Math Geosci
https://doi.org/10.1007/s11004-018-9754-x

Variable Selection in Compositional Data Analysis Using Pairwise Logratios

Michael Greenacre[1,2]

Received: 22 September 2017 / Accepted: 5 June 2018

Abstract In the approach to compositional data analysis originated by John Aitchison, a set of linearly independent logratios (i.e., ratios of compositional parts, logarithmically transformed) explains all the variability in a compositional data set. Such a set of ratios can be represented by an acyclic connected graph of all the parts, with edges one less than the number of parts. There are many such candidate sets of ratios, each of which explains 100% of the compositional logratio variance. A simple choice consists in using additive logratios, and it is demonstrated how to identify one set that can serve as a substitute for the original data set in the sense of best approximating the essential multivariate structure. When all pairwise ratios of parts are candidates for selection, a smaller set of ratios can be determined by automatic selection, but preferably assisted by expert knowledge, which explains as much variability as required to reveal the underlying structure of the data. Conventional univariate statistical summary measures as well as multivariate methods can be applied to these ratios. Such a selection of a small set of ratios also implies the choice of a subset of parts, that is, a subcomposition, which explains a maximum percentage of variance. This approach of ratio selection, designed to simplify the task of the practitioner, is illustrated on an archaeometric data set as well as three further data sets in an "Appendix". Comparisons are also made with existing proposals for selecting variables in compositional data analysis.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11004-018-9754-x) contains supplementary material, which is available to authorized users.

✉ Michael Greenacre
michael.greenacre@upf.edu

[1] Department of Economics and Business, Universitat Pompeu Fabra, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain

[2] Barcelona Graduate School of Economics, Ramon Trias Fargas 25-27, 08005 Barcelona, Spain

Published online: 09 July 2018

Keywords Compositional data · Logratio transformation · Logratio analysis · Logratio distance · Multivariate analysis · Ratios · Subcompositional coherence · Univariate statistics · Variable selection

1 Introduction

Compositional data are samples of non-negative multivariate 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 (CODA; Aitchison 1986). The original totals, whatever they were, are generally not of interest—rather, the relative values, collectively called a composition, are relevant for understanding the structure of the data set. The components of a composition are called its parts. If a subset of the parts is considered and the data are re-expressed relative 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 often not of interest. For example, in the case of time budget data where activities such as sleeping, eating, leisure, and work, transport are recorded during a 24-hour day, there is usually no reason to drop an activity and re-express the remaining ones relative to the total without that activity, with the possible exception of removing the sleeping hours in order to study activities during waking hours relative to their respective total. And if additional activities are added, these would generally be subdivisions of existing activities, for example, subdividing the leisure activities.

This article concentrates rather on compositional data where subcompositions or extended compositions naturally occur when there is a large pool of possible parts to choose from in a particular study, for example, geochemical data in geology, or fatty acid data in ecology. In these cases, the proportions depend on the particular choice of parts made by the researcher and this choice can vary between studies.

The act of transforming a set of values into its set of relative values by dividing by their total is called closure (the term normalization is sometimes used). It is exactly because the compositional values associated with the parts change after reclosure (or 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 means and correlation coefficients. These summary statistics make no sense when comparing different studies unless such studies have used exactly the same set of parts. In multivariate analysis of compositional data, it has been recognized, at least since Aitchison (1982), that a valid approach to compositional data is to analyze ratios of parts whose values do not depend on the choice of the set of parts. Basing the analysis on ratios is an approach possessing the property of subcompositional coherence: results and conclusions concerning a subset of parts do not change when some other parts are excluded or added (Aitchison 1992, 2005; van den Boogaart and Tolosana-Delgado 2013). One of the consequences of using ratios is that all the data values are required to be strictly positive, which is what is assumed henceforth.

The logarithmic transformation of the ratios is important, because ratios are compared multiplicatively rather than additively; hence logratios are used.

Logratio analysis (LRA; Aitchison 1990; Aitchison and Greenacre 2002) is a subcompositionally coherent variant of principal component analysis (PCA) that displays the reduced dimensional structure of all logratios of the parts in the form of a biplot. The form of LRA in the abovementioned publications can be qualified as unweighted LRA, where all parts are given the same weight in the analysis, both in the definition of total variance and in the dimension reduction. Subsequently, Lewi (2005), Greenacre and Lewi (2009) and Greenacre (2010a, 2011b) showed the benefits of weighting the parts by their average proportions, which are a reasonable default set of weights in the absence of any additional knowledge about the measurement errors of the parts. Without this weighting, the rare parts can engender large ratios, and can dominate the analysis and overshadow the smaller ratios engendered by the frequent parts. This is exactly the problem with unweighted LRA that is highlighted by Baxter et al. (2005), which led them to propose to down-weight the influence of variables having high relative variation rather than giving them equal weight. This weighted form of LRA is identical to spectral mapping, a method that has long been used in drug development and biostatistical applications (Lewi 1976, 1980, 1989; Wouters et al. 2003). Hron et al. (2017) also introduce weights to account for measurement errors or to adjust the role of parts for substantive reasons.

Another reason for the weighting is that it leads to the analysis respecting the principle of distributional equivalence (Benzécri 1973; Greenacre 2016), which can be enunciated as follows: if two parts occur in the same relative amounts (i.e., part A is a constant multiple of part B), then they can be amalgamated into a single part without affecting the structure of the samples. Expressing this in an opposite way, if one part is arbitrarily split into, for example, ten equal parts (i.e., the part values are divided by ten and the original part replaced by the ten "subparts"), then the unweighted analysis, which is generally used at present, will count the ten repeated subparts ten times their original value in the computation of the total variance, and thus influence all the results. By contrast, the weighted analysis will be unaffected, since the parts are distributionally equivalent. This highly desirable property, thanks to the weighting, is proved by Greenacre and Lewi (2009). In what follows, LRA will by default refer to the weighted form of the analysis, whereas the unweighted form will be qualified specifically as unweighted LRA.

There are many papers on Aitchison's approach to CODA: to mention only a few key publications, Aitchison (1986, 1994), Aitchison et al. (2000), Aitchison and Egozcue (2005) and Pawlowsky-Glahn et al. (2007, 2015), and the multi-authored book edited by Pawlowsky-Glahn and Buccianti (2011). The authors who take inspiration from Aitchison's ground-laying work, insisting that the criterion of subcompositional coherence, for example, be steadfastly followed, will be referred to as the "CODA school". The fundamental idea of this approach is the logratio transformation, and several types have been proposed, examples of which are the additive logratio (ALR), the centered logratio (CLR) and the isometric logratio (ILR), each with advantages and disadvantages (definitions of these are given in Sect. 2). Of all these, only ALRs have a simple interpretation in practice, since they are constructed from ratios of parts and do not involve ratios of geometric means of parts, as is the case with the other transforma-

tions. In this paper it is hoped to show that by slightly relaxing the strict requirements of the CODA school's approach in a controlled and measurable way, then sets of simple ratios of parts can be used, which are easier for the practitioner to interpret, and—for the purpose of interpreting the compositional data set and understanding its structure—serve the same purpose as the more complex ones.

In the same spirit as the present article, a relaxation of the strict requirement of subcompositional coherence 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 CODA school requires methods for analyzing compositional data to obey strictly the principle of subcompositional coherence—for example, Pawlowsky-Glahn et al. (2007) state unequivocally that the condition of subcompositional coherence is fulfilled by any statistical method applied to compositions, and Bacon-Shone (2011) state that subcompositional coherence is one of the principles that characterize CODA. The rigorous adherence to this principle excludes methods that are close to being coherent, for example correspondence analysis (CA), which is theoretically linked to LRA by the Box-Cox transformation $f(x) = (1/\alpha)(x^{\alpha} - 1)$, for $x \geq 0$ and $\alpha > 0$ (Box and Cox 1964). Here the x variables are either the original compositional data values (for unweighted LRA) or the contingency ratios (for weighted LRA), which are essentially the compositional data divided by their respective part means—Greenacre (2009, 2010a) gives details. The chi-square distance in CA (e.g., Greenacre 2016) approximates the logratio distance in LRA, coming closer and closer to it as the power parameter α of the Box-Cox transformation reduces to 0. and in the limit, the two methods are identical. Moreover, CA handles data zeros with ease, which is one of the main problems of the logratio approach. In this sense, CA provides a satisfactory alternative to LRA for identifying dimensions underlying a compositional data set (Greenacre 2011b) and its departure from subcompositional coherence 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 relaxes 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 pragmatic spirit, proposing alternative simpler approaches to CODA that come measurably close to the strict requirements of the CODA school.

In summary, the approach proposed in this paper is to replace the original compositional data by a set of carefully selected ratios, with the following requirements: (1) the ratios are easily interpretable by the practitioner; (2) the set of ratios explain either all or a very high percentage of the variability contained in the original data set and/or come measurably close to the geometry of the original data set and so preserve its essential multivariate structure; and (3) the ratios can be validly reported as univariate statistics, with conventional summary measures, which can be compared with the same ratios in similar studies.

Section 2 on Materials and Methods describes the data used as an application and lays the methodological foundations for what is to come. The fundamental measure of total variability in a compositional data set, the sample logratio variance, is defined,

in both unweighted and weighted forms. Two areas of methodology are essential to support this pragmatic approach, namely network graph representation and the form of multivariate regression called redundancy analysis (RDA). A stepwise process is described for selecting ratios, and how their usefulness in preserving the multivariate data structure can be measured. The practical benefit of part weighting in multivariate analysis of compositional data is again stressed.

In Sect. 3, several results are reported. First, it is demonstrated that a set of ALRs can be measurably close to the complete set of logratios in more than one sense and is thus of potential use to the practitioner tackling a compositional data set. Then, more generally, when all pairwise logratios are considered as candidates, a stepwise process of ratio selection can be implemented, measuring how much logratio variance is preserved in their selection, and showing how close their implied inter-sample distances are to the distances based on all logratios. The performance of the chosen set of simple logratios is compared to that of CLRs and ILRs.

Section 4 concludes with a discussion that highlights the pragmatic role played by simple logratios, and contrasts their use with present CODA practice, which generally involves more complex transformations, admittedly with interesting theoretical properties, but lacking an easy interpretation for the practitioner of CODA.

An "Appendix" provides supplementary material: first, the application of this approach to three additional compositional data sets further justifies the use of simple logratios; second, some theoretical material on Procrustes analysis; third, a comparison with a recently published approach on variable selection by Martín-Fernández et al. (2018); and finally, a simulation study.

2 Materials and Methods

2.1 Data

To illustrate the proposed pragmatic approach to CODA, an archaeometric data set due to Baxter et al. (1990) will be used throughout. This consists of the compositions of $m = 11$ oxides in a set of $n = 47$ Roman glass cups from an archaeological site in eastern England, where oxygen is combined with elements silicon (Si), aluminum (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) and are also provided electronically 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 avoids the side issue of data zeros.

2.2 Total Variability of a Compositional Data Set

The total variability in a compositional data set, following Aitchison (1983, 1986), is measured by the (sample) logratio variance. This measure is so fundamental that several equivalent definitions will be presented to highlight different properties. Sup-

pose that the data are in a samples-by-parts matrix $\mathbf{X}$ ($n \times m$), where the rows of $\mathbf{X}$ 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) logratio variance, following Aitchison (1986), consists of first defining the logarithms of the ratios of all ½$m(m-1)$ pairs of parts, that is, expanding the columns of $\mathbf{X}$ into an $n \times$ ½$m(m-1)$ matrix of logratios $\mathbf{Z}$ and then computing the grand total of the column variances, noting that the variances will always be defined as the sum of squared deviations from the mean divided by n, not by $(n-1)$

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

where the notation $\sum_{j<j'}$ indicates the double summation over all ½$m(m-1)$ unique pairs of the index. Specifically, Aitchison defines the square $m \times m$ variation matrix $\mathbf{T}$ with elements $\tau_{jj'}$ = the variance of the (j, j')th logratio, and, referring to the half triangle of this matrix, like the summation $\sum_{j<j'}$ above, claims that the sum of all the ½$m(m-1)$ logratio variances must provide some measure of total variability for the composition. Notice further that Pawlowsky-Glahn et al. (2007) change this definition by dividing (1) by m—in fact, they sum over all variances in the variation matrix, thus counting all variances twice, and then divide by $2m$.

The quantity in Eq. (1) can be equivalently defined in terms of all the pairwise-squared differences between the logratios, as follows

$$\begin{aligned}
\text{TotVar (Aitchison)} &= \frac{1}{n^2} \sum_{i<i'} \sum_{j<j'} \left(z_{i,jj'} - z_{i',jj'}\right)^2 \\
&= \frac{1}{n^2} \sum_{i<i'} \sum_{j<j'} \left(\log \frac{x_{ij}}{x_{ij'}} - \log \frac{x_{i'j}}{x_{i'j'}}\right)^2 \qquad (2) \\
&= \frac{1}{n^2} \sum_{i<i'} \sum_{j<j'} \left(\log \frac{x_{ij}}{x_{ij'}} \frac{x_{i'j'}}{x_{i'j}}\right)^2. \qquad (3)
\end{aligned}$$

Notice that Eq. (2) includes all the possible sets of ALRs. A set of ALRs is the set of logratios with respect to a specified part k. Denoting such a set by ALR:k, its set of values for the ith sample consists of the $(m-1)$ values

$$\text{ALR:}k(i) = \log\left(x_{ij}/x_{ik}\right), \quad j = 1, \ldots m, \quad j \neq k. \tag{4}$$

The equivalent formulation in Eq. (3) shows the sum of squares of the logarithmically transformed cross-product ratios based on all unique pairs of rows and pairs of columns of the data matrix.

As shown by Greenacre and Lewi (2009), weighting the parts has many important advantages, apart from adhering to the principle of distributional equivalence, described in Sect. 1. Parts occurring in low proportions need to be down-weighted because they induce large components of logratio variance and will dominate any

analysis of the complete set of logratios. The rare oxide of the element Mn in the present glass cups data set, which occurs only with values 0.01, 0.02 and 0.03%, illustrates this argument perfectly (Baxter et al. 2005). In the absence of knowledge about measurement error of the different parts, a reasonable 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 CA and spectral mapping (Greenacre 2016; Wouters et al. 2003). 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. The column weights are $c_1, c_2, \ldots, c_m$, where $c_j =$ jth part mean, also summing to 1. Thus, parts with low means, which generally have high variance in their logratios, will be down-weighted in the analysis. For other tables of positive data all on the same scale, and thus suitable for analysis using logratios, for example, counts, the row weights can also vary, or different row weights might be prescribed for some substantive reason, for example, to correct for sampling bias. The definitions of weighted logratio variance corresponding to the unweighted ones in Eqs. (1) and (2), additionally including the varying column weights and possibly varying row weights too, are thus respectively as follows

$$\text{TotVar} = \sum_{j<j'} c_j c_{j'} \sum_i r_i \left(z_{i,jj'} - \bar{z}_{jj'}\right)^2 \quad \text{where } z_{i,jj'} = \log \frac{x_{ij}}{x_{ij'}} \text{ and } \bar{z}_{jj'} = \sum_i r_i z_{i,jj'} \tag{5}$$

$$\begin{aligned} \text{TotVar} &= \sum_{i<i'} r_i r_{i'} \sum_{j<j'} c_j c_{j'} \left(z_{i,jj'} - z_{i',jj'}\right)^2 = \sum_{i<i'} \sum_{j<j'} r_i r_{i'} c_j c_{j'} \left(\log \frac{x_{ij}}{x_{ij'}} - \log \frac{x_{i'j}}{x_{i'j'}}\right)^2 \\ &= \sum_{i<i'} \sum_{j<j'} r_i r_{i'} c_j c_{j'} \left(\log \frac{x_{ij}}{x_{ij'}} \frac{x_{i'j'}}{x_{i'j}}\right)^2. \end{aligned} \tag{6}$$

The weighted versions in Eqs. (5) and (6), where Eq. (6) represents a totally symmetric formulation with respect to rows and columns, will be maintained as the definition of the total logratio variance. This total variance will be qualified with the adjective "unweighted" when $c_j = 1/m$ for all j. When, in addition, $r_i = 1/n$ for all i, the definition is identical to Aitchison's in Eq. (1) divided by m^2, or the one by Pawlowsky-Glahn et al. (2007) divided by m. Notice that both Aitchison's definition as well as that of Pawlowsky-Glahn et al. increase quadratically and linearly, respectively, with the number of parts, whereas Eq. (6) is an averaging over all elements in the compositional data set, and can be compared across matrices of different sizes.

The (weighted) logratio distance between two samples i and i' appears in the formulation in Eq. (6)

$$d_{ii'} = \sqrt{\sum_{j<j'} c_j c_{j'} \left(\log \frac{x_{ij}}{x_{ij'}} - \log \frac{x_{i'j}}{x_{i'j'}}\right)^2} = \sqrt{\sum_{j<j'} c_j c_{j'} \left(\log \frac{x_{ij}}{x_{ij'}} \frac{x_{i'j'}}{x_{i'j}}\right)^2}, \tag{7}$$

so that the logratio variance can also be written as the weighted sum of squares of all the inter-sample distances

$$\text{TotVar} = \sum_{i<i'} r_i r_{i'} d_{ii'}^2. \tag{8}$$

A third equivalent definition of the logratio variance uses the matrix $\mathbf{Y}$ of centered logratios (CLRs; Aitchison 1986): $y_{ij} = \log(x_{ij}/\prod_j x_{ij}^{c_j}) = \log(x_{ij}) - \sum_j c_j \log(x_{ij})$, that is, the rows of the log-transformed data matrix are centered with respect to their respective weighted row means. Then the total variance is the weighted average of the variances of the m columns of $\mathbf{Y}$

$$\text{TotVar} = \sum_{j=1}^{m} c_j \sum_{i=1}^{n} r_i (y_{ij} - \bar{y}_j)^2, \quad \text{where } \bar{y}_j = \sum_{i=1}^{n} r_i y_{ij} \text{ and usually } r_i = \frac{1}{n}. \tag{9}$$

Since the log-transformed data matrix $\mathbf{X}$, $\log(\mathbf{X}) = [\log(x_{ij})]$, is first centered row-wise by the respective weighted row means to obtain the matrix $\mathbf{Y}$, it follows that the elements $y_{ij} - \bar{y}_j$ in Eq. (9), which are centered by the column means, are a double-centering of the matrix $\log(\mathbf{X})$ using the column and row weights c_j and r_i, respectively. Then these double-centered values are squared and summed using the column and row weights again to obtain the total logratio variance. In matrix notation, where the column and row weights are gathered in vectors $\mathbf{c}$ and $\mathbf{r}$ respectively, the CLR matrix is

$$\mathbf{Y} = \log(\mathbf{X})\,(\mathbf{I}_m - \mathbf{1}_m \mathbf{c}^{\mathrm{T}})^{\mathrm{T}}, \tag{10}$$

where $\mathbf{I}_m$ is the $m \times m$ identity matrix and $\mathbf{1}_m$ is the m vector of ones. Then the double-centered matrix is the log-transformed matrix $\log(\mathbf{X})$ pre- and post-multiplied by the respective centering matrices

$$\mathbf{S} = \mathbf{Y} - \mathbf{1}_n \mathbf{r}^{\mathrm{T}} \mathbf{Y} = (\mathbf{I}_n - \mathbf{1}_n \mathbf{r}^{\mathrm{T}}) \log(\mathbf{X})\,(\mathbf{I}_m - \mathbf{1}_m \mathbf{c}^{\mathrm{T}})^{\mathrm{T}}. \tag{11}$$

The logratio variance in Eqs. (5), (6), (8) and (9) is the weighted sum of squares of the elements s_{ij} of $\mathbf{S}$

$$\text{TotVar} = \sum_i \sum_j r_i c_j s_{ij}^2 = \text{trace}(\mathbf{D}_r \mathbf{S} \mathbf{D}_c \mathbf{S}^{\mathrm{T}}). \tag{12}$$

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

A fourth equivalent definition of the logratio variance can be obtained by defining a set of ILRs known as balances, which are a recursive partitioning of the parts that can be represented in a binary tree (dendrogram), with each balance defining a log contrast in a subset of parts, that is, the log of the ratio of two geometric means. For example, a set of balances, called pivot balances (Hron et al. 2017) can be defined as

proportional to ratios of individual parts to geometric means of the remaining parts, in a chosen ordering, in its unweighted form as follows

$$C \log\left(\frac{x_{ij}}{\left(\prod_{k=j+1}^{m} x_{ij}\right)^{1/(m-j)}}\right) = C\left(\log(x_{ij}) - \frac{1}{m-j}\sum_{k=j+1}^{m}\log(x_{ij})\right) \quad \text{for } j = 1,\ldots,m-1, \tag{13}$$

where the constant $C = \sqrt{\frac{m-j}{m-j+1}}$.

Pivot balances can be drawn as a dendrogram where the first part is split off against the rest, then among the rest, the first part is again split off against the rest, and so on in a nested fashion. But any dendrogram of the parts can define a set of balances. One way of proving the total variance decomposition across a set of balances is to resort to the familiar decomposition of total sum of squares in Ward clustering, discriminant analysis and multivariate analysis of variance (Krzanowski 2000), namely that the total sum of squares (TSS) is decomposed into two components, the between-groups sum of squares (BSS) and within-groups sum of squares (WSS): that is, 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 WSS is in turn decomposed according to subsequent binary splits and each time a BSS is computed. Eventually, once the partitioning arrives at two single parts, which have zero WSS, the sequence of BSS has decomposed the total variance. The number of possible sets of balances, which define labelled, non-ranked dendrograms of m parts, is equal to $(2m-2)!/(2^{m-1}(m-1)!)$—Murtagh 1984; Bóna 2006—which for the archaeometric data set is a very large number: $20!/(2^{10} \times 10!) = 654{,}729{,}075$, over 650 million.

2.3 Graph Representation of Logratios

It will be useful to represent logratios of parts as connections between the parts in a network. In graph theory, the parts are called vertices (or nodes) and the connections edges. For example, the CODA school's approach analyses all ½$m(m-1)$ pairwise logratios, and these can be represented in the graph of Fig. 1a, called a complete graph because every pair of vertices is connected by an edge. This is also a directed graph, where the edges are arrows pointing from the denominator of the ratio to the numerator. In graph theory, such a complete directed graph is called a tournament, the analogy being in sport (e.g., football) where every team is in a game against every other team (Harary and Palmer 1973; Bóna 2006).

A set of $m-1$ ALRs is displayed in Fig. 1b where the oxide of Si (silicon dioxide, or silica) is the denominator. When sets or subsets of ratios are dealt with in later sections, the representation of these in a graph will enhance understanding, and certain graph-theoretic results will be useful.

An acyclic connected graph (also called a spanning tree) is one that connects all m parts, but contains no cycle. These two concepts can be defined formally as: (1) a connected graph is one where all vertices are reachable by edges; (2) an acyclic

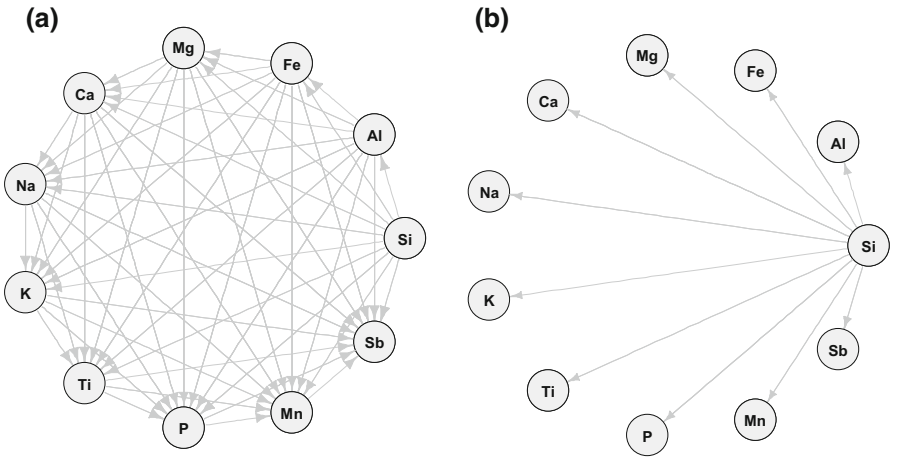

Fig. 1 Graph representations of networks for 11 vertices defined by **a** all 55 pairwise logratios, **b** the set of 10 additive logratios (ALRs) with Si as the denominator, indicated by the arrows emanating from Si to the others

graph is one without closed circuits, that is, when following the edges of the graph from one vertex to any other vertex, no vertices can be visited twice. It can be proved by contradiction that such a graph has exactly $(m-1)$ edges—if it has less edges, it cannot connect all parts, and if it has more edges then there must be a cycle. Exactly $(m-1)$ ratios are defined by an acyclic connected graph, and none of them can be obtained from the others, that is, their logratios form a linearly independent set. For example, sets of ALRs such as the one in Fig. 1b form acyclic connected graphs. From graph theory, the number of acyclic connected graphs of m elements is known—it is the Cayley number, equal to m^{m-2} (Bóna 2006), which, in the archaeometric data set, is another huge number: $11^9 = 2{,}357{,}947{,}691$, over 2 billion (note that the edges are not ranked in this counting, just like the nodes of the dendrograms were not ranked in determining the number of balances). Hence, if a set of ratios is to be chosen, an efficient procedure is required.

2.4 Redundancy Analysis and Procrustes Analysis

RDA is a variant of multivariate regression, where there are m 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 PCA of instrumental variables. Gittins (1985) gives a thorough treatment and equates the term redundancy with explained variance, which is exactly the use made of it here. In its simplest form, given a $n \times m$ matrix of responses $\mathbf{Z}$ and an $n \times p$ matrix of explanatory variables $\mathbf{W}$ (which serve to explain variance in each of the columns of $\mathbf{Z}$), dimension reduction is performed not on $\mathbf{Z}$ itself but rather on its projection $\mathbf{Z}^*$ onto the space defined by $\mathbf{W}$: $\mathbf{Z}^* = \mathbf{W}\,(\mathbf{W}^{\mathrm{T}}\mathbf{W})^{-1}\mathbf{W}^{\mathrm{T}}\mathbf{Z}$. 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

residual part in **Z**–**Z***, which is uncorrelated with **W** and thus unexplained by the explanatory variables. The **vegan** package (Oksanen et al. 2015) 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 **Z** of responses, or on the square distance matrix between samples that is implicit in the analysis of **Z**. In the present application, RDA will be used to measure how much logratio variance is explained by a subset of logratios used as explanatory variables to explain all the logratios. Notice the distinction between "contained variance" and "explained variance": the former refers to the numerical amount of variance contributed to the total variance by a logratio, whereas the latter refers to the proportion of variance the logratio explains in the regression sense. A logratio will have a generally low contained variance, since it is one of many logratios contributing additively to the total variance, whereas its explained variance can be high when it is correlated with many of the other logratios.

The **vegan** package also includes Procrustes analysis, in two functions, **procrustes()** and **protest()**. Procrustes analysis (Gower and Dijksterhuis 2004) is used to measure the difference between two multidimensional configurations, matching one to the other by rotation, translation and rescaling—in other words, it measures how close their multivariate structures are. In the present application, Procrustes analysis will be used to measure how close a configuration based on a subset of logratios is to the configuration based on all the logratios. A convenient measure of the matching of two configurations is provided by the Procrustes correlation (Legendre and Legendre 2012). The matrix formulation of Procrustes analysis and the definition of the Procrustes correlation is given in the "Appendix".

2.5 Selection of Logratios

The goal is to obtain a smaller, simpler set of parts or part ratios that satisfactorily represents the whole compositional data set by explaining a large percentage of its variance and preserving its multivariate data structure as closely as possible. Variable selection in the context of interval-scale data in PCA has already been studied. For example, Krzanowski (1987) made the point that dimension reduction can reduce a data set to a smaller number of components, but all the original variables need to be retained. His proposal was to reduce the number of variables using Procrustes rotation to match the multidimensional configuration produced by a subset of variables to that of the full set. The lowest Procrustes loss, equivalent to the highest Procrustes correlation, would then indicate the best subset of variables that preserves the multivariate data structure. Dijksterhuis et al. (2002) applied this approach to a 15-variable data set in food research, enumerating all $2^{15} - 2 = 32{,}766$ proper subsets and evaluating the Procrustes loss for each one. These authors additionally come to the important conclusion that the subsets corresponding to a range of lowest loss values should be considered instead of the single best subset producing the lowest loss value. Thus, the practitioner can intervene to choose a subset that might be slightly sub-optimal but more meaningful substantively.

Variable selection in the compositional data context can proceed similarly, but there are 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 such as highest logratio variance explained, and/or highest Procrustes correlation, after which the data are reclosed and analysis continues in the logratio framework using this subcomposition. If the second path is chosen, then a subset of logratios is found in the same way and these are treated afterwards just like regular variables, or this subset can suggest the subcomposition that contains the selected ratios. If a subset of parts are selected, these will imply using all the pairwise logratios amongst them. For reasons of parsimony, selection of simple logratios will be preferred here, and it will be demonstrated that a set of logratios can preserve the essential multivariate structure without resorting to the approach based on CLRs and ILRs that involve all parts and all logratios. Simple ratios are compatible across studies, are easily interpretable and their choice can also be guided by the practitioner who has expert knowledge about the data and its context. Whichever way the 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, firstly, in terms of how much the chosen ratios explain the logratio variance, using RDA, and secondly, by how close the distances based on the chosen ratios approximate the logratio distances, using Procrustes analysis.

An automatic way of identifying a good set of ratios can proceed in a stepwise fashion, trying in the first step every logratio as an explanatory variable in explaining the logratio, and selecting the one with the highest percentage of variance explained. This ratio is then fixed as the first logratio and then the second best logratio 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: $\mathrm{A/C} = \mathrm{A/B} \times \mathrm{B/C}$ (in other words, a cycle linking A, B and C is created in the graph). 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 an m-part compositional data set is $m - 1$, and all the parts will have appeared in at least one logratio after $m - 1$ steps of the above procedure, the variance explained will be 100%.

Dependence might arise via several ratios of parts. Here the graph representation can be very enlightening; for example, Fig. 2a represents five independent ratios by solid arrows from the denominator to the numerator in each. The two ratios Si/Mn and Ca/P, depicted by dashed arrows in Fig. 2b, are dependent on the others since they close a circuit. For example, the ratio Ca/P 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 (it is additive/subtractive in the logratios). Thus, $\mathrm{Ca/P} = \mathrm{Ti/P} \times \mathrm{Si/Ti} \times (\mathrm{Si/Ca})^{-1}$. The only way to add independent ratios to this particular network (i.e., linearly independent logratios) would thus be to add new parts. In the case that all 11 oxides are in the network but only 10 ratios define edges, this implies that the graph must necessarily be connected and acyclic. Algorithmically, dependence is easy to detect in the choice of ratios and does not present a problem, since a ratio

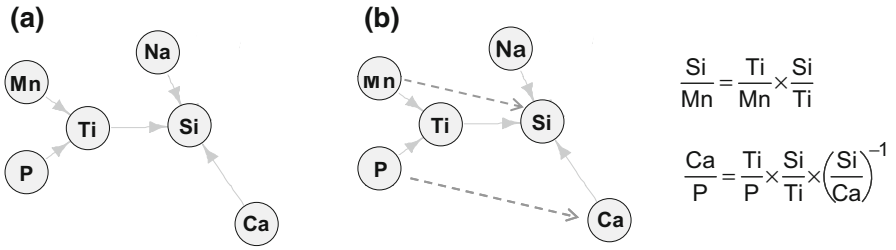

Fig. 2 Illustration of acyclic graphs and dependent ratios: **a** An acyclic connected graph of six parts and five ratios, where the arrows point towards the respective numerators of the ratios; **b** Two additional ratios, Si/Mn and Ca/P, that are dependent on the existing ones because they form cycles

that is dependent on previously selected ones will not add any additional explained variance.

3 Results

3.1 Additive Logratios

The simple and particular case of the ratios created by an ALR transformation in Eq. (4) is considered first. In an m-part compositional data set, there are m potential sets of ALRs, each of which is depicted graphically by arrows emanating from one of the parts to all the other $m-1$ parts (the set of ALRs with Si as denominator is shown in Fig. 1b). Each set of ALRs was used as explanatory variables in a RDA to measure how well they explain the logratio data set and, as expected, each set explains 100% of the total logratio variance. But some have higher Procrustes matching to the full logratio configuration—see Table 1, where the elements are ordered according to the Procrustes correlation. The set of ALRs using Si has the highest Procrustes correlation, equivalent to very low Procrustes loss, and this very high correlation suggests that these ALRs constitute a set of ratios that can adequately represent the complete data set, as will be demonstrated now.

Using this best set of ALRs with respect to Si, denoted by ALR:Si, and comparing the logratio distances in (7) with the interpoint distances based on this set of ALRs, the scatterplot in Fig. 3a is obtained. Each of these sets of distances is computed in a 10-dimensional space, the space of all 55 logratios and the space of the 10 ALRs, respectively. The optimal two-dimensional biplots representations of these respective analyses are given in Fig. 4a, b, respectively. The configurations of samples appear very similar, and the only noticeable difference is that the percentage of variance explained on the first axis is much higher in the ALR analysis of Fig. 4b, because the total variance of the ALRs is lower (this difference in percentages will be referred to again in Sect. 3.4). Figure 3b compares the inter-sample distances in these respective two-dimensional solutions, and Fig. 3c does the same when using only the first dimension. 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 becomes even higher, so extremely little of the important

Table 1 Results for ALRs using each part in turn as the reference one in the denominator

Part	R^2	Weight	Procrustes correlation
Si	1	0.7237	0.9975
Mg	1	0.0046	0.9539
Al	1	0.0194	0.9043
Na	1	0.1825	0.8439
Sb	1	0.0036	0.7689
K	1	0.0040	0.7017
Ca	1	0.0567	0.6989
Fe	1	0.0031	0.6704
Ti	1	0.0007	0.6564
P	1	0.0005	0.6187
Mn	1	0.0001	0.5692

R^2 is the part of total logratio variance explained; weight is the average proportion of the reference part, used in the weighted analysis; Procrustes is the Procrustes correlation that measures similarity between the multidimensional geometry of the samples in the ALR space and that of the samples using all logratios (in both cases the space is ten-dimensional)

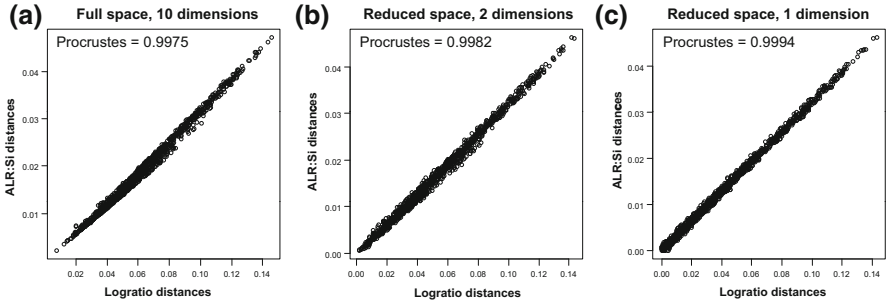

Fig. 3 Comparison of logratio distances with distances based on ALRs with respect to Si; **a** the full ten-dimensional space; **b** reduced two-dimensional space; and **c** reduced one-dimensional space. Notice that the scales are different on the two axes, since the ALR distances are shorter than those based on the full set of logratios

logratio distance variance contained in the major LRA dimensions is lost by using this set of ALRs. In fact, later in Sect. 3.4, when the issue of signal and noise is treated, it will be determined that there is most likely only one dimension in this data set that can be considered significant, in the sense that it is incompatible with random variation.

All in all, for all practical purposes, the ALR analysis of 10 ratios appears to be equivalent to the full-blooded LRA of all 55 ratios, and has a simpler interpretation, constructed from a small set of logratios of parts. In this example, the set of ALR:Si appeared the most useful, but similar results were found using other sets of ALRs too, such as ALR:Mg and ALR:Al, which also have high Procrustes correlations (Table 1). In the next section, results will demonstrate that even fewer than $m - 1$ logratios can

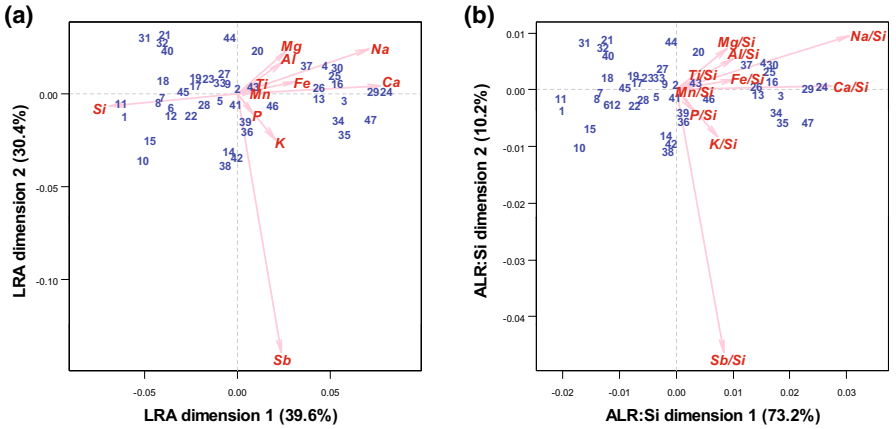

Fig. 4 **a** Logratio analysis (LRA) of the full data set (contribution biplot); **b** Principal component analysis (PCA) of the additive logratios (ALRs) with respect to Si

adequately represent the full set of logratios, and thus simplify the practitioner's task even further.

As an interesting side remark, it is noteworthy that the set of ALRs that best represents the original data is when the oxide of Si (i.e., SiO_2) is used in the denominator of the ratios (Fig. 1b). This recalls one of the oldest methods for identifying relationships between oxides in a geological context, the variation diagram or Harker diagram (Harker 1909), which plots oxides of elements against SiO_2—see https://brocku.ca/earthsciences/people/gfinn/petrology/variatn.htm and Cortés (2009).

3.2 Stepwise Ratio Selection

The stepwise procedure starts by selecting, from the $\frac{1}{2} \times 11 \times 10 = 55$ logratios in this example, the one that explains the most logratio variance, using RDA. The logratio 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 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. 5 represents the set of ratios in its acyclic graph connecting all the parts, where the numbers on 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, respectively). These statistics may be validly compared with the same ratios in other archaeometric 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

Table 2 Sequence of logratios of mineral oxides entering in a stepwise search, explaining the logratio variance of the whole compositional data set

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

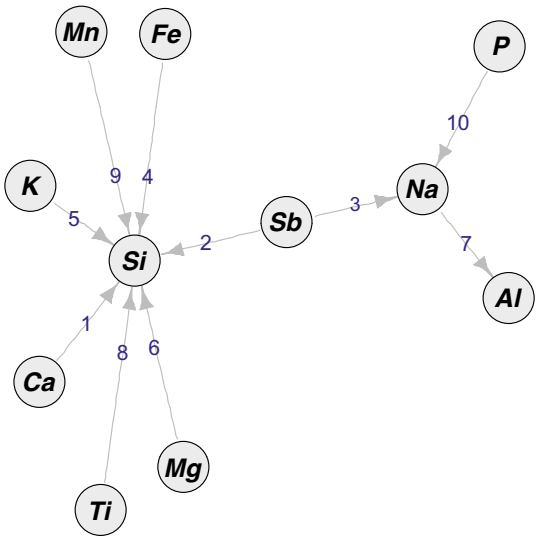

Fig. 5 Graph of the ten ratios chosen in a stepwise procedure to explain maximum logratio variance at each step, with numbers indicating their order of selection

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 been 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 the present implementation of the algorithm, these ties were broken by choosing the ratio that, when added to the list of ratios at that step, maximized the improvement in the Procrustes correlation. 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.

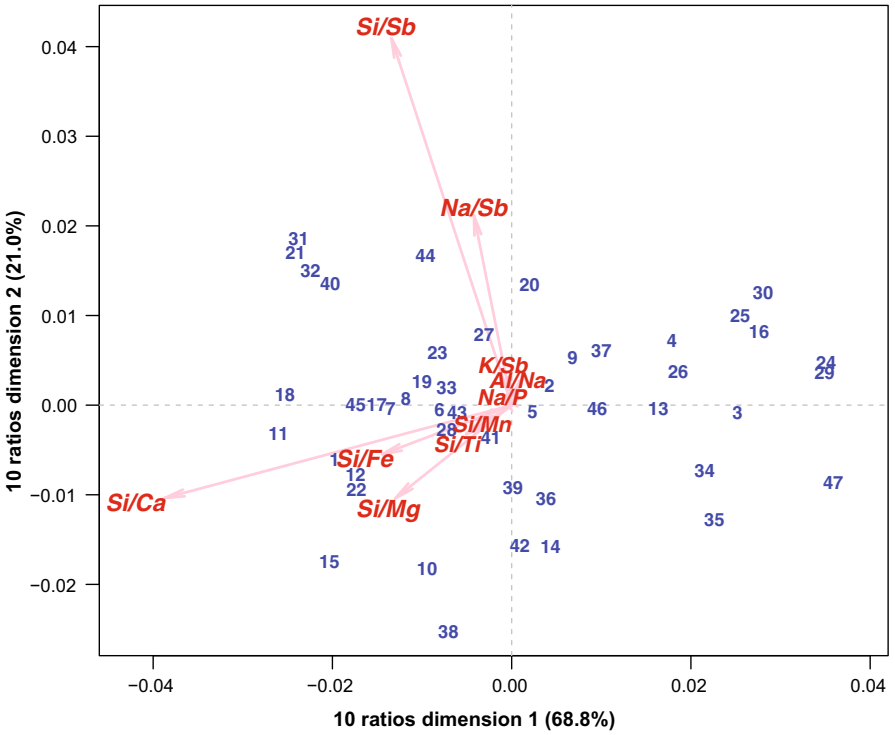

Fig. 6 PCA contribution biplot of the ten ratios shown in Fig. 5

The logratio biplot of all parts in Fig. 4a sheds light on the choice of the ratios. This incorporates the weighting of the parts, as before, and the contribution biplot scaling (Greenacre 2013), which defines each part's coordinate on a dimension as the square root of the part's proportional contribution to that dimension's explained variance. The Si versus Ca opposition appears as 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 the second ratio chosen, to include Sb which is the most important contributor on the second axis and thus defining ratios that have low correlation with the first ratio.

Now, using the 10 identified logratios in Table 2, which explain 100% of the logratio variance, Fig. 6 shows the PCA contribution biplot—notice that the definitions of the ratios involving Si are inverted compared with Fig. 4b, with Si appearing in the numerator. The resemblance with the configuration of samples in Fig. 4a is again apparent, and once again (as seen in the ALR results of Fig. 4b) the percentage of variance explained on the first axis is much higher—the first dimension of Fig. 7 explains 68.8% of the variance, whereas in Fig. 4a, it has practically the same interpretation and explains only 39.6%. The Procrustes correlation for the 47 glass cups between the two-dimensional configurations of Figs. 4a and Fig. 6 is 0.966.

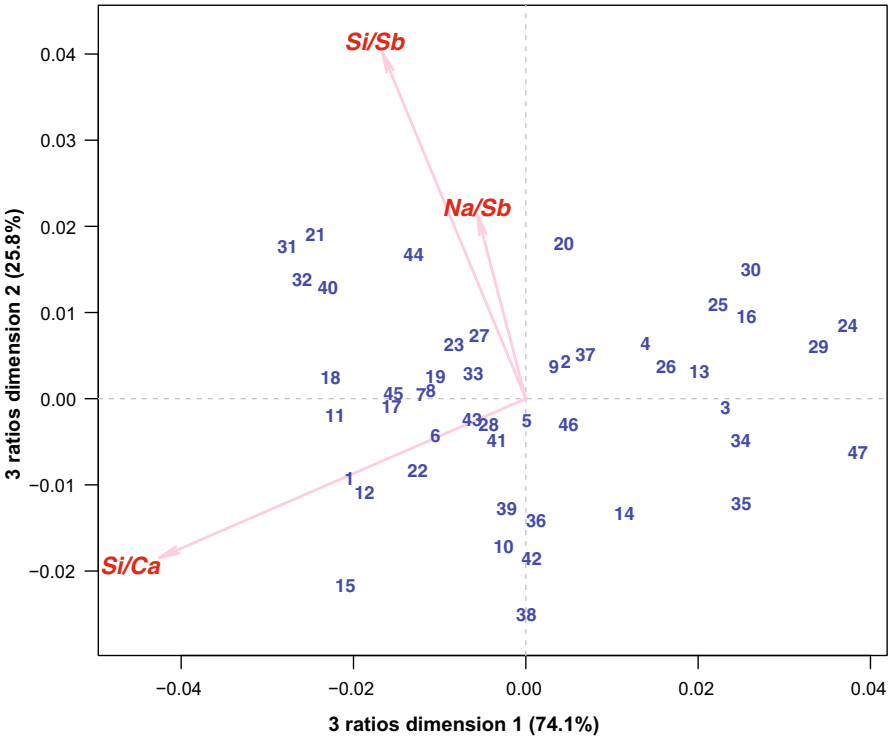

Fig. 7 PCA contribution biplot of the first three ratios Si/Ca, Si/Sb and Na/Sb found in the stepwise ratio-selection procedure

3.3 Choosing a Reduced Set of Ratios

In the contribution biplot of Fig. 6, 3 ratios stand out from the rest: Si/Ca, Si/Sb and Na/Sb, exactly the first 3 ratios that entered the stepwise process, explaining 86.4% of the logratio variance. Figure 7 shows the biplot using just these three ratios, and it hardly differs from Fig. 6. 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 accurately, with a Procrustes correlation of 0.950 between the samples in Fig. 7 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 labels of the other seven oxides, with a Procrustes correlation of 0.988.

3.4 Signal Versus Noise: Significant Dimensionality

In some of the previous analyses, reduced dimensional solutions have also been compared. This is important because, after all, these solutions are what will be interpreted by the practitioner, assuming that the minor dimensions reflect the noise variance. The present data have a relatively low variance, and the question can be asked how many dimensions of the data set can be judged to be statistically significant, that is, not a result of random variation. Applying a permutation test in the style of Greenacre (2016) to the LRA solution, where the null hypothesis is that the compositions are all equal to the mean composition, the first dimension was found to be highly significant ($p < 0.001$), but for all the other dimensions, the p values were high and non-significant. The test statistics are the eigenvalues of the dimensions. The conclusion is thus drawn that the data are not a simple random variation around the mean, but that there is one dimension in these data that is statistically justifiable. Therefore, one-dimensional solutions should be compared, not two-dimensional ones or even full-dimensional ones (ten-dimensional in this case). Figure 8 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 an alternative visual way to interpret 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. 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 logratios that are included in the analysis. This situation is no different from augmenting a data set in a PCA 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.

If one accepts that it is really only the first dimension of the LRA which is worth interpreting, and that the remaining dimensions are compatible with random variation, then a single logratio of Si/Ca accounts for this dimension quite accurately. This implies that the whole data set can perhaps 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 only informative parts, ordinating the glass cups in terms of log(Si/Ca), whereas the other nine oxides represent non-informative random information, appearing as additional columns and serving only to inflate the total variance.

3.5 Balances

The option of choosing balances is now compared with the choice of simple ratios. Balances were defined in two ways: first, as a stepwise divisive (descending) procedure where at each step one part was contrasted against the remainder at that step (see

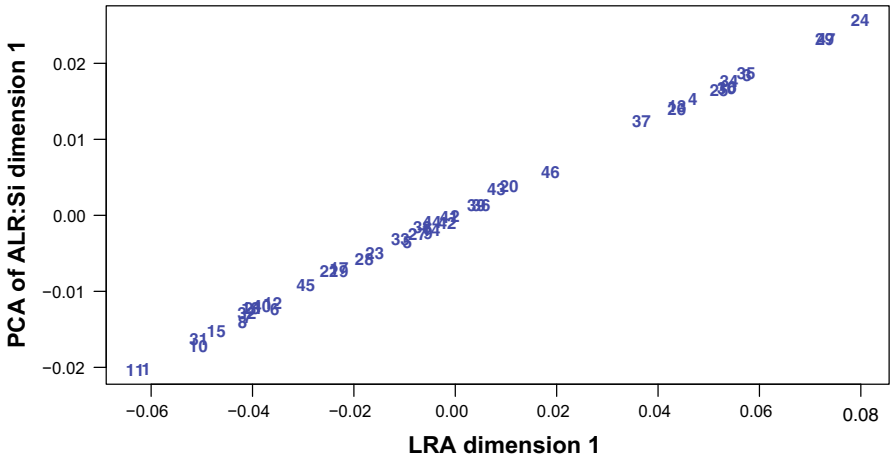

Fig. 8 Almost perfect agreement of coordinates of 47 glass cups on first axis of LRA (i.e., their coordinates on the first axis of Fig. 4a) and first axis of PCA of ALRs with respect to Si (i.e., their coordinates on the first axis of Fig. 4b)—correlation is 0.999. Notice again that the ALR coordinates are much smaller than those of the LRA

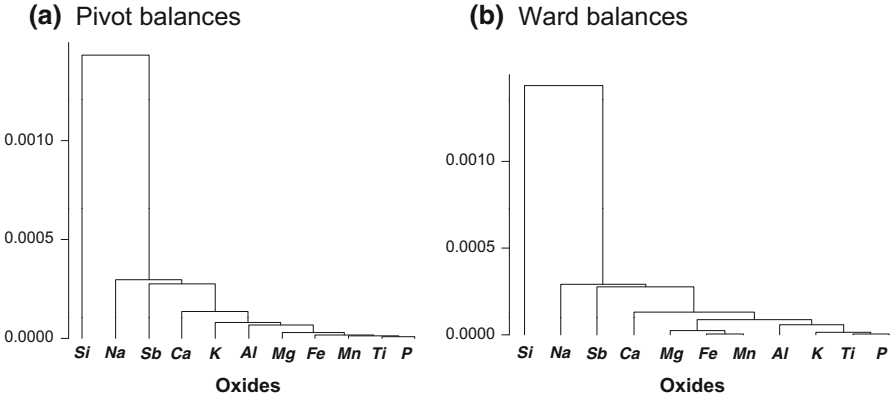

Fig. 9 Dendrograms associated with two balances that define isometric logratios (ILRs) for the glass cup data: **a** a chain balance, **b** the Ward balance

Eq. 13), and the optimal part in the numerator is chosen at each node using the same criterion of explained logratio variance—that is, the optimal pivot balances (Fig. 9a); and second, an agglomerative (ascending) procedure by weighted Ward clustering (Fig. 9b), in which the Ward balances are optimal in their stepwise ascending sense. In both cases, the sum of the heights of the ten nodes is exactly the total logratio variance, equal to 0.002339 for the glass cup data, and hence define a decomposition of the total variance.

In Table 3, the cumulative percentages of logratio variance explained by the balances defined at the nodes when moving down each tree are respectively listed. In both cases, the sequence of simple logratios in Table 2 gives cumulative percentages greater than

Table 3 Sequence of optimal pivot balances and ward balances, and associated cumulative percentages of logratio variance explained

Pivot balances		Ward balances	
Numerator part	Cumulative R^2 (%)	Definition	Cumulative R^2 (%)
Si	61.3	Same as pivot	61.3
Na	73.9	Same as pivot	73.9
Sb	85.7	Same as pivot	85.7
Ca	91.3	Same as pivot	91.3
K	94.6	Mg, Fe, Mn / Al, K, Ti, P	95.1
Al	97.5	Al / K, Ti, P	97.6
Mg	98.6	Mg / Fe, Mn	98.8
Fe	99.3	K / Ti, P	99.4
Mn	99.7	Fe / Mn	99.7
Ti/P	100.0	Ti / P	100.0

For pivot balances, the oxide in the numerator is listed. The first four Ward balances are identical to the corresponding pivot balances, after which the oxides in the numerator and denominator are listed

those based on the balances. Thus, at least in this example, a simple sequence of well-chosen logratios, involving just two parts at a time, can be superior to balances which involve all the parts, starting as early as the first balance.

To demonstrate the stability of the logratio log(Si/Ca) compared to the ILR balances, the following simulation exercise was performed. The other nine oxides (excluding Si and Ca) were ordered according to their contributions to total variance, from the lowest to the highest: Mn, P, Ti, K, Al, Mg, Fe, Sb and Na. In this order, the values in their respective columns were successively randomized one at a time, the data set reclosed, and then: (1) the logratio that best accounts for the total variance was recorded; and (2) the Ward hierarchical clustering was performed. The results are given in the "Appendix" and can be summarized as follows. For every successive randomization, the logratio log(Si/Ca) remained the most important one, even when the values of all nine other elements were randomly permuted. The Ward clusters remained fairly stable until the fourth element was randomized, and from then the clusters changed substantially as further elements were randomized. This demonstrates that any attempt to define ILR balances that involve the full set of parts is highly affected by the presence of random parts in the data, whereas the search for an important non-random logratio remains stable and unchanging. Since the present data set, with low total variance, almost surely contains parts compatible with random variation, this puts a question mark on any ILR balance defined on the full set of parts.

A recent paper by Martín-Fernández et al. (2018) proposes an "advance" on the selection of principal balances, by undertaking a divisive clustering algorithm with an exhaustive search, applied on a ten-part geochemical data set, the Aar Massif data from the book by Van den Boogaart and Tolosana-Delgado (2013). This algorithm is very time-consuming compared to the present logratio selection approach, and the

authors claim it is feasible for up to 15-part compositions. Mert, Filzmoser and Hron (2015) give a sub-optimal algorithm to determine ILR balances that can be used for data sets with many parts. These authors all try to find a sequential binary partitioning of the parts, but their criterion is not explained variance in the sense used in this article, but rather "contained variance", that is, the part of variance contributed by the ILRs to the total (see the previous explanation of this distinction in Sect. 2.4). This is a weaker criterion than explained variance, since it measures the part of variance of a balance in total isolation of the other balances. Similarly, a logratio has a certain part of contained variance, one of the many additive contributions to total variance, but may explain a high proportion of total variance because of its correlations with other logratios. If a logratio, or an ILR, contributes the highest part of contained variance to the total variance, this does not necessarily imply that it explains the highest percentage of variance in a regression sense.

Using the same ten-part Aar Massif geochemical data set, the variances explained by the ILR balances in Martín-Fernández et al. (2018) were computed (these authors did not compute these explained variances, but rather the part contributions to variance, that is, the contained variances as described above). Then the present approach of stepwise search for simple logratios was performed, along with their explained variances, and this much faster algorithm, with its easily understandable solution, compared very favorably with the ILRs, and even sometimes slightly outperformed them. For example, the optimal principal balance, defined as the ratio of the geometric mean of 4 parts to the geometric mean of 5 parts (thus involving 9 of the total of 10 parts) had an explained variance of 70.7%, whereas the optimal simple pairwise logratio (involving only 2 parts) had an explained variance of 69.1%. Logratios of amalgamations are also investigated and shown to perform almost as well as the ILR balances. The full results are given in the "Appendix".

As far as speed and scalability of the present algorithm is concerned, on a regular laptop, the execution time for finding the complete list of $J-1$ optimal logratios was 1.3 s for $J=10$ (a 10-part composition), 14 s for $J=20$, 41.8 s mins for $J=30$ and 2.09 min for $J=40$. Since the usual objective would be to look for a shorter list of the best logratios, these times would be proportionally reduced: for example, to find the best 10 logratios in a 100-part composition, the execution time was 2.94 min. When the stepwise process is conducted so that at each step the competing set of top variance-explaining logratios is listed, one of which is selected by the practitioner based on substantive knowledge, then the time for a single step becomes relevant. For example, for a large 200-part composition, the time taken to find the best logratios in the first step was 51.9 s, which means that the interactive process could proceed with each step taking less than a minute.

3.6 Summary of Results

What has been learnt about the glass cup data, using the logratio approach advocated here, can be summarized as follows.

First, the ALRs with respect to Si explain 100% of the logratio variance and reproduce the logratio distances very accurately. Second, using a stepwise selection process

from the pool of logratios, a set of logratios can be found that similarly explains 100% of the logratio variance and also has a very high Procrustes correlation with the complete logratio structure. Third, a small set of logratios can adequately represent the data in the sense of conserving its multivariate structure; for example, to reproduce faithfully the two-dimensional solution obtained by LRA of the full 11-part data set, only 3 logratios are needed. Fourth, based on a permutation test, it is highly likely that there is only one statistically significant dimension in this compositional data set, the remainder being compatible with random variation. Fifth, the logratio of Si/Ca coincides with this single dimension and the conclusion can be that all the other parts are just contributing random variation to the data. Sixth, because the smaller subsets of ratios contain less of the noise variance, their component analyses show first dimensions with much higher percentages of variance explained; for example, the analysis of the 10 ALRs has a first dimension that appears much stronger than that of the LRA, which analyses 55 logratios, even though both analyses are taking place in 10-dimensional spaces.

4 Conclusions

The main objective of this article is to demonstrate that a selection of simple ratios, logarithmically transformed, can account for all or most of the logratio variance in a compositional data set. These ratios can preserve the essential multivariate structure of the data, and can thus be used for univariate or multivariate analysis as a substitute for the original data. In the example presented here, one set of ALRs has essentially served the same purpose as all the pairwise logratios. Three additional examples in the "Appendix" further support the use of ALRs and serve as counter-examples to the assertion that ALRs are not worth considering.

Pawlowsky-Glahn et al. (2007) dismiss the ALR transformation, saying that it is frequently used in many applied sciences but should be avoided. The ALR transformation was the original transformation used by Aitchison (1986), but it is not used in the book by van den Boogaart and Tolosana-Delgado (2013), where it is stated that ALR deforms distances, angles and shapes. Other authors from the CODA school similarly reject this transformation, for at least the following two reasons (Pawlowsky-Glahn et al. 2007): first, that it is not symmetrical in its components, and second, that it defines coordinates in an oblique basis, which affects distances if the usual Euclidean distance is computed from ALR coordinates.

The critique that ALRs are not symmetrical in their components is difficult to understand, and no reason why components need to be symmetrical has been offered—it has rather been demonstrated here that a well-chosen set of ALRs can satisfactorily substitute the original data set, explain the totality of the logratio variance and have the practical advantage of being easy to interpret.

The second criticism of the ALR transformation is that it will affect the logratio distances and deform the space of the samples, that is, it will poorly represent the inter-sample distances, but again those who express this offer no measurement of this deformation. Since the sum in the definition of Eq. (7) of logratio distance is over $\frac{1}{2}m(m-1)$ logratios, whereas the ALR distances (i.e., weighted Euclidean distances

between the ALR-transformed sample points) involve summing over only $m-1$ of them, it follows that all the inter-sample distances will be changed. But the removal of all the redundant logratios linearly dependent on the set of ALRs does not necessarily affect the multivariate structure—the application in Sect. 3.1 as well as those in the "Appendix" show that there is very little measurable deformation in the samples' relative positions in the space of the ALRs.

Hence, a conclusion of this paper is that ALRs should not be rejected, but rather explored for their effectiveness in representing the compositional data content. These results recall the recommendation of Aitchison (1994) who advocated the simplicity of the ALR transformation, followed by the application of the appropriate, standard multivariate procedures to the ALR-transformed vectors.

It is clear that when a reduced set of logratios is selected, either ALRs or a set of stepwise-selected logratios, only a submatrix of the complete covariance structure is used in the analysis. But the fact that the inter-sample distances are approximately preserved (up to an overall scaling constant) means that a PCA of the reduced data set will be minimally affected and closely resemble that of all pairwise logratios, as has been demonstrated in the examples presented here.

Univariate analysis of the logratios, or of 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 otherwise comprehensive book on CODA, edited by Pawlowsky-Glahn and Buccianti (2011), contains almost no mention of univariate analysis, except a passing reference by Lovell et al. (2011) to a paper by Filzmoser et al. (2009), who use the ILR transformation to arrive at a set of $m-1$ variables that replace the original data set. The claimed advantage of ILRs is that they are based on orthonormal contrasts and reconstruct exactly the logratio distances. This is an elegant mathematical property and has some benefit when a complete set of ILR balances is defined and used as an alternative set of coordinates for the whole data set. However, the lack of these ideal properties in simple logratios does not impede their use in serving practically the same purpose, giving a close approximation to these distances. The interpretability of individual ILRs remains a problem, because each ILR involves many parts (potentially all), making it virtually impossible to interpret (van den Boogaart and Tolosana-Delgado 2013). In a paper read at the CODA workshop in 2003, Aitchison alluded to the complications posed by ILRs, saying that we should not let pure mathematical ideas drive us into making statistical modelling more complicated than necessary (Aitchison 2003). 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), and are also linearly dependent, which can be problematic when doing statistical analyses (van den Boogaart and Tolosana-Delgado 2013).

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 heterogeneous in their composition, particularly in their ratios of soda (Na_2O)-to-potash (K_2O) and lime (CaO)-to-magnesium (MgO). If required for substantive reasons, these ratios can

be forced into the first two steps of the present, 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 logratios, of Na/K and Ca/Mg, explain only 16.6% of the logratio variance. The automatic selection that follows immediately brings in the logratio 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.

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—for example, Kraft et al. (2015). The reporting of a selected set of ratios, their distributions and confidence intervals on means, for example, should be the norm rather than the exception, since these quantities are comparable across studies. CLRs and ILRs are not convenient variables for the biochemist to interpret, as opposed to the simplicity of logratios. Biochemists do, however, favor aggregations of parts, for example summing together all the poly-unsaturated or saturated fatty acids, and making ratios of these sums: $\sum$poly-unsaturated FAs/$\sum$saturated FAs—Kraft et al. (2015) report such a ratio. Similarly, in geochemistry, interest may be on elements that behave together, for instance, combining alkilies such as K_2O and Na_2O.

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 logratio variance, the between-group logratio variance would serve as the variance to be explained by the selected ratios. The best ratio would then constitute the logratio that explains the most between-group variance, and so on. Similarly, in the simpler case of a multiple regression situation, with a single response variable, a stepwise search on all pairwise ratios as explanatory variables can be conducted in the same way.

Another novel aspect of this paper is the representation of the chosen ratios as a directed graph (see Figs. 1, 2 and 5). This visualization can be further enhanced by relating the areas of vertex circles to the respective means, for example, of the parts, and the edge lengths and/or widths to results related to the ratios, for example, additional variance explained or increment in the Procrustes correlation.

Hron et al. (2013) present another way of selecting subcompositions in CODA based on the part contributions to variance made by the CLRs. They first remove the CLR with the lowest contribution to variance (i.e., lowest contained variance), form the subcomposition without it, recompute the CLRs and carry on in this stepwise removal of parts, defining a test statistic for stopping the procedure. The problem with this strategy, like the approach of Martín-Fernández et al. (2018), is that intercorrelations are not taken into account, and the CLR with the lowest contribution to total variance is not necessarily the one that explains the least variance in the data set.

For the present archaeological data set, the table of part contributions to total variance is published by Greenacre and Lewi (2009) for both unweighted and weighted variances. This shows that the unweighted contributions place element Mn as the highest-contributing part—this is the part with only three values of 0.03, 0.02 and 0.01%, engendering large logratios and the highest logratio variance of all 11 parts. This element would be retained by the approach of Hron et al. (2013). By contrast, the

weighted contributions place Mn as the least-contributing part, and would be eliminated—quite correctly, in this case—by the stepwise procedure.

While the approach presented here is based on choosing logratios that explain maximum variance, the chosen set of logratios also implies the choice of a subcomposition, formed by the parts that are present in the logratios. Alternatively, a forward stepwise procedure reminiscent of Hron et al.'s (2013) backward algorithm described above is to start by choosing the logratio explaining the most variance, as before, which defines a two-part composition. Then a third part is added which increases the explained variance the most, and so on, until a subcomposition is obtained that reproduces the multivariate structure very closely, according to Procrustes analysis. This selection of a subcomposition can be used to reduce the number of parts before applying the exhaustive search procedure of Martín-Fernández et al. (2018).

Apart from the three applications briefly summarized in the "Appendix", one specific data set has been examined in detail throughout this article, with the caveat that the results might not always turn out as successfully as in the applications considered here. Nevertheless, these examples stand to counteract the strict and more complex requirements of the CODA school, for example, that variables defined on a compositional data set have to be orthonormal contrasts, and to promote this simpler approximative approach, which is reminiscent of the "close to subcompositional coherence" idea of Greenacre (2011a). This simpler and more parsimonious approach should 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 approaches provided so far in the CODA literature.

Once a set of logratios has been selected that represents the multivariate structure of the compositional data set satisfactorily, any of the well-known multivariate statistical procedures can be applied, such as PCA (as shown here), but also discriminant analysis, RDA (when additional explanatory variables are available), canonical correlation analysis (when correlational relationships with other continuous variables are to be explored), or canonical CA (when the compositional data are regarded as predictors of a set of categorical responses). The logratios can also be used in formal modeling, either as responses or explanatory variables, with the assurance that they are easy to understand and to interpret.

As a final remark, the logratio-selection approach proposed in this paper, for compositional data, applies equally well to any positive ratio-scale multivariate data, as long as all data values are in the same units of measurement, for example, counts, weights (e.g., in grams) or measurements (e.g., in centimeters). Applications of this approach to more general types of data can be found in Lewi (1989; applied to biological activity spectra, whence the name "spectral map" was coined), Wouters et al. (2003; gene expression data), Greenacre and Lewi (2009; counts in linguistics) and Greenacre (2010b; morphometric data on fish).

Acknowledgements This work is dedicated to the memory of John Aitchison who passed away in December 2016 and whom I met when he gave a seminar in Girona, Catalonia, in 2000. He started his talk with a slide containing a single blank triangle, following which, it was like the scales fell from my eyes.

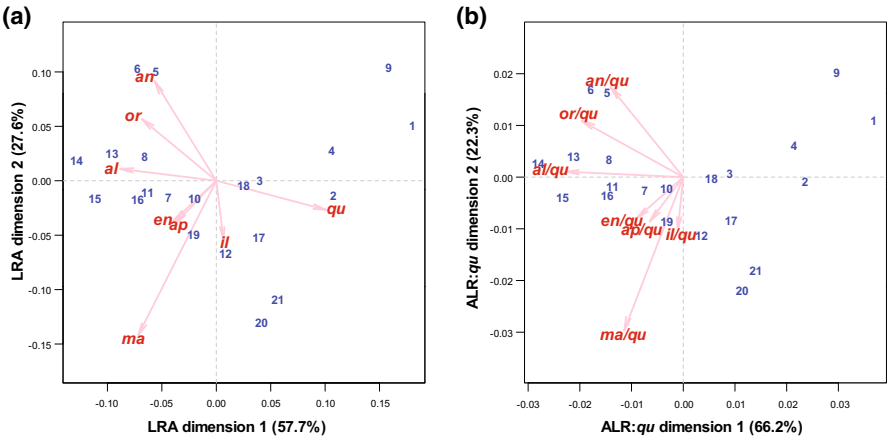

Fig. 10 Two-dimensional biplots of mineral compositions: **a** LRA biplot; **b** PCA biplot of optimal set of ALRs

Appendix

A.1 Three Additional Data Sets

Three more data sets are analyzed, to demonstrate the benefit of using ALRs as a substitute for the full compositional data set. Two of these compositional data sets are taken from Aitchison (2005) and the third one is considered by Greenacre (2016) in the context of CA. For each data set, the sets of ALRs are computed, using each part in turn as the reference in the denominator. The set of ALRs that lead to inter-case distances that best match the logratio distances, using the Procrustes correlation as the criterion, is identified.

Data Set 1 (Aitchison 2005)
Minerals compositions: 21 samples, 8 minerals
qu: Quartz ***or***: orthoclase ***al***: albite ***an***: anorthite
en: Enstatite ***ma***: magnetite ***il***: ilmenite ***ap***: apatite

The ALRs with respect to quartz (qu) give the best agreement—the Procrustes correlation (between full space configurations) is equal to 0.995. Figure 10 shows the two-dimensional LRA based on all 28 logratios alongside the PCA of the 7 ALRs, showing the almost identical configurations of sample points.

Data Set 2 (Aitchison 2005)
Activity pattern of a statistician: 20 days, 6 activities
te = Teaching; ***co*** = consultation; ***ad*** = administration;
re = Research; ***ot*** = other wakeful activities; ***sl*** = sleep

The ALRs with respect to sleep (sl) give the best agreement—the Procrustes correlation (between full space configurations) is equal to 0.960. Figure 11 shows the two-dimensional LRA based on all 15 logratios alongside the PCA of the 5 ALRs, showing the highly similar configurations of sample points. The first dimension of the

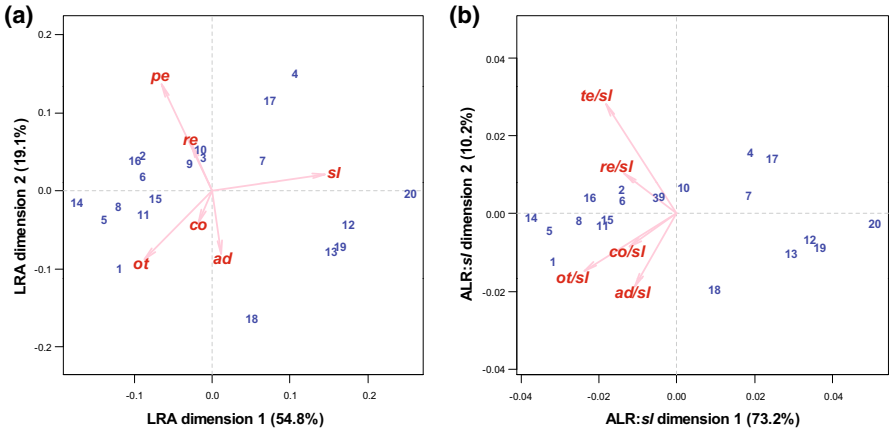

Fig. 11 Two-dimensional biplots of activity patterns of statisticians: **a** LRA biplot; **b** PCA biplot of optimal set of ALRs

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

This data set consists of groups of marine organisms collected in three different seasons. The ALRs with respect to fatty acid 16:0 give the best agreement to the multivariate structure—the Procrustes correlation (between full space configurations) is equal to 0.989. Figure 12 shows the two-dimensional LRA based on all 300 logratios alongside the PCA of the 24 ALRs, showing the similar groupings of the three seasonal subsets of data, separated by the ALR analysis just as well as by 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 (Fig. 12).

A.2 Procrustes Analysis and Procrustes Correlation

The following matrix formulation summarizes the computations required:

Suppose $\mathbf{F}_1$ ($n_1 \times p$) and $\mathbf{F}_2$ ($n_2 \times p$) are two matrices of coordinates defining two configurations of the same labelled points in separate p-dimensional spaces. Both matrices are column-centered (i.e., column means are zero). Then the following steps lead to the Procrustes correlation.

1. Normalize both matrices: $\mathbf{F}_1^* = \mathbf{F}_1/\sqrt{\text{trace}(\mathbf{F}_1^{\text{T}}\mathbf{F}_1)}$, $\mathbf{F}_2^* = \mathbf{F}_2/\sqrt{\text{trace}(\mathbf{F}_2^{\text{T}}\mathbf{F}_2)}$
2. Compute cross-product matrix: $\mathbf{S} = \mathbf{F}_1^{*\text{T}}\mathbf{F}_2^*$
3. Perform singular value decomposition (SVD): $\mathbf{S} = \mathbf{U}\mathbf{D}_\alpha\mathbf{V}^{\text{T}}$
4. Procrustes rotation matrix: $\mathbf{Q} = \mathbf{V}\mathbf{U}^{\text{T}}$

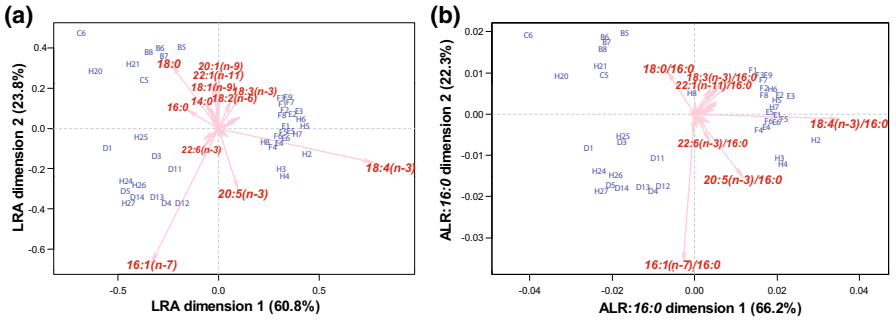

Fig. 12 Two-dimensional biplots of fatty acids: **a** LRA biplot; **b** PCA biplot of optimal set of ALRs. In respective biplots, the labels of fatty acids and fatty acid ratios close to the center (i.e., with low contributions to the solution) have been omitted to improve legibility

5. Sum of squared errors between normalized coordinates after rotation of the second matrix:

$$E = \text{trace}[(\mathbf{F}_1^* - \mathbf{F}_2^*\mathbf{Q})^{\mathrm{T}}(\mathbf{F}_1^* - \mathbf{F}_2^*\mathbf{Q})]$$

6. Procrustes correlation: $r = \sqrt{1 - E}$

A.3 Comparison of the Present Logratio Approach with the Principal Balances of Martín-Fernández et al. (2018)

Martín-Fernández et al. (2018) developed an algorithm for a stepwise selection of ILR balances, by successively partitioning the parts using an exhaustive search at each step of this divisive algorithm. They apply their method to the ten-part Aar Massif geochemical data set from the book by Van den Boogart and Tolosana-Delgado (2013), and their approach uses unweighted parts, which is the present practice of the CODA school. A major difference between their approach and the one in the present article is they do not use variance explained in the sense used here, but rather "variance contained" in, or "variance contributed" to the logratio variance (although they sometimes do use the term "variance explained", but they mean "variance contained"). This is a weaker criterion than the variance explained one that is proposed in the present study, because a part of variance contributed by a logratio or a balance is a measure in isolation from the remainder of the variability in the rest of the data set (see Sect. 3.5 of the article for more explanation). Thus, in order to compare our results with those of Martín-Fernández et al. (2018), the explained variances have had to be computed for the sequence of ILR balances published in that paper (Table 4, columns 3 and 4). In addition, the simpler approach of selecting logratios proposed in the present study was executed (Table 4, columns 1 and 2, see Fig. 13 for a graph of these ratios). As a yet further comparison, the simple logratios of amalgamated parts, using the same partitioning sequence as the ILR balances, were also computed and their explained variances computed—these can be termed "amalgamation balances"

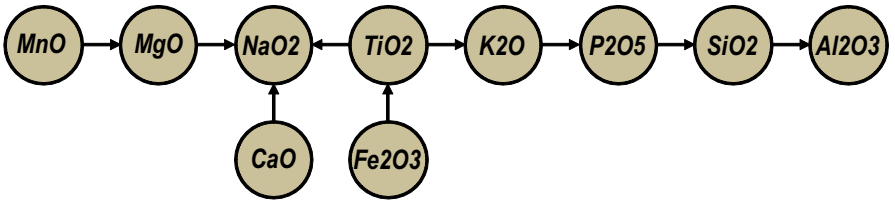

Fig. 13 Graph of set of logratios in first column of Table 4

Table 4 Cumulative explained variances of sequences of simple logratios, ILR balances, amalgamation balances and principal components

Simple logratios[a]		ILR balances[b]		Amalgamation balances[c]	Principal components[d]	
Ratios	CumVar	PB	CumVar	CumVar	Dimension	CumVar
Na_2O/MgO	0.6906	O1	0.7067	0.6901	1	0.7122
P_2O_5/K_2O	0.8932	O6	0.8940	0.8741	2	0.9027
SiO_2/P_2O_5	0.9338	O3	0.9317	0.9292	3	0.9455
Na_2O/TiO_2	0.9656	O2	0.9507	0.9510	4	0.9721
K_2O/TiO_2	0.9866	O8	0.9757	0.9729	5	0.9901
Na_2O/CaO	0.9931	O7	0.9944	0.9938	6	0.9968
MgO/MnO	0.9979	O5	0.9982	0.9966	7	0.9985
Al_2O_3/SiO_2	0.9993	O4	0.9994	0.9989	8	0.9995
TiO_2/Fe_2O_3	1.0000	O9	1.0000	0.9997	9	1.0000

[a]Logratios using the ratio-selecting method of this article, with cumulative explained variances
[b]ILR balances published by Martín-Fernández et al. (2018) using their labeling of the balances (PB = principal balance) and cumulative explained variances [these have been computed, since they were not published by Martín-Fernández et al. (2018)]
[c]Amalgamation balances using the same partitions of the parts as for the ILR balances, but using sums in the ratios, not geometric means, with cumulative explained variances
[d]Optimal principal components (i.e., dimensions of logratio analysis (LRA) or PCA of centered logratios, with cumulative explained variances)

(Table 4, fifth column). Finally, the variances explained by the principal component axes (i.e., dimensions of the unweighted LRA of the data), which are the optimal explained variances, are reproduced (Table 4, columns 6 and 7). Note that these last explained variances are the only ones where the definition of variance explained is equivalent to variance contained.

The results are also presented graphically in Fig. 14 in the style of Table 3 of Martín-Fernández et al. (2018). The results have been graphed in two separate figures for clarity. In both, the PCA sequence of cumulative explained variances is shown to give common reference points. In the left-hand figure, the first ILR, involving nine out of the ten parts, is higher by 1.5 percentage points compared to the first logratio Na_2O/MgO, involving only two parts. At steps 3, 4 and 5, the simple logratio sequence is superior to the ILR sequence, after which the two sequences converge. In the right-hand figure, the ILR balance sequence is superior to the amalgamation

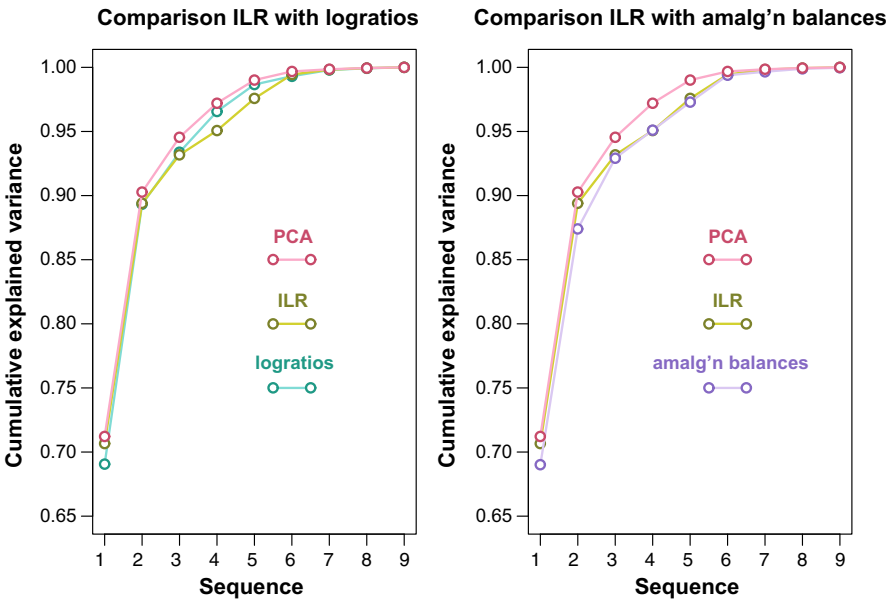

Fig. 14 Plots of cumulative variances in Table 4. The optimal values, obtained by PCA, are shown in both plots as a reference for comparison

balance sequence for the first two steps, but afterwards, they are practically identical. Notice that the amalgamation balance sequence does not necessarily reach exactly 100% variance explained, but in this example, it reaches 99.97% variance explained using 9 balances, lacking only 0.03%.

In conclusion, this is another example where the sequence of simple logratios seems perfectly adequate to explain the variance of the whole compositional data set. They are comparable to the ILR sequence in terms of explained variance, sometimes even outperforming it, and are much easier to compute and interpret. Using amalgamations instead of geometric means is an alternative way of defining balances, and these also have an easier interpretation in practice.

A.4 Simulation Study of the Ward Dendrogram as Parts are Sequentially Randomized

The idea of this simulation is to study how the dendrogram from the Ward clustering breaks down as parts are sequentially randomized (i.e., columns are randomly permuted) to simulate growing random noise in the data set. The values of each part (i.e., oxide in the Roman glass cups data set) are permuted in turn, the data reclosed and the Ward clustering repeated. The order of the parts randomized is from the part with the least part of variance to that of the highest part (the parts being randomized are shown in the boxes next to the dendrograms). Figure 15 is read in horizontal steps, and after three parts are randomized, the structure is still fairly stable, but starts to

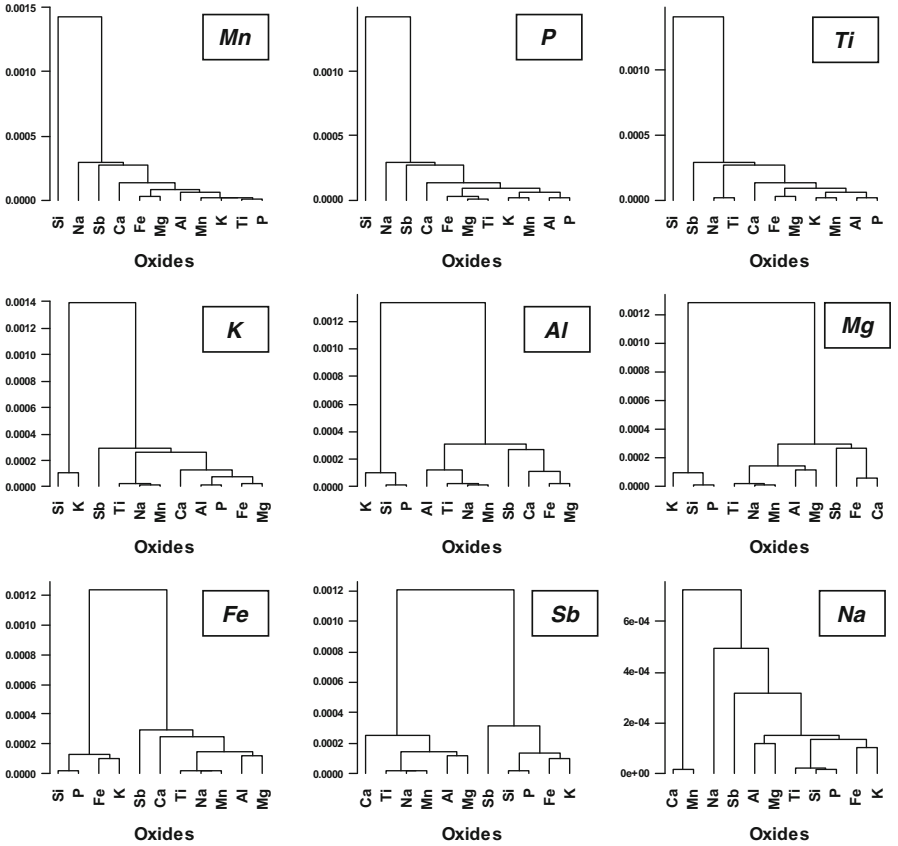

Fig. 15 Sequence of weighted Ward clusterings of the glass cup data as elements shown in boxes are successively randomized in the compositional data matrix

break down from the fourth part being randomized onwards. The element Si is kept fixed throughout, but by the last randomization, the whole data set has been effectively converted to noise.

References

Aitchison J (1982) The statistical analysis of compositional data (with discussion). J R Stat Soc B 44:139–177
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 (1992) On criteria for measures of compositional difference. Math Geol 24:365–379
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, pp 73–81

Aitchison J (2003) Compositional data analysis: where are we and where should we be heading? In: Proceedings of the compositional data analysis workshop, CoDaWork'03, Girona, Spain. CD-format, ISBN 84-8458-111-X
Aitchison J (2005) A concise guide to compositional data analysis. http://ima.udg.edu/Activitats/CoDaWork05/A_concise_guide_to_compositional_data_analysis.pdf. Accessed 29 May 2018
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
Aitchison J, Greenacre MJ (2002) Biplots for compositional data. J R Stat Soc Ser C (Appl Stat) 51:375–392
Aitchison J, Barceló-Vidal C, Martín-Fernández JA, Pawlowsky-Glahn V (2000) Logratio analysis and compositional distance. Math Geol 32:271–275
Bacon-Shone J (2011) A short history of compositional data analysis. In: Pawlowsky V, Buccianti A (eds) Compositional data analysis: theory and applications. Wiley, Chichester, pp 3–11
Baxter MJ, Cool HEM, Heyworth MP (1990) Principal component and correspondence analysis of compositional data: some similarities. J Appl Stat 17:229–235
Baxter MJ, Beardah CC, Cool HEM, Jackson CM (2005) Compositional data analysis of some alkaline glasses. Math Geol 37:183–196
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, 2nd edn. World Scientific Publishing, Singapore
Box GEP, Cox DR (1964) An analysis of transformations. J Roy Stat Soc Ser B 26:211–252
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
Dijksterhuis G, Frøst MB, Byrne DV (2002) Selection of a subset of variables: minimisation of Procrustes loss between a subset and the full set. Food Qual Prefer 13:89–97
Filzmoser P, Hron K, Reimann C (2009) Univariate statistical analysis of environmental (compositional) data: problems and possibilities. Sci Total Environ 407:6100–6108
Gittins R (1985) Canonical analysis: a review with applications in ecology. Springer, New York
Gower JC, Dijksterhuis GB (2004) Procrustes problems. Oxford University Press, Oxford
Greenacre MJ (2009) Power transformations in correspondence analysis. Comput Stat Data Anal 53:3107–3116
Greenacre MJ (2010a) Logratio analysis is a limiting case of correspondence analysis. Math Geosci 42:129–134
Greenacre MJ (2010b) Biplots in practice. BBVA Foundation, Bilbao. www.multivariatestatistics.org. Accessed 29 May 2018
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: theory and applications. Wiley, Chichester, pp 104–113
Greenacre MJ (2013) Contribution biplots. J Comput Graph Stat 22:107–122
Greenacre MJ (2016) Correspondence analysis in practice, 3rd edn. Chapman & Hall/CRC, Boca Raton
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
Hron K, Filzmoser P, Donevska S, Fišerová E (2013) Covariance-based variable selection for compositional data. Math Geosci 45:487–498
Hron K, Filzmoser P, de Caritat P, Fišerová E, Gardlo A (2017) Weighted pivot coordinates for compositional data and their application to geochemical mapping. Math Geosci 49:777–796
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
Krzanowski WJ (1987) Selection of variables to preserve multivariate data structure, using principal components. Appl Stat 36:22–33
Krzanowski WJ (2000) Principles of multivariate analysis: a user's perspective. Oxford University Press, Oxford
Legendre P, Legendre L (2012) Numerical ecology, 3rd edn. Elsevier, Amsterdam
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
Lewi PJ (1989) Spectral map analysis. Factorial analysis of contrasts, especially from log ratios. Chemometr Intell Lab 5:105–116
Lewi PJ (2005) Spectral mapping, a personal and historical account of an adventure in multivariate data analysis. Chemometr Intell Lab 77:215–223
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: theory and applications. Wiley, Chichester UK, pp 193–207
Martín-Fernández JA, Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2018) Advances in principal balances for compositional data. Math Geosci 50:273–298
Mert MC, Filzmoser P, Hron K (2015) Sparse principal balances. Stat Model 15:159–174
Murtagh F (1984) Counting dendrograms: a survey. Discrete Appl Math 7:191–199
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. Accessed 11 June 2018
Pawlowski-Glahn V, Buccianti A (eds) (2011) Compositional data analysis. Wiley, Chichester
Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2007) Lecture notes on compositional data analysis. http://dugi-doc.udg.edu/bitstream/handle/10256/297/CoDa-book.pdf?sequence=1. Accessed 11 June 2018
Pawlowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) Modeling and analysis of compositional data. Wiley, Chichester
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
R core team (2015) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/
van den Boogaart KG, Tolosana-Delgado R (2013) Analyzing compositional data with R. Springer, 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