplicial 'dimension', whereas the constancy of blue, red and yellow (third, fourth and fifth values) in $\boldsymbol{\beta}_2$ shows a similar stability of this subcomposition in the second dimension.

Finally, we have been using the classical form of the biplot, now often referred to as the linear biplot since the definition of non-linear biplots by Gower and Harding (1988). In non-linear

biplots the biplot axes are replaced by curved trajectories and can also be calibrated. This richer but more complex biplot can possibly identify a wider class of relationships in compositional data, but its potential still needs to be fully explored.

Acknowledgements

We express thanks to John Birks and Richard Reyment for valuable discussion in the earlier stages of this work. Cajo ter Braak and John Gower gave useful comments on an earlier version of this paper that was submitted for publication. Rosemarie Nagel and Kic Udina made valuable comments and pointed out the historical references to Johann Wolfgang von Goethe and James Clerk Maxwell respectively, whose colour theories were both based on triangular co-ordinates. Michael Greenacre's research is supported by Spanish Ministry of Science and Technology grant BFM2000-1064.

Appendix A: Equivalence of log-ratio and centred log-ratio biplots

Here we prove the result that is stated in Section 3. Suppose that we collect the logarithms $l_{ij} \equiv \log(x_{ij})$ of the compositional data in the matrix $\mathbf{L}$ $(n \times p)$. Then the matrix of all log-ratios $\log(x_{ij}/x_{ij'}) = l_{ij} - l_{ij'}$ (for $j < j'$) is equal to $\mathbf{LE}_p$, where $\mathbf{E}_p$ is the $p \times \frac{1}{2}p(p-1)$ differencing matrix with 0s in each column except for a 1 and -1 in two rows. The matrix of centred log-ratios $\log\{x_{ij}/(x_{i1}\ldots x_{ip})^{1/p}\}$ is equal to $\mathbf{LC}_p$, where $\mathbf{C}_p$ is the $p \times p$ idempotent centring matrix $\mathbf{I} - (1/p)\mathbf{1}\mathbf{1}^{\mathrm{T}}$. Examples of the differencing and centring matrices are, for $p = 4$,

$$\mathbf{E}_4 = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & -1 & -1 \end{pmatrix}, \qquad \mathbf{C}_4 = \begin{pmatrix} \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & -\frac{1}{4} & -\frac{1}{4} & \frac{3}{4} \end{pmatrix}.$$

Consider the matrix $\mathbf{LC}_p$ of centred log-ratios first. A biplot of this matrix as described in Section 2 would centre with respect to column means as in equation (5), i.e. premultiply by $\mathbf{C}_n$, $\mathbf{Z} = \mathbf{C}_n\mathbf{LC}_p$, and then proceed as before with the SVD as in equation (2). The matrix $\mathbf{Z}$ is thus the double-centred matrix of log(compositions), with elements $z_{ij} = l_{ij} - l_{i\cdot} - l_{\cdot j} + l_{\cdot\cdot}$ where the dot subscript indicates averaging over the corresponding index. Suppose that $\mathbf{Z}$ has SVD $\mathbf{Z} = \mathbf{U\Gamma V}^{\mathrm{T}}$. The fact that $\mathbf{Z}$ is double centred implies that the elements of each singular vector in $\mathbf{U}$ and $\mathbf{V}$ are centred: $\mathbf{C}_n\mathbf{U} = \mathbf{U}$ and $\mathbf{C}_p\mathbf{V} = \mathbf{V}$.

Consider now the matrix $\mathbf{LE}_p$ of pairwise log-ratios. This matrix, again centred with respect to column means, gives $\mathbf{Y} = \mathbf{C}_n\mathbf{LE}_p$ and leads to a biplot which depicts the individuals and each (j, j') ratio pair $(j < j')$. Suppose that $\mathbf{Y}$ has SVD $\mathbf{Y} = \mathbf{A\Psi B}^{\mathrm{T}}$, where $\mathbf{B}$ has $\frac{1}{2}p(p-1)$ rows.

These two biplots are directly related through the SVDs as follows. Firstly, the form matrices of $\mathbf{Z}$ and $\mathbf{Y}$ are identical, apart from an overall scale factor,

$$\mathbf{ZZ}^{\mathrm{T}} = \mathbf{C}_n\mathbf{LC}_p\mathbf{C}_p\mathbf{L}^{\mathrm{T}}\mathbf{C}_n = \mathbf{C}_n\mathbf{LC}_p\mathbf{L}^{\mathrm{T}}\mathbf{C}_n = \mathbf{U\Gamma}^2\mathbf{U}^{\mathrm{T}},$$

$$\mathbf{YY}^{\mathrm{T}} = \mathbf{C}_n\mathbf{LE}_p\mathbf{E}_p^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}\mathbf{C}_n = p\mathbf{C}_n\mathbf{LC}_p\mathbf{L}^{\mathrm{T}}\mathbf{C}_n = \mathbf{U}(p\mathbf{\Gamma}^2)\mathbf{U}^{\mathrm{T}}$$

since $\mathbf{E}_p\mathbf{E}_p^{\mathrm{T}} = p\mathbf{C}_p$. Thus the singular values differ by a constant scale factor of $\sqrt{p}$, $\mathbf{\Psi} = \mathbf{\Gamma}\sqrt{p}$, and the left singular vectors are identical in the two SVDs, $\mathbf{A} = \mathbf{U}$. In contrast, the scalar products of the columns, which provide the covariances in the two biplots, have the following connection:

$$\mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \mathbf{C}_p\mathbf{L}^{\mathrm{T}}\mathbf{C}_n\mathbf{LC}_p = \mathbf{V\Gamma}^2\mathbf{V}^{\mathrm{T}}.$$

Premultiplying and postmultiplying by $\mathbf{E}_p^{\mathrm{T}}$ and $\mathbf{E}_p$ respectively and using the fact that the columns of $\mathbf{E}_p$ are centred, $\mathbf{C}_p\mathbf{E}_p = \mathbf{E}_p$, we obtain

$$\mathbf{Y}^{\mathrm{T}}\mathbf{Y} = \mathbf{E}_p^{\mathrm{T}}\mathbf{L}^{\mathrm{T}}\mathbf{C}_n\mathbf{C}_n\mathbf{L}\mathbf{E}_p = (\mathbf{E}_p^{\mathrm{T}}\mathbf{V})\mathbf{\Gamma}^2(\mathbf{E}_p^{\mathrm{T}}\mathbf{V})^{\mathrm{T}},$$

i.e. the right singular vectors of $\mathbf{B}$ are proportional to the corresponding differences between rows of $\mathbf{V}$. Since $(\mathbf{E}_p^{\mathrm{T}}\mathbf{V})^{\mathrm{T}}(\mathbf{E}_p^{\mathrm{T}}\mathbf{V}) = \mathbf{V}^{\mathrm{T}}\mathbf{E}_p\mathbf{E}_p^{\mathrm{T}}\mathbf{V} = \mathbf{V}^{\mathrm{T}}(p\mathbf{C}_p)\mathbf{V} = p\mathbf{V}^{\mathrm{T}}\mathbf{V} = p\mathbf{I}$ it follows that $\mathbf{B} = \mathbf{E}_p^{\mathrm{T}}\mathbf{V}/\sqrt{p}$ and we verify again that $\mathbf{\Psi} = \mathbf{\Gamma}\sqrt{p}$.

With the above notation it is easy to show that, in general, a matrix $\mathbf{Y}$ (column centred or not) has form matrix $\mathbf{Y}\mathbf{Y}^{\mathrm{T}}$, whereas the form matrix of its column differences $\mathbf{Y}\mathbf{E}_p$ is $p\mathbf{Y}\mathbf{C}_p\mathbf{Y}^{\mathrm{T}}$. Thus the form matrices agree (up to the scale value p) if $\mathbf{Y}$ is row centred, but also if $\mathbf{Y}$ has constant row sums since row centring would then just involve subtracting a constant from every matrix element. Thus a regular principal component analysis of a matrix of compositional data also has the property that links are optimal representations of the column differences.

Appendix B: Linear biplot calibration

Suppose that we want to calibrate the biplot axis which passes through two column points A and B, with given co-ordinates (a_1, a_2) and (b_1, b_2) on the first two dimensions of the biplot. Denote the projection of the origin of the biplot onto the biplot axis by the point (o_1, o_2). Suppose that the mean difference in the values of $B - A$ (calculated from the data) is equal to m.

Now the squared distance from A to B is equal to $d^2 = (b_1 - a_1)^2 + (b_2 - a_2)^2$ and the length of 1 unit on the biplot axis is thus $s = 1/d$. By simple trigonometry, the co-ordinates (o_1, o_2) are equal to

$$o_1 = \frac{a_1(b_2 - a_2)^2 - a_2(b_1 - a_1)(b_2 - a_2)}{d^2},$$

$$o_2 = \frac{a_2(b_1 - a_1)^2 - a_1(b_1 - a_1)(b_2 - a_2)}{d^2}$$

and the tick mark for value t has co-ordinates (t_1, t_2):

$$t_1 = o_1 + \frac{s(t - m)(b_1 - a_1)}{d} = o_1 + \frac{(t - m)(b_1 - a_1)}{d^2},$$

$$t_2 = o_2 + \frac{s(t - m)(b_2 - a_2)}{d} = o_2 + \frac{(t - m)(b_2 - a_2)}{d^2}.$$

As an example, for the blue–red link in the form biplot of Fig. 6, the co-ordinate values are the standard co-ordinates of the apexes of the red and blue points respectively: $(a_1, a_2) = (-0.7839, 0.1201)$ and $(b_1, b_2) = (0.5878, 0.1042)$, and the mean value of log(blue/red), m, equals 0.6153. The link distance is equal to 1.372 and the unit distance on the biplot axis is thus $1/1.372 = 0.7290$. The origin projected onto the biplot axis, corresponding to the mean value, has co-ordinates $(o_1, o_2) = (0.0013, 0.1110)$ and the tick mark for the log-ratio value of 1.1, for example, has co-ordinates

$$t_1 = 0.0013 + (1.1 - 0.6153)(0.5878 + 0.7839)/1.372^2 = 0.3545,$$

$$t_2 = 0.1110 + (1.1 - 0.6153)(0.1042 - 0.1201)/1.372^2 = 0.1069.$$

These formulae can be used to calibrate a ray as well by taking the point A as the origin and thus setting $(a_1, a_2) = (0, 0)$.

The calibration described above applies to the biplot axes in log-ratio units. To calibrate directly in ratio units, as in Fig. 6, the value of each tick mark on the ratio scale is first transformed by the natural logarithm and then the above procedure is followed. For example, to place the tick mark for the ratio 5.0 on the blue–red axis, i.e. five times as much blue as red, we calculate $\log(5.0) = 1.609$ and then proceed as before, substituting 1.609 for 1.1 in the above calculations to obtain co-ordinates of the tick mark as $(t_1, t_2) = (0.7260, 0.1026)$.

References

Aitchison, J. (1986) *The Statistical Analysis of Compositional Data*. London: Chapman and Hall.

———(1992) On criteria for measures of compositional difference. *Math. Geol.*, **22**, 223–226.

———(1999) Logratios and natural laws in compositional data analysis. *Math. Geol.*, **31**, 563–589.

———(2001) Simplicial inference. In *Algebraic Structures in Statistics and Probability* (eds M. A. G. Viana and D. St P. Richards), pp. 1–22. Providence: American Mathematical Society.

Benzécri, J.-P. (1973) *L'Analyse des Données*, vols I, II. Paris: Dunod.

Bradu, D. and Gabriel, K. R. (1978) The biplot as a diagnostic tool for models of two-way tables. *Technometrics*, **20**, 47–68.

David, M., Campiglio, C. and Darling, R. (1974) Progresses in R- and Q-mode analysis: correspondence analysis and its application to the study of geological processes. *Can. J. Earth Sci.*, **11**, 131–146.

Eckart, C. and Young, G. (1936) The approximation of one matrix by another of lower rank. *Psychometrika*, **1**, 211–218.

Gabriel, K. R. (1971) The biplot-graphic display of matrices with application to principal component analysis. *Biometrika*, **58**, 453–467.

———(1978) Analysis of meteorological data by means of canonical decomposition and biplots. *J. Appl. Meteorol.*, **11**, 1072–1077.

———(1981) Biplot display of multivariate matrices for inspection of data and diagnosis. In *Interpreting Multivariate Data* (ed. V. Barnett), pp. 147–173. New York: Wiley.

Gabriel, K. R. and Odoroff, C. L. (1990) Biplots in biomedical research. *Statist. Med.*, **9**, 469–485.

Goethe, J. W. (1810) *Zur Farbenlehre*. Tübingen. (Available from `http://www.colorsystem.com/projekte/engl/14goee.htm`.)

Gower, J. C. and Hand, D. (1996) *Biplots*. London: Chapman and Hall.

Gower, J. C. and Harding, S. (1988) Non-linear biplots. *Biometrika*, **73**, 445–455.

Greenacre, M. J. (1984) *Theory and Applications of Correspondence Analysis*. London: Academic Press.

———(1993) Biplots in correspondence analysis. *J. Appl. Statist.*, **20**, 251–269.

———(2001) Analysis of matched matrices. *Working Report 539*. Department of Economics and Business, Universitat Pompeu Fabra, Barcelona. (Available from `http://www.econ.upf.es/cgi-bin/onepaper?539`.)

Greenacre, M. J. and Underhill, L. G. (1982) Scaling a data matrix in low-dimensional Euclidean space. In *Topics in Applied Multivariate Analysis* (ed. D. M. Hawkins), pp. 183–268. Cambridge: Cambridge University Press.

Hardy, G. H. (1908) Mendelian proportions in a mixed population. *Science*, **28**, 49–50.

Krauskopf, K. B. (1979) *Introduction to Geochemistry*. New York: McGraw Hill.

Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998) Measures of difference for compositional data and hierarchical clustering. In *Proc. 4th A. Conf. International Association for Mathematical Geology* (eds A. Baccianti, G. Nardi and R. Potenza), pp. 526–531. Naples: De Frede.

Maxwell, J. C. (1860) On the theory of compound colours. *Phil. Trans.*, **150**, 57–84. (Available from `http://www.colorsystem.com/projekte/engl/19maxe.htm`.)

Miesch, A. T., Chao, E. C. T. and Cuttitta, F. (1966) Multivariate analysis of geochemical data on tektites. *J. Geol.*, **74**, 673–691.

Teil, H. and Cheminée, J. L. (1975) Application of correspondence factor analysis to the study of major and trace elements in the Erta ale Chain (Afar, Ethiopia). *Math. Geol.*, **7**, 13–30.