Distributional Equivalence and Subcompositional Coherence in the Analysis of Compositional Data, Contingency Tables and Ratio-Scale Measurements

Michael Greenacre

Universitat Pompeu Fabra, Barcelona, Spain

Paul Lewi

Janssen Pharmaceutica, Vosselaar, Belgium

Summary. We consider two fundamental properties in the analysis of two-way tables of positive data: the principle of distributional equivalence, one of the cornerstones of correspondence analysis of contingency tables, and the principle of subcompositional coherence, which forms the basis of compositional data analysis. For an analysis to be subcompositionally coherent, it suffices to analyse the ratios of the data values. A common approach to dimension reduction in compositional data analysis is to perform principal component analysis on the logarithms of ratios, but this method does not obey the principle of distributional equivalence. We show that by introducing weights for the rows and columns, the method achieves this desirable property and can be applied to a wider class of methods. This weighted log-ratio analysis is theoretically equivalent to "spectral mapping", a multivariate method developed almost 30 years ago for displaying ratio-scale data from biological activity spectra. The close relationship between spectral mapping and correspondence analysis is also explained, as well as their connection with association modelling. The weighted log-ratio methodology is used here to visualize frequency data in linguistics and chemical compositional data in archaeology.

Keywords: association models; biplot; correspondence analysis; weighted log-ratio analysis; singular value decomposition; spectral mapping.

The first author acknowledges research support from the Fundación BBVA in Madrid as well as partial support by the Spanish Ministry of Science and Technology, grant BST06-6587.

1. Introduction

There are a number of techniques available for the multidimensional analysis of tables of nonnegative data, for example, principal component analysis, correspondence analysis and, in the special case of compositional data, various methods based on analysing ratios between components. Our objective in this paper is to examine the foundational principles on which such methods are constructed and to show how the methods are related, both from a theoretical and practical point of view. In the course of our description we shall focus on a method based on a weighted form of log-ratio analysis, also called the "spectral map", which has all the favourable properties one might wish for when analysing positive ratio-scale data, its main inconvenience being the difficulty in handling data zeros.

Correspondence analysis (Benzécri, 1973; Greenacre, 1984, 2007; Lebart, Morineau and Warwick, 1984) is one of a family of methods based on the singular value decomposition, and has become a standard method for graphically displaying tables of nonnegative data. The method is particularly popular in the social and environmental sciences for analyzing frequency data (see, for example, Greenacre and Blasius (1994) and ter Braak (1985) respectively). As emphasised by Benzécri, who originally developed correspondence analysis (CA) as a method for exploring large tables of counts in linguistics, a fundamental property of CA is the so-called *principle of distributional equivalence*: "Our first principle is that of distributional equivalence" (Benzécri, 1973: vol. I, p. 23). This principle can be stated in a simplified form as follows: if two columns (resp., two rows) have the same relative values, then merging them does not affect the distances between rows (resp., columns).

For example, consider the data in Table 1, the counts of the 26 letters of the alphabet in 12 different English texts, pairs of which are written by the same author (these data are from dataset 'author' provided originally in S-PLUS (2005) and included in the correspondence analysis **ca** package by Nenadić and Greenacre (2007) for R (R Development Core Team, 2007). As we shall show later, although there are very small differences in relative frequencies of letters between texts, it is nevertheless possible to discriminate between the six authors, mainly due to differences in the use of consonants. Since the vowels have distributions across the texts which are almost identical, it is possible to merge their counts into one category called "vowels". The principle of distributional equivalence ensures that the distances between texts (chi-square distances in CA) are hardly changed by merging these almost "distributionally equivalent" categories, and in the limit when the distributions are identical, these distances would remain unaffected. For more details about distributional equivalence and a proof in the context of CA and related methods that follow this principle, see Benzécri (1973), Escofier (1978), Greenacre (1984: Section 4.1.17), Bavaud (2002) and Greenacre (2007: pp.37–38).

Compositional data analysis (Aitchison, 1986) is concerned with data vectors of (strictly) positive values summing to one, that is with the unit-sum constraint, or *closure*. This methodology has become popular in the physical sciences, especially geology and chemistry. For example, chemical samples are typically analyzed into constituent components by weight or by volume, expressed as proportions of the total sample. One of the founding principles of compositional data analysis is that of *subcompositional coherence*. For example, suppose that a chemical sample has inorganic and organic components, and that scientist A is investigating all of these components, whereas scientist B is investigating just the organic components of the same samples. B's data constitute a subcomposition where proportions have been calculated relative to total organic material; that is, the values in the subcomposition have been "re-closed" to add up to 1. Subcompositional coherence means that any relationships found by scientist B concerning the relationships between components of the subcomposition should be the same as scientist A's, unaffected by the fact that B is looking at a reduced data set. In our geometric framework we shall make this concept more precise by saying that measures of association or measures of dissimilarity between components, for example correlations or distances, are unaffected by considering subcompositions. One way to guarantee subcompositional coherence is to analyse ratios of components, which are unaffected by forming subcompositions.

For example, consider the data in Table 2 from Baxter, Cool and Heyworth (1990) on the percentages by weight of 11 elements in a sample of Roman glass cups found in archeological sites in Colchester. The dominating element is Silicon (Si) and one might choose to make an analysis of the other 10 elements by themselves, re-closing their weights as percentages of the non-Silicon part in each sample. Clearly, a measurement of relationship, for example a correlation, between two elements such as phosphorus (P) and potassium (K) should be invariant to whether we analyse the 10 elements alone or the full composition including Silicon. But the usual linear correlation coefficient would change in the subcomposition, hence the need for an alternative approach. Now the ratio P/K of phosphorus to potassium remains unchanged whether it is part of the full composition or the subcomposition, so any measure of difference or association between P and K that depends only on these ratios across the samples will be invariant: for example, var[log(P/K)] = var[log(P)–log(K)], the variance of the differences in their logarithms, would be the same in the full composition and a subcomposition.

Aitchison (1980, 1983) defined a variant of principal component analysis for compositional data, based on logarithmically transforming component ratios, called *log-ratios*. Kazmierczak (1988) demonstrated several graphical properties of this method, which he called "logarithmic analysis". The biplot version of this display has several interesting properties, summarized by Aitchison and Greenacre (2002): for example, it is equivalent to analyze all the log-ratios for

pairs of components within samples or to analyse the logarithms of the components relative to their geometric mean for the sample. However, although this "log-ratio biplot" has subcompositional coherence, it does not obey the principle of distributional equivalence. This is unfortunate for compositional data analysis, because if two components were always occurring in the same proportion in every sample, then the analysis should be unaffected by considering these two components amalgamated into one. In other words, in our glass cups example above, if the ratio P/K were constant across the samples, then we should be able to amalgamate their values into one value without changing the measure of distance between the glass cups. Distributional equivalence also means that any part of the composition can be broken down into subparts, all in proportion to the original part, without affecting the distances between cups.

In this paper we will show that by introducing weights into Aitchison's log-ratio analysis (LRA), in the same spirit that CA weights the rows and columns of a data table, the method does indeed achieve distributional equivalence. In the particular case when the weights are proportional to the margins of the table, this method of data visualization turns out to be equivalent to *spectral mapping*, developed by Lewi (1976, 1980), in the specific context of the analysis of biological activity spectra. In fact, the same issue of analyzing relative values rather than their original absolute values is present in this biomedical context as well as several other areas of research, outside the realm of compositional data. For example, in the analysis of contingency tables vectors of relative frequencies, or profiles, are visualized in CA, while odds and odds ratios are analyzed in association modelling. In the analysis of biometric measurements, for example measurements on animal skulls for purposes of classification, we are not interested so much in the overall level of the measurements, or "size", but rather in their relative values, or "shape". In this latter case, the principle of distributional equivalence is again of importance: if one measurement is the sum total of smaller component measurements and if the component measurements are always in the same proportion across the individuals, then we should be able to retain just the sum, omitting its components (or retain the components, omitting the sum), without affecting our measure of distance between individuals.

In the course of our explanation we will use the two data matrices given in Tables 1 and 2 to show how the weighted LRA functions, how its results are interpreted and how it compares to CA, in the context of frequency and compositional data.

2. Weighted log-ratio analysis

We consider a general matrix $\mathbf{N}$ ($I \times J$) of positive values $n_{ij} > 0$, with row totals, column totals and grand total denoted by n_{i+}, n_{+j} and n respectively. Denote by $\mathbf{L}$ the matrix of natural logarithms of the frequencies, $l_{ij} = \log(n_{ij})$. In the case of compositional data, where $n_{i+}=1$ for all i, Aitchison's "relative variation diagram" (Aitchison, 1980) consists of double-centring the matrix $\mathbf{L}$ with respect to averages of the rows and columns, followed by a singular value decomposition (SVD) to obtain least-squares matrix approximations and maps depicting rows and columns in a low-dimensional subspace. The same result can be achieved by row-centring $\mathbf{L}$ and then applying a regular principal component analysis (PCA) with column-centring but no column-normalization. Aitchison and Greenacre (2002) describe the properties of the biplots that are obtained from the above SVD, specifically the *form biplot* that favours the display of distances between samples (rows), and the *covariance biplot* that favours the display of the components (columns), explained in more detail below.

Applying this unweighted form of LRA to Baxter's cup data in Table 2, we obtain the form biplot in Figure 1. This map shows three diagonal bands of points which are due to the element manganese (Mn), which takes on only three different values in the data set, all very small: 0.01 (35 cups), 0.02 (10 cups) and 0.03 (2 cups). These values, reported to two decimal places on a percentage scale, engender large differences on the logarithmic scale and in all log-ratios; for example, amongst themselves there are differences as high as threefold. Hence manganese turns out to have the highest variance than any other component in the data set, while having the lowest percentage by weight. As a consequence, this rare component dominates the solution, as can be seen in Figure 1: the cups with percentage values 0.01, 0.02 and 0.03 (samples 3 and 25 have the highest values, 0.03%) project onto three separate locations on the Mn biplot axis (remember that the scale is logarithmic). The three resulting bands are lining up with the other high variance component antimony (Sb) – see the first column of Table 3, described more fully later.

One possible course of action is to omit an over-influential component such as manganese and analyse the remaining components as a subcomposition. Another option, which we present here and which we believe to be more appropriate because it retains all the data, is to down-weight its influence in the graphical display by introducing weights in the analysis. In CA the inherent weights are row and column sums relative to the grand total: $r_i = n_{i+}/n$ and $c_j = n_{+j}/n$, which are called *masses*. For a table of frequencies, the masses would be the row and column proportions, while if we applied CA to a matrix of compositional data, the row masses would be equal to a constant $1/I$ and the column masses would be the average proportions of the components across the samples. Using these weights in the glass cups application would mean attributing

importance to the components proportional to their average weights, effectively down-weighting the influence of the manganese component. Notice that using the margins of the table to define weights implicity assumes that all data are on the same scale.

The argument we present below is valid for any chosen set of row or column weights; for example, in the case of compositional data one might have information about the precision of measurement, which could be used to define weights for the columns, and different row weights could be defined to correct for disproportionate sampling. When we consider amalgamating rows or columns, however, we assume that the weights are additive, that is the weight of two columns, for example, that are merged into one by summation, is the sum of the two weights.

Let **r** be the vector of row weights, **c** the vector of column weights and $\mathbf{D}_r$ and $\mathbf{D}_c$ the corresponding diagonal matrices. The only condition on the weights is that they be positive and – purely for notational convenience – be closed to sum to 1. We shall discuss later the special case when we choose weights proportional to the table margins, as is the practice in CA. Otherwise, we follow very closely the CA approach: the row and column weights are introduced first into the double-centring stage, so that centring is with respect to weighted averages, and then – more importantly – into the matrix approximation stage, so that fitting is by weighted least squares. As a direct result of the weighting, if we agglomerate distributionally equivalent columns, and similarly agglomerate their weights, then the principle of distributional equivalence is satisfied (this result is proved in Section 3).

We now summarize the four-step algorithm for performing a weighted LRA, including the definitions of the various maps of the rows and columns. This methodology applies to any matrix of positive data, transformed to logarithms in the $I \times J$ matrix **L**, and using any sets of row and column weights, **r** and **c**, which are positive values summing to 1. Since our main interest will be in the weights defined by the relative marginal totals as in CA, we use the term "mass" throughout for the weights.

Step 1. Double-centre the matrix **L** with respect to its weighted row and column averages, the order of centring being invariant. That is, calculate the weighted averages of the rows of **L**, using the column masses to weight each column element: $l_{i\cdot} = \Sigma_j\, c_j\, l_{ij}$ (i=1,$\cdots$,I) and then subtract these averages from all the elements in the corresponding row, $l_{ij} - l_{i\cdot}$ (this is "weighted row-centring"). Then perform "weighted column-centring" by calculating weighted averages of the columns, using the row masses to weight each element: $\Sigma_i\, r_i\, (l_{ij} - l_{i\cdot})$ (j=1,$\cdots$,J), and then subtract these averages from all the elements in the corresponding columns. The result of this operation is a double-centred matrix with elements $a_{ij} = l_{ij} - l_{i\cdot} - l_{\cdot j} + l_{\cdot\cdot}$, where the dot subscript indicates

weighted averaging over the corresponding subscript. In matrix notation, this double-centring can be written as (where **I** is the identity matrix and **1** the vector of ones of appropriate order):

$$\mathbf{A} = (\mathbf{I} - \mathbf{1r}^{\mathsf{T}})\mathbf{L}(\mathbf{I} - \mathbf{c1}^{\mathsf{T}})^{\mathsf{T}} \qquad (1)$$

.

Step 2. To prepare the matrix for a weighted SVD, multiply a_{ij} by $(r_i\, c_j)^{1/2}$, that is multiply the rows and columns by the square roots of their respective masses:

$$\mathbf{S} = \mathbf{D}_r^{½}\, \mathbf{A}\, \mathbf{D}_c^{½}$$

Step 3. Perform the SVD of this transformed matrix:

$$\mathbf{S} = \mathbf{U\Gamma V}^{\mathsf{T}} \quad \text{where } \mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{V}^{\mathsf{T}}\mathbf{V} = \mathbf{I}$$

where the singular values down the diagonal of $\mathbf{\Gamma}$ are in descending order: $\gamma_1 \geq \gamma_2 \geq \cdots > 0$.

Step 4. Calculate the *standard coordinates* (Greenacre, 1984) by dividing the rows of the matrix of left singular vectors by $r_i^{½}$, and the rows of the matrix of right singular vectors by $c_j^{½}$:

(row standard) $\mathbf{X} = \mathbf{D}_r^{-½}\, \mathbf{U}$ (column standard) $\mathbf{Y} = \mathbf{D}_c^{-½}\, \mathbf{V}$

The *principal coordinates* for the rows and columns are the standard coordinates scaled by the singular values:

(row principal) $\mathbf{F} = \mathbf{X\Gamma} = \mathbf{D}_r^{-½}\, \mathbf{U\Gamma}$ (column principal) $\mathbf{G} = \mathbf{Y\Gamma} = \mathbf{D}_c^{-½}\, \mathbf{V\Gamma}$

In general, the coordinates can be written $\mathbf{D}_r^{-½}\mathbf{U\Gamma}^{\alpha}$ (for the rows) and $\mathbf{D}_c^{-½}\mathbf{V\Gamma}^{\beta}$ (for the columns), the above options being α and β equal to 1 or 0 for principal and standard coordinates respectively. Notice how the masses are used to pre-transform the matrix in step 2 and post-transform the resultant singular vectors in step 4, which engenders a weighted (or generalized) SVD on the centred matrix **A** (for a description of the generalized SVD see Greenacre, 1984: Appendix 1).

As in all methods of this type, we can choose to represent either of two so-called *asymmetric maps*:

(i) Use **F** and **Y** to represent the rows and columns respectively – this map is also called "row-principal" or "row-metric-preserving (RMP)" (Gabriel, 1971), with $\alpha = 1, \beta = 0$.

(ii) Use **X** and **G** to represent the rows and columns respectively – this asymmetric map is called "column-principal", or "column-metric-preserving (CMP)", with $\alpha = 0, \beta = 1$.

For representing the points in a two-dimensional map, for example, use the first two columns of the respective coordinate matrices defined above.

Both asymmetric maps are biplots in the true sense of the term (Gabriel, 1971), characterized by the condition $\alpha + \beta = 1$, where row–column scalar products approximate the elements of the double-centred matrix **A**. When the data are in the usual cases-by-variables format, Aitchison and Greenacre (2002) call the RMP biplot a *form biplot* and the CMP biplot a *covariance biplot*. A popular alternative map, especially in CA, is the *symmetric map* where both rows and columns are represented in principal coordinates **F** and **G** respectively ($\alpha = 1, \beta = 1$). The symmetric map is, strictly speaking, not a biplot (see, for example, Greenacre, 1993), but Gabriel (2002) shows that the scalar-product approximations are not substantially degraded in most cases.

The description of the weighted LRA method so far allows for any weighting system on the rows and the columns. In many situations, in the absence of additional information, the row and column margins of the original data table provide an excellent default weighting system, which is the one we shall use here in our applications. Thus, in the analysis of Table 2, the element manganese will be considerably down-weighted in the least-squares fitting of the plane of our biplot solution. Figure 2 shows the corresponding form biplot for Table 2, verifying that the role played by manganese has diminished dramatically. Although the element antimony (Sb) appears to be an outlier, its role is also not so strong owing to its low mass in the analysis. The outlying positions of points with low masses is a phenomenon that occurs in CA as well, and is partly due to the scaling of the asymmetric map. Greenacre (2007: Chapter 13) proposes an alternative biplot, called the "standard biplot", where the points in standard coordinates (the components in this case) are multiplied by the square roots of their masses, in which case the lengths of the vectors are directly related to their contributions to the solution. In any case, to understand numerically the true role of each component in the solution, the contributions of each component can be calculated, as is done regularly in CA (see, for example, Greenacre, 2007: chapter 11). Table 3 shows the percentage contributions of the 11 elements to the two-dimensional maps of Figures 1 and 2. In the unweighted analysis the contribution by manganese (Mn) to the variance of the two-dimensional map is the highest (39.48%), while it drops to one of the lowest in the weighted analysis (0.37%). On the other hand, the most common element silicon (Si) contributes 7.11% to the unweighted map, and when its very high weight is incorporated in the analysis its contribution rises to 21.05%. Notice that the very large weight given to silicon, which is on average 72.31% by weight of the glass cups, does not increase its contribution exorbitantly, because the point Si is now much closer to the centroid (weighted average), and a point's contribution is equal to its mass times squared distance to the centroid. Hence, the weighting is important in centring the data as well.

Points that are displayed in principal coordinates are approximating distances between the rows or columns of the original data matrix. For example, in Figure 2 where the rows are represented in principal coordinates, the true underlying (squared) distance function between rows i and i' is:

$$d^2_{ii'} = \sum_j c_j \left(\log \frac{n_{ij}}{\bar{g}(\mathbf{n}_i)} - \log \frac{n_{i'j}}{\bar{g}(\mathbf{n}_{i'})} \right)^2 \qquad (2)$$

where $\bar{g}(\mathbf{n}_i) = n_{i1}^{c_1} n_{i2}^{c_2} \cdots n_{iJ}^{c_J}$ is the (weighted) geometric mean of the *i*-th row. In the same way as was shown by Aitchison and Greenacre (2002) in the unweighted case, the distance (2) may be expressed equivalently in terms of the ½$J(J-1)$ log-ratios between unique pairs of columns:

$$d^2_{ii'} = \sum\sum_{j<j'} c_j c_{j'} \left(\log \frac{n_{ij}}{n_{ij'}} - \log \frac{n_{i'j}}{n_{i'j'}} \right)^2 \qquad (3)$$

where the (*j*,*j'*)-th term is weighted by the product $c_j c_{j'}$ of the weights.

With a slight re-arrangement within the parenthesis, this squared distance (3) is identical again to:

$$d^2_{ii'} = \sum\sum_{j<j'} c_j c_{j'} \left(\log \frac{n_{ij}}{n_{i'j}} - \log \frac{n_{ij'}}{n_{i'j'}} \right)^2 \qquad (4)$$

showing that log-ratios can be considered between pairs of values in the same column rather than across columns. Another alternative form of the weighted LRA distance function in (3) or (4) is in terms of the logarithms of odds-ratios for the four cells defined by row indices *i*,*i'* and column indices *j*,*j'*:

$$d^2_{ii'} = \sum\sum_{j<j'} c_j c_{j'} \left(\log \frac{n_{ij}}{n_{i'j}} \frac{n_{i'j'}}{n_{ij'}} \right)^2 \qquad (5)$$

Zero distance between a pair of rows means that all ratios are equal, that is the rows have the same relative values, or *profile*: $n_{ij}/n_{i+} = n_{i'j}/n_{i'+}$. Thus, if the distance between rows *i* and *i'* is short in the display, and assuming that the display is an accurate representation of the data, this indicates that the rows are approximately proportional to one another, just as in CA. If the data are compositional with the unit-sum constraint, this would imply approximate equality in their compositions. Similarly, if two column points *j* and *j'* displayed in principal coordinates are close together, this would indicate similar column profiles. For compositional data similar column profiles would mean that – although the overall levels of two components are different – they have similar "peaks" and "troughs" across the samples (for example, component *j* occurs approximately twice as much as component *j'* in all samples).

Any of the equivalent forms (2) – (5) of the squared distance between rows applies similarly to distances between columns; for example, formula (5) can be rewritten for columns as:

$$d^2_{jj'} = \sum\sum_{i<i'} r_i r_{i'} \left(\log \frac{n_{ij}}{n_{ij'}} \frac{n_{i'j'}}{n_{i'j}} \right)^2 \qquad (6)$$

This form shows that the values in parentheses would be unaffected by defining subcompositions of the columns, followed by rowwise closure, since the ratios $n_{ij}/n_{ij'}$ for subcomponents j and j' in each row would remain the same as their values in the full composition. The weights r_i would be equal to $1/I$ in both composition and subcomposition, so this illustrates the subcompositional coherence property mentioned earlier in terms of invariance of the distance. In the more general case of a contingency table, however, the margins of a subset of rows would differ from those of the complete table and induce changes in the masses r_i, which would affect the distance function. In this case a version of weighted LRA could be used which maintains the original masses of the table in all analyses of subtables, as Greenacre and Pardo (2006) have proposed in the case of CA.

To express the total variance in the table, we can calculate the weighted sum of squared distances of the rows (or columns) to their centroid. In LRA, however, the centroid is of no practical interest – it is rather the row-to-row and column-to-column distances and directions that are interpreted, since these approximate the log-ratios. The measure of total variance can thus be equivalently expressed in a more relevant form as $\sum\sum_{i<i'} r_i r_{i'} d^2_{ii'}$ or $\sum\sum_{j<j'} c_j c_{j'} d^2_{jj'}$, called the "geometric variability" by Cuadras and Fortiana (1998). Bavaud (2002) calls the ability to express the total variance in this equivalent way, summing over all pairs of squared interpoint distances, as "Huygens weak principle".

All the properties of the unweighted LRA described by Aitchison and Greenacre (2002) carry over to the weighted version described here, the only difference being in the centring of the matrix and the weighted approximation, giving more or less weight to the elements of the double-centred matrix according to the row and column margins.

3. Principle of distributional equivalence

We now prove that the weighted LRA map obeys the principle of distributional equivalence. Suppose that two columns j and j' have the same profile, that is the ratios $n_{ij}/n_{ij'}$ are identical for all rows i. Without loss of generality we can assume that these are the first two columns, $j = 1$ and $j' = 2$, and that these ratios are equal to a constant K, so that $n_{i1} = K\, n_{i2}$. Let us now amalgamate these two columns into one column with values equal to $n_{i1} + n_{i2} = (K+1)\, n_{i2}$ ($i = 1,\ldots, I$), and column mass $c_1 + c_2$. The distances between columns are unaffected by this merger, since we have just replaced two column points at the same position by one with mass equal to the sum

of the previous two masses. The more challenging property to prove is that the distances between rows are unaffected. In the distance formula (3) for weighted LRA all terms with log-ratios not involving columns 1 and 2 are unaffected by the merger, so we just need to consider terms involving columns 1 and 2 before and after they are combined. Before the merger, the first term of (3), for (j, j') = (1, 2), is equal to 0 since the ratios are equal and have zero difference. The other terms involving log-ratios with columns 1 and 2 can be written as:

$$\sum_{j'=3}^{J} c_1 c_{j'} \left(\log \frac{n_{i1}}{n_{ij'}} - \log \frac{n_{i'1}}{n_{i'j'}} \right)^2 + \sum_{j'=3}^{J} c_2 c_{j'} \left(\log \frac{n_{i2}}{n_{ij'}} - \log \frac{n_{i'2}}{n_{i'j'}} \right)^2$$

$$= \sum_{j'=3}^{J} c_1 c_{j'} \left(\log \frac{Kn_{i2}}{n_{ij'}} - \log \frac{Kn_{i'2}}{n_{i'j'}} \right)^2 + \sum_{j'=3}^{J} c_2 c_{j'} \left(\log \frac{n_{i2}}{n_{ij'}} - \log \frac{n_{i'2}}{n_{i'j'}} \right)^2$$

$$= \sum_{j'=3}^{J} (c_1 + c_2) c_{j'} \left(\log \frac{n_{i2}}{n_{ij'}} - \log \frac{n_{i'2}}{n_{i'j'}} \right)^2 \qquad (6)$$

because the factor K disappears in the subtraction of the log-ratios. After the merger, columns 1 and 2 are eliminated and a new column is formed by adding the previous columns 1 and 2. The terms in the distance function corresponding to log-ratios with respect to this new column are:

$$\sum_{j'=3}^{J} (c_1 + c_2) c_{j'} \left(\log \frac{(1+K)n_{i2}}{n_{ij'}} - \log \frac{(1+K)n_{i'2}}{n_{i'j'}} \right)^2$$

$$= \sum_{j'=3}^{J} (c_1 + c_2) c_{j'} \left(\log \frac{n_{i2}}{n_{ij'}} - \log \frac{n_{i'2}}{n_{i'j'}} \right)^2 \qquad (7)$$

where again the factor $(1+K)$ cancels out from the log-ratio differences. Since (6) and (7) are identical, the distances between the rows are shown to be unaffected by the merging of these columns, so the principle of distributional equivalence is satisfied.

4. Application to non-compositional data: spectral mapping

The methodology described in Section 2 applies just as well to positive data that are not necessarily compositional, for example contingency tables or any data measured on a ratio scale. Lewi (1976) independently developed this method, the "spectral mapping" for the analysis and visualization of biological activity spectra. These spectra define an $I \times J$ table of biological activities of a set of I compounds as observed in a battery of J tests. Later Lewi (1980) proposed weights monotonically related to the table margins, since more importance is given to more potent compounds (compounds that are highly active in all or most tests) and to tests that are

more sensitive (tests that produce higher activities from all or most compounds. In this weighted form of spectral mapping, also known as spectral map analysis (SMA), Lewi also found that the marginal "masses" of the table constitute good default weights in the analysis of the double-centred table, where the double-centring removes the component of potency and sensitivity of tests.

Following the work of Lewi (1998), this weighting applies equally well to count data: for example, applying these weights to the rows and columns of the letter counts in Table 1, ratios would be weighted higher when the overall counts are higher. As shown in the distance formulations (3) and (4), one can think of the log-ratios row-wise or column-wise: either the ratios between counts of different letters within the same text are visualized, or the ratios between counts for the same letter across the texts. Figure 3 shows the resulting symmetric weighted LRA (or SMA) map where both texts and letters are represented in principal coordinates. The symmetric map has the advantage that the row and column points can be plotted on the same scale (compare with Figure 2, where it was necessary to scale up the row coordinates to represent the rows on the same scale as the columns), and both configurations have a distance interpretation. The most surprising result of this display is the proximity of the pairs of texts by the same author – one might think that letter counts would not discriminate well between authors, but this map shows otherwise. In fact, a permutation test shows that no other allocation of the 12 row labels (amongst over 10000 possible allocations) gives a lower sum of the six "within-author" distances than the labelling of the configuration in Figure 3 – in this sense the authors are discriminated in the map with a *P*-value less than 0.0001.

Gabriel (1972) showed how the biplot represents differences between variables as the vectors joining them. These *links*, i.e. vectors joining pairs of letters in this example, represent logarithms of ratios of two letters. In the case of compositional data, Aitchison and Greenacre (2002) showed that points that lie in straight lines are an indication of constant "log-contrasts". This property carries over to the general case of the present example. For example, in Figure 3 the letters *k*, *y* and *x* are closely aligned, and Table 4 shows the ratios of *k* and *y* with respect to *x* and the corresponding log-ratios. Figure 4 plots log(*y*/*x*) versus log(*k*/*x*) and there is a clear linear relationship (correlation = 0.93). The weighted regression equation, using the row (book) weights r_i, has a slope of 0.80 and an intercept of 1.34. This implies the model:

$$\log(y/x) = 0.80 \log(k/x) + 1.34$$

or
$$\log(y) - 0.20 \log(x) - 0.80 \log(k) = 1.34 \quad (8)$$

i.e.
$$y = 3.81\, x^{0.2} k^{0.8} \quad (9)$$

On the left of (8) is a linear combination of logarithms of the three letters, with coefficients adding up to 0, hence the term *log-contrast*. Their equivalent multiplicative form, exemplified by (9) has index powers on both sides of the equation having the same sum (1 in this case). In many applications constant log-contrasts such as (8) have a clear substantive meaning and are associated with equilibrium relationships, for example in geology and population genetics (Aitchison, 1980). In the present linguistic context of English texts it is not known if the above equilibrium relationship between the letters *k*, *x* and *y* has any particular substantive relevance, but the relationship is certainly apparent in this data set.

5. Relationship to correspondence analysis

The SVDs on which the weighted LRA (SMA) and CA are based are closely connected. Let us first summarize the matrices being decomposed in each case. We have already seen that the spectral map double-centres the matrix $\mathbf{L} = \log(\mathbf{N})$, using weights proportional to the table margins (CA masses) $\mathbf{A} = (\mathbf{I} - \mathbf{1r}^\mathsf{T})\mathbf{L}(\mathbf{I} - \mathbf{c1}^\mathsf{T})^\mathsf{T}$ (see formula (1)). Then **A** is decomposed using a weighted SVD. Since any constant row- or column-effect added to the elements of **L** will be removed by the double-centring, let us define $\mathbf{L}^*$ as the matrix of logarithms of the so-called Pearson contingency ratios, denoted by q_{ij}:

$$l^*_{ij} = \log(q_{ij}) = \log\left(\frac{n_{ij}}{n_{i+}n_{+j}/n}\right) = \log(n_{ij}) - \log(n_{i+}) - \log(n_{+j}) + \log(n) \qquad (10)$$

so that **A** can be written equivalently as: $\mathbf{A} = (\mathbf{I} - \mathbf{1r}^\mathsf{T})\mathbf{L}^* (\mathbf{I} - \mathbf{c1}^\mathsf{T})^\mathsf{T}$. The contingency ratios are the observed values divided by the "expected" values, where expected value is defined as that obtained if the profiles of the rows (or of the columns) agree perfectly with the average profiles defined by the table margins (the terms observed and expected are used in the context of contingency tables, where the expected value is under the independence hypothesis, but we extend their usage here to all tables of positive numbers). Lewi (1998) aptly terms the contingency ratios as the *double-closure* of the original table, since the (weighted) row and column sums of the matrix **Q** of contingency ratios are all equal to 1.

Now CA, which has many equivalent definitions, can be defined as the double-centring with respect to weighted averages (using the masses as weights) of the matrix **Q**, followed by the weighted SVD. We have the following well-known approximation, using a first-order Taylor approximation:

$$\log(q_{ij}) = \log(1 + q_{ij} - 1) \approx q_{ij} - 1$$

when q_{ij} –1 is small. Since double-centring of $\mathbf{Q}–\mathbf{1}\mathbf{1}^{\mathrm{T}}$ yields the same matrix as double-centring of **Q**, it follows that weighted LRA (SMA) and CA tend to the same solution as q_{ij} –1 tends to 0, that is as "observed" values tend to "expected" ones. In practical terms, whenever variance (called *inertia* in CA) in a matrix is low, the two methods will give approximately the same results. In the case of both practical examples considered here, the variance is indeed low, especially for the letter counts of Table 1. Figure 5 shows the CA symmetric map of Table 2 and it is indeed quite similar to Figure 3, even the amounts and percentages of inertia on each dimension are similar in value. While CA has several interesting graphical properties of its own, such as optimal scaling and maximizing correlation between rows and columns (see, for example, Greenacre (2007: chapter 7)), it does not have subcompositional coherence, nor does it have the model diagnostic features of the weighted log-ratio map – for example, the letters *k*, *x* and *y* are no longer lined up in Figure 5.

5. Relationship to association modelling

Association modelling (Goodman 1968, 1983) for contingency tables is concerned with models for the probability π_{ij} that a case falls into the (*i*,*j*)-th cell of the table. Specifically, the so-called RC(M) association model, where R stands for "row", C for "column" and "M" for the number of bilinear terms in the model, can be written as:

$$\pi_{ij} = \alpha_i \beta_j e^{\phi_1 \mu_{i1} \nu_{j1} + \ldots + \phi_M \mu_{iM} \nu_{jM}} \tag{11}$$

where α_i, β_j, ϕ_m, μ_{im}, ν_{jm} are parameters of the model (i=1,…,I; j=1,…,J; m=1,…,M) with various identification constraints. In logarithmic form this is:

$$\log(\pi_{ij}) = \log(\alpha_i) + \log(\beta_j) + \sum_{m=1}^{M} \phi_m \mu_{im} \nu_{jm} \tag{12}$$

If M = min{I–1, J–1} the model is called "saturated", since it will fit the data perfectly. Usually values M = 1 or 2 are used, the model is fitted by maximum likelihood to the data, and then hypothesis testing allows decisions to be made about how many terms are needed to fit the data, or whether some parameters are equal. Such tests are valid for contingency tables established from a random sample of n individuals on whom two categorical variables are observed. Notice that the RC(M) model estimates the cell probabilities π_{ij}, which are strictly positive, but the data can have zero values.

The parametric model (12) has a form very similar to the data decomposition in the weighted LRA (SMA) and the CA described previously, which can be written respectively as follows, where $p_{ij} = n_{ij}/n$:

Weighted LRA (SMA): $$\log(p_{ij}) = \sum_j c_j \log(p_{ij}) + \sum_i r_i \log(p_{ij}) + \sum_{m=1}^{M} \tilde{\phi}_m \tilde{\mu}_{im} \tilde{\nu}_{jm} \quad (13)$$

CA: $$\log(p_{ij}) = \log(p_{i+}) + \log(p_{+j}) + \log(1 + \sum_{m=1}^{M} \bar{\phi}_m \bar{\mu}_{im} \bar{\nu}_{jm}) \quad (14)$$

$$\approx \log(p_{i+}) + \log(p_{+j}) + \sum_{m=1}^{M} \bar{\phi}_m \bar{\mu}_{im} \bar{\nu}_{jm}$$

where the approximation for CA holds if the summation in (14) is small, i.e., when the data is close to independence (low inertia). The essential difference between these three methods is thus the way the row and column "main effects" and "interaction terms" are estimated. In weighted LRA the weighted row and column averages of the logarithms of observed probabilities estimate the main effects and the interaction terms are obtained by a weighted SVD of the residuals. In CA the row and column sums estimate the (multiplicative) main effects and the interaction terms are obtained by a weighted SVD of the residuals. In association modelling, main effects and interaction terms are estimated simultaneously, for a given "dimensionality" *M*, by maximum likelihood. The similarity between (12) and (13) suggest that association modelling, using the marginal proportions as weights, and weighted LRA will give approximately the same answers, which is indeed the case. We fitted the RC(2) model to the author data using the LEM program (Vermunt, 1997) and the results differ only very slightly from those reported in Figure 5.

6. Discussion

In this article we have shown how the introduction of row and column weights improves both the theoretical properties and practical application of log-ratio analysis. With the convention that weights be added if rows or columns are merged, weighted LRA maps, alias spectral map analysis, obey the principle of distributional equivalence. The chi-square distance in CA and the weighted log-ratio distance are not the only distances that obey this principle. Escofier (1978) shows that the Hellinger distance also has this property: using previous notation, the Hellinger distance (squared) between rows *i* and *i'* is:

$$d_{ii'}^2 = \sum_j \left(\sqrt{\frac{n_{ij}}{n_{i+}}} - \sqrt{\frac{n_{i'j}}{n_{i'+}}} \right)^2$$

(see also Cuadras, Cuadras and Greenacre (2006)). It can also be shown, in a similar way as in Section 3, that a weighted form of normalized PCA is also distributionally equivalent. For example, for a table **N** of non-negative data, normalize the columns *j* by dividing by any appropriate scale-dependent quantity s_j such as the standard-deviation, sum, maximum or range. Then, again using column weights c_j applicable to the problem, define the squared distance between rows as:

$$d_{ii'}^2 = \sum_j c_j \left(\frac{n_{ij}}{s_j} - \frac{n_{i'j}}{s_j} \right)^2$$

Notice here that the data elements can be considered transformed by a single scale value $\sqrt{c_j} / s_j$, but the two parts of this quotient play different roles in the analysis: the s_j normalize the columns to make the columns comparable (the columns could be ratio-scale variables or components in compositional data), while the c_j are used in the centring and weight least-squares fitting of the normalized data. This weighted, normalized PCA has distributional equivalence but not subcompositional coherence.

Bavaud (2002, 2004) defines a broad class of distances based on the contingency ratios (which he aptly calls "independence quotients"), where all distances in this class obey the distributional equivalence principle (Bavaud calls these distances "aggregation invariant"). However, this class does not include the weighted LRA distance (2) but an alternative where the denominators in (2) are the row means rather than their weighted geometric means; in other words, Bavaud's log-transformed data are centred by the log of the mean rather than the mean of the logs.

SMA was developed originally by Lewi (1976) for the analysis of biological activity spectra in the context of drug development. This method has been used extensively in biomedical research, for example Wouters *et al.* (2003) apply it to gene expression data from microarrays and compare it with principal component analysis and CA. In this application context the rationale for the weighting of the rows and columns of the log-transformed data has been to take into account the higher importance of potent compounds and sensitive tests, as explained in Section 4, but the weighting makes sense in the analysis of contingency tables and compositional data as well. As in the case of Table 1, we often find that there is larger relative error in data of lower value, so that weighting the log-ratios takes the precision of measurement into account in this particular way. In the CA of a contingency table, the rationale is similar, since under the assumption of independence, the variability of the contingency ratio for the (i,j)-th cell is approximately $1/(r_i c_j)$, which justifies the weighting in the least-squares formulation by $r_i c_j$, approximately normalizing of the contribution of each row-column term.

In the case of count or abundance data n_{ij}, weighted LRA has the disadvantage of being applicable to strictly positive data only, which rules it out for many social science applications and most ecological applications where data matrices contain many zero frequencies. At a low-level occurrence of zero data $n_{ij} = 0$, one can apply the transformation $C + n_{ij}$ for a positive constant C that depends on the context. In the case of the author data, which had only one zero count, we simply replaced the zero with the value ½. In the case of compositional data and other measurement data, zero values can be replaced by some fraction of the detection limit followed

by an additive or multiplicative adjustment of the remaining values (see Martín-Fernández *et al.*, 2003, for an investigation of the problem of zero values in this context, as well as Beardah *et al.*, 2003, for an extensive practical study of zero treatment strategies as well as a comparison of several alternatives to LRA in compositional data analysis). Apart from this drawback, the method has very similar properties to CA, with several additional benefits such as subcompositional coherence and the model diagnostic properties. Thus, in the case of strictly positive data matrices, weighted LRA alias SMA may be judged superior to CA from a theoretical point of view. In the usual context of CA applications, mostly contingency tables in the social sciences, subcompositional coherence can sometimes be relevant, as explained by Greenacre and Pardo (2006) who describe how a variant of CA can be used to analyse subtables of rows and/or columns of a contingency table. In this so-called *subset CA* the masses of the full table are maintained and the proportions are not closed in the analysis of the subtable, thus giving a "subset coherent" version of CA.

References

Aitchison, J. (1980). Relative variation diagrams for describing patterns of variability in compositional data. *Mathematical Geology 22*, 487–512.

Aitchison, J.P. (1983). Principal component analysis of compositional data. *Biometrika 70*, 57–65.

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

Aitchison, J. & Greenacre, M.J. (2002). Biplots of compositional data. *Applied Statistics 51*, 375–392.

Baxter, M.J., Cool, H.E.M. and Heyworth, M.P. (1990). Principal component and correspondence analysis of compositional data: some similarities. *Journal of Applied Statistics 17*, 229–235.

Bavaud, F. (2002). Quotient dissimilarities, Euclidean embeddability, and Huygens' weak principle. In H.A.L. Kiers *et al.* (eds), *Data Analysis, Classification and Related Methods*, pp. 195–202. New York: Springer.

Bavaud, F. (2004). Generalized factor analyses for contingency tables. In D. Banks *et al.* (eds), *Classification ,Clustering, and Data Mining Applications*, pp. 597–606. New York: Springer.

Beardah, C.C., Baxter, M.J., Cool, H.E.M., and Jackson, C.M. (2003). Compositional data analysis of archaeological glass: problems and possible solutions. *Proceedings of the Compositional Data Analysis Workshop, CODAWORK'03*, Girona, Spain.
URL http://ima.udg.edu/Activitats/CoDaWork03/paper_baxter_Beardah2.pdf

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

Cuadras, C., Cuadras, D. and Greenacre, M.J. (2006). A comparison of methods for analyzing contingency tables. *Communications in Statistics – Simulation and Computation 35*, 447–459.

Cuadras, C. and Fortiana, J. (1998). Visualizing categorical data with related metric scaling. In J. Blasius and M.J. Greenacre (eds), *Visualization of Categorical Data*, pp. 112–129. San Diego: Academic Press.

Escofier, B. (1978). Analyse factorielle et distances répondant au principe d'équivalence distributionelle. *Revue de Statistique Appliquée 26*, 29–37.

Gabriel, K.R. (1971). The biplot-graphical display with applications to principal component analysis. *Biometrika 58,* 453–467.

Gabriel, K.R. (1972). Analysis of meteorological data by means of canonical decomposition and biplots. *Journal of Applied Meteorology 11,* 1071–1077.

Gabriel, K. R. (2002). Goodness of fit of biplots and correspondence analysis. *Biometrika 89,* 423–436.

Goodman, L.A. (1968). The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables, with or without missing entries. *Journal of the American Statistical Association 63*, 1091–1131.

Goodman, L.A. (1985). The analysis of cross-classified data having ordered and/or unordered categories: association models, correlation models, and asymmetry models for contingency tables with or without missing entries. *The Annals of Statistics 13*, 10 – 98.

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

Greenacre, M.J. (1993). Biplots in correspondence analysis. *Journal of Applied Statistics 20,* 251–269.

Greenacre, M.J. (2007). *Correspondence Analysis in Practice. Second Edition.* London: Chapman & Hall / CRC.

Greenacre, M.J. and Blasius, J., eds (1994). *Correspondence Analysis in the Social Sciences.* London: Academic Press.

Greenacre, M.J. and Pardo, R. (2006). Subset correspondence analysis: visualizing relationships among a selected set of response categories from a questionnaire survey. *Sociological Methods and Research 35*, 193–218.

Kazmierczak, J.B. (1988). Analyse logarithmique: deux exemples d'application. *Revue de Statistique Appliquée 33*, 13–24.

Lebart, L., Morineau, A. and Warwick, K. (1984). *Multivariate Descriptive Statistical Analysis.* New York: Wiley.

Lewi, P.J. (1976). Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. *Arzneim. Forsch. (Drug Res.) 26*, 1295–1300.

Lewi, P.J. (1980). Multivariate data analysis in APL. In G.A. van der Linden (ed.), *Proceedings of APL-80 Conference*, pp. 267–271. Amsterdam: North-Holland.

Lewi, P.J. (1998). Analysis of contingency tables. In B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. de Jong, P.J. Lewi, J. Smeyers-Verbeke (eds.), *Handbook of Chemometrics and Qualimetrics: Part B*, Chapter 32, pp. 161–206. Amsterdam: Elsevier.

Martín-Fernández, J.A., Barceló-Vidal, C. And Pawlowsky-Glahn, V. (2003). Dealing with zeros and missing values in compositional data sets. *Mathematical Geology 35*, 253–278.

Nenadić, O. and Greenacre, M.J. (2007). Correspondence analysis in R, with two- and three-dimensional graphics: the ca package. *Journal of Statistical Software 20(3).* URL http://www.jstatsoft.org/v20/i03/.

R Development Core Team (2007). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

S-PLUS, Version 7 (2005). Insightful Corporation, Seattle, WA.URL http://www.insightful.com.

Ter Braak, C.J.F. (1985). Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. *Biometrics 41*, 859–873.

Vermunt, J.K. (1997). LEM: a general program for the analysis of categorical data. Department of Methodology and Statistics, Tilburg University.

Wouters, L., Göhlmann, H.W., Bijnens, L., Kass, S.U., Molenberghs, G. and Lewi, P.J. (2003). Graphical exploration of gene expression data: a comparative study of three multivariate methods. *Biometrics 59*, 1131–139.

Table 1 Letter counts in 12 samples of texts from books by six different authors (R Development Core Team, 2005).

Abbrev.	*a*	*b*	*c*	*d*	*e*	*f*	*g*	*h*	*i*	*j*	*k*	*l*	*m*
TD-Buck	550	116	147	374	1015	131	131	493	442	2	52	302	159
EW-Buck	557	129	128	343	996	158	129	571	555	4	76	291	247
Dr-Mich	515	109	172	311	827	167	136	376	432	8	61	280	146
As-Mich	554	108	206	243	797	164	100	328	471	4	34	293	149
LW-Clark	590	112	181	265	940	137	119	419	514	6	46	335	176
PF-Clark	592	151	251	238	985	168	152	381	544	7	39	416	236
FA-Hem	589	72	129	339	866	108	159	449	472	7	59	264	158
Is-Hem	576	120	136	404	873	122	156	593	406	3	90	281	142
SF7-Faul	541	109	136	228	763	126	129	401	520	5	72	280	209
SF6-Faul	517	96	127	356	771	115	189	478	558	6	80	322	163
Pen3-Holt	557	97	145	354	909	97	121	479	431	10	94	240	154
Pen2-Holt	541	93	149	390	887	133	154	463	518	4	65	265	194

Abbrev.	*n*	*o*	*p*	*q*	*r*	*s*	*t*	*u*	*v*	*w*	*x*	*y*	*z*
TD-Buck	534	516	115	4	409	467	632	174	66	155	5	150	3
EW-Buck	479	509	92	3	413	533	632	181	68	187	10	184	4
Dr-Mich	470	561	140	4	368	387	632	195	60	156	14	137	5
As-Mich	482	532	145	8	361	402	630	196	66	149	2	80	6
LW-Clark	403	505	147	8	395	464	670	224	113	146	13	162	10
PF-Clark	526	524	107	9	418	508	655	226	89	106	15	142	20
FA-Hem	504	542	95	0	416	314	691	197	64	225	1	155	2
Is-Hem	516	488	91	3	339	349	640	194	40	250	3	104	5
SF7-Faul	471	589	84	2	324	454	672	247	71	160	11	280	1
SF6-Faul	483	617	82	8	294	358	685	225	37	216	12	171	5
Pen3-Holt	417	477	100	3	305	415	597	237	64	194	9	140	4
Pen2-Holt	484	545	70	4	299	423	644	193	66	218	2	127	2

Abbreviations: TD (Three Daughters), EW (East Wind) – Buck (Pearl S. Buck)

Dr (Drifters), As (Asia) – Mich (James Michener)

LW (Lost World), PF (Profiles of Future) – Clark (Arthur C. Clarke)

FA (Farewell to Arms), Is (Islands) – Hem (Ernest Hemingway)

SF7 and SF6 (Sound and Fury, chapters 7 and 6) – Faul (William Faulkner)

Pen3 and Pen2 (Bride of Pendorric, chapters 3 and 2) – Holt (Victoria Holt)

Table 2 Percentage compositions of 47 Roman glass cups (Baxter et al 1990).

Cups	*Si*	*Al*	*Fe*	*Mg*	*Ca*	*Na*	*K*	*Ti*	*P*	*Mn*	*Sb*
1	75.2	1.84	0.26	0.47	5.00	16.3	0.44	0.06	0.04	0.01	0.36
2	72.4	1.80	0.28	0.46	5.89	18.2	0.44	0.06	0.04	0.01	0.33
3	69.9	2.08	0.40	0.57	6.33	19.5	0.54	0.09	0.06	0.03	0.44
4	70.2	2.23	0.41	0.60	6.10	19.5	0.42	0.08	0.05	0.01	0.34
5	73.0	2.16	0.35	0.51	5.66	17.3	0.44	0.07	0.05	0.01	0.37
6	74.2	2.02	0.33	0.51	5.34	16.5	0.52	0.07	0.05	0.01	0.35
7	74.2	1.80	0.25	0.39	5.35	17.1	0.44	0.06	0.04	0.01	0.31
8	74.4	1.74	0.27	0.42	5.41	16.8	0.49	0.06	0.05	0.01	0.31
9	72.8	1.81	0.30	0.66	5.86	17.6	0.40	0.07	0.04	0.01	0.33
10	74.8	1.71	0.22	0.35	5.48	16.3	0.42	0.06	0.05	0.01	0.51
11	75.0	1.74	0.22	0.32	5.03	16.8	0.43	0.05	0.05	0.01	0.30
12	73.8	1.93	0.31	0.42	4.94	17.6	0.43	0.05	0.04	0.01	0.38
13	70.3	1.94	0.30	0.44	6.31	19.5	0.57	0.07	0.05	0.01	0.39
14	72.7	1.74	0.25	0.37	5.90	17.8	0.50	0.06	0.05	0.01	0.53
15	74.3	1.88	0.30	0.40	4.76	17.3	0.41	0.05	0.04	0.01	0.48
16	70.2	2.23	0.42	0.56	6.65	18.7	0.61	0.09	0.06	0.02	0.35
17	73.1	1.90	0.29	0.41	5.13	18.2	0.45	0.05	0.04	0.01	0.31
18	73.7	1.78	0.23	0.32	4.98	18.1	0.45	0.06	0.04	0.01	0.27
19	73.3	1.89	0.30	0.41	5.37	17.8	0.42	0.07	0.04	0.01	0.30
20	71.7	1.75	0.27	0.42	6.04	19.0	0.41	0.06	0.05	0.01	0.24
21	73.7	1.80	0.25	0.36	5.15	17.9	0.45	0.06	0.04	0.01	0.18
22	73.1	1.82	0.23	0.32	5.13	18.4	0.46	0.06	0.04	0.01	0.38
23	73.0	1.90	0.27	0.44	5.48	17.9	0.52	0.07	0.05	0.01	0.28
24	68.8	2.03	0.38	0.51	7.02	20.0	0.59	0.07	0.06	0.02	0.40
25	70.2	2.11	0.42	0.59	6.53	19.0	0.53	0.08	0.06	0.03	0.33
26	70.5	2.11	0.39	0.56	6.18	19.1	0.57	0.07	0.05	0.02	0.37
27	72.7	1.96	0.30	0.50	5.58	17.9	0.52	0.07	0.05	0.02	0.28
28	73.1	1.78	0.26	0.42	5.48	17.9	0.46	0.06	0.05	0.01	0.36
29	69.3	2.21	0.45	0.54	6.87	19.4	0.57	0.10	0.06	0.02	0.41
30	70.2	2.25	0.43	0.54	6.77	18.7	0.54	0.09	0.06	0.02	0.31
31	74.4	1.94	0.26	0.46	5.07	17.0	0.47	0.07	0.05	0.01	0.18
32	73.9	1.90	0.26	0.46	5.04	17.6	0.45	0.07	0.04	0.01	0.20
33	72.6	1.81	0.27	0.41	5.48	18.5	0.37	0.07	0.05	0.01	0.31
34	69.9	1.87	0.32	0.46	6.34	19.8	0.58	0.07	0.06	0.02	0.49
35	69.7	2.04	0.36	0.48	6.20	19.8	0.56	0.07	0.06	0.01	0.58
36	72.3	2.08	0.36	0.53	5.47	18.0	0.58	0.08	0.06	0.01	0.49
37	70.5	2.00	0.33	0.59	5.83	19.8	0.42	0.09	0.05	0.01	0.33
38	72.3	1.71	0.21	0.36	5.27	18.8	0.48	0.06	0.07	0.01	0.63
39	72.2	2.02	0.34	0.51	5.36	18.4	0.54	0.08	0.05	0.01	0.46
40	73.8	1.88	0.26	0.45	5.12	17.6	0.45	0.07	0.05	0.01	0.21
41	72.4	1.92	0.29	0.48	5.45	18.4	0.51	0.07	0.05	0.02	0.38
42	72.6	2.00	0.33	0.46	5.41	17.7	0.75	0.08	0.08	0.01	0.54
43	71.6	1.90	0.27	0.48	5.32	19.4	0.47	0.06	0.05	0.01	0.35
44	72.3	2.03	0.30	0.48	5.41	18.6	0.50	0.07	0.05	0.01	0.21
45	73.4	1.93	0.24	0.37	5.18	17.8	0.55	0.06	0.04	0.01	0.30
46	71.7	2.02	0.42	0.53	5.73	18.3	0.62	0.10	0.06	0.02	0.39
47	69.3	2.04	0.40	0.50	6.85	19.5	0.62	0.08	0.06	0.02	0.57
mean	72.31	1.94	0.31	0.46	5.66	18.24	0.50	0.07	0.05	0.01	0.36

Table 3 Percentage contributions by components in unweighted and weighted log-ratio maps, where the weights are given by the column means of Table 2. In the unweighted analysis the rare components Mn and Sb dominate, while in the weighted analysis more components contribute to the solution, including the most frequent one, Si.

	unweighted	***weighted***
Si	7.11	21.05
Al	2.57	2.76
Fe	2.15	4.34
Mg	2.94	3.44
Ca	0.51	25.93
Na	2.89	22.33
K	0.23	2.20
Ti	1.92	0.53
P	0.80	0.37
Mn	39.48	0.37
Sb	39.39	16.68

Table 4 Ratios and log-ratios between letter counts for *y*, *k* and *x*

Book	***y/x***	***k/x***	***ln(y/x)***	***ln(k/x)***
TD-Buck	30.0	10.4	3.401	2.342
EW-Buck	18.4	7.6	2.912	2.028
Dr-Mich	9.8	4.4	2.281	1.472
As-Mich	40.0	17.0	3.689	2.833
LW-Clark	12.5	3.5	2.523	1.264
PF-Clark	9.5	2.6	2.248	0.956
FA-Hem	155.0	59.0	5.043	4.078
Is-Hem	34.7	30.0	3.546	3.401
SF7-Faul	25.5	6.5	3.237	1.879
SF6-Faul	14.3	6.7	2.657	1.897
Pen3-Holt	15.6	10.4	2.744	2.346
Pen2-Holt	63.5	32.5	4.151	3.481

Figure 1 Unweighted log-ratio biplot of Baxter data, showing rows in principal coordinates and columns in standard coordinates (form biplot). Row coordinate values have been multiplied by 10. The two-dimensional solution explains 70.0% of the total variance.

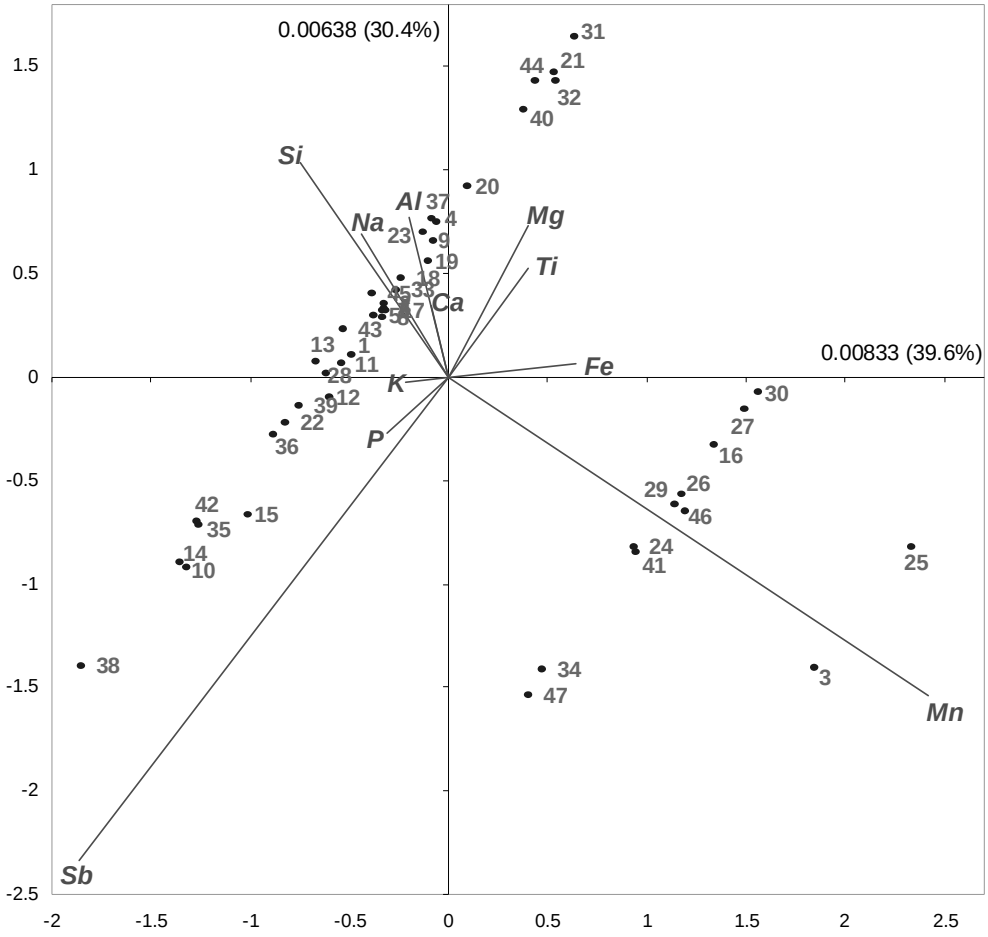

Figure 2 Weighted log-ratio biplot of Baxter data, showing rows in principal coordinates and columns in standard coordinates (form biplot). Row coordinate values have been multiplied by 50. The two-dimensional solution explains 79.7% of the total variance.

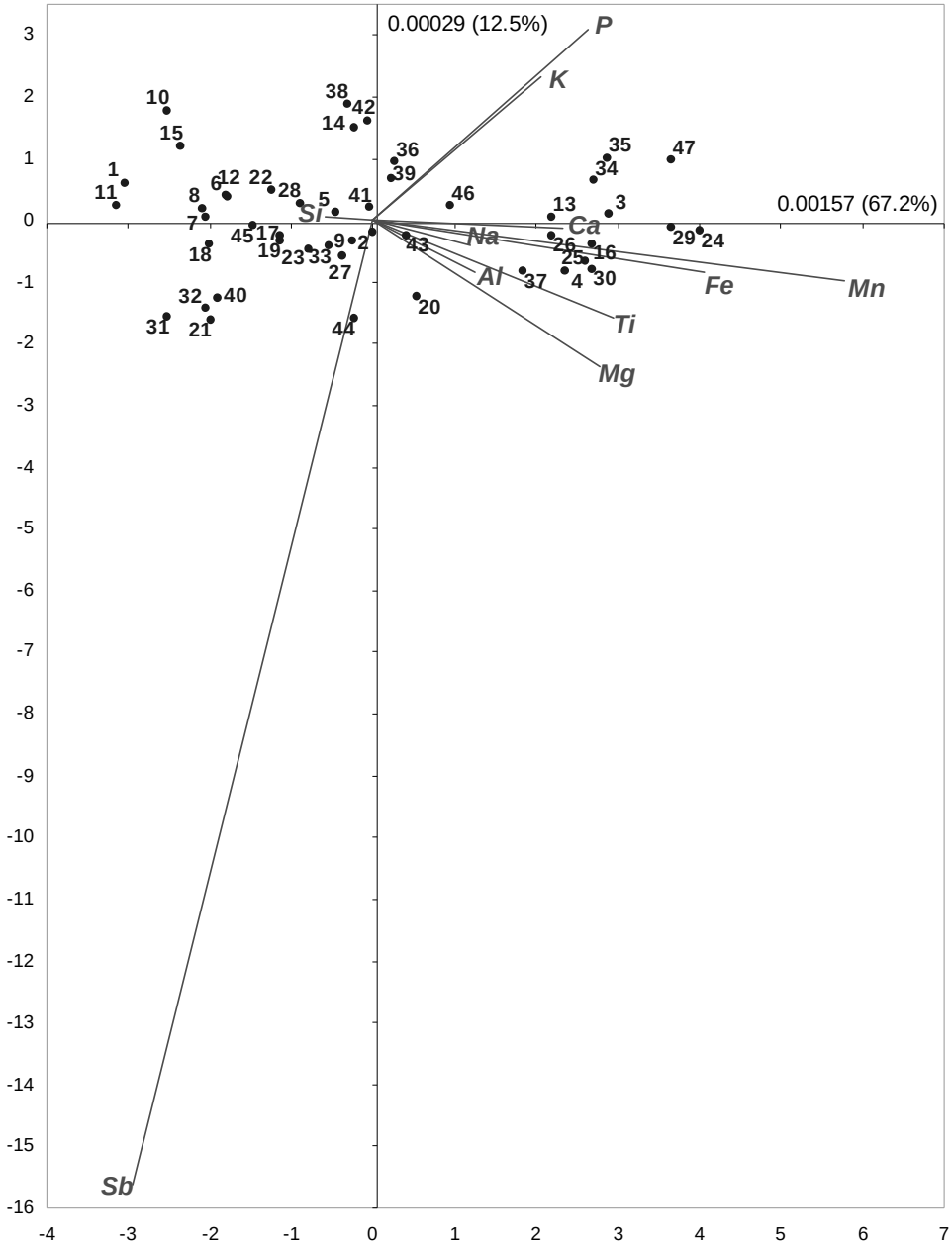

Figure 3 Weighted log-ratio map of author data, showing both rows and columns in principal coordinates (symmetric map). The two-dimensional solution explains 59.5% of the total variance.

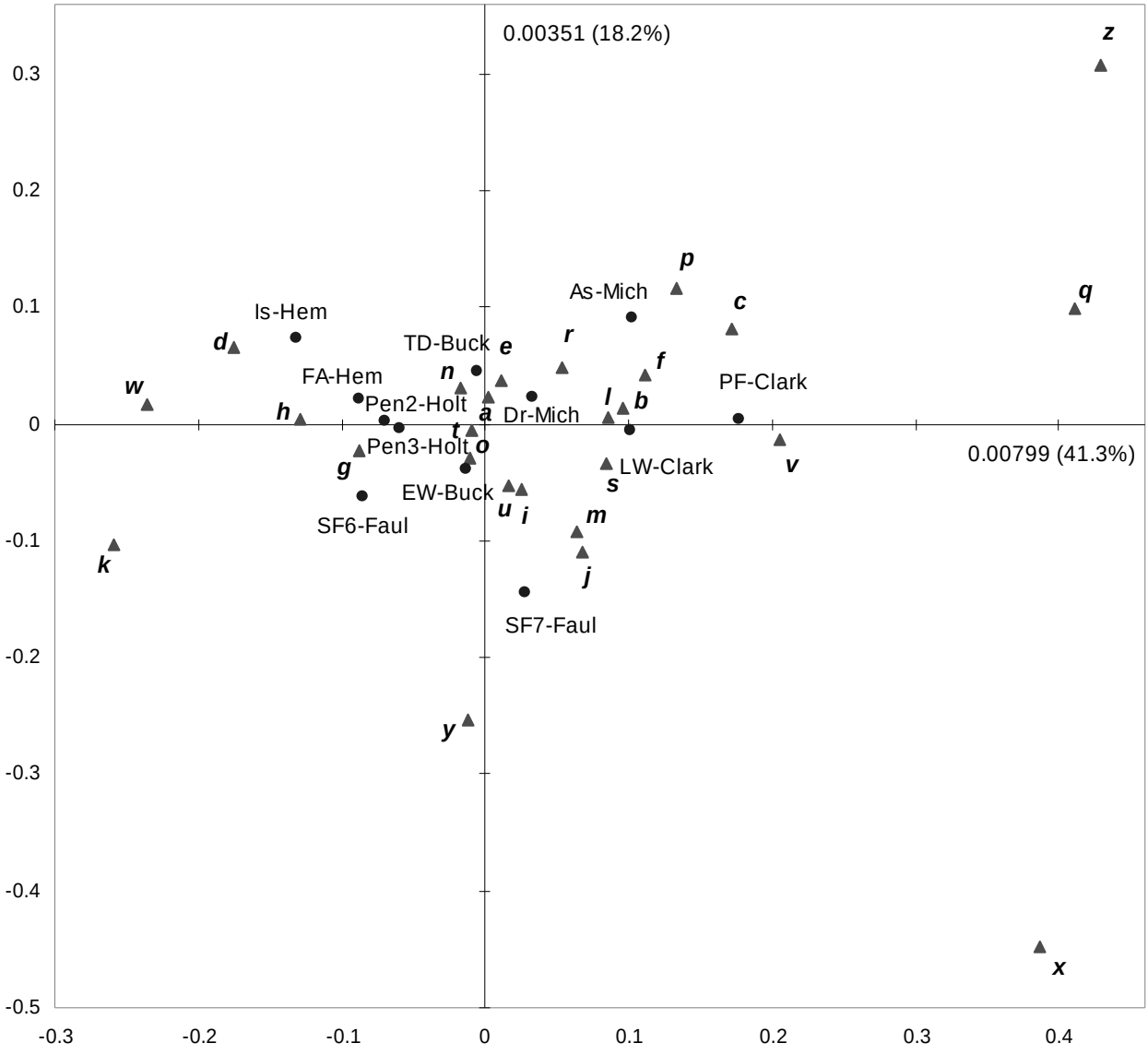

Figure 4 Scatterplot of log-ratios in Table 4, showing the relationship diagnosed by the lining up of letters *k*, *x* and *y* in the weighted log-ratio map of Figure 3. The regression line indicated has slope 0.80 and intercept 1.34.

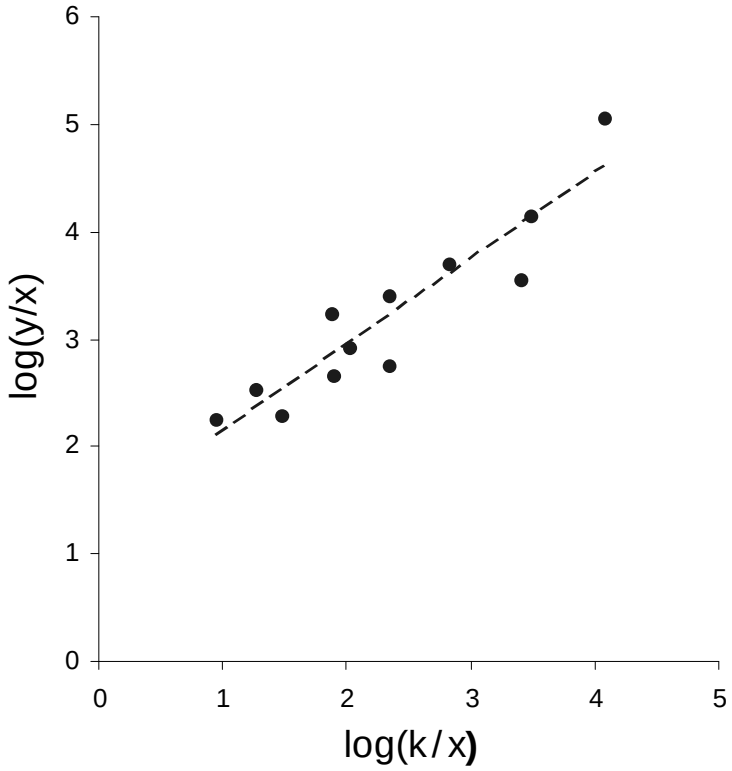

Figure 5 CA map of author data, showing both rows and columns in principal coordinates (symmetric map). The two-dimensional solution explains 60.6% of the total inertia.