Measuring Subcompositional Incoherence

Michael Greenacre

Universitat Pompeu Fabra, Barcelona, Spain

Summary. Subcompositional coherence is a fundamental property of Aitchison's approach to compositional data analysis, and is the principal justification for using ratios of components. We maintain, however, that lack of subcompositional coherence, that is 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. The measure that we propose is based on the distance measure between components. We show that the two-part subcompositions, which are the most sensitive to subcompositional incoherence, can be used to establish a distance matrix which can be directly compared with the pairwise distances in the full composition. The closeness of these two matrices can be quantified using a stress measure that is common in multidimensional scaling, providing a measure of subcompositional incoherence. Furthermore, we strongly advocate introducing weights into this measure, where rarer components are weighted proportionally less than more abundant components. The approach is illustrated using power-transformed correspondence analysis, which has already been shown to converge to logratio analysis as the power transform tends to zero.

***Keywords*:** chi-square distance, correspondence analysis, logratio distance, multidimensional scaling, stress, subcompositional coherence

Acknowledgment: 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.

1. Introduction

In his seminal *Biometrika* paper John Aitchison (1983) stated:

> "A desirable feature of any form of compositional data analysis is an ability to study subcompositions, that is subvectors rescaled to give unit sum. One important requirement is an ability to quantify the extent to which a subcomposition retains a picture of the variability of the whole composition."

The property of subcompositional coherence is indeed one of the cornerstones of Aitchison's approach to compositional data analysis: results should be the same for components in a full composition as in any subcomposition, where the subcomposition has been closed again to give unit sum, or "reclosed". An example that is often given of subcompositional *incoherence* is that the correlation coefficient between two components in a (reclosed) subcomposition is not the same as that for the same two components in the full composition. Using ratios as the basic input data for analysis solves this paradox and the logratio transformation has become a standard procedure to guarantee subcompositional coherence.

For ease of exposition we shall often refer to subcompositional coherence simply as coherence. Coherence is an absolute property which a procedure either possesses or not. But if it does not, that is if it is incoherent, we maintain that there are levels of incoherence that can be usefully measured and exploited. For example, what if our method was 'close' to being coherent – would that not be useful if in the process we fixed up other problems, such as the treatment of zeros in the data? As a context for our investigation, we have chosen the area of visualization of compositional data in the form of maps, in the style of principal component analysis (PCA) and multidimensional scaling (MDS), because these are based on the concept of distance and distance is one of the most fundamental aspects of multivariate analysis.

The logratio approach to PCA of compositional data originates in the papers of Aitchison (1983, 1986, 1990), which we call logratio analysis, abbreviated as LRA. Simply stated, LRA is the principal component analysis (PCA) of a matrix of positive compositional

data – assumed to be closed row-wise – after logarithmically transforming the data and 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. Interestingly, even though the rows and columns are different entities (samples and components) LRA treats them totally symmetrically and the results would be identical if the matrix were transposed.

A different approach, also symmetric with respect to rows and columns, 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, even with zeros. (In fact, it is its ability to handle zeros, even lots of zeros in very sparse tables, that has made CA so popular in environmental and archeological research). 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 account of CA, see Greenacre 2007, 2008a).

Greenacre (2008b) has shown that LRA and CA are actually part of a common family parameterized by a power transformation – a summary of these findings aimed at compositional data analysts is given by Greenacre (2008c). Putting this result simply, if you power up your compositional data by a power α, reclose row-wise (although closure is entirely optional here), and then perform a regular CA of the transformed data, with a rescaling of the solution by $1/\alpha$, then this procedure converges exactly at the LRA solution as the power parameter α tends to 0. In fact, this is nothing else but the Box-Cox transformation in disguise (Box and Cox, 1964) – see Greenacre (2008b). This means that we can come arbitrarily close to Aitchison's LRA by

performing a CA: numerically, there is hardly any difference between the CA just described using $\alpha = 0.001$, for example, and LRA. Now while LRA is coherent, CA is not. But it follows intuitively from the limiting result mentioned above, and we shall indeed show this to be true, that CA comes closer and closer to being coherent as the power parameter approaches 0.

Since CA can handle zeros in a completely natural way, whereas LRA can not, benefit can be gained by using power-transformed CA instead of LRA and coming "close enough" to coherence for all practical purposes. This is the background to our need to be able to measure coherence and study its behavior in different scenarios.

2. Logratio and chi-square distances for compositions and subcompositions

As intimated in the introduction we adopt a distance-based approach where the concept of between-component distance will be fundamental. Notice that we are not interested here in between-sample distance since the property of coherence applies to the relationships between components. For our purposes coherence will mean that distances calculated between the components in the full composition will be identical in the subcomposition. Since we will be generally concerned with Euclidean type distances, which are embeddable in an inner product space, this distance-based property of coherence will mean that all the classical statistics such as variance, correlation and covariance will also be coherent.

Suppose that the compositional data table of I samples (rows) and J components (columns) is denoted by $\mathbf{X}$ ($I \times J$). The two equivalent definitions of Aitchison's logratio distance of relevance to us here, between two components j and j', are as follows (Aitchison, 1983, 1986), expressed in squared distance form:

$$d^2_{jj'} = \frac{1}{I}\sum_{i=1}^{I}\left[\log\left(\frac{x_{ij}}{g(\mathbf{x}_j)}\right) - \log\left(\frac{x_{ij'}}{g(\mathbf{x}_{j'})}\right)\right]^2 \qquad (1)$$

where $g(\mathbf{x}_j)$ is the geometric mean of the j-th column corresponding to the j-th component (i.e., $\log(g(\mathbf{x}_j))$ is the arithmetic average of $\log(x_{ij})$, i=1,…,I). The alternative definition is in terms of all pairwise 'odds-ratios' across all pairs of samples:

$$d^2_{jj'} = \frac{1}{I^2}\sum_{i<i'}\sum\left[\log\left(\frac{x_{ij}x_{i'j'}}{x_{ij'}x_{i'j}}\right)\right]^2 \qquad (2)$$

Notice that compared to Aitchison's original definition, which is in general use, we have averaged the squared terms over the samples, so that the distance is not sample-size dependent – this is the form of the distance that is compatible with the chi-square distance in CA, which is also averaged over samples. Although definition (1) involves centering each $\log(x_{ij})$ with respect to the average $(1/I)\Sigma_i \log(x_{ij})$, definition (2) shows that the distance is actually independent of this centering – this is another reason for using distance as the fundamental concept for judging and measuring coherence. Definition (2) also shows quite clearly that the logratio distance is coherent: if any subcomposition involving components j and j' is considered and reclosed row-wise, the ratios row-wise $x_{ij}/x_{ij'}$ remain identical, and so (2) remains the same.

In CA it is the chi-square distance that defines distance between columns. First the column profiles are calculated by dividing the elements of each column j by their sum x_{+j}. Then the sum of squared distances between profile elements is calculated, weighted inversely by the profile of the row sums. Since for **X** these row sums are all 1, the marginal row profile has constant values $(1/I)$, hence the squared chi-square distance between columns j and j' is:

$$\chi^2_{jj'} = \sum_{i=1}^{I}\left[\frac{x_{ij}}{x_{+j}} - \frac{x_{ij'}}{x_{+j'}}\right]^2 \Big/ (1/I) \;=\; I\sum_{i=1}^{I}\left[\frac{x_{ij}}{x_{+j}} - \frac{x_{ij'}}{x_{+j'}}\right]^2 \qquad (3)$$

Clearly, the chi-square distance is incoherent, but from the results of Greenacre (2008b, 2008c) mentioned previously it follows that the chi-square distance on the power-transformed data tends to the logratio distance as the power parameter α tends to 0. The convergence of CA to LRA is a direct result of the Box-Cox transformation $(1/\alpha)(x^\alpha - 1)$ which tends to $\log(x)$ as α tends to 0. To illustrate this convergence empirically in the case of the chi-square distance,

Table 1 shows four versions of a subset of distances calculated on the 11 components (mostly oxides) of the 47×11 compositional data set on Roman glass cups published by Baxter, Cool and Heyworth (1990), reproduced by Greenacre and Lewi (2008: Table 2). The chi-square distances are at top right, then reading clockwise the chi-square distances based on a double square root transformation (α = ¼), then a power transformation close to zero (α = 0.001) and finally the logratio distances. Figure 1 shows the maximum absolute difference between the chi-square distances and the logratio distances for 1000 different CAs, starting with α = 1 (untransformed CA) and descending in steps of 0.001, i.e., 0.999, 0.998, and so on, until α = 0.001. This effectively shows that one can get as close as one likes to coherence by lowering the value of α towards 0. The concept of coherence is more, however, than just showing that the chi-square distance converges to the logratio distance – it actually concerns the behavior of the distance function on subcompositions, as treated in the next section.

3. A measure of subcompositional coherence

Coherence is the invariance of the statistical procedure, in this case the distance computation which affects all our subsequent multivariate analyses, when applied to subsets of components that are reclosed. Since we know that CA is incoherent, let us see to what extent it is by calculating the chi-square distances for different subsets of the components of the Roman glass cup data set. The chi-square distances for the full 11-part composition serve as a reference to which we will compare the chi-square distances for every relevant subset of components: the $\binom{11}{2} = 55$ subsets of size 2, the $\binom{11}{3} = 165$ subsets of size 3, and so on, until the $\binom{11}{10} = 11$ subsets of size 10. For example, the top left table of Table 1 shows the chi-square distances between the first five components of the full composition. If we select these five components and then reclose then to form a five-part subcomposition, the chi-square distances

turn out as the first table in Table 2. This table is remarkably similar to the original chi-square distances in Table 1, and their maximum absolute difference is only 0.00066. This is because we have included in the subcomposition some of the highest components, so that the reclosure does not affect the values too much. However, if we consider the last five elements, which happen to be amongst the rarest, the second distance table in Table 2 is obtained, which is much further away from the original ones (maximum absolute difference = 0.0368).

So far, to compare two distance matrices we have simply used the maximum absolute difference, a quantity with a scale which is hard to get to grips with because it depends on the scale of distance. In the MDS literature there are several well-known normalized measures for quantifying the fit of one distance matrix to another, called measures of 'stress'. Of these we have selected the so-called 'stress formula 1' (see, for example, Borg and Groenen 2005):

$$\text{stress} = \sqrt{\frac{\sum\sum_{j<j'}(d_{jj'} - \delta_{jj'})^2}{\sum\sum_{j<j'} d_{jj'}^2}} \tag{4}$$

where d denotes the target distances in the full composition and δ the distances in the subcomposition. The denominator serves to normalize the sum of squared differences in the numerator, and the stress value is often multiplied by 100 and thought of as a percentage of badness of fit. For the two subcompositions analyzed in Table 2, the stress values are reported as 0.00245 (i.e., 0.245%) and 0.06574 (i.e., 6.574%). To get an idea how this deviation from coherence varies across subsets of different sizes, Figure 2 plots the average stress against subset size for regular CA and repeats this for chi-square distances from two power-transformed CAs – this demonstrates what we hinted at before, namely that CA becomes closer and closer to coherence as the power parameter decreases.

In addition, this shows what might have been suspected before: subcompositions of size 2 are the 'worst case scenario' for deviation from coherence, since they are the most affected by reclosure. In other words, if we can bring the stress of subcompositions of size 2 acceptably low enough then we are guaranteeing that all other subcompositions will be at least more coherent on average. This is a very convenient result, because all the pairwise distances from

two-part subcompositions can be placed in a square distance matrix, which can then be compared directly with the pairwise distances in the full composition using just one overall stress measure. Notice that this calculation is different to the one used to calculate average stress for two-part subcompositions in Figure 2 – there we averaged stress values calculated for each subcomposition, where in each of the 55 cases stress formula (4) consisted of a single term in the numerator and denominator; whereas here the stress formula will have numerator and denominator equal to the sums of those 55 numerators and denominators respectively. Table 3 gives three examples, showing just the last five out of the 11 components, for $\alpha = 1$, 0.25 and 0.001 – the distances on the left are computed in the full composition, and the distances on the right are those obtained by forming each subcomposition corresponding to the row-column pairs. Again we witness the convergence as α decreases. Figure 4 shows a continuous version of the stress as a function of α. If a 1% level of stress were acceptable as being 'close enough' to coherence, then the power transform with $\alpha = 0.106$ would be appropriate.

4. To weight or not to weight

So far we have treated each component equally, as is general practice in compositional data analysis, even in the paper on logratio biplots by Aitchison and Greenacre (2002). However, Greenacre and Lewi (2008) have brought to attention the necessity for and benefits of weighting the components when doing LRA. Convenient weights are the so-called "masses" in CA, namely the marginal averages of the components – thus a rare component with low average value in the data set is downweighted compared to the abundant components. Although this appears to be an issue only when analyzing the data, for example visualizing the compositional distances in a subspace of reduced dimension, it is also an issue when measuring stress, as we now demonstrate.

We have just come to the conclusion that a power-transformed CA of these data with power parameter $\alpha = 0.106$ would reduce the incoherence of CA to 1% , but let us look at this 1% lack of coherence in a bit more detail. The stress measure is a sum of positive numbers for each cell in an 11×11 table – Figure 4 shows a graphical display where the contribution of each of these values is indicated by the area of a circle. It is immediately obvious that this incoherence, albeit small, is almost totally due to the element Mn (manganese). In previous analyses of these data Mn has already been singled out by Greenacre and Lewi (2008) as a problem, because it takes on only three small values: 0.03%, 0.02% and 0.01% (i.e., 0.0003, 0.0002 and 0.0001 on a proportion scale), engendering large values on the ratio and logratio scale. Their proposal to weight the components in proportion to their marginal averages eliminates the influence of this rare but outlying component. Our stress measure of incoherence can also be easily modified to take the 'abundance' of each component into account in the measure, in which case Mn would not feature so prominently. Then the measure would be measuring incoherence weighted by the average level of each component, with incoherence in higher-abundance components being taken into account more than incoherence in rare components. This weighted stress measure is then:

$$\text{weighted stress} = \sqrt{\frac{\sum\sum_{j<j'} c_j c_{j'} (d_{jj'} - \delta_{jj'})^2}{\sum\sum_{j<j'} c_j c_{j'} d_{jj'}^2}} \qquad (5)$$

where c_j denotes the weight of the j-th component, usually taken to be equal to its marginal average proportion. The lower curve in Figure 3 traces out weighted stress as a function of the power parameter – it is considerably lower than the unweighted curve at the top, and now even regular untransformed CA is seen to have less than 1% incoherence overall. Figure 5 shows the contribution-to-weighted-stress plot for regular CA – Mn is no longer an important contributor, the highest contributions to incoherence come from two distances involving calcium, Ca to Si (silica) and Ca to Na (sodium).

5. Discussion and conclusions

The main proposal of this paper is to measure subcompositional coherence by the stress between the inter-component distance matrix calculated using the full composition and the matrix of pariwise component distances computed from all the two-part subcompositions. From the results of the previous section and from the discussion of Greenacre and Lewi (2008), we strongly advise to supplement this proposal with the weighting of the components proportional to their average value in the data set. We have seen in the example of the Roman glass compositional data set that regular CA owes most of its incoherence (when measured without weights) to one problematic component that is rare. Weighting eliminates this problem and then we see that CA is, in fact, very close to coherence. Application of this idea to a wider spectrum of compositional data sets will show to what extent CA, with or without power transformations, can be used as an alternative to LRA. It will also allow alternative methods, such as PCA with or without standardization, to be judged with respect to their subcompositional coherence properties. Incidentally, we measured the subcompositional incoherence for the present data set, using the Euclidean distance with and without standardization of the components. The weighted stress measures are 0.3442 (34.42%) and 0.1828 (18.28%) respectively – if one compares these values with those for CA shown in Figure 3, one realizes how high these measures are and how far away from coherence PCA is. There is also a quirk in the two-part compositions in PCA, due to the centering with respect to component means. Since the pair of closed values has the property $x_{ij'} = 1 - x_{ij}$, the two centered values have the property $y_{ij'} = - y_{ij}$, and thus also have the same variance, s_j say, and it can be easily deduced that the unstandardized Euclidean distance between components j and j' is a constant multiple of the standard deviation, $2\sqrt{n-1}s_j$, while the standardized Euclidean distance is a constant $2\sqrt{n-1}$ for all two-part subcompositions. The correlation between the components of any two-part subcomposition is –1, independent of the data. It seems that PCA

on unstandardized or standardized data is out of the question for compositional data analysis if one places importance on the principle of subcompositional coherence[1].

Coming back to CA with possible power transformations, where the implicit chi-square distance appears to be close to coherence, an obvious benefit is that for CA with nonzero power parameters, zeros in the data can still be analyzed – hence this holds promise for the analysis of compositional data with zeros, which is a perennial problem with the logratio transformation (see, for example, Martín-Fernández, Barceló-Vidal and Pawlowsky-Glahn 2003).

Greenacre and Lewi (2008) already showed that a regular CA of these data and a weighted LRA gave almost the same two-dimensional biplot, so the fact that CA is almost coherent (using weighted stress) fits in with this result. It is already known that CA gives similar results to association modeling (Goodman 1968) when the variance in the data is low (for example, see Cuadras, Cuadras and Greenacre 2006) and that weighted LRA has strong theoretical similarities to association modeling (see Greenacre and Lewi 2008). Here low variance means that the observed data are close to their expected values based on the table margins. It follows that CA and weighted LRA will give similar results in such a low variance situation where the samples are very similar to one another, which is the case of the present example and often the case in archeological data. But when the variance is high, which is often the case for geological and geochemical data where there can be many data zeros, the power family of CAs will show greater differences across the range of power transformations. It remains to be shown whether we can use a power transformation to come close enough to coherence while being able to analyze zeros as zeros without having to resort to replacing them artificially with some small positive number. But, at least, we now have a tool to measure incoherence to be able to judge how close we are to subcompositional coherence in different situations.

[1] The performance of 10-part subcompositions should be the most favourable for evaluating PCA, but the incoherence is large even for these: for this example, the average stress for all 10-part subcompositions was calculated as 0.1371 (13.71%) for unstandardized PCA and 0.0425 (4.25%) for standardized PCA. Average weighted stresses are 0.1906 (19.06%) and 0.0940 (9.40%) respectively. Compare these to regular (untransformed) CA, which for the 11 10-part subcompositions of these data has average unweighted and weighted stresses of 0.0029 (0.29%) and 0.0021 (0.21%) respectively.

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) 416 p

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

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

Borg I, Groenen P (2005) Modern multidimensional scaling, second edition. Springer, New York

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

Chessel D, Dufour AB, Thioulouse J (2004) The ade4 package: one-table methods. R News 4: 5–10

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

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, 280 p

Greenacre M (2008a) La práctica del análisis de correspondencias. Fundación BBVA, Madrid, 384 p

Greenacre M (2008b) Power transformations in correspondence analysis. Comp 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 (2008c) Logratio analysis is a limiting case of correspondence analysis. Submitted to Math Geosc

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

Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V (2003) Dealing with zeros and missing values in compositional data sets. Math Geol 35: 253–278

Table 1: Three sets of chi-square distances based on CAs with different power transformations (starting at top right with power $\alpha = 1$, the regular untransformed CA), and finally at bottom left, the logratio distances from LRA. Parts of each 11×11 table of distances are shown, as well as the maximum absolute difference between the distances in the full table and their corresponding logratio distances.

$\alpha = 1$ (untransformed CA)

	Si	***Al***	***Fe***	***Mg***	***Ca***	...
Si	0.0000	0.0920	0.2259	0.1850	0.1241	...
Al	0.0920	0.0000	0.1441	0.1261	0.0855	...
Fe	0.2259	0.1441	0.0000	0.1280	0.1472	...
Mg	0.1850	0.1261	0.1280	0.0000	0.1387	...
Ca	0.1241	0.0855	0.1472	0.1387	0.0000	...
.	.	.	.	.	.	...
.	.	.	.	.	.	...
.	.	.	.	.	.	...

Max abs diff = 0.0797

$\alpha = 0.25$

	Si	***Al***	***Fe***	***Mg***	***Ca***	...
Si	0.0000	0.0909	0.2207	0.1878	0.1209	...
Al	0.0909	0.0000	0.1404	0.1282	0.0850	...
Fe	0.2207	0.1404	0.0000	0.1190	0.1468	...
Mg	0.1878	0.1282	0.1190	0.0000	0.1404	...
Ca	0.1209	0.0850	0.1468	0.1404	0.0000	...
.	.	.	.	.	.	...
.	.	.	.	.	.	...
.	.	.	.	.	.	...

Max abs diff = 0.0142

$\alpha = 0$ (LRA)

	Si	***Al***	***Fe***	***Mg***	***Ca***	...
Si	0.0000	0.0913	0.2209	0.1882	0.1213	...
Al	0.0913	0.0000	0.1403	0.1279	0.0849	...
Fe	0.2209	0.1403	0.0000	0.1168	0.1471	...
Mg	0.1882	0.1279	0.1168	0.0000	0.1404	...
Ca	0.1213	0.0849	0.1471	0.1404	0.0000	...
.	.	.	.	.	.	...
.	.	.	.	.	.	...
.	.	.	.	.	.	...

Max abs diff = 0

$\alpha = 0.001$

	Si	***Al***	***Fe***	***Mg***	***Ca***	...
Si	0.0000	0.0913	0.2209	0.1882	0.1213	...
Al	0.0913	0.0000	0.1403	0.1280	0.0849	...
Fe	0.2209	0.1403	0.0000	0.1168	0.1471	...
Mg	0.1882	0.1280	0.1168	0.0000	0.1404	...
Ca	0.1213	0.0849	0.1471	0.1404	0.0000	...
.	.	.	.	.	.	...
.	.	.	.	.	.	...
.	.	.	.	.	.	...

Max abs diff = 0.000042

Figure 1: Maximum absolute difference between chi-square distances from power-transformed CA and the logratio distances, for powers from 1 to 0.001 (calculations made for 1000 values of the power α = 1, 0.999, 0.998, …, 0.001).

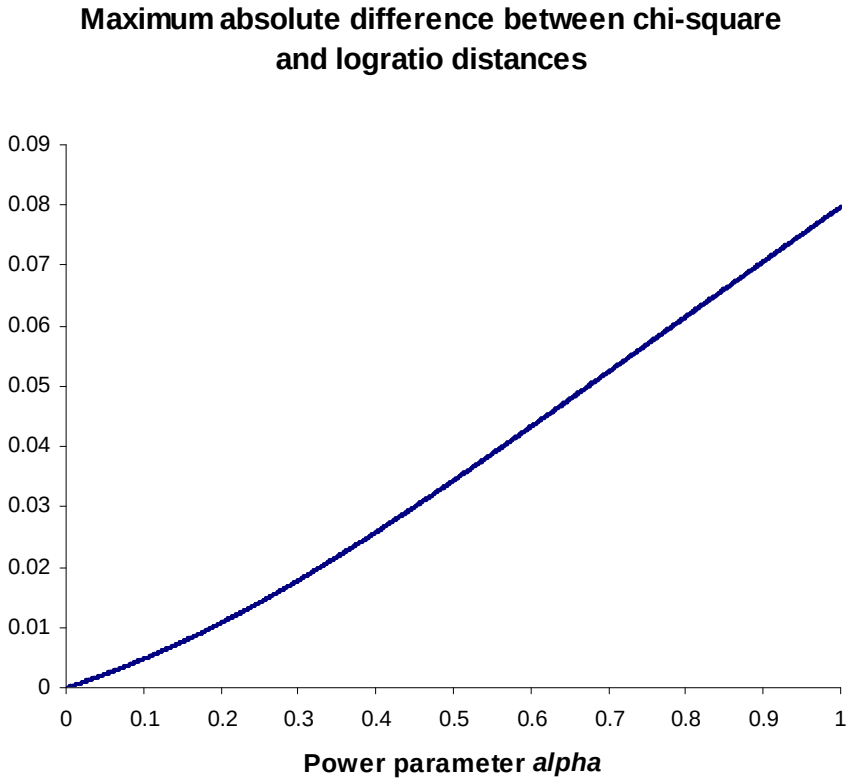

Table 2: Two sets of chi-square distances based on CAs of subcompositions of size 5.

Subset 1

	Si	***Al***	***Fe***	***Mg***	***Ca***
Si	0.0000	0.0922	0.2264	0.1849	0.1247
Al	0.0922	0.0000	0.1445	0.1256	0.0857
Fe	0.2264	0.1445	0.0000	0.1280	0.1472
Mg	0.1849	0.1256	0.1280	0.0000	0.1385
Ca	0.1247	0.0857	0.1472	0.1385	0.0000

Max abs diff = 0.00066

Stress = 0.00245

Subset 2

	K	***Ti***	***P***	***Mn***	***Sb***
K	0.0000	0.1562	0.1235	0.3396	0.2648
Ti	0.1562	0.0000	0.1505	0.3339	0.3152
P	0.1235	0.1505	0.0000	0.3407	0.2527
Mn	0.3396	0.3339	0.3407	0.0000	0.4351
Sb	0.2648	0.3152	0.2527	0.4351	0.0000

Max abs diff = 0.03682

Stress = 0.06574

Figure 2: 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'.

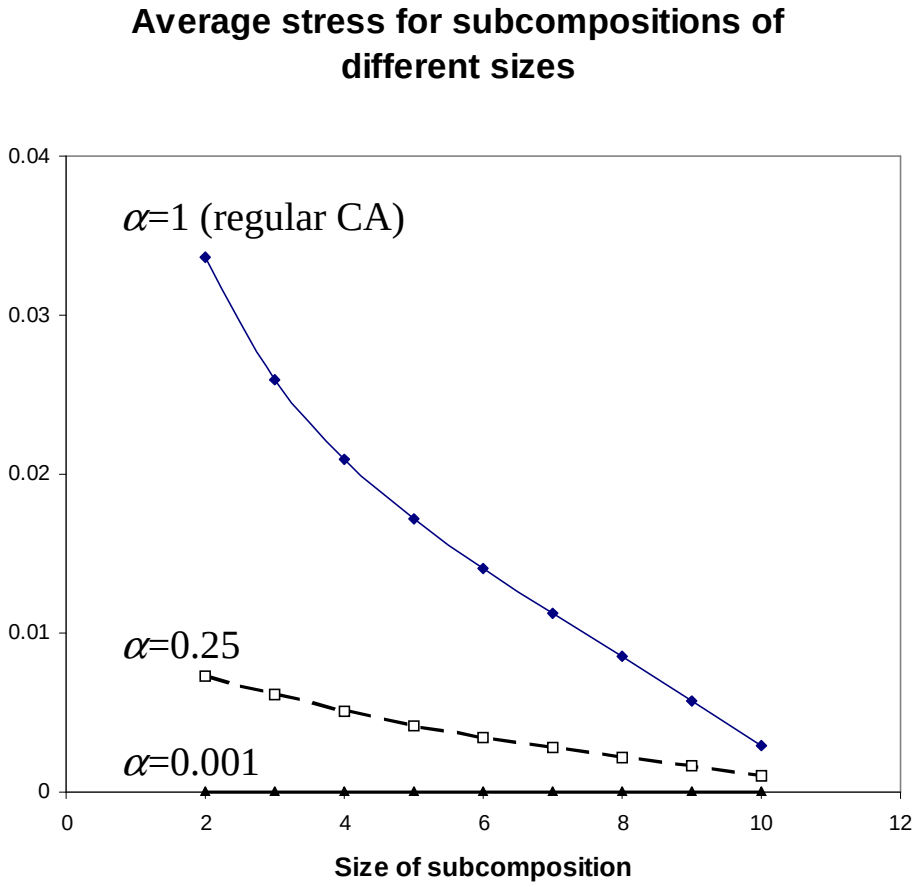

Table 3: Inter-component chi-square distances for the regular CA and two power-transformed CAs (α = 0.25 and 0.001), showing on the left the distances for the full composition and on the right the corresponding distances based on two-part subcompositions. Only the last five components are shown, but the maximum absolute differences and the stress values are computed for the whole 11×11 matrix of distances in each case.

Full composition, untransformed CA(α=1)

		K	Ti	P	Mn	Sb
	:	:	:	:	:	:
K	...	0.0000	0.1573	0.1217	0.3704	0.2611
Ti	...	0.1573	0.0000	0.1615	0.3500	0.3191
P	...	0.1217	0.1615	0.0000	0.3739	0.2407
Mn	...	0.3704	0.3500	0.3739	0.0000	0.4719
Sb	...	0.2611	0.3191	0.2407	0.4719	0.0000

Two part subcompns, untransformed CA(α=1)

		K	Ti	P	Mn	Sb
	:	:	:	:	:	:
K	...	0.0000	0.1586	0.1274	0.3358	0.2647
Ti	...	0.1586	0.0000	0.1527	0.3030	0.3182
P	...	0.1274	0.1527	0.0000	0.3095	0.2677
Mn	...	0.3358	0.3030	0.3095	0.0000	0.4196
Sb	...	0.2647	0.3182	0.2677	0.4196	0.0000

Max abs diff = 0.07415

Stress = 0.06441

Full composition, transformed CA(α=0.25)

		K	Ti	P	Mn	Sb
	:	:	:	:	:	:
K	...	0.0000	0.1534	0.1242	0.3072	0.2678
Ti	...	0.1534	0.0000	0.1543	0.2957	0.3206
P	...	0.1242	0.1543	0.0000	0.3142	0.2531
Mn	...	0.3072	0.2957	0.3142	0.0000	0.4178
Sb	...	0.2678	0.3206	0.2531	0.4178	0.0000

Two part subcompns, transformed CA(α=0.25)

		K	Ti	P	Mn	Sb
	:	:	:	:	:	:
K	...	0.0000	0.1534	0.1248	0.2946	0.2699
Ti	...	0.1534	0.0000	0.1526	0.2830	0.3213
P	...	0.1248	0.1526	0.0000	0.2991	0.2581
Mn	...	0.2946	0.2830	0.2991	0.0000	0.4053
Sb	...	0.2699	0.3213	0.2581	0.4053	0.0000

Max abs diff = 0.01514

Stress = 0.02114

Full composition, transformed CA(α=0.001)

		K	Ti	P	Mn	Sb
	:	:	:	:	:	:
K	...	0.0000	0.1530	0.1246	0.2907	0.2703
Ti	...	0.1530	0.0000	0.1526	0.2816	0.3218
P	...	0.1246	0.1526	0.0000	0.2985	0.2574
Mn	...	0.2907	0.2816	0.2985	0.0000	0.4047
Sb	...	0.2703	0.3218	0.2574	0.4047	0.0000

Two part subcompns, transformed CA(α=0.001)

		K	Ti	P	Mn	Sb
	:	:	:	:	:	:
K	...	0.0000	0.1530	0.1246	0.2906	0.2703
Ti	...	0.1530	0.0000	0.1526	0.2815	0.3218
P	...	0.1246	0.1526	0.0000	0.2985	0.2575
Mn	...	0.2906	0.2815	0.2985	0.0000	0.4046
Sb	...	0.2703	0.3218	0.2575	0.4046	0.0000

Max abs diff = 0.000059

Stress = 0.000108

Figure 3: Stress between chi-square distances calculated in two-part subcompositions and the corresponding chi-square distances in the full composition for the Roman glass cup data, for power transformations α = 1, 0.999, 0.998, …, 0.001. The power parameter corresponding to a stress of 0.01 (1%) has value 0.106, as indicated. The weighted stress takes into account the average level of the components, discussed later.

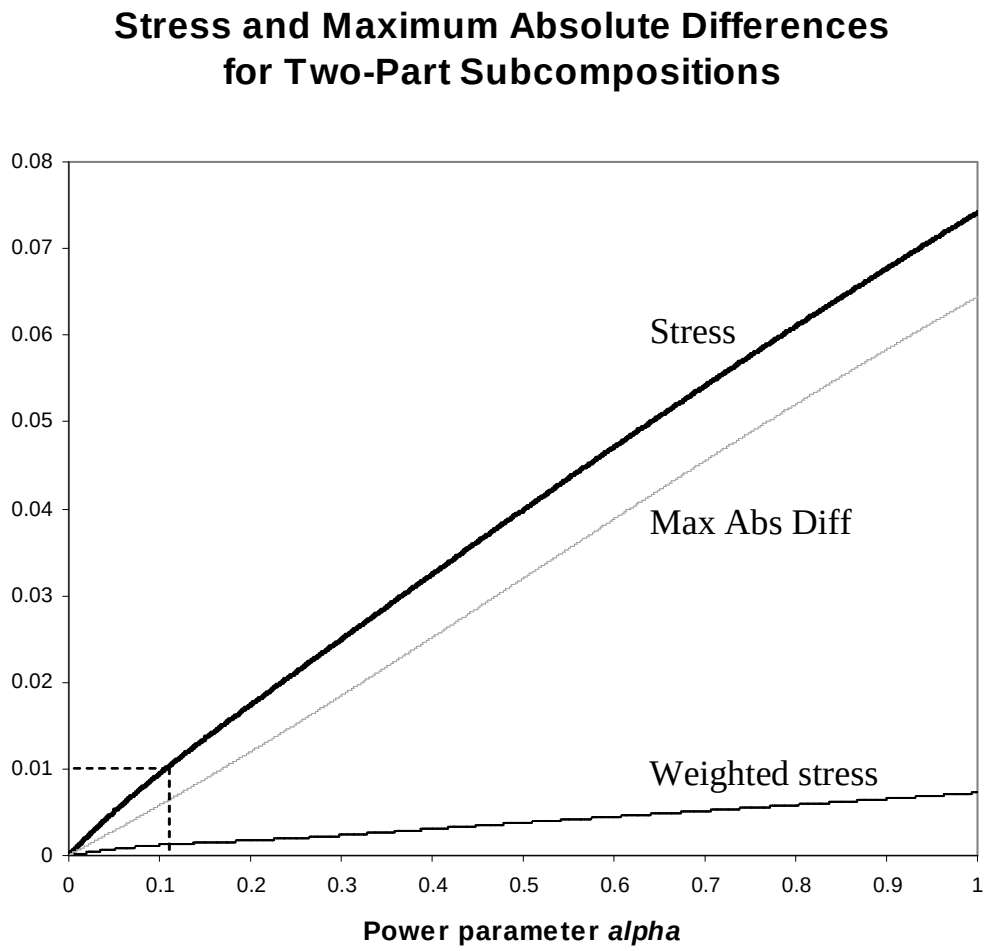

Figure 4: Values that constitute the stress measure for measuring incoherence in the CA with power transformation $\alpha = 0.106$. The area of the circles is proportional to the contribution to stress (function `table.dist` in the R package `ade4` – by Chessel, Dufour and Thioulouse 2004). The lack of coherence is concentrated almost entirely in the Mn (manganese) oxide component.

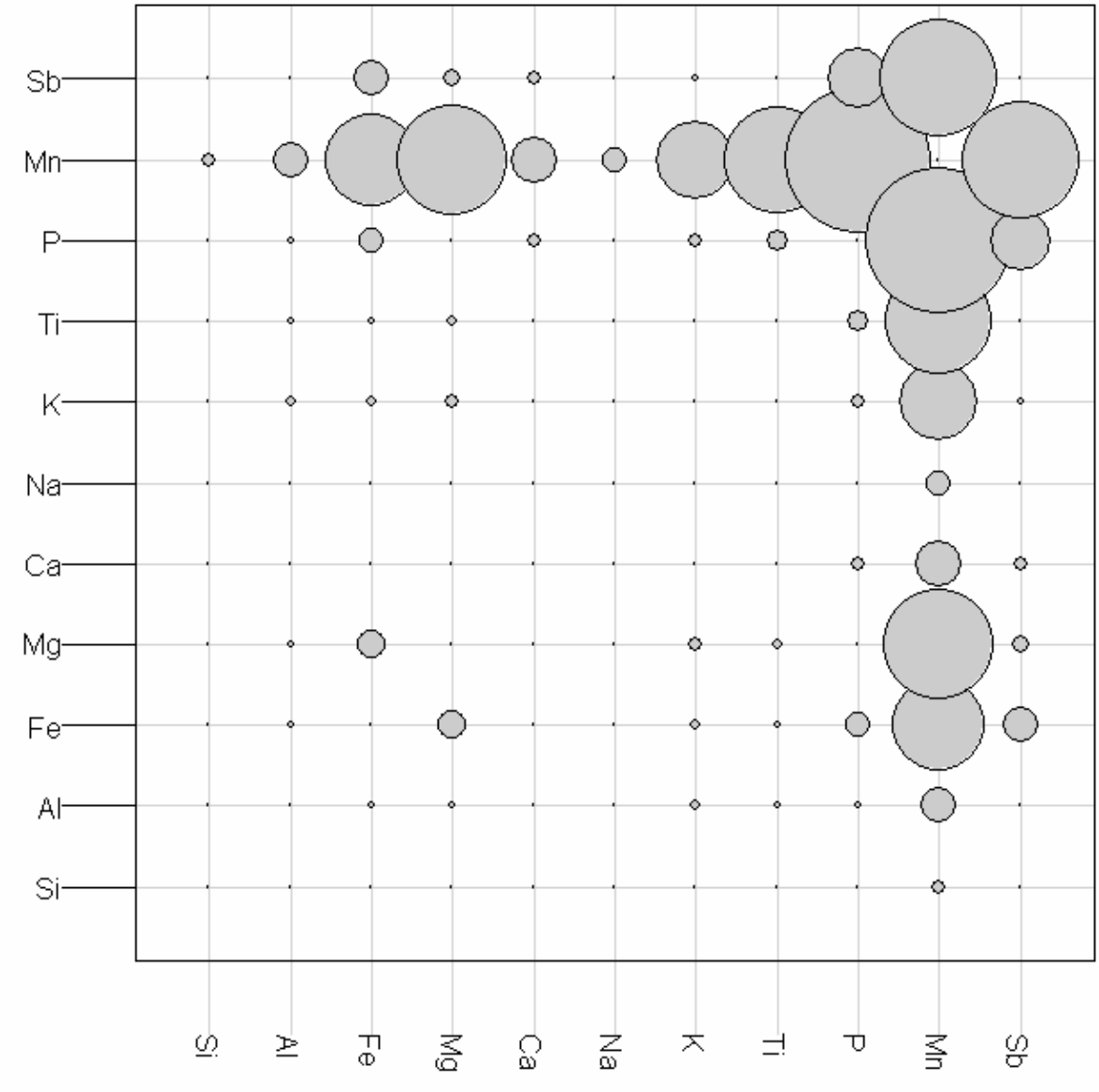

Figure 5: Values that constitute the weighted stress measure for measuring incoherence in a regular CA. The area of the circles is proportional to the contribution to weighted stress.