Measuring subcompositional incoherence*

MICHAEL GREENACRE[1]
[1] Departament d'Economia i Empresa – Universitat Pompeu Fabra, Spain, michael.greenacre@upf.edu

Subcompositional coherence is a fundamental property of Aitchison's approach to compositional data analysis, and is the principal justification for using ratios of components. For dimension reduction of a matrix of compositional data, either an unweighted (Aitchison & Greenacre 2002) or weighted (Greenacre & Lewi 2009; Greenacre 2010a: chapter 7) form of log-ratio analysis can be used, and these are both subcompositionally coherent. Many alternative methods that might be applied to compositional data are subcompositionally *incoherent*, but some can be judged to be less incoherent than others. In other words, either for a particular data set, or in general, a method might actually be quite subcompositionally "robust" in that its results for a subcomposition are quite close to its results for the same components as part of a full composition.

So we propose that lack of subcompositional coherence, that is subcompositional incoherence, can be measured in an attempt to evaluate whether any given technique is close enough, for all practical purposes, to being subcompositionally coherent. This opens up the field to alternative methods, which might be better suited to cope with problems such as data zeros and outliers, while being only slightly incoherent.

1. A measure of subcompositional incoherence

There are several potential ways for measuring subcompositional incoherence (hereafter referred to simply as incoherence), but – since the concept of distance between samples and distances between components is fundamental to compositional data analysis – the measure that we propose is based on the distance measure between components.

Although not demonstrated conclusively, it seems intuitive that two-component subcompositions, after reclosure, would be the most sensitive to incoherence. That is, two components will have a certain inter-component distance when part of a full composition, but would have a severely changed distance apart when isolated in a two-component composition (apart from the log-ratio approach, of course, which maintains the distance fixed, since it will be based on the ratio of the two components).

Closeness of the distance matrix $\boldsymbol{\Delta}$ based on the two-component composition and the distance matrix $\mathbf{D}$ between the two components in the full composition can be quantified using a stress measure that is common in multidimensional scaling. Of the available stress measures available, we have selected the so-called stress formula 1 (Borg and Groenen, 2007):

$$\text{stress} = \sqrt{\frac{\sum\sum_{j<j'}(d_{jj'} - \delta_{jj'})^2}{\sum\sum_{j<j'} d_{jj'}^2}} \qquad (1)$$

2. An illustration using correspondence analysis

As an example we measure the incoherence of correspondence analysis (CA) as an increasingly strong power-transformation is applied to the data. It has already been shown that CA of power-

* The full paper on which this talk is based has been accepted for publication in *Mathematical Geosciences*. The manuscript of the final accepted version can be viewed at http://www.econ.upf.edu/en/research/onepaper.php?id=1106

Egozcue, J.J., Tolosana-Delgado, R. and Ortego, M.I. (eds.)
ISBN: 978-84-87867-76-7

transformed data converges to log-ratio analysis (LRA) as the power tends to zero (Greenacre 2009, 2010b). This means that one can come arbitrarily close to LRA and thus arbitrarily close to subcompositional coherence, using CA.

The data are compositional oxide measurements of 47 Roman glass cups reported by Baxter, Cool and Heyworth (1990). This data set has 11 components, and we first consider the effect on the inter-component distances for subcompositions of different sizes, from 10 down to 2. Figure 1 summarizes the results: stress is higher for subcompositions of smaller size, and also stress decreases as the power transformation is made stronger (i.e., α smaller). Notice that for the tiny value of $\alpha = 0.001$ the stress is practically zero, due to the convergence of CA to LRA.

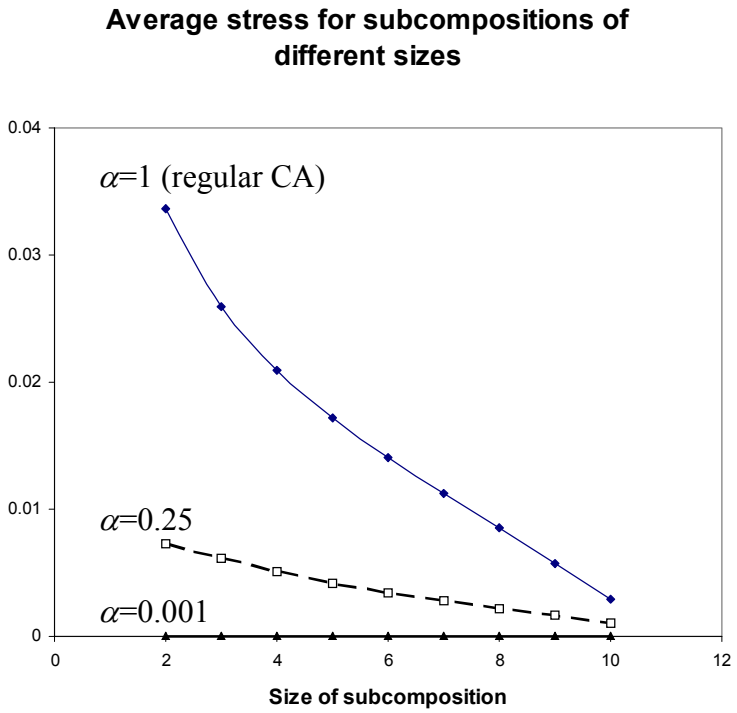

Figure 1. Average stress between chi-square distances calculated in subcompositions of different sizes and corresponding chi-square distances in the full composition, for regular CA and two power-transformed CAs, $\alpha = 0.25$ and $\alpha = 0.001$. In the last case there is almost no subcompositional incoherence. Subcompositions of size 2 are seen to be the worst case.

3. Discussion

Two aspects are important: component weighting and data zeros.

It has been shown (Greenacre and Lewi, 2009) how differential weighting of the components can dramatically improve LRA. Furthermore, regular CA is much less incoherent when compared to weighted LRA. Principal component analysis (PCA) is found to be much more incoherent by our measure.

When there are data zeros, the convergence of CA to LRA does not apply when there are zeros in the data, or it applies in the sense that CA would diverge at the limit just as with LRA it is impossible to compute the ratios. However, a power transformation of the data (with zeros intact) can be sought such that the CA comes the closest to subcompositional coherence while not destabilizing the analysis. The CA solution is monitored while reducing the power α successively as long as the incoherence descends. Then at some value of the power the effect of the zeros "kicks in" and the incoherence starts to rise again, as illustrated in Figure 2. It is this value of α that should be used to transform the data, giving the least incoherent power-transformed CA, conserving the zeros in the data with no need to replace them.

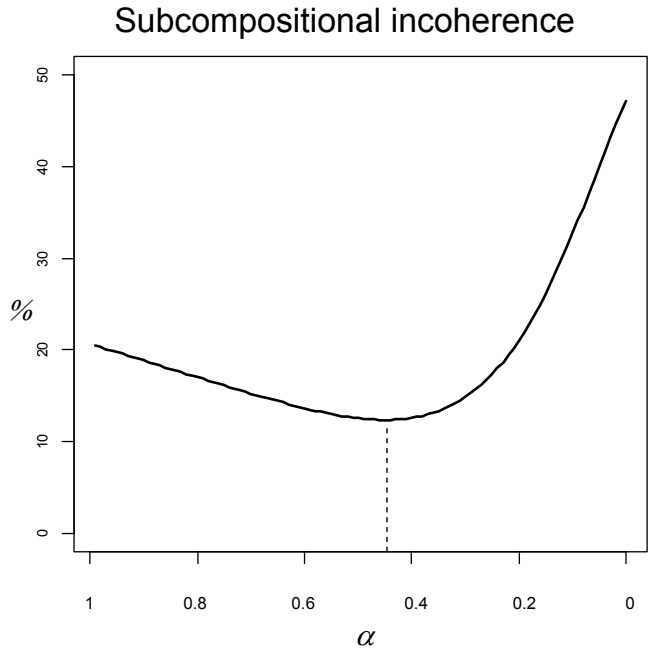

Figure 2. Subcompositional incoherence for power-transformed CA of a data set containing zeros, for values of the power α between 1 (regular CA) and 0.001 (almost identical to LRA). A value of 0.46 gives minimum incoherence, which might suggest a square root transformation, for example, which is close to the minimum.

References

Aitchison, J. & Greenacre, M. (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

Borg, I. and Groenen, P. (2005). *Modern Multidimensional Scaling, 2nd Edition*. Springer, New York.

Greenacre, M. (2009). Power transformations in correspondence analysis. *Computational Statistics and Data Analysis* 53, 3107–3116.

Greenacre, M. & Lewi, P. J. Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio scale measurements. *Journal of Classification* 26, 29–54.

Greenacre, M. (2010a). *Biplots in Practice*. BBVA Foundation, Madrid. Available for free download at http://www.multivariatestatistics.org.

Greenacre, M. (2010b). Log-ratio analysis is a limiting case of correspondence analysis. *Mathematical Geosciences* 42, 129–134.