Math Geosci (2010) 42: 129–134
DOI 10.1007/s11004-008-9212-2

SHORT NOTE

Log-Ratio Analysis Is a Limiting Case of Correspondence Analysis

Michael Greenacre

Received: 6 August 2008 / Accepted: 31 October 2008 / Published online: 27 January 2009

Abstract It is common practice in compositional data analysis to perform the log-ratio transformation in order to preserve sub-compositional coherence in the analysis. Correspondence analysis is an alternative approach to analyzing ratio-scale data and is often contrasted with log-ratio analysis. It turns out that if one introduces a power transformation into the correspondence analysis algorithm, then the limit of the power-transformed correspondence analysis, as the power parameter tends to zero, is exactly the log-ratio analysis. Depending on how the power transformation is applied, we can obtain as limiting cases either Aitchison's unweighted log-ratio analysis or the weighted form called "spectral mapping". The upshot of this is that one can come as close as one likes to the log-ratio analysis, weighted or unweighted, using correspondence analysis.

Keywords Compositional data analysis · Contingency ratios · Distributional equivalence · Log-ratios · Power transformation · Spectral mapping

1 Introduction

Log-ratio analysis (LRA) is one of the methods of choice for distance-based analysis of compositional data, leading to visualizations of samples and components in a reduced space (see Aitchison 1983, 1986, 1990; Aitchison and Greenacre 2002). Simply stated, LRA is the principal component analysis (PCA) of a compositional table (where row elements sum to 1) after logarithmically transforming the table and

Electronic supplementary material The online version of this article (http://dx.doi.org/10.1007/s11004-008-9212-2) contains supplementary material, which is available to authorized users.

M. Greenacre (✉)
Departament d'Economia i Empresa, Universitat Pompeu Fabra, Ramon Trias Fargas, 25-27, Barcelona, 08005, Spain
e-mail: michael@upf.es

centering each row of the log-transformed values by its respective row mean. Since the first step of the ensuing PCA is to center the columns of the table, it is said that the log-transformed table is double-centered—the dimension-reduction step is then performed using the singular value decomposition.

A different approach is to use correspondence analysis (CA), a method applicable to any table of nonnegative numbers, as long as they are all on the same ratio-scale of measurement, and hence suitable for compositional data as well. The table is first centered with respect to the "expected values" based on the row and column margins of the table, a term that is borrowed from contingency table analysis. The rows and columns are weighted proportional to these marginal values—in the case of compositional data samples (rows) would have the same weights but components (columns) would be weighted proportionally to their average in the data set. The subsequent dimension-reduction step is similar to that of PCA apart from the row and column weighting factors (for a recent practical account of CA, see Greenacre 2007, 2008a). Several authors, including the present one, have contrasted LRA and CA as if they were competing methodologies. Beardah et al. (2003) and Greenacre and Lewi (2008) apply both methods to the same data set and compare their results. Moreover, Greenacre and Lewi (2008) find that introducing weights into LRA, in exactly the same way as in CA, improves the theoretical and practical properties of LRA. Furthermore, they point out that this weighted form of LRA has existed since the mid 1970s under the name of "spectral mapping", developed in the pharmaceutical industry for analyzing activity spectra of new chemical compounds (Lewi 1976, 1980).

Putting aside these interesting details about parallel developments of log-ratio methodologies, the purpose of the present paper is to point out to compositional data analysts that LRA, in both its unweighted form (à la Aitchison) and weighted form (à la Greenacre and Lewi), is actually directly linked to CA through the Box–Cox power transformation. This discovery has a number of implications for compositional data analysis, the essence of which is that there exists a family of methods parameterized by a power-transformation of the original compositional data: when this power is equal to 1 the ensuing method is exactly CA, and when this power tends to zero the limiting method is exactly LRA. In between we have a continuum of interesting special cases, for example, square root and double square root transformations, but the main point is that these two apparently unrelated and competing methods are really members of a wider common family.

The rest of this short note summarizes the main results and then discusses the wider implications for the analysis of compositional data.

2 Box–Cox Power Transformations

The Box–Cox transformation (Box and Cox 1964) with power parameter α is defined as follows

$$\begin{aligned} f(x) &= (1/\alpha)\big(x^{\alpha} - 1\big), \quad \alpha > 0 \\ &= \log(x), \quad \alpha = 0. \end{aligned} \tag{1}$$

In fact, $f(x) \to \log(x)$ as $\alpha \to 0$.

This transformation is often used in statistics to symmetrize the distribution of a response variable in a regression model to satisfy the model assumptions (Hinkley 1975). Power transformations are routinely performed in many applied contexts. For example, in the analysis of frequency data, assuming the counts follow a Poisson distribution, the square root transformation is used to stabilize the variance (Bartlett 1936). In ecological research, abundance data are almost always highly overdispersed and some ecologists routinely apply a fourth-root transformation before proceeding with statistical analysis (Field et al. 1982). Here we import these ideas into the analysis of compositional data to produce families of methods where the power parameter α is considered on a continuous scale from 1 to the limit of 0.

3 From Correspondence Analysis to Log-Ratio Analysis

CA has many equivalent definitions, but we focus on two of them here for our purpose, one based on the differences between observed and expected values in the table and the other on their ratios. Suppose that $\mathbf{X}$ is the $I \times J$ table of compositional data, then divide $\mathbf{X}$ by its grand total x_{++} to obtain the so-called "correspondence matrix" $\mathbf{P} = (1/x_{++})\mathbf{X}$. Let the row and column marginal totals of $\mathbf{P}$ be the vectors $\mathbf{r}$ and $\mathbf{c}$, respectively—these are the weights, or "masses", associated with the rows and columns. Let $\mathbf{D}_r$ and $\mathbf{D}_c$ be the diagonal matrices of these masses. The two definitions are based on the following equivalent expressions for the same matrix $\mathbf{S}$

Definition 1

$$\mathbf{S} = \mathbf{D}_r^{-1/2}\big(\mathbf{P} - \mathbf{r}\mathbf{c}^\top\big)\mathbf{D}_c^{-1/2}. \tag{2}$$

Definition 2

$$\mathbf{S} = \mathbf{D}_r^{1/2}\big(\mathbf{I} - \mathbf{1}\mathbf{r}^\top\big)\big(\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1}\big)\big(\mathbf{I} - \mathbf{1}\mathbf{c}^\top\big)^\top\mathbf{D}_c^{1/2}. \tag{3}$$

In Definition 1, the expected values in the matrix $\mathbf{r}\mathbf{c}^\top$ are subtracted from the data: $\mathbf{P} - \mathbf{r}\mathbf{c}^\top$, with elements $p_{ij} - r_i c_j$, whereas in Definition 2 they are divided into the data: $\mathbf{D}_r^{-1}\mathbf{P}\mathbf{D}_c^{-1}$, with elements $p_{ij}/(r_i c_j) \equiv q_{ij}$, called the "contingency ratios". The (weighted) double-centering of the data is apparent in the second definition. CA then continues by performing a singular value decomposition of the matrix $\mathbf{S}$, which provides the row and column coordinates for the joint display of samples and components.

The following results are a direct consequence of the Box–Cox transformation (1):

Result 1 Perform a power transformation of the original data $x_{ij}(\alpha) = x_{ij}^\alpha$. Then perform CA on the transformed data matrix $\mathbf{X}(\alpha)$; i.e., recompute the correspondence matrix, margins (masses), and so on, as in (2). Divide the new matrix $\mathbf{S}$ by α before applying the singular value decomposition. This procedure converges exactly to unweighted LRA as α tends to 0. At the limit each margin of the correspondence matrix has equal elements.

Result 2 Perform a power transformation of the contingency ratios $q_{ij}(\alpha) = [p_{ij}/(r_i c_j)]^{\alpha}$. Then perform the double-centering of $\mathbf{Q}(\alpha)$ and row and column weighting in (3) (maintaining the original masses $\mathbf{r}$ and $\mathbf{c}$) to obtain the new $\mathbf{S}$, which is again divided by α before applying the singular value decomposition. This procedure converges exactly to weighted LRA as α tends to 0.

Further technical details and proofs can be found in Greenacre (2008b), where power transformations, in general, are treated from a CA point of view. Some examples comparing LRA and CA can be found in Greenacre and Lewi (2008), including a compositional data analysis example of archeological data (from Baxter et al. 1990), where the benefit of the weighting in LRA is demonstrated.

4 Discussion: Implications for Compositional Data Analysis

This theoretical link between LRA, both unweighted and weighted, and CA has many practical consequences. LRA is known to have the property of subcompositional coherence, whereas CA does not. It follows that as the power transformation parameter α tends to 0, the "power-transformed CA" must be coming progressively closer to being subcompositionally coherent. It may be that with a certain power transformation we are close enough for all practical purposes to subcompositional coherence. For nonzero power parameters, zeros in the data can still be analyzed, so this holds promise for the analysis of compositional data with zeros, which is a perennial problem with the log-ratio transformation (Martín-Fernández et al. 2003). As shown by Greenacre and Lewi (2008), CA has the property of distributional equivalence, whereas unweighted LRA does not. Distributional equivalence is an important property of ratio-scale data analysis, and can be simply stated in the context of compositional data analysis as follows: If two components have the same relative values (i.e., component A always occurs twice as much as component B in all samples), then merging them into one component does not affect the distances between samples. Introducing weights into the definition of LRA, as long as the weights are added when columns are merged, leads to LRA having the property of distributional equivalence. In fact, the whole power family based on Definition 2 and the corresponding Result 2 has the property of distributional equivalence. This is another (theoretical) reason why the weighted LRA approach is preferable to the unweighted one, apart from the (practical) reason that the weighting gives less importance in the analysis to components with small proportions that often have high variances on the log-ratio scale.

Finally, it is already well-known that association modeling (Goodman 1968) gives similar results to CA when the variance in the data is low (Cuadras et al. 2006). Here low variance means that the observed data are close to their expected values based on the table margins. Weighted LRA has strong theoretical similarities to association modeling and gives practically identical results to CA in such a low variance situation where the samples are very similar to one another, which is often the case in archeological data, for example. But when the variance is high the family of methods based on Result 2 will show greater differences across the range of power transformations,

which is often the case for geological and geochemical data. This power family, with CA at the one extreme and weighted LRA at the other, provides a flexible new range of analytic tools for compositional data analysis, where the power parameter may be chosen to optimize an additional objective function of the user's choice.

Acknowledgements Michael Greenacre's research is supported by the Fundación BBVA in Madrid, Spain. Partial support by the Spanish Ministry of Education and Science, grant MEC-SEJ2006-14098 is also hereby acknowledged.

Appendix: Videos Demonstrating Correspondence Analysis Converging to Log-Ratio Analysis

In the Supplementary Material, two videos are included which demonstrate the convergence of power-transformed CA to either unweighted or weighted LRA, depending on the way the power transformation is applied (Results 1 and 2, respectively, of Sect. 3). The data used here are from Baxter et al. (1990), reproduced in Greenacre and Lewi (2008), concerning an archeological sample of 47 Roman glass cups for which the compositions of 11 oxides are measured. The first video shows the power transformation of the original data (Result 1), as the power parameter descends from 1 (CA) to 0 (unweighted LRA)—notice how dominated the unweighted LRA is by the rarest component, Mn (manganese), which has only three observed values of 0.01, 0.02, and 0.03%, and thus very high log-ratios. The second video shows the power transformation of the contingency ratios (Result 2)—notice how similar CA is to the weighted LRA, where the weighting reduces the influence of rare components, with hardly any difference as the power parameter descends from 1 (CA) to 0 (weighted LRA).

References

Aitchison J (1983) Principal component analysis of compositional data. Biometrika 70(1):57–65

Aitchison J (1986) The statistical analysis of compositional data. Chapman & Hall, London. Reprinted in 2003 with additional material by Blackburn Press

Aitchison J (1990) Relative variation diagrams for describing patterns of compositional variability. Math Geol 22(4):487–511

Aitchison J, Greenacre M (2002) Biplots for compositional data. J R Stat Soc Ser C (Appl Stat) 51(4):375–392

Bartlett MS (1936) The square root transformation in analysis of variance. Suppl J R Stat Soc 3:68–78

Baxter MJ, Cool HEM, Heyworth MP (1990) Principal component and correspondence analysis of compositional data: some similarities. J Appl Stat 17(2):229–235

Beardah CC, Baxter MJ, Cool HEM, Jackson CM (2003) Compositional data analysis of archaeological glass—problems and possible solutions. Paper presented at Compositional Data Analysis Workshop, Girona, Spain. Available at http://ima.udg.es/Activitats/CoDaWork03/paper_baxter_Beardah2.pdf

Box GEP, Cox DR (1964) An analysis of transformations (with discussion). J R Stat Soc Ser B 35:473–479

Cuadras C, Cuadras D, Greenacre M (2006) A comparison of methods for analyzing contingency tables. Commun Stat Simul Comput 35(2):447–459

Field JG, Clarke KR, Warwick RM (1982) A practical strategy for analysing multispecies distribution patterns. Mar Ecol Prog Ser 8:37–52

Goodman LA (1968) The analysis of cross-classified data: independence, quasi-independence, and interactions in contingency tables, with or without missing entries. J Am Stat Assoc 63:1091–1131

Greenacre M (2007) Correspondence analysis in practice. Chapman & Hall/CRC Press, London
Greenacre M (2008a) La práctica del análisis de correspondencias. Fundación BBVA, Madrid
Greenacre M (2008b) Power transformations in correspondence analysis. Comput Stat Data Anal (to appear). Economics Working Paper 1044, Universitat Pompeu Fabra (2007). Available at URL http://www.econ.upf.edu/en/research/onepaper.php?id=1044
Greenacre M, Lewi P (2008) Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. J Classif (to appear). Economics Working Paper 908, Universitat Pompeu Fabra (2005). Available at URL http://www.econ.upf.edu/en/research/onepaper.php?id=908
Hinkley D (1975) On power transformations to symmetry. Biometrika 62(1):101–111
Lewi PJ (1976) Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. Arzneim Forsch (Drug Res) 26:1295–1300
Lewi PJ (1980) Multivariate data analysis in APL. In: van der Linden GA (ed) Proceedings of APL-80 conference. North-Holland, Amsterdam, pp 267–271
Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V (2003) Dealing with zeros and missing values in compositional data sets. Math Geol 35(3):253–278