Universitat Pompeu Fabra Barcelona

Department of Economics and Business

Economics Working Paper Series

Working Paper No. 1554

Towards a pragmatic approach to compositional data analysis

Michael Greenacre

January 2017

Towards a pragmatic approach to compositional data analysis

Michael Greenacre
Department of Economics and Business
Universitat Pompeu Fabra,
Ramon Trias Fargas 25-27
08005 Barcelona
Spain

Email: michael.greenacre@upf.edu

Abstract: Compositional data are nonnegative data with the property of closure: that is, each set of values on their components, or so-called parts, has a fixed sum, usually 1 or 100%. The approach to compositional data analysis originated by John Aitchison uses ratios of parts as the fundamental starting point for description and modeling. I show that a compositional data set can be effectively replaced by a set of ratios, one less than the number of parts, and that these ratios describe an acyclic connected graph of all the parts. Contrary to recent literature, I show that the additive log-ratio transformation can be an excellent substitute for the original data set, as shown in an archaeological data set as well as in three other examples. I propose further that a smaller set of ratios of parts can be determined, either by expert choice or by automatic selection, which explains as much variance as required for all practical purposes. These part ratios can then be validly summarized and analyzed by conventional univariate methods, as well as multivariate methods, where the ratios are preferably log-transformed.

Keywords: compositional data, log-ratio transformation, log-ratio analysis, log-ratio distance, multivariate analysis, ratios, subcompositional coherence, univariate statistics.

1. Introduction

Compositional data are sets of non-negative data that have been expressed relative to a fixed total (usually as proportions summing to 1 or percentages summing to 100%), and their analysis is called *compositional data analysis* (Aitchison 1986). The original totals, whatever they were, are not of interest − rather, the relative values, collectively called a *composition*, are relevant for summarizing and statistical analysis. The components of a composition are called its *parts*. If a subset of the parts are considered and the data are re-expressed with respect to the new subtotals, this is called a *subcomposition*. In several situations, where the total of the original data is the same for all samples, considering a subcomposition is usually not an issue. For example, in the case of time budget data where activities such as sleeping, eating, leisure, work, transport, etc., are recorded during a 24-hour day, there is usually no point in dropping an activity and re-expressing the parts relative to the total without that activity. Similarly, concentrations in parts per million (ppm), for example, are analysed as such, without re-expression as proportions. In this article I concentrate on compositional data where subcompositions or extended compositions are possible, for example, geochemical data, or fatty acid data in ecology, in other words where the proportions depend on the particular choice of parts made by the researcher.

The act of converting a set of values into its set of relative values by dividing by the total, is called *closure*. The term *normalization* is also used (which unfortunately has an alternative meaning in statistical theory); for example, it is said that "the data are normalized", "the data are closed", "some parts are excluded, or added, and the data are renormalized, or reclosed", etc… It is exactly because the compositional values associated with the parts change, after renormalization, that makes compositional data unique, and needing special approaches.

In spite of the compositional values being dependent on the particular mix of parts chosen by the user, parts of a composition are still often summarized by statistical measures such as the mean and correlation coefficient. Clearly, these summary statistics make no sense when comparing different studies, unless studies being compared have used exactly the same set of parts. In multivariate analysis of compositional data, it has long been recognized that a valid approach to compositional data is to analyse ratios of parts, which are invariant to the choice of the set of parts. Basing the analysis on ratios is an approach with the property of *subcompositional coherence*: the relationships between parts is invariant to the particular mix of compositional parts in a data set, not changing when some other parts are excluded or added (see, for example, Aitchison 2005, sect. 1.7; van den Boogaart 2013, sect. 2.2.3). The logarithmic transformation is important, because ratios are compared multiplicatively, and log-ratio analysis (Aitchison 1990, Aitchison and Greenacre 2002) is a subcompositionally coherent variant of principal component analysis that displays the reduced dimensional structure of all log-ratios of the parts. Subsequently, Greenacre and Lewi (2009) showed the benefits of weighting the parts by their average proportions, leading to a weighted form of log-ratio analysis that is identical to *spectral mapping* , a method that has been used in drug development and biostatistical applications (Lewi 1976, 1980; Wouters et al. 2003).

There are many papers on Aitchison's approach to compositional data analysis, abbreviated as CoDa: to mention but a few key publications, Aitchison 1994, Aitchison et al. 2000, Aitchison and Egozcue 2005, Pawlowsky-Glahn, Egozcue and Tolosana-Delgado 2007, and the seminal book edited by Pawlowsky-Glahn and Buccianti (2011). I will refer to these authors who take inspiration from Aitchison's foundational work as the "CoDa school". The fundamental idea of their approach is the log-ratio transformation, and three log-ratio transformations have been proposed: the *additive log-ratio* (ALR), the *centred log-ratio* (CLR) and the *isometric log-ratio* (ILR), each with their advantages and disadvantages. Of these three, only ALRs have a clear practical interpretation, since they are constructed from simple ratios of parts and do not involve ratios of geometric means of parts, as is the case with CLRs and ILRs. I hope to show that, by slightly relaxing the strict requirements of the CoDa school's approach in a controlled and measurable way, then other sets of simple ratios of parts can be used, which are easier for the practitioner to interpret, and – for all practical purposes – serving the same purpose as the classic ones.

A relaxation of the strict requirement of compositional coherence, in the same spirit as the present article, has already been proposed by Greenacre (2011a), who defined a measure of subcompositional incoherence, that is how far any given method applied to a compositional data set is from the ideal of subcompositional coherence. The requirement of the CoDa school that methods should be strictly subcompositionally coherent excludes methods that are close to being coherent, for example correspondence analysis (CA), which is known to be theoretically linked to log-ratio analysis (LRA) by the Box-Cox transformation (Greenacre 2009, 2010a). The chi-square distance in CA approximates the log-ratio distance in LRA, coming closer and closer to it as the power parameter of the Box-Cox transformation reduces to 0 and in the limit the two methods are identical. Moreover, since CA handles data zeros with ease, which is one of the main problems of Aitchison's log-ratio approach. In this sense, CA provides the next best choice to LRA to identify dimensions underlying a compositional data set (Greenacre 2011b). Moreover, its closeness to being compositionally coherent can be measured, as proposed by Greenacre (2011a). This relaxation of a strict requirement, while checking how much the method deviates from that requirement in a particular application, is no different from current statistical practice that uses theory involving the strict assumption of the normal distribution, for example, but then relaxing this assumption to perform a hypothesis test after checking that the data do not deviate too much from normality. This article continues in the same spirit, proposing alternative simpler approaches to CoDa that come measurably close to the strict requirements of the CoDa school.

An initial section (Sect. 2) lays the methodological foundation for what is to come. Two areas of methodology are essential to support my approach, namely network graph representation and the form of multivariate regression called redundancy analysis. I will stress once again the practical usefulness in multivariate CoDa of part weighting, where parts occurring in low proportions should receive lower weight than their higher proportion counterparts.

In Sect. 3 I will start with a simple application that restores the ALR transformation back to the stage of CoDa. Pawlowsky-Glahn, Egozcue and Tolosana-Delgado (2007) dismiss the ALR transformation, saying that it "is frequent in many applied sciences and should be avoided". I will discuss their reasons for this statement and show empirically, by contrast, that the ALR can be very "close" to the complete set of log-ratios in more than one sense and thus of potential use to the practitioner tackling a compositional data set. I will also summarize the well-known theoretical niceties of the CLR and ILR transformations, being equivalent to the full set of log-ratios, but explain why I think that they are not as useful for the practitioner as using a set of simple ratios.

In Sect. 4 I will discuss ratio selection, a CoDa version of variable selection, leading to a small set of ratios, measuring how much "information" in the form of variance is lost in their selection, and showing how close their implied inter-sample distances are to the log-ratio distances based on all pairwise ratios. Using a few ratios that are essentially equivalent, for all practical purposes, to the original data set and which have a simple and intuitive interpretation, can considerably ease the task of the applied researcher. Given the lack of subcompositional coherence in the parts themselves, it seems obvious that when it comes to reporting univariate statistics, these should rather be in terms of ratios of parts. In certain research areas, for example in fatty acid analyses in studies of the marine food web, some specific ratios are indeed proposed as indicators of certain substantive phenomena – see, for example, Kraft et al. (2015). In my opinion, reporting a selected set of ratios, their distributions and confidence intervals on means, for example, should be the norm rather than the exception, since these are the only quantities that are comparable across studies.

As an illustration I will use throughout the following sections a data set on the compositions of 11 oxides in a set of 47 Roman glass cups from an archaeological site in eastern England (Baxter, Cool and Heyworth 1990), where oxygen is combined with elements silicon (Si), aluminium (Al), iron (Fe), magnesium (Mg), calcium (Ca), sodium (Na), potassium (K), titanium (Ti), phosphorus (P), manganese (Mn) and antimony (Sb). These oxides will always be referred to and labelled by these abbreviations of the elements. The data are reproduced in Table 2 of Greenacre and Lewi (2009), who highlight the difference between unweighted and weighted log-ratio analysis of these data, demonstrating the benefits of the latter. This data set is provided electronically here as supplementary material. This is a highly suitable data set, since it consists of parts with widely varying average proportions. It also has no zero values, which facilitates comparison with the case where all log-ratios are considered and avoids the side issue of data zeros.

2. Some theoretical concepts

2.1 Total variance of a compositional data set

The total variance in a compositional data set, following Aitchison (1983, 1986), is measured by the total log-ratio variance. This measure is so fundamental that I shall present several equivalent definitions of this concept, serving to highlight different properties. Suppose that the data are in a samples-by-parts matrix **X** ($n \times p$), where the rows of **X** sum to a constant, which can be set to 1 without loss of generality (hence, the data are proportions and the rows are compositions). Then the (unweighted) log-ratio variance, defined by Aitchison (1983), consists of first defining the logarithms of the ratios of all $\frac{1}{2}p(p-1)$ pairs of parts, that is expanding the columns of **X** into a $n \times \frac{1}{2}p(p-1)$ matrix of log-ratios **Z** and then computing the grand total of the column variances:

$$\text{TotVar (unweighted)} = \sum_{j<j'} \frac{1}{n} \sum_i \left(z_{i,jj'} - \bar{z}_{jj'}\right)^2 \quad \text{where } z_{i,jj'} = \log \frac{x_{ij}}{x_{ij'}} \text{ and } \bar{z}_{jj'} = \frac{1}{n} \sum_i z_{i,jj'} \qquad (1)$$

The notation $\sum_{j<j'}$ indicates the double summation over all $\frac{1}{2}p(p-1)$ unique pairs of the index. This quantity can be equivalently defined in terms of all the pairwise squared differences between the log-ratios, as follows:

$$\text{TotVar (unweighted)} = \frac{1}{n^2} \sum_{i<i'} \sum_{j<j'} \left(z_{i,jj'} - z_{i',jj'}\right)^2 = \frac{1}{n^2} \sum_{i<i'} \sum_{j<j'} \left(\log \frac{x_{ij}}{x_{ij'}} - \log \frac{x_{i'j}}{x_{i'j'}} \right)^2$$

$$= \frac{1}{n^2} \sum_{i<i'} \sum_{j<j'} \left(\log \frac{x_{ij}}{x_{ij'}} \frac{x_{i'j'}}{x_{i'j}} \right)^2 \qquad (2)$$

The last version on the right hand side of Eq. (2) shows the sum of squares of the logarithmically transformed odds-ratios based on all unique pairs of rows and columns of the data matrix.

As shown by Greenacre and Lewi (2009), weighting the parts has many important advantages. Parts occurring in low proportions need to be down-weighted because they induce large log-ratio variances and will dominate any analysis of the complete set of log-ratios (the rare oxide of the element Mn in the glass cups data set, which occurs only with values 0.01%, 0.02% and 0.03%, illustrates this argument perfectly) . A good default weighting system for a table of positive values, all on the same scale, is to weight the rows and columns proportionally to their marginal totals, as is the case in correspondence analysis and spectral mapping . For compositional data, the row sums are all 1, so the row weights (which should sum to 1) are $r_i = 1/n$, constant across samples, and the column weights are $c_j = j^{\text{th}}$ part mean. For other tables of positive data all on the same scale, for example counts, the row weights can also vary. The definitions of (weighted) log-ratio variance corresponding to the unweighted ones in Eqs. (1) and (2) are thus respectively as follows:

$$\text{TotVar} = \sum\nolimits_{j<j'} c_j c_{j'} \sum\nolimits_i r_i \left(z_{i,jj'} - \bar{z}_{jj'}\right)^2 \quad \text{where } z_{i,jj'} = \log\frac{x_{ij}}{x_{ij'}} \text{ and } \bar{z}_{jj'} = \sum\nolimits_i r_i z_{i,jj'} \qquad (3)$$

$$\text{TotVar} = \sum\nolimits_{i<i'} r_i r_{i'} \sum\nolimits_{j<j'} c_j c_{j'} \left(z_{i,jj'} - z_{i',jj'}\right)^2 = \sum\nolimits_{i<i'} \sum\nolimits_{j<j'} r_i r_{i'} c_j c_{j'} \left(\log\frac{x_{ij}}{x_{ij'}} - \log\frac{x_{i'j}}{x_{i'j'}}\right)^2$$

$$= \sum\nolimits_{i<i'} \sum\nolimits_{j<j'} r_i r_{i'} c_j c_{j'} \left(\log\frac{x_{ij}}{x_{ij'}}\frac{x_{i'j'}}{x_{i'j}}\right)^2 \qquad (4)$$

Because of the demonstrated benefits of the weighting, I will maintain the weighted version in Eqs. (3) and (4) as the definition of the log-ratio variance, and qualify it with the adjective "unweighted" when referring to Aitchison's original definition in Eq. (1) or its equivalent in Eq. (2).

A third equivalent definition of the log-ratio variance uses the matrix **Y** of centered log-ratios (CLRs): $y_{ij} = \log(x_{ij} / \prod\nolimits_j x_{ij}^{c_j}) = \log(x_{ij}) - \sum\nolimits_j c_j \log(x_{ij})$, that is the rows of the log-transformed centered with respect to their weighted row means. Then the total variance is the weighted average of the variances of the *p* columns of **Y**:

$$\sum_{j=1}^{p} c_j \sum_{i=1}^{n} r_i (y_{ij} - \bar{y}_j)^2 \text{ , where } y_{ij} = \log(x_{ij}) - \sum_{j=1}^{p} c_j \log(x_{ij}) \text{ and } \bar{y}_j = \sum_{i=1}^{n} r_i y_{ij} = \frac{1}{n}\sum_{i=1}^{n} y_{ij} \qquad (5)$$

which can be expressed in vector notation and another equivalent form in terms of inter-CLR squared distances:

$$\sum_{j=1}^{p} c_j \frac{1}{n} \sum_{i=1}^{n} (y_{ij} - \bar{y}_j)^2 = \frac{1}{n} \sum_{j=1}^{p} c_j \,\|\, \mathbf{y}_j - \bar{\mathbf{y}} \,\|^2 = \frac{1}{n} \sum_{j=1}^{p} c_j \,\|\, \mathbf{y}_j - \bar{\mathbf{y}} \,\|^2 = \frac{1}{n} \sum \sum\nolimits_{j<j'} c_j c_{j'} \,\|\, \mathbf{y}_j - \mathbf{y}_{j'} \,\|^2 \quad (6)$$

where $\mathbf{y}_j$ is the *j*-th column of the CLR matrix **Y**.

Since the log-transformed data matrix **X**, log(**X**) = [log(x_{ij})], is first centered by their respective weighted row means to obtain the matrix **Y**, it follows that the elements $y_{ij} - \bar{y}_j$, which are centered by the column means, are a double-centering of the matrix log(**X**) using the column and row weights c_j and $r_i = 1/n$ respectively. Then these double-centered values are squared and summed using the column and row weights again to obtain the total log-ratio variance. In matrix notation, where the column and row weights are gathered in vectors **c** and **r** respectively, the CLR matrix is

$$\mathbf{Y} = \log(\mathbf{X})(\mathbf{I} - \mathbf{1}\mathbf{c}^{\mathsf{T}})^{\mathsf{T}} \tag{7}$$

and the double-centered matrix is

$$\mathbf{S} = \mathbf{Y} - \mathbf{1}\mathbf{r}^{\mathsf{T}}\mathbf{Y} = (\mathbf{I} - \mathbf{1}\mathbf{r}^{\mathsf{T}})\log(\mathbf{X})(\mathbf{I} - \mathbf{1}\mathbf{c}^{\mathsf{T}})^{\mathsf{T}} \tag{8}$$

Then the total log-ratio (LR) variance in Eqs. (3) and (4) is the weighted sum of squares of **S**:

$$\text{LR variance} = \text{trace}(\mathbf{D}_r \mathbf{S}\, \mathbf{D}_c \mathbf{S}^{\mathsf{T}}). \tag{9}$$

where $\mathbf{D}_r$ and $\mathbf{D}_c$ are the respective diagonal matrices of the weights.

A fourth equivalent definition of the log-ratio variance can be obtained by defining a set of ILRs known as *balances*. Balances are defined by a sequential binary partition of the parts, the simplest being part 1 versus the remaining 2,..., p, then part 2 versus the remaining 3,..., p, and so on, until part $p-1$ versus part p, represented in the dendrogram of Fig. 1(a). One way of proving the total variance decomposition across the balances is to resort to the well-known result attributed by Benzécri (1973) to the Dutch mathematician Christiaan Huygens (1629–1695) – this result is used in multivariate analysis of variance to calculate the decomposition of total sum-of-squares (TSS) into between-groups sum-of -squares (BSS) and within-groups sums-of-squares (WSS): TSS = BSS + WSS. Contrasting two groups in a partition (of the parts in this case), the two group centroids define a between-group sum-of-squares. Each of these group's within sum-of-squares is in turn decomposed according to subsequent binary splits and each time a between-group sum-of-squares is computed. Eventually, once the partitioning arrives at two single parts, which have no within sums-of-squares, the sequence of between-group sums-of-squares has decomposed the total variance. This procedure was executed in two ways: first, as a stepwise divisive (descending) procedure where at each step one part was contrasted against the remainder at that step, which I will call a *chain balance* (Fig. 1a), and an agglomerative (ascending procedure by weighted Ward clustering (Fig. 1b), which I will call the *Ward balance*. In both cases the sum of the heights of the 10 nodes is exactly the total log-ratio variance, equal to 0.002339 for the glass cup data, and hence define a decomposition of the total variance. I will return to the CLR and ILR transformations later, but suffice it to say for the moment that although their properties might be of theoretical interest, the CLRs and ILRs are not convenient for the practitioner to interpret. In my experience, mainly with compositional fatty acid data in ecology, no researcher finds a geometric mean, let alone a ratio that involves a geometric mean, an interesting summary statistic. However, practitioners do favor aggregating parts, for example summing together all the saturated or unsaturated fatty acids, and making ratios of these sums: $\sum$saturated FAs/$\sum$unsaturated FAs. Similarly in geochemistry, interest may be on elements that behave together, for instance combining alkilies such as K_20 and Na_2O. However, log-ratios of sums do not fit into the rigors of the CoDa school's approach. I will show in Sect. 4 that practitioners' wishes for more interpretable variables can be accommodated in a pragmatic approach that has only slight (and measurable) incompatibility with the CoDa school's strict framework.

2.2 Graph representation of log-ratios

It will be very useful to represent log-ratios of parts as links between the parts in a network. In graph theory the parts are called *vertices* and the links *edges*. For example, the CoDa school's approach analyses all $\frac{1}{2}p(p-1)$ pairwise log-ratios, and these can be represented in the graph of Fig. 2a, called a *complete graph* because every pair of vertices is connected by an edge. In Fig. 2b, the set of $p-1$ additive log-ratios are displayed, where the oxide of **Si** is the denominator and edges in the form of arrows pointing towards the numerator – this is called a *directed graph*.

Actually, Fig. 2a should also be directed, with its ½$p(p - 1)$ edges as arrows, in which case the complete directed graph is called a *tournament* – see, for example, Harary and Palmer (1973) or Bóna (2006). When I deal with subsets of ratios in Sects. 3 and 4, the representation of these in a graph will enhance understanding, and certain graph-theoretic results will be very useful.

2.3 Redundancy analysis and the vegan package in R

Redundancy analysis (RDA) is a variant of multivariate regression, where there are p responses instead of just one, as in the standard regression model. The name originates in a paper of Wollenberg (1977) but the method was first defined by Rao (1964), who called it "principal component analysis of instrumental variables". Gittins (1985) gives a thorough treatment and equates the term "redundancy" with explained variance, which is exactly the use I will make of it here. In its simplest form, given a $n \times p$ matrix of responses $\mathbf{Z}$ (which are appropriately transformed and would be suitable for PCA) and an $n \times m$ matrix of explanatory variables $\mathbf{W}$ (which serve to explain variance in each of the columns of $\mathbf{Z}$), then dimension reduction is performed not on $\mathbf{Z}$ itself but rather on its projection $\mathbf{Z}^*$ onto the space defined by $\mathbf{W}$: $\mathbf{Z}^* = \mathbf{W}(\mathbf{W}^\mathsf{T}\mathbf{W})^{-1}\mathbf{W}^\mathsf{T}\mathbf{Z}$. The total variance in $\mathbf{Z}$ is then split into two orthogonal components: the part in $\mathbf{Z}^*$ that is explained by $\mathbf{W}$ and the part in $\mathbf{Z} - \mathbf{Z}^*$ that is uncorrelated with $\mathbf{W}$ and thus unexplained by the explanatory variables. The **vegan** package in R (R Core Team 2015) implements RDA in the function **`rda`**, but also in another function **`adonis`**, which has the advantage of operating either on a rectangular matrix $\mathbf{Z}$ of responses, or on the square distance matrix implicit in the analysis of $\mathbf{Z}$, for example the matrix of Euclidean interpoint distances, as in PCA. As we have seen in Eq. (6), for example, the total variance can always be expressed equivalently in terms of a matrix of squared interpoint distances. Because of **`adonis`**'s ability to analyze distance or dissimilarity matrices directly as responses, I have used it throughout in what follows, while noting that identical results are obtained whether the original matrix **Z** is given or the distance matrix between rows of **Z**.

3. Resurrecting the additive log-ratio transformation

The additive log-ratio (ALR) transformation was the original transformation used by Aitchison (1986) but in the book van den Boogaart and Tolosana-Delgado (2013, p. 44) the authors say they "will not use it in this book" (p. 44) and ALR "should not be used in case that distances, angles, and shapes are involved, as it deforms them" (p.45). Other authors from the CoDa school similarly reject this transformation, for at least the following two reasons (quoted from Pawlowsky-Glahn et al. 2007): "the alr is not symmetrical in its components"; it defines "coordinates in an oblique basis, something that affects distances if the usual Euclidean distance is computed from the alr coordinates". I will deal with these in turn.

The critique that it is not symmetrical in its components is a feeble one – it is at least a set of simple ratios that the practitioner can understand, and I will show that it can do the job it is intended for, namely represent the variance in the original data and not distort the results. Because the total log-ratio variance is known to be contained in a (p–1)-dimensional space, it can be easily demonstrated that any set of ALRs, which are by definition p–1 in number and containing every part at least once, will explain the totality of the log-ratio variance. There is a large component of redundancy in the complete set of log-ratios, which are differences in the logarithms with many repetitions, and are fully explained when regressed on a set of log-ratios that contain each part at least once. Notice the distinction between explaining the variance as opposed to containing the log-ratio variance – it is clear that a set of p–1 ALRs must have less variance than the set of all ½p(p – 1) log-ratios. But some sets of ALRs contain more variance than others, as I will show now.

There are p potential sets of ALRs, each of which can be shown by arrows emanating from one of the parts in Fig. 2b to all the other p–1 parts (the set of ALRs from Si is shown in that figure). Each set of ALRs was used as explanatory variables in a redundancy analysis to measure how well they explain the log-ratio data set, and as expected each set explains 100% of the total log-ratio variance. But some sets contain more inherent variance than others – see Table 1. It is clear that the set of ALRs using Si as the reference contains the most variance, almost 90% of the total. As will be shown soon, it is an excellent set of ratios representative of the complete data set.

The second criticism of the ALR transformation is that it will affect the log-ratio distances and "deform" the space of the samples, that is poorly represent the inter-sample distances. These distances are the so-called Aitchison distances, but defined here in their weighted form, between two cases (rows) i and i', as follows (cf. (4)):

$$d_{ii'} = \sqrt{\sum_{j<j'} c_j c_{j'} \left(\log \frac{x_{ij}}{x_{ij'}} - \log \frac{x_{i'j}}{x_{i'j'}} \right)^2} \qquad (10)$$

Since the sum in Eq. (10) is over ½$p(p-1)$ log-ratios whereas the ALR distances involve summing over only $p-1$ of them, it is clear that all the distances will be affected, but what is not clear is if there is any serious "deformation" and if this has any practical consequences on the results. Using the "best" set of ALRs (i.e., containing most variance, 89.3%), with respect to Si, and comparing the log-ratio distances in Eq. (10) with the interpoint distances based on this set of ALRs, the scatterplot in Fig. 3a is obtained. Because I will show the two-dimensional solutions later and even surmise that these data only have one significant dimension, the same comparison is made of two- and one-dimensional LRA solutions with the corresponding two- and one-dimensional solutions of the analysis of the 10 ALRs based on Si, in Fig.s 3b and 3c respectively. The reproduction of the distances in each case can clearly be seen to be excellent and refutes the general assertion that the distances are "deformed(...)if the usual Euclidean distance is computed from the alr coordinates", at least in this example. As an alternative approach that compares configurations rather than inter-point distances, the last column in Table 1 shows the Procrustes correlation, also computed in **vegan** using the function **`Procrustes`**, which compares the $(p-1)$-dimensional configuration from the log-ratio analysis (LRA) of the data set and the $(p-1)$-dimensional configurations based on the PCAs of the respective sets of ALRs. There are several very high Procrustes correlations, the highest being for the ALRs with respect to Si. This suggests that the ALRs are preserving more of the "signal" in the data than might have been expected, since we know that all data sets contain a high proportion of unexplained "noise" variation. Computing the Procrustes correlations that compare the two-dimensional and one-dimensional solutions (for one dimension this is just the regular correlation coefficient), the agreement is even higher, so almost none of the important log-ratio distance variance contained in the major LRA dimensions is lost by using this set of ALRs.

This last idea of comparing reduced-dimensional solutions is important because, after all, in a multivariate CoDa this is what will be interpreted by the practitioner, assuming that the minor dimensions reflect the "noise" variance. The two-dimensional solution based on the "full" LRA and the two-dimensional solutions based on the sets of ALRs with respect to Si are shown in the form of biplots in Fig. 4. These are contribution biplots (Greenacre 2013), where the lengths of the arrows are directly related to the corresponding variable's contribution to the two-dimensional solution. The configuration of the case points (labelled 1 to 47) of Fig. 4b, based on the ALRs with respect to Si, compared to that in the LRA biplot in Fig. 4a, shows a strong resemblance, as expected from the high Procrustes correlation of 0.9982 (see Fig. 3b). This shows again that these ALRs are sufficient to give practically the same result as the LRA (in Sect. 4 I will show that even less log-ratios are required to achieve this goal). In addition, notice that the first dimension in Fig. 4b has a much higher percentages of variance on the first axes, compared to Fig. 4a, which suggests the following possible reason.

These data have a relatively low variance, and the question can be asked what the non-random dimensionality of the data set is. Applying a permutation test in the style of Greenacre (2016, chapter 30) to the LRA solution, the first dimension was found to be highly significant (p<0.001), but for all the other dimensions the p-values were very high and non-significant. The conclusion is that there is really only one non-random dimension in these data. Therefore, we should be comparing one-dimensional solutions, not two-dimensional ones. Figure 5 shows the simple scatterplot of the 47 cases (the glass cups) according to the first coordinate in the LRA plot (horizontal axis of Fig. 4a) and the first axis of the ALR analysis (horizontal axis of Fig. 4b). This is the vizualization of the correlation of 0.9994 in Fig. 3c. If one accepts the permutation test that there is only one significant dimension, then the ALR analysis has detected it almost perfectly, but no less perfectly whatever the "signal" in the data might be. This suggests the explanation why the first dimension in the LRA analysis has a much lower percentage of variance on the first dimension, because this analysis has a higher total variance, containing the "noise" variance augmented by a multiplicity of additional redundant log-ratios that are included in the analysis. This situation is no different from PCA, for example, if a data set with p variables is augmented with new variables that are linear combinations of the old ones, for example differences between pairs of variables. The dimensionality and information content is the same, but the more these new variables are added, the more the total variance in the data set increases and the percentages of variance on the major axes decrease.

To summarize what has been learnt about the glass cup data in this section, it seems that:

(i) the ALR analysis reproduces log-ratio distances based on all pairwise ratios very accurately, sacrificing a very small part of distance variance, less than 0.5%;

(ii) there may be only one non-random dimension in these compositional data, the remainder are compatible with random variation;

(iii) an analysis based on simple ALRs has almost exactly identified this first dimension, but also identifies other dimensions accurately;

(iv) the (p−1)-variate ALR analysis does not contain as much noise variance as the ½p(p − 1)-variate LRA, so the first dimension appears relatively stronger in the ALR analysis, compared to the LRA one.

All in all, the ALR analysis appears to be equivalent to a full-blooded LRA for all practical purposes, and has a simpler interpretation, constructed from a small set of log-ratios of parts. In this example the set of ALRs based on Si appeared the most useful, but I found very similar results using other sets of ALRs too, always monitoring their variance explained, variance contained, Procrustes correlation and distance plots in comparison to the Aitchison approach. In the next section I will show that even fewer than $p - 1$ log-ratios can be adequate to represent the full set of log-ratios, and thus simplify the practitioner's task even more.

As an interesting side result, it is noteworthy that the set of ALRs that emerges the closest to the CoDa log-ratio "ideal" is when the oxide of Si (i.e., SiO_2) is used in the denominator of the ratios (Fig. 2b). This recalls the definition of the Harker diagram for identifying relationships between oxides in a geological context: "The oldest method is the variation diagram or Harker diagram which dates from 1909, and plots oxides of elements against SiO_2" (quoted from **https://brocku.ca/earthsciences/people/gfinn/petrology/variatn.htm** – original reference is Harker (1909), see also Cortés (2009)).

4. Variable selection in compositional data analysis

I see variable selection in CoDa taking two possible paths: selecting parts or selecting ratios. If the first path is chosen, then a subset of parts is found based on some optimality criterion, after which the data are reclosed and analysis continues in the log-ratio framework using this subcomposition. If the second path is chosen, then a subset of log-ratios is found in the same way and these are treated afterwards just like regular variables, or they suggest a reduced set of parts for further analysis. I will concentrate on the selection of log-ratios, in the belief that simple ratios can solve practitioners' data analytic problems without resorting to the "all log-ratios" approach based on CLRs or ILRs, or even the p–1 variable set of ALRs as in Sect. 3. Simple ratios are compatible across studies, and their choice can also be guided by the practitioner who has expert knowledge about the data and its context. Whatever the way ratios are chosen, by expert knowledge, by statistical criteria or a combination of both, the relationship of the ratios to the original data set can be measured, for example how much the ratios explain the log-ratio variance, how much of the log-ratio variance they contain and how close the distances based on ratios approximate the log-ratio distances.

4.1 Choosing a set of ratios

An automatic way of identifying a "good" set of ratios can proceed in a stepwise fashion, trying in the first step every individual log-ratio as an explanatory variable in explaining the log-ratio variance in a redundancy analysis, and selecting the one with the highest percentage of variance explained. This ratio is then fixed as the first log-ratio and then the second best log-ratio in combination with the first is sought, then fixed, and so on, similar to stepwise regression. Care must be taken to choose ratios that are "independent" of the ones already chosen: for example, if A/B and B/C have already been selected, then A/C is no longer a candidate for selection, since it depends on the others: A/C = A/B ×B/C, or on the log-scale, log(A) −log(C) is the sum of, and thus linearly dependent on, log(A) −log(B) and log(B) −log(C). Since the dimensionality of a p-

part compositional data set is $p - 1$, and if all the parts have appeared in at least one log-ratio after $p - 1$ steps of the above procedure, the variance explained will be 100%, as was the case for each set of ALRs in Sect. 3.

If ratio selection is done stepwise "by hand" by an expert, it will be necessary to check that the chosen ratio of two parts is not dependent on the ones chosen previously. This is easy to check in the case of three parts, as just described, for which at most two ratios are admissible, but becomes more difficult when the dependence might be via several ratios of parts. Here the graph representation can be very enlightening; for example, Fig. 6a represents five ratios, showing their exact definition by indicating solid arrows from the denominator to the numerator in each case (i.e., Si/Ca, Si/Na, Si/Ti, Ti/P, Ti/Mn). No other ratios of the parts represented here are admissible, since they will necessarily depend on the existing ratios. Two such inadmissble candidate ratios are shown by dashed arrows in Fig. 6b, Si/Mn and Ca/P. The ratio Si/Mn is like the A, B, C example mentioned above, and following the arrows the dependence relationship can be seen as Si/Mn = Ti/Mn × Si/Ti. The ratio Ca/P also closes a circuit and so can be obtained by following the path from P to Ca, where an arrow that goes in the reverse direction implies inverting the corresponding ratio. This is a type of vector geometry but multiplicative/divisive rather than additive/subtractive (but it is additive/subtractive in the log-ratios). Thus, Ca/P = Ti/P × Si/Ti × $(\mathrm{Si/Ca})^{-1}$. The only way to add ratios to this network would thus be to add new parts. It is clear that the network of multiplicatively independent ratios (by which is meant linearly independent log-ratios), cannot have any connected cycles. The requirement that all 11 parts appear in the network but only 10 ratios thus implies that they must be connected and *acyclic* (i.e., no cycles). From graph theory, the number of acyclic connected networks of p elements (also called *spanning trees*) is known – it is the Cayley number, equal to p^{p-2} (Bóna 2006), which in this example is 11^9 = 2 357 947 691, over 2 billion. (If this seems a lot, consider that the number of balances, which define dendrograms such as Fig. 1, is equal to $p! \times (p-1)!$ /

2^{p-1} = 141 455 160 000, over 141 billion). Clearly, an efficient way of choosing ratios is required and a stepwise procedure is proposed here.

This procedure starts by selecting, from the the ½×11×10 = 55 log-ratios in this example, the one that explains the most log-ratio variance. Using the **`adonis`** function again, the log-ratio of Si/Ca turned out to be the best, explaining 61.5% of the variance. The second best is Si/Sb, explaining an additional 12.6%, so the variance explained is now 74.1%, then Na/Sb which brings the variance explained up to 86.4%, and so on. The sequence of ratios and their accumulated explained variances are given in Table 2, and Fig. 7 represents the set of ratios in its acyclic graph connecting all the parts, where the edges show their order of entry in the stepwise procedure. In addition, Table 2 reports the medians of these ratios, as well as their reference ranges based on the estimated 0.025 and 0.975 quantiles (i.e., 2.5% and 97.5% percentiles). These statistics may be validly compared with the same ratios in other archaeological studies, whether the list of oxides is extended or not, since the ratios are invariant to the parts chosen by the researcher. The importance of Si as an element of high degree (i.e., the number of links to it, which is 7) is seen once more, even in this stepwise procedure.

What is not clear from the stepwise selection is that at some steps there can be more than one ratio competing for entry, giving the same additional benefit of variance explained. For example, in Table 2, the third ratio chosen was Na/Sb, explaining an additional 12.3% of the variance (increasing from 74.1% to 86.4%), but exactly the same increase would have been obtained if Si/Na or Ca/Na had entered. The important aspect of this third step is the entry of Na, which can be in a ratio with either Si, Ca or Sb. In my algorithm these ties were broken by choosing the ratio that, when added to the list of ratios at that step, maximized the Procrustes fit of the distances to the log-ratio distances. In this third step, this ratio turned out to be Na/Sb. It is also at this point that an expert could intervene to choose one of the "competing" ratios that has some relevant substantive meaning and interpretation in the context of the data.

The log-ratio biplot in Fig. 4a sheds light on the choice of the ratios. This incorporates the weighting of the parts (Greenacre and Lewi 2009; Greenacre 2009, 2010, 2011b), and the contribution biplot scaling (Greenacre 2013). Clearly the Si vs. Ca opposition is the most important along the first axis, which clarifies the choice of the first ratio as Si/Ca. It is no surprise either that Si/Sb is then chosen, to include Sb which is the most important contributor on the second axis. In the light of what was said in the previous section about the second dimension in Fig. 4a being compatible with random variation, it could be that the ratio Si/Sb is not worth retaining, but for the moment the stepwise procedure is designed only to replicate, as closely as possible, the log-ratio distances based on all the parts, whether the contained relationships are non-random or not.

Now, using only the 10 identified log-ratios in Table 2, Fig. 8a shows the contribution biplot – notice that the definitions of the ratios are inverted compared with Fig. 4, with Si appearing in the numerator. The resemblance with Fig. 4a is clear, and once again (as we saw for the ALR results in Fig. 4b) the percentage of variance explained on the first axis is much higher – the first dimension of Fig. 8a explains 67.2 % of the variance, whereas in Fig. 4a it has practically the same interpretation and explains only 39.6 %. Although Fig. 8b shows the distances based on these ratios to be less concordant with the log-ratio distances, compared to the ALR distances in Fig. 3, the Procrustes correlation for the 47 glass cups between the two-dimensional configurations of Fig. 4a and Fig. 8b is nevertheless a high 0.966.

4.2 Choosing a reduced set of ratios

In the contribution biplot of Fig. 8a three ratios stand out from the rest: Si/Ca, Si/Sb and Na/Sb, exactly the first three ratios that entered the stepwise process described in Sect. 4.1, explaining 86.4 % of the log-ratio variance. Figure 9a shows the biplot using just these three ratios, and it hardly differs from Fig. 8a. The configuration of the 47 glass cups is practically the same, and now the percentage of explained variance for this three-dimensional example is 74.1 % + 25.8 % = 99.9 %, with only 0.1 % lost on the remaining third dimension. Just these three ratios capture the first two dimensions of the original LRA analysis very accurately, with a Procrustes correlation of 0.950 between the samples in Fig. 9a and the LRA of Fig. 4a. Instead of the three ratios, a LRA can be performed on the subcomposition comprising the four oxides of Si, Na, Ca and Sb that make up these ratios. This LRA operates on the six pairwise ratios between the four oxides, including three ratios that are "redundant". The result is almost identical to Fig. 4a if one simply deletes the other seven oxides, with a Procrustes correlation of 0.988, but percentages of variance on the two dimensions of 75.2% and 15.5%. This approach of using the selection of ratios to lead to the selection of a subcomposition may be more acceptable to the CoDa school.

As a final step towards extreme parsimony, if one accepts that it is really only the first dimension of the LRA that is non-random, then a single ratio Si/Ca accounts for this dimension very accurately, as shown in the scatterplot of Fig. 9b. Here the ratio has been inverted and log(Ca/Si) is plotted against the first dimension of Fig. 4a, with an astounding correlation of 0.953. This begs the question that perhaps the non-random component of the whole data set can be reduced to the study of the subcomposition of only two parts, that is a single ratio between the oxides of calcium and silicon. In other words, this data set may be compatible with the situation where oxides of calcium and silicon are the informative parts, ordinating the glass cups in a non-random way in terms of log(Ca/Si), whereas the other nine oxides are non-informative random values added as additional columns, merely inflating the total variance.

5. Discussion and conclusion

The main point of this article is to show that a simple choice of ratios, logarithmically transformed, can account for all or most of the log-ratio variance in a compositional data set, and can be used for univariate or multivariate analysis as a substitute for the original data. In the example presented here, one set of ALRs, for example, has been shown to serve the same purpose as a LRA that involves all pairwise log-ratios, and comes to the same conclusion about the data using variables that are easier to interpret. I am not proving that this will always be the case, but this example serves as a counter-example to the assertion that ALRs should be avoided. I would rather say that ALRs can always be explored in CoDa for their effectiveness in summarizing the compositional data content. In case this single example is considered not sufficient evidence, the supplementary material contains three additional data sets and the results when ALRs are considered as substitutes for the full set of log-ratios, as in Sect. 3. Two of these compositional data sets are taken from Aitchison (2005) and the third one is considered by Greenacre (2016) in the context of correspondence analysis. In each of these three examples a set of ALRs gave excellent results, with Procrustes correlations 0.995, 0.960 and 0.989 respectively between the configuration based on the ALRs and the configuration based on the full set of log-ratios. These results echo the words of Aitchison (1994, p.76), who says that "a simple methodology for compositional data analysis" is suggested as follows: "Transform each composition (...) to its (vector of additive log-ratios...), then apply the appropriate, standard multivariate procedures to the logratio vectors." Notice Aitchison's use of the term "simple methodology", with which I fully agree.

Furthermore, I have shown that a reduced set of log-ratios, explaining part of the log-ratio variance, can be found to represent the essential content of the data set considered here. In the glass cup data set, three log-ratios were more than sufficient, and if one accepts the hypothesis that there is really only one non-random dimension in these data, then only one log-ratio was necessary to represent that component.

Univariate analysis of the log-ratios, or the ratios themselves, is particularly relevant since ratios are subcompositionally coherent and comparable across studies, whereas univariate statistics based on the original parts are not. The definitive book on CoDa, edited by Pawlowsky-Glahn and Buccianti (2011), contains almost no mention of univariate analysis of compositional data, except a passing reference by Lovell et al. (2011) to a paper by Filzmoser, Hron and Reimann (2009), who use the isometric log-ratio (ILR) transformation to arrive at a set of $p-1$ variables that replace the original data set. These new ILRs are defined as proportional to ratios of parts to geometric means of parts as follows (the constant of proportionality is not relevant here)

$$\log\left(\frac{x_{ij}}{(\prod_{k=j+1}^{p} x_{ij})^{1/(p-j)}}\right) = \log(x_{ij}) - \frac{1}{p-j}\sum_{k=j+1}^{p}\log(x_{ij}) \quad \text{for } j = 1,\dots, p-1 \qquad (11)$$

(balances described earlier in Sect. 2 are a special case of ILRs). The claimed advantage of ILRs is that they are orthogonal and reconstruct exactly the log-ratio distances. But as shown here, the lack of orthogonality in simple log-ratios hardly impedes their use in serving practically the same purpose, giving an excellent approximate representation of these distances. The problem with ILRs is that they have no easy interpretative meaning for the practitioner and remain of theoretical interest only – as van den Boogaart and Tolosana-Delgado (2013, p. 45) themselves say: "each coordinate might involve many parts (potentially all), which makes it virtually impossible to interpret [ILRs] in general". The same can be said of CLRs, which are a very useful computational short-cut to performing LRA on all pairwise ratios, but are by themselves of no practical usefulness as interpretable variables, are not subcompositionally coherent (Pawlowsky-Glahn et al. 2007, page 19), and are also linearly dependent , which can be "a source of problems when doing statistical analyses" (van den Boogaart and Tolosana-Delgado 2013, p. 42).

A specifically chosen set of independent part ratios, as I propose here, can come sufficiently close in practical terms to using all log-ratios, and has an easier interpretation, especially if guided by experts who are familiar with the data context. The set of ratios can be nicely visualized in an acyclic graph, which facilitates interpretation and understanding, clearly demonstrated in the glass cups application. Ratios also provide univariate statistics that can be validly summarized by regular statistical measures of centrality such as the mean and median, and dispersion measures such as standard deviation, margins of error and quantiles. These ratios, preferably log-transformed, can even be correlated or combined in multivariate analyses such as regression and principal component analysis, with the assurance that they are subcompositionally coherent. In a particular field, for example the archaeology of ancient glass where the set of parts (oxides) is fairly similar across studies, one can imagine a set of ratios becoming a benchmark for easier comparison of data sets.

Based on expert knowledge, a selection of ratios can be made that have a substantive interpretation, or expert knowledge can be combined with automatic statistical selection. For example, Tanimoto and Rehren (2008) consider the composition of glasses from the late bronze age and point out some elements that are "rather heterogeneous in their composition, particularly in their ratios of soda (Na_2O) to potash (K_2O) and lime (CaO) to magnesium (MgO)". If required, these ratios can be "forced" into the first two steps of the present algorithm, after which the same stepwise procedure can be performed searching among the other ratios. In the present glass cup data set it turns out that those two log-ratios, of Na/K and Ca/Mg, explain only 16.6% of the log-ratio variance. The automatic selection that follows immediately brings in the log-ratio of Si/Ca (or equivalently Si/Na), which increases the variance explained dramatically to 74.1% . A sequence of ratios then follows, bringing in a similar sequence of elements as in Table 2, and reaching 100% with 10 ratios, as before.

The same ratio-selecting approach can be followed if the compositional data set involves a comparison of groups, for example comparing two groups of glass cups according to the archaeological periods they come from. Instead of the total log-ratio variance, the between-group log-ratio variance would serve as the variance to be explained by the selected ratios. The best ratio would then constitute the log-ratio that explains the most between-group variance, and so on.

I firmly believe that the concept of subcompositionally coherent log-ratios is the key idea in compositional data analysis. I have shown that not all $\frac{1}{2}p(p-1)$ log-ratios are necessary to represent a compositional data set, and at most $p-1$ are necessary, and these form an acyclic graph. There are thus $\frac{1}{2}p(p-1)-(p-1)=\frac{1}{2}(p-1)(p-2)$ redundant log-ratios that depend linearly on the others and are merely inflating the total variance. Furthermore, only a few log-ratios are necessary to preserve the relevant part of the variance and correctly represent the major features of the data set. The ability of these ratios to represent the complete data set can be measured and assessed in many different ways: variance explained, variance contained, Procrustes correlation, distance plots, etc. One specific example has been used throughout, with the caveat that it might not always turn out as successfully as in the application considered here, where a small set of log-ratios was particularly effective in replacing the full set of log-ratios. Nevertheless, this example stands to counteract the strict and more complicated requirements of the CoDa school, and to promote this simpler "approximative" approach, which is reminiscent of the idea of Greenacre (2011a). So I hope that this simpler approach will now be explored for its usefulness in the analysis of other compositional data sets. Reducing a compositional data set to a few ratios is a major simplification for the practitioner and provides a pragmatic alternative to the restrictive solutions provided so far in the recent CoDa literature.

References

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

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 (1994) Principles of compositional data analysis. In Anderson TW, Olkin I, Fang KT (eds) Multivariate Analysis and its Applications. Institute of Mathematical Statistics, Hayward, California, pp.73−81

Aitchison J (2005) A Concise Guide to Compositional Data Analysis. URL: http://ima.udg.edu/Activitats/CoDaWork05/A_concise_guide_to_compositional_data_analysis.pdf (last accessed 31 December 2016)

Aitchison J, Barceló-Vidal C, Martín-Fernandez JA, Pawlowsky-GlahnV (2000). Logratio analysis and compositional distance. Math Geol 32: 271−275

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

Aitchison J, Egozcue JJ (2005) The statistical analysis of compositional data: Where are we and where should we be heading?" Math Geol 37:829–850

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

Benzécri J-P (1973) Analyse des Données, Tôme II; Analyses des Correspondances. Dunod, Paris

Bóna, M. (2006) A Walk Through Combinatorics: An Introduction to Enumeration and Graph Theory. Second Edition. World Scientific Publishing, Singapore

Cortés J (2009) On the Harker variation diagrams; a comment on “The statistical analysis of compositional data. Where are we and where should we be heading?” by Aitchison and Egozcue (2005). Math Geosc 41: 817−828

Filzmoser P, Hron K, Reimann C (2009) Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Science of the Total Environment 407: 6100−6108

Gittins R (1985) Canonical Analysis: a Review with Applications in Ecology. Springer, New York

Greenacre MJ (2009) Power transformations in correspondence analysis. Comp Stat Data Anal 53: 3107–3116

Greenacre MJ (2010a) Log-ratio analysis is a limiting case of correspondence analysis. Math Geosc 42: 129–134

Greenacre MJ (2010b) Biplots in Practice. BBVA Foundation, Bilbao. Free download from www.multivariatestatistics.org

Greenacre MJ (2011a) Measuring subcompositional incoherence. Math Geosc: 43, 681–693

Greenacre MJ (2011b) Compositional data and correspondence analysis. In: Pawlowski-Glahn V, Buccianti A (eds) Compositional Data Analysis. Wiley, Chichester UK, pp.104−113

Greenacre MJ (2013) Contribution biplots. J Comp Graph Stat 22: 107–122

Greenacre MJ (2016) Correspondence Analysis in Practice. Third edition. Chapman & Hall / CRC, Boca Raton, Florida

Greenacre MJ, Lewi PJ (2009) Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. J Classif 26: 29–64

Harary F, Palmer EM (1973) Graphical Enumeration. Academic Press, New York

Harker A (1909) Natural History of the Igneous Rocks. Methuen, London

Kraft A, Graeve M, Janssen D, Greenacre MJ, Falk-Petersen S (2015) Arctic pelagic amphipods: lipid dynamics and life strategy. J Plank Res 37:790−807

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

Lovell D, Müller W, Taylor J, Zwart A, Helliwell C (2011) Proportions, percentges, ppm: do the molecular biosciences treat compositional data right? In: Pawlowski-Glahn V, Buccianti A (eds) Compositional Data Analysis. Wiley, Chichester UK, pp.193−207

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H (2015). vegan: Community Ecology Package. R package version 2.3-2. https://CRAN.R-project.org/package=vegan

Pawlowski-Glahn V, Egozcue JJ, Tolosana-Delgado R (2007) Lecture Notes on Compositional Data Analysis. URL: http://dugi-doc.udg.edu/bitstream/handle/10256/297/CoDa-book.pdf?sequence=1 (last accessed 31 December 2016)

Pawlowski-Glahn V, Buccianti A (eds) (2011) Compositional Data Analysis. Wiley, Chichester UK

R core team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Rao CR (1964) The use and interpretation of principal component analysis in applied research. Sankhya A 26: 329–358

Tanimoto S, Rehren T (2008) Interactions between silicate and salt melts in LBA glassmaking. J Archaeol Sci 35: 2566–2573 van den Boogaart KG, Tolosana-Delgado R (2013) Analyzing Compositional Data with R. Springer-Verlag, Berlin.

Wollenberg AL (1977) Redundancy analysis – an alternative for canonical analysis. Psychometrika 42: 207–219

Wouters L, Göhlmann HW, Bijnens L, Kass SU, Molenberghs G, Lewi PJ (2003) Graphical exploration of gene expression data: a comparative study of three multivariate methods. Biometrics 59: 1131–1139

Table captions

Table 1: Results for ALRs using each part in turn as the reference one in the denominator. R^2 is the part of total log-ratio variance explained; variance is the amount of variance contained in the ALRs, used to order the list in descending order; %variance is the percentage of the total log-ratio variance (which is equal to 0.002339) contained in the ALRs; weight is the average proportion of the reference part, used in the weighted analysis; Procrustes is the Procrustes correlation that measures similarity between the geometry of the ALRs and the geometry of all log-ratios.

Table 2: Sequence of log-ratios of mineral oxides entering in a stepwise search, explaining the log-ratio variance of the whole compositional data set.

	R^2	variance	%variance	weight	Procrustes
Si	1	0.002089	89.3%	0.7237	0.9975
Na	1	0.000956	40.9%	0.1825	0.8439
Sb	1	0.000325	13.9%	0.0036	0.7689
Ca	1	0.000700	30.0%	0.0567	0.6989
Al	1	0.000153	6.5%	0.0194	0.9043
Mg	1	0.000150	6.4%	0.0046	0.9539
Fe	1	0.000135	5.8%	0.0031	0.6704
K	1	0.000116	5.0%	0.0040	0.7017
Ti	1	0.000022	0.9%	0.0007	0.6564
P	1	0.000016	0.7%	0.0005	0.6187
Mn	1	0.000015	0.7%	0.0001	0.5692

Table 1

Ratio	Cumulative Explained Variance	Median	95% Reference Range
1. Si/Ca	61.5%	13.3	10.1-15.0
2. Si/Sb	74.1%	206.5	120.4-403.5
3. Na/Sb	86.4%	53.3	32.1-93.6
4. Si/Fe	93.6%	244.3	163.8-340.8
5. Si/K	96.6%	151.8	112.3-181.9
6. Si/Mg	98.4%	157.8	117.3-230.0
7. Al/Na	99.2%	0.106	0.092-0.122
8. Si/Ti	99.5%	1043	726-1485
9. Si/Mn	99.8%	7260	2505-7497
10. Na/P	100.0%	358.0	273.3-455.0

Table 2

Figure captions

Figure 1: The dendrograms associated with two balances that define ILRs (isometric log-ratios) for the glass cup data: (a) a chain balance, (b) the Ward balance.

Figure 2: Graph representations of networks defined by (a) all pairwise log-ratios, (b) the additive log-ratios (ALRs) with Si as the denominator, indicated by the arrow emanating from Si to the others.

Figure 3: Comparison of log-ratio distances with distances based on ALRs with respect to Si (a) the full 10-dimensional space; (b) reduced 2-dimensional space; and (c) reduced 1-dimensional space.

Figure 4: (a) Log-ratio analysis (LRA) of the full data set (contribution biplot); (b) Principal component analysis (PCA) of the ALRs with respect to Si.

Figure 5: Almost perfect agreement of coordinates of 47 glass cups on first axis of LRA (i.e., their coordinates on the first axis of Figure 4a) and first axis of PCA of ALRs with respect to Si (i.e., their coordinates on the first axis of Figure 4b) – correlation is 0.999.

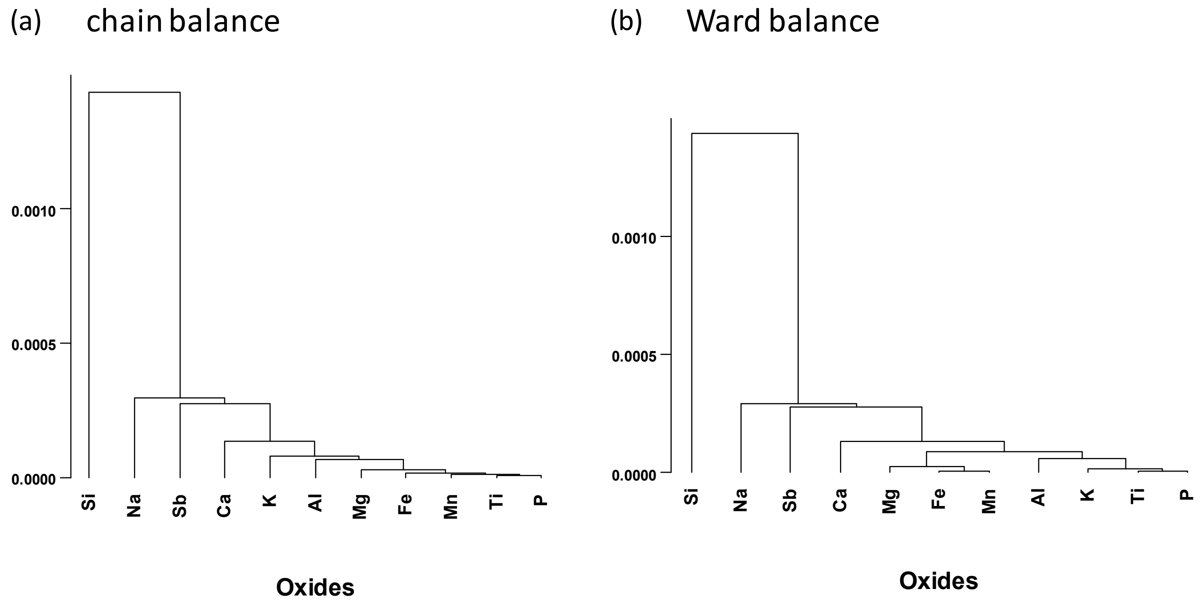
Figure 6: Illustration of acyclic graphs and dependent ratios: (a) An acyclic connected graph of six parts and five ratios, where the arrows point towards the respective numerators of the ratios;

(b) Two additional ratios, Si/Mn and Ca/P, that are dependent on the existing ones because they form cycles.

Figure 7: The graph of the 10 ratios chosen in a stepwise procedure to explain maximum log-ratio variance at each step, with numbers indicating their order of choice.

Figure 8: (a) The PCA contribution biplot of the 10 ratios shown in Figure 7; (b) The log-ratio distances compared to the distances defined between the 10 chosen log-ratios.

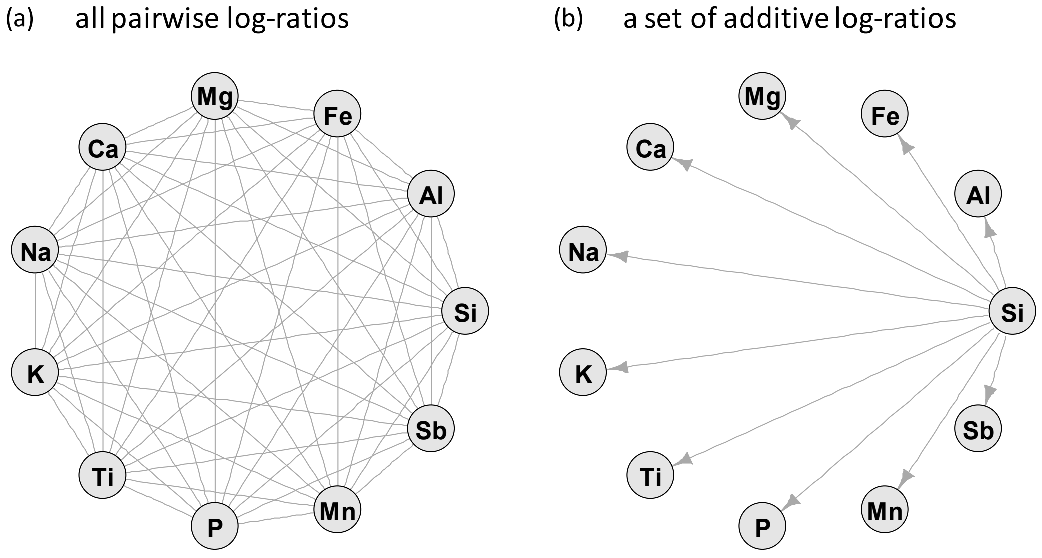
Figure 9: (a) PCA contribution biplot of the 3 "best" ratios Si/Ca, Si/Sb and Na/Sb from the stepwise ratio-selection procedure; (b) Concordance between the first coordinate of the LRA of Figure 4a and the log-ratio log(Ca/Si) – correlation is 0.953.



(a) chain balance

(b) Ward balance

Figure 1



(a) all pairwise log-ratios

(b) a set of additive log-ratios

Figure 2

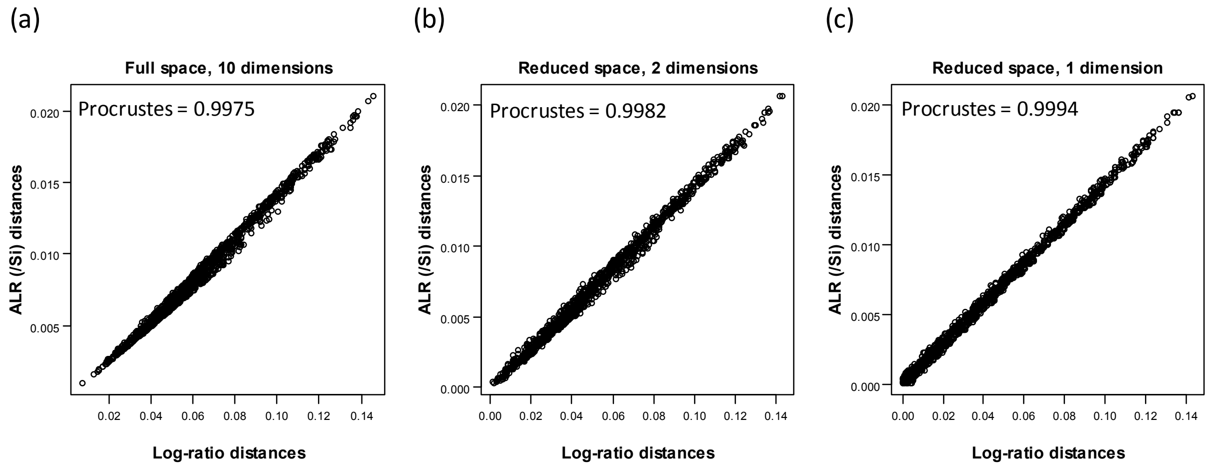

(a)
Full space, 10 dimensions
Procrustes = 0.9975
ALR (/Si) distances
0.005
0.010
0.015
0.020
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Log-ratio distances
(b)
Reduced space, 2 dimensions
Procrustes = 0.9982
ALR (/Si) distances
0.000
0.005
0.010
0.015
0.020
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Log-ratio distances
(c)
Reduced space, 1 dimension
Procrustes = 0.9994
ALR (/Si) distances
0.000
0.005
0.010
0.015
0.020
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
Log-ratio distances

Figure 3

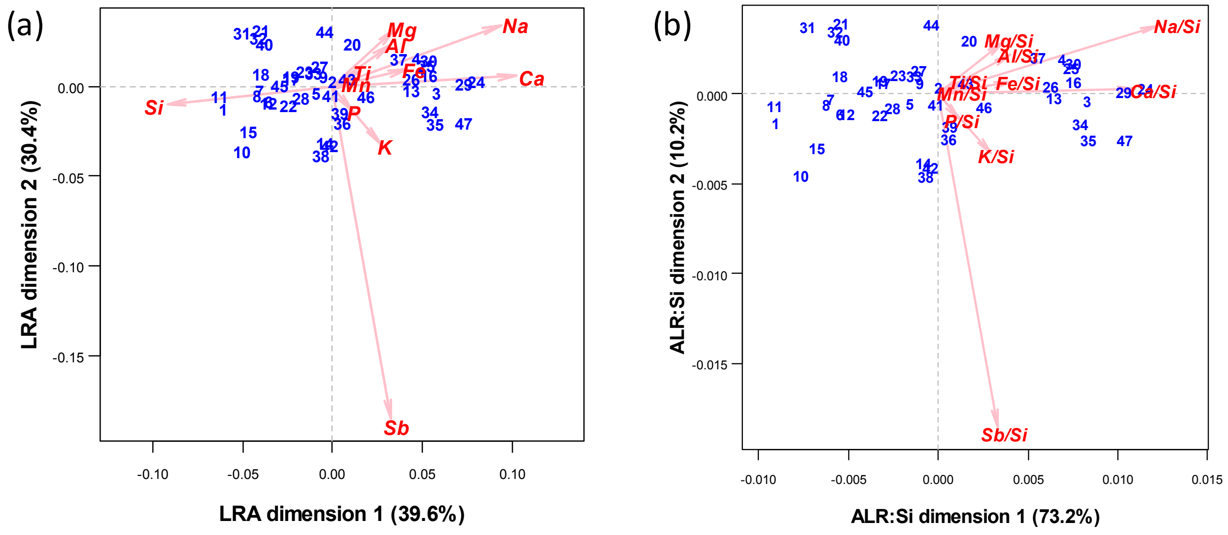

(a)
LRA dimension 1 (39.6%)
LRA dimension 2 (30.4%)
-0.10
-0.05
0.00
0.05
0.10
-0.15
Si
Na
Ca
Mg
Al
Ti
Fe
Mn
P
K
Sb
(b)
ALR:Si dimension 1 (73.2%)
ALR:Si dimension 2 (10.2%)
-0.010
-0.005
0.000
0.005
0.010
0.015
-0.015
Na/Si
Ca/Si
Mg/Si
Al/Si
Ti/Si
Fe/Si
Mn/Si
P/Si
K/Si
Sb/Si

Figure 4

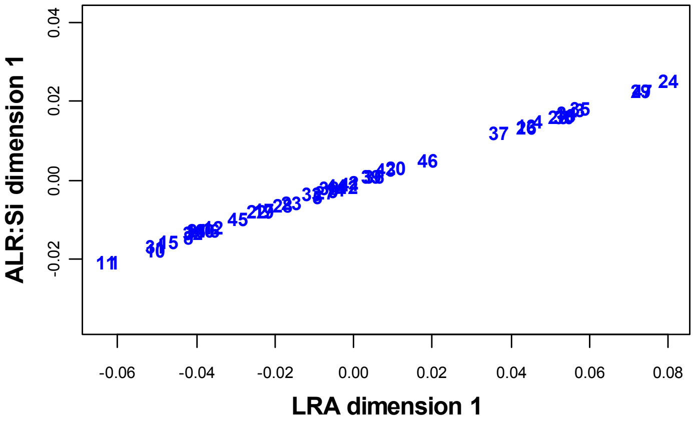

ALR:Si dimension 1
LRA dimension 1
-0.06
-0.04
-0.02
0.00
0.02
0.04
0.06
0.08
-0.02
0.00
0.02
0.04
11
31
5
10
45
46
37
24

Figure 5

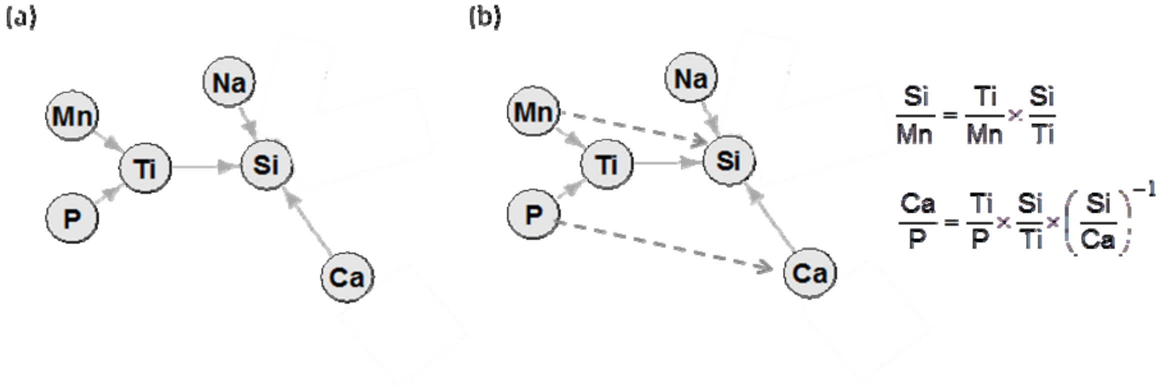

(a)
Na
Mn
Ti
Si
P
Ca
(b)
Na
Mn
Ti
Si
P
Ca
$\frac{Si}{Mn} = \frac{Ti}{Mn} \times \frac{Si}{Ti}$
$\frac{Ca}{P} = \frac{Ti}{P} \times \frac{Si}{Ti} \times \left(\frac{Si}{Ca} \right)^{-1}$

Figure 6

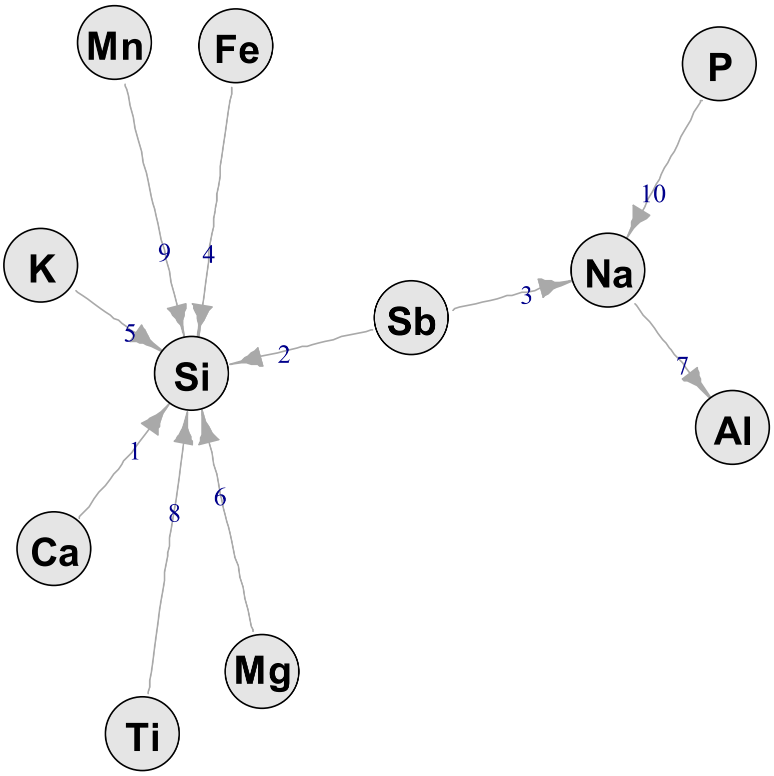


Figure 7

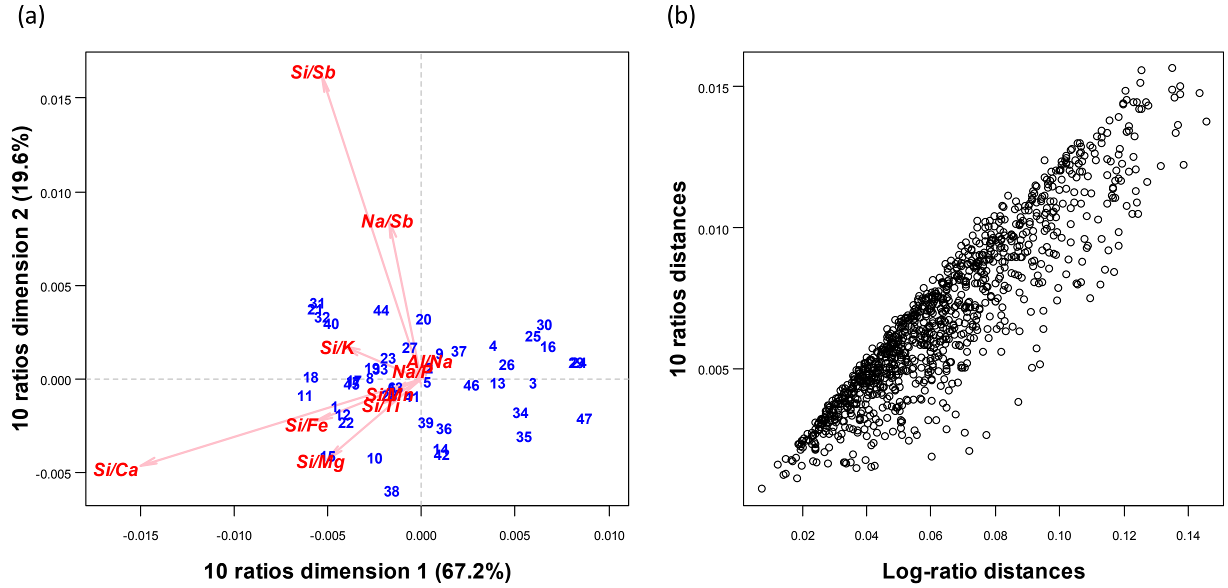

(a)
(b)
Si/Sb
Na/Sb
Si/K
Al/Na
Na/P
Si/Mn
Si/Ti
Si/Fe
Si/Mg
Si/Ca
10 ratios dimension 2 (19.6%)
10 ratios dimension 1 (67.2%)
0.015
0.010
0.005
0.000
-0.005
-0.015
-0.010
-0.005
0.000
0.005
0.010
31
21
32
40
44
20
30
25
16
27
4
37
9
23
26
19
33
18
8
5
46
13
3
11
41
1
12
22
34
47
39
36
35
14
42
15
10
38
10 ratios distances
Log-ratio distances
0.015
0.010
0.005
0.02
0.04
0.06
0.08
0.10
0.12
0.14

Figure 8

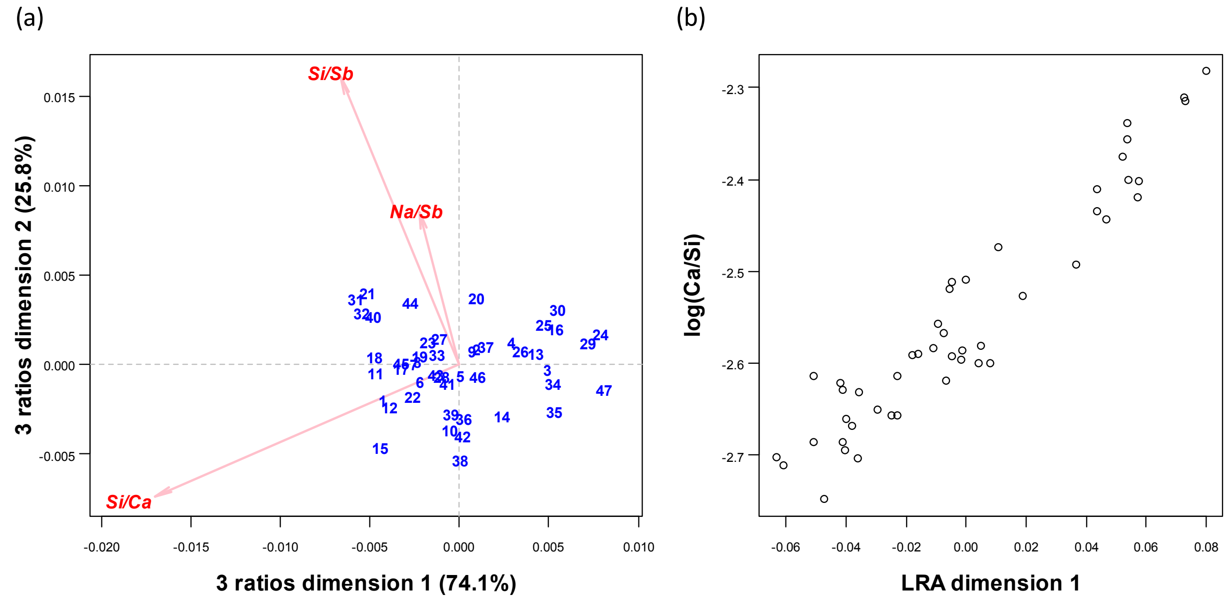

(a)
(b)
Si/Sb
Na/Sb
Si/Ca
3 ratios dimension 1 (74.1%)
3 ratios dimension 2 (25.8%)
-0.020
-0.015
-0.010
-0.005
0.000
0.005
0.010
0.015
LRA dimension 1
log(Ca/Si)
-0.06
-0.04
-0.02
0.00
0.02
0.04
0.06
0.08
-2.3
-2.4
-2.5
-2.6
-2.7

Figure 9

Supplementary material

Three compositional data sets are considered, with *n* rows (cases) and *p* columns (parts).

For each data set the sets of additive log-ratios (ALRs) are computed, using each part in turn as the reference part in the denominator. The set of ALRs that lead to inter-case distances that best match the log-ratio distances, using the Procrustes correlation as the criterion, is identified. Two figures are shown for each example: (a) the two-dimensional configuration obtained using LRA, that takes into consideration all pairwise log-ratios; (b) the two-dimensional configuration obtained using the set of ALRs. The Procrustes correlation between the full-space ($p - 1$ dimensional) configurations is also reported in each case.

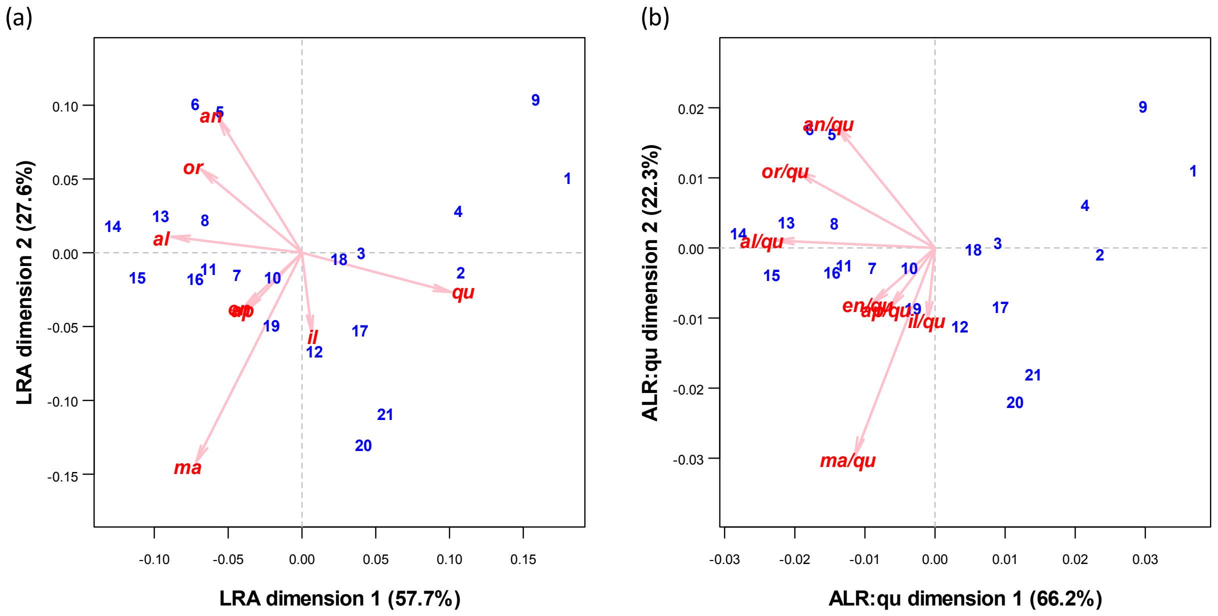
In all these examples, the assertion in Section 3 that a set of ALRs can effectively represent the log-ratio variance is validated, with an almost perfect resemblance between the two-dimensional configurations.

Data set 1 (Aitchison 2005, Table 4.7.1)

Minerals compositions: 21 samples, 8 minerals

qu: quartz ***or***: orthoclase ***al***: albite ***an***: anorthite

en: enstatite ***ma***: magnetite ***il***: ilmenite ***ap***: apatite



Procrustes correlation (between full space configurations) = 0.995

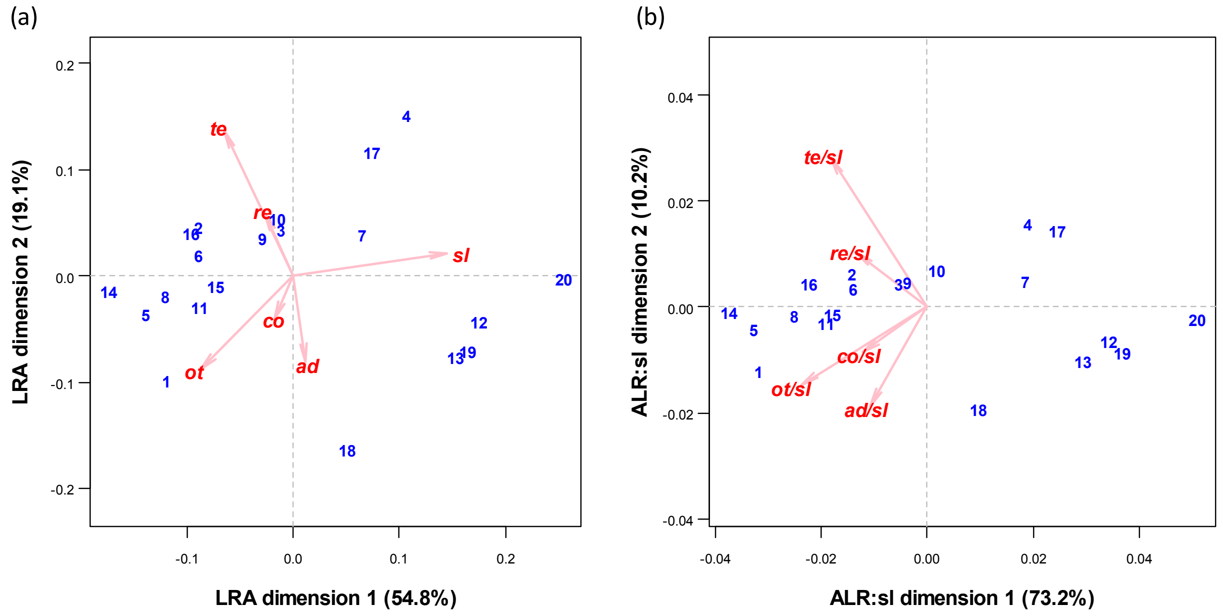
Remark: the relative positions of the case points are practically identical

Data set 2 (Aitchison 2005, Table 1.1.6)

Activity pattern of a statistician: 20 days, 6 activities

te = teaching; ***co*** = consultation; ***ad*** = administration;

re = research; ***ot*** = other wakeful activities; ***sl*** = sleep

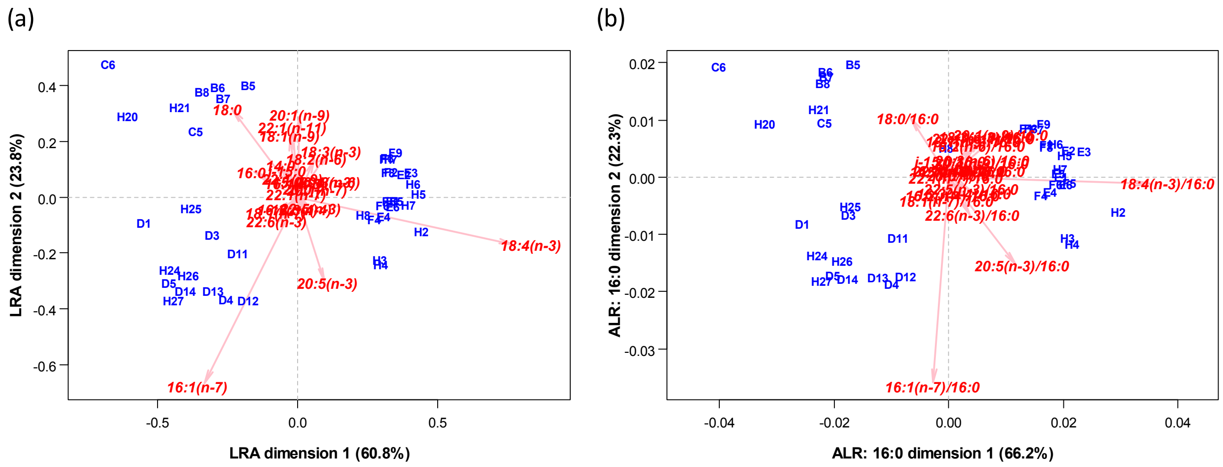


Procrustes correlation (between full space configurations) = 0.960

Remark: the first dimension of the ALR analysis accounts for a much higher percentage of variance, similar to the glass cup example in the main text, suggesting that there is only one relevant dimension and that the LRA analysis is inflated with redundant variance.

Data set 3 (see Greenacre 2016, Appendix E)

Fatty acid data: 42 samples, 25 fatty acids with nonzero values



Procrustes correlation (between full space configurations) = 0.989

Remark: this data set separates groups of marine organisms collected in three different seasons, and the ALR analysis separates the groups just as well as the LRA. The four ratios that stand out in the contribution biplot on the right are made up of the four parts prominently radiating out from the centre in the LRA on the left, expressed relative to the more centrally located fatty acid 16:0.