Weighted Logratio Biplots, Correspondence Analysis and Spectral Maps

Michael Greenacre[1] **and Paul Lewi**[2]
[1]Universitat Pompeu Fabra, Barcelona, Spain; *michael@upf.es*
[2]Center for Molecular Design, Janssen Pharmaceutica , Belgium; *PLEWI@PRDBE.jnj.com*

Abstract

Starting with logratio biplots for compositional data, which are based on the principle of subcompositional coherence, and then adding weights, as in correspondence analysis, we rediscover Lewi's spectral map and many connections to analyses of two-way tables of non-negative data. Thanks to the weighting, the method also achieves the property of distributional equivalence.

Keywords: biplot; compositional data; contingency tables; correspondence analysis; distributional equivalence; logratio transformation; singular value decomposition; spectral map.

1 Introduction

Compositional data analysis is concerned with the analysis of positive data with row sums (or column sums) equal to a constant, usually 1 if the data are proportions or 100 if they are percentages. One of the reigning principles in Aitchison's (1980, 1983, 1986) approach is that of *subcompositional coherence*, that is the invariance in the analysis to considering subcompositions by themselves or as part of the full composition. With this principle in mind, ratios of values in the data matrix are considered, since these are invariant to taking subcompositions. Aitchison and Greenacre (2002) defined a biplot based on the logarithms of ratios, or *logratios*, which has many interesting properties and links to the modelling of two-way tables. As we will show, however, this biplot can be perturbed by low-value components. A remedy is to introduce weights for each component, proportional to their means, in the same way as rows and columns are weighted proportional to margins in correspondence analysis. This leads to what we call a "weighted logratio biplot" (Greenacre 2002), a method shown to have all the properties of the logratio biplot previously defined by Aitchison and Greenacre, but in addition obeying the principle of *distributional equivalence*, one of the cornerstones of correspondence analysis. Furthermore, the method can be applied to any matrix of positive data, and turns out to be identical to Lewi's *spectral map*, defined originally by Lewi (1976) in the context of analysing biological activity spectra. Wouters *et al.* (2003) compare principal component analysis, correspondence analysis and spectral mapping on the same data.

2 Weighted logratio biplot

Suppose that $\mathbf{N} = [n_{ij}]$ denotes an $I \times J$ table of positive data, with row totals, column totals and grand total denoted by n_{i+}, n_{+j} and n respectively. Let $r_i = n_{i+}/n$ and $c_j = n_{+j}/n$ be the row and column sums relative to the grand total, which we call *masses*. Let $\mathbf{r}$ be the vector of row masses, $\mathbf{c}$ the vector of column masses and $\mathbf{D}_r$ and $\mathbf{D}_c$ the corresponding diagonal matrices. Denote by $\mathbf{L}$ the matrix of logarithms of the frequencies, $l_{ij} = \log(n_{ij})$. Aitchison's "relative variation diagram" (Aitchison 1980, 1983) 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 with column-centring but no column-normalization.

Applying this algorithm to Baxter's data on the chemical composition of 47 Roman glass cups from Colchester (Baxter, Cool and Heyworth 1990), we obtain the map in Figure 1, reported by the same authors. This map shows three diagonal bands of points which are due to the values of the element manganese (Mn) which takes on only three different values in the data set, all very small: 0.01, 0.02 and 0.03. These values, reported (and presumably measured) to two decimal places on a percentage scale, engender large differences in the logratios; for example, amongst themselves there are differences as high as threefold, hence manganese has a much higher variance than any other component in the data set. As a consequence, this rare component dominates the solution, as can be seen in Figure 1, with samples 3 and 25 having the highest values (0.03%).

In order to downgrade the role of such rare components with high logratio variance, an idea is imported from correspondence analysis to use the row and column masses as weights. In this "weighted logratio" approach, the row and column masses are introduced first into the double-centring stage, so that centring is with respect to weighted averages, and then into the matrix approximation stage, so that fitting is by weighted least squares. This simple modification of the algorithm, giving differential importance to the rows and columns in the centring and fitting, rectifies the above problem and also bestows on the method the principle of distributional equivalence, which is a fundamental property of correspondence analysis.

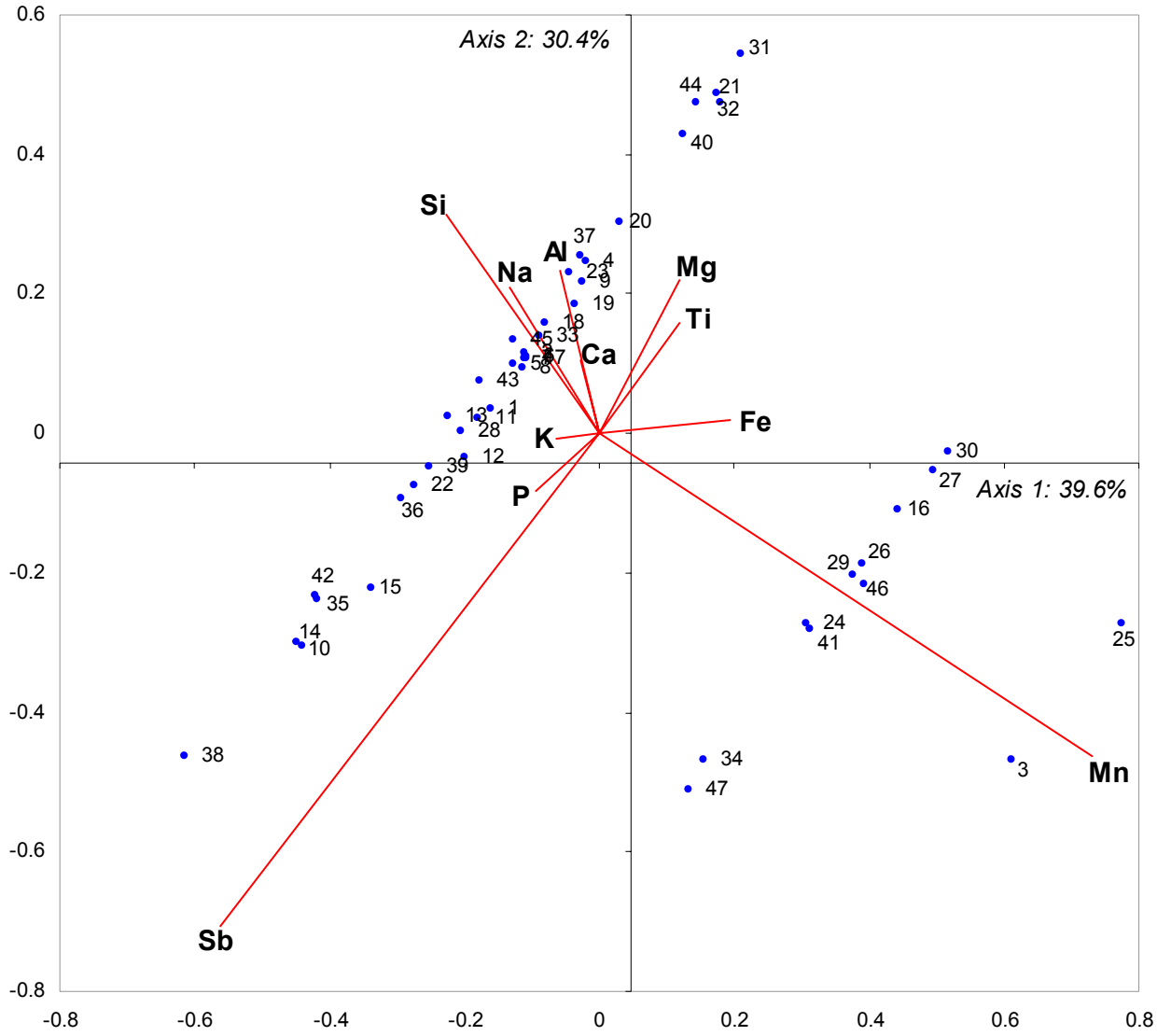

Figure 1: Unweighted logratio biplot of Baxter data, showing rows in principal coordinates (form biplot).

The computational steps to find the coordinates of the rows and columns in the weighted logratio map are as follows:

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: $l_{i\cdot} = \sum_j c_j \, l_{ij}$ (i=1,⋯,I), subtracting these averages from all the elements in the corresponding row. The resultant matrix, with general element $l_{ij} - l_{i\cdot}$, is now centred with respect to weighted averages of the columns, using the row masses to weight each element: $\sum_i r_i \, (l_{ij} - l_{i\cdot})$ (j=1,⋯,J), subtracting these averages from all the elements in the corresponding columns. The result of this operation is the double-centred matrix with elements $z_{ij} = l_{ij} - l_{i\cdot} - l_{\cdot j} + l_{\cdot\cdot}$, 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):

$$\mathbf{Z} = (\mathbf{I} - \mathbf{1}\mathbf{r}^{\mathsf{T}})\mathbf{L}(\mathbf{I} - \mathbf{c}\mathbf{1}^{\mathsf{T}})$$

.

Step 2. Multiply z_{ij} by $(r_i\ c_j)^{1/2}$, that is multiply the rows and columns by the square root of their respective masses (this transformation induces the weighted least-squares approximation):

$$\mathbf{S} = \mathbf{D}_r^{1/2}\,\mathbf{Z}\,\mathbf{D}_c^{1/2}$$

Step 3. Perform the SVD of this transformed matrix:

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

and the singular values down the diagonal of $\boldsymbol{\Gamma}$ 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 $r_i^{1/2}$, and the rows of the matrix of right singular vectors by $c_j^{1/2}$:

$$\text{(row standard)}\ \ \mathbf{X} = \mathbf{D}_r^{-1/2}\mathbf{U} \qquad\qquad \text{(column standard)}\ \ \mathbf{Y} = \mathbf{D}_c^{-1/2}\mathbf{V}$$

The row and column *principal coordinates a*re the standard coordinates scaled by the singular values:

$$\text{(row principal)}\ \ \mathbf{F} = \mathbf{X}\boldsymbol{\Gamma} = \mathbf{D}_r^{-1/2}\mathbf{U}\boldsymbol{\Gamma} \qquad\qquad \text{(column principal)}\ \ \mathbf{G} = \mathbf{X}\boldsymbol{\Gamma} = \mathbf{D}_c^{-1/2}\mathbf{V}\boldsymbol{\Gamma}$$

Notice how the masses are used to pre-transform the matrix and post-transform the resultant singular vectors, which is effectively performing a weighted (or generalized) SVD (Greenacre 1984: Appendix). As in all methods of this type, we can choose to represent either of two so-called *asymmetric maps*, representing the respective rows and columns by either (i) **F** and **Y** – this map is also called "row-principal" or "row-metric-preserving (RMP)", or (ii) **X** and **G** – this asymmetric map which is "column-principal" or "column-metric-preserving (CMP)". Both asymmetric maps are biplots in the true sense of the term (Gabriel 1971), where row-column scalar products approximate the elements of the double-centred matrix **Z**. When the data are in the usual row by column format of cases (or samples) by variables, Aitchison and Greenacre (2002) call the RMP biplot a *form biplot* and the CMP biplot a *covariance biplot*. A popular alternative map, especially in correspondence analysis, is the *symmetric map*, using principal coordinates **F** and **G** for both rows and columns. 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.

3 Application and properties of weighted logratio biplot

The weighted logratio biplot of Baxter's glass cup data is given in Figure 2. Notice that the effect of the rare element Mn has been downplayed in the map and we obtain a visualization of the data that is distributed more evenly over all the elements. The first axis shows the contrast between the most frequent element, Silicon (Si) and all the others. Notice that because of the weighting, the origin of the biplot is at the centroid of the elements, where the elements are weighted by their respective average percentages. The second axis shows a contrast between antimony (Sb), along with phosphorous and potassium (P and K) to a lesser extent, and the other elements that are positive on the vertical axis. If one is not interested in the Silicon/non-Silicon contrast, which is accentuated by our weighting of the elements proportional to their averages, then the projection onto axes 2 and 3 would be of primary interest (in the workshop presentation of this example, a three-dimensional view of the points will be shown, with rotation of the configuration).

All the properties of the unweighted logratio map listed by Aitchison and Greenacre (2002) carry over to the present weighted version: the multidimensional scaling properties in terms of representing ratios of elements by lines connecting the respective pairs of points, and the diagnosis of equilibrium models for subcompositions of elements that lie on an approximate straight line. In addition, thanks to the weighting, the map now obeys the *principle of distributional equivalence*, which is one of the cornerstones of correspondence analysis. Suppose two columns j and j' (e.g., elements or compounds) have the same relative values (in correspondence analysis we say that they have the same *profile*), that is the ratios $n_{ij}/n_{ij'}$ are identical for all i. Then they can be amalgamated into one column with values equal to the sum of respective elements without any change to the analysis (this property is proved by Greenacre (2002)). In

compositional data analysis this property would be highly desirable, assuring invariance with respect to combining components that are essentially identical across samples apart from their overall levels.

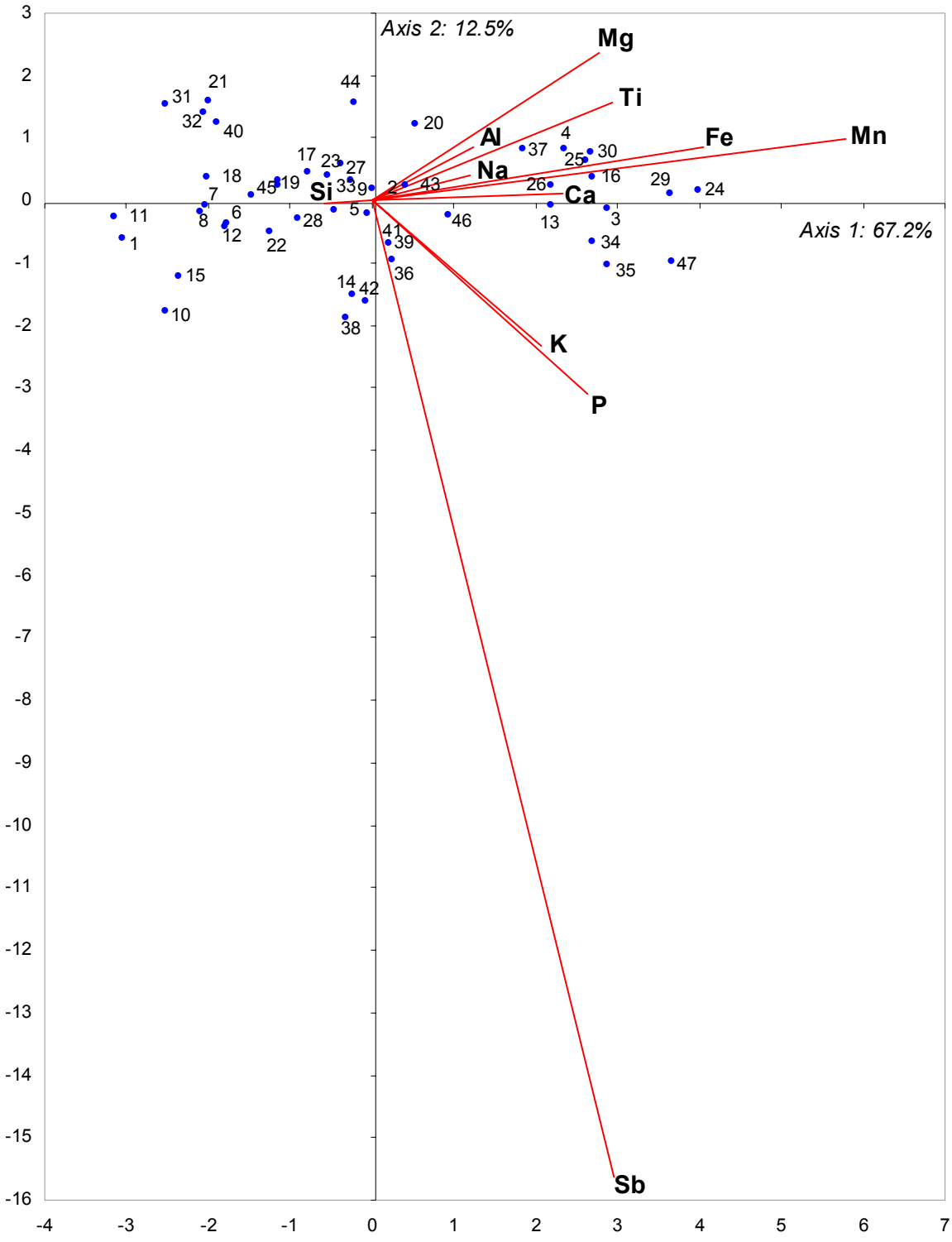

Figure 2: Weighted logratio biplot of Baxter data, showing rows in principal coordinates (form biplot).

4 Beyond compositional data: the spectral map

Everything described above applies equally to any table of positive numbers. In fact, the methodology defined in Section 2 as a weighted version of the logratio biplot of Aitchison and Greenacre (2002) is identical to the so-called *spectral map* developed by Lewi (1976) in applications to biological activity

spectra, and more recently by Wouters *et al* (2003) to gene expression data. The spectral map can also be applied to contingency tables and is thus a direct competitor to correspondence analysis as a visualization method for such data – see Lewi (1998). It has the disadvantage of being applicable to positive data only, which rules it out for many social science applications and most ecological applications where the data matrix contains zero frequencies. As in the analysis of contingency tables by logarithmic models, zeros could be replaced by some acceptable positive value, for example half the detection limit (i.e., ½ in the case of count data). Apart from this drawback, the method has very similar properties to correspondence analysis and may even be judged superior from a theoretical viewpoint since correspondence analysis does not possess subcompositional coherence.

When the data have low variability, correspondence analysis and the weighted logratio biplot are almost identical. This fact stems from the approximation:

$$\log\left(\frac{p_{ij}}{r_i c_j}\right) = \log\left(1 + \frac{p_{ij}}{r_i c_j} - 1\right) \approx \frac{p_{ij}}{r_i c_j} - 1 \qquad \text{when } p_{ij} \approx r_i c_j$$

where $p_{ij} = n_{ij} / n$ is the value of the (i,j)-th cell relative to the grand total, and where the ratios $p_{ij} /(r_i c_j)$ are called *contingency ratios* in the context of contingency tables, i.e. the ratios of "observed" to "expected" frequencies. The weighted logratio biplot / spectral map involves a double-centring and weighted SVD of the left-hand expression, while correspondence analysis involves the double-centring and weighted SVD (with the same weights) of the right-hand expression. When p_{ij} is close to $r_i c_j$, i.e. $p_{ij} /(r_i c_j) - 1$ is small, the two approaches are practically the same, but if there is high variance in the data, so that "observed" and "expected" values are widely different, the approaches give different results. This observation holds equally well for compositional data.

Acknowledgements

The authors acknowledge the support of the Fundación BBVA and Janssen Pharmaceutica, respectively.

References

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

Aitchison, J. (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.

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

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

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

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

Greenacre, M.J. (2002). Ratio maps and correspondence analysis. Research report 598, Departament d'Economia i Empresa, Universitat Pompeu Fabra, submitted for publication.

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. (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.

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.