ANNUAL REVIEWS

Annual Review of Statistics and Its Application

Compositional Data Analysis

Michael Greenacre

Department of Economics and Business, Universitat Pompeu Fabra, and Barcelona Graduate School of Economics, Barcelona 08005, Spain; email: michael.greenacre@upf.edu

Annu. Rev. Stat. Appl. 2021. 8:271–99

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

https://doi.org/10.1146/annurev-statistics-042720-124436

Keywords

amalgamation, coherence, composition, correspondence analysis, count data, logratios, regression, redundancy analysis, subcomposition

Abstract

Compositional data are nonnegative data carrying relative, rather than absolute, information—these are often data with a constant-sum constraint on the sample values, for example, proportions or percentages summing to 1% or 100%, respectively. Ratios between components of a composition are important since they are unaffected by the particular set of components chosen. Logarithms of ratios (logratios) are the fundamental transformation in the ratio approach to compositional data analysis—all data thus need to be strictly positive, so that zero values present a major problem. Components that group together based on domain knowledge can be amalgamated (i.e., summed) to create new components, and this can alleviate the problem of data zeros. Once compositional data are transformed to logratios, regular univariate and multivariate statistical analysis can be performed, such as dimension reduction and clustering, as well as modeling. Alternative methodologies that come close to the ideals of the logratio approach are also considered, especially those that avoid the problem of data zeros, which is particularly acute in large bioinformatic data sets.

1. INTRODUCTION

Unit-sum constraint: the condition on a nonnegative multivariate sample $x_1, x_2, \ldots, x_J$ that $\sum_j x_j = 1$

Closing the data: the operation of dividing nonnegative data by their total

Subcomposition: a subset of parts after closure

Compositional data analysis (CoDA) is the analysis of nonnegative multivariate data where the absolute values of the data carry only relative meaning.[1] For example, compositional data are often data with a constant-sum constraint: That is, the values for each multivariate sample are either observed as summing to a constant, usually 1 or 100%, or are expressed as values relative to a total that is irrelevant to the research objective, for example, relative counts in samples of different sizes, or the composition of a chemical sample. Compositional data are observed in many fields (Aitchison 2005), most prominently: geochemistry (e.g., mineral compositions), ecology (e.g., relative abundances of species), biochemistry (e.g., fatty acid proportions), morphology (e.g., the shapes of living organisms), sociology (e.g., time budgets), geography (e.g., proportions of land use), political science (e.g., voting proportions), marketing (e.g., brand shares), and recently genomics and microbiome research (e.g., proportions of operational taxonomic units—see, for example, Li 2015, Tsilimigras & Fodor 2016, Quinn et al. 2018). In most applications the original totals, or sizes, on which the relative values are computed, are irrelevant, with a few exceptions. Notable exceptions are in marine biology, where different organisms are classified and counted in samples of equal volume, and similarly in botany, where plant species are counted in samples of equal area—in these cases, the total sample counts would be meaningful, distinguishing between samples of varying levels of species richness, and this total could be related to the relative abundances of the species. Similarly, voting percentages for political parties could be related to the size of each constituency's electorate.

Data with the unit-sum constraint are assumed here, so a J-component multivariate sample of nonnegative $x_1, x_2, \ldots, x_J$, with the property $\sum_j x_j = 1$, is called a composition. The components of a composition are called its parts, and because of the unit-sum constraint, the multivariate samples exist in a mathematical simplex: 3-part compositions are inside a triangle (see **Figure 1*b***), 4-part compositions are inside a tetrahedron, and so on for higher-dimensional simplexes. Zero values in a sample would force it to lie on a side of the simplex. Compositions are often obtained from data in the form of counts or positive measurements, such as physical measures of size (grams, liters, centimeters, etc.), by dividing them by their respective totals, a process called closure or normalization. Relative abundances of species in ecology, for example, are normalized, or closed, versions of the original abundance count data. The closure operator is denoted by $C(\,)$, so if $n_1, \ldots, n_J$ are the original counts or measurements, then $C(n_1, \ldots, n_J)$ are the relative values $x_1, \ldots, x_J$ with unit sum ($x_j = n_j / \sum_{j'} n_{j'}$), forming a composition. When the original totals have some relevance to a study, then the issues of size and shape in the statistical analysis are important, where size is related to the total and shape to the composition (Greenacre 2017). If a subset of parts is selected and the values of that subset are closed again, i.e., reclosed, this results in a subcomposition.

The problem of spurious correlations that result from closure has been known for over a century (Pearson 1897), and even compositional data based on random counts exhibit important correlations due to the sum constraint (Mosimann 1962, Aitchison 1981). Moreover, any observed composition is inevitably a subcomposition of a potentially larger composition. In biochemistry, for example, suppose that 30 fatty acids (FAs) are measured in a set of samples in laboratory A, and the same samples are reanalyzed in laboratory B but only 20 of those FAs are identified. Then, since the data are closed with respect to different totals, the compositional values of the 20 FAs common to the laboratories will be artificially different and not just different due to measurement error.

[1] For a short history of CoDA, see Bacon-Shone (2011). For books on CoDA, see Tolosana-Delgado & van den Boogaart (2011), Pawlowsky-Glahn & Buccianti (2011), van den Boogaart & Tolosana-Delgado (2013), Greenacre (2018), and Filzmoser et al. (2018).

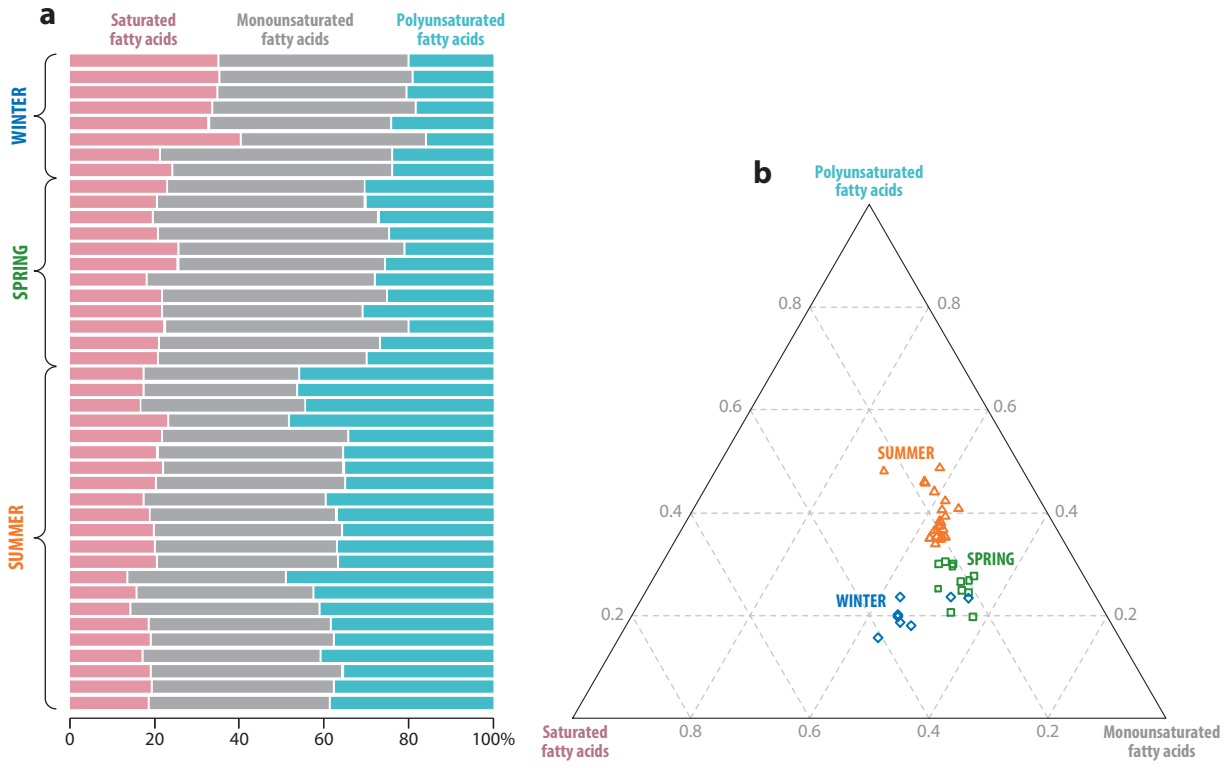

Figure 1

Three-part fatty acid compositional data from a sample of 42 marine copepods, observed in three different seasons (these data are more fully introduced in Section 3). The FAs are amalgamated into three major categories: saturated fatty acids (SFA), monounsaturated fatty acids (MUFA) and polyunsaturated fatty acids (PUFA). (*a*) A compositional bar chart of the data, in percentages, using function `BAR` in the R package `easyCODA`. (*b*) A triangular coordinate representation of the same data, as proportions, using function `ternaryplot` in the package `vcd`. Some minor cosmetic adjustments were made to panels *a* and *b* from the original output to improve legibility and to comply with print specifications.

In the approach to CoDA by Aitchison (1982, 1986, 2008), such paradoxes are eliminated by not considering the original values of the compositional parts, but rather their ratios, since the ratio between part A and part B remains constant irrespective of what other parts are present, before or after closure. Hence, univariate statistics on ratios can be validly compared between studies that have parts in common.

This property of the ratios is known as subcompositional coherence (Aitchison 1986) because ratios are unaffected by eliminating (or adding) some parts and reclosing—they are, quite simply, coherent. The original compositional data themselves are not coherent (i.e., incoherent) since the values of a subset of parts would change after closing to have unit sum. Hence, summary statistics such as means, variances, and correlations, as well as methods such as regression and analysis of variance, are incoherent when computed on the parts in a compositional data set. Ratios are generally compared multiplicatively, so the logarithmic transformation converts the ratios on a multiplicative scale to an additive scale (i.e., from a ratio scale to an interval scale). Thus, the logratio transformation takes the compositional data out of the simplex into real vector space, with an additive scale, thereby complying with most standard statistical methodologies, but with special issues around the interpretability of logratios in analytical results.

Subcompositional coherence: the property that relationships between parts are unaffected by forming subcompositions after closure

Pairwise logratio (LR): logratio of a pair of parts

Additive logratios (ALRs): $J-1$ LRs with respect to a fixed part

This review continues as follows. Section 2 defines several logratio transformations and describes their properties. Section 3 defines the total logratio variance, which quantifies the total variability in a compositional data set. Section 4 defines logratio distances, both between samples and between parts. Section 5 shows how logratio-transformed data can be visualized using biplots and cluster analysis. Section 6 deals with variable selection, to find small subsets of logratios or subcompositions that approximate the full composition, thereby reducing the number of variables to be considered. Section 7 treats compositional data as either response variables or explanatory variables in statistical models. Section 8 gives a brief exposé of possible advantages of weighting the parts. Section 9 treats the analysis of high-dimensional compositional data regularly found in bioinformatic applications, including the use of correspondence analysis as an alternative approach. The review concludes with a summary and suggestions of future issues.

2. LOGRATIO TRANSFORMATIONS

There are several types of logratio transformations, summarized in **Table 1** and dealt with one at a time in this section.

The simplest example of a logratio is the log-transformed ratio of two parts of a composition, or pairwise logratio, denoted throughout this review by LR. For a J-part composition $\{x_1, x_2, \ldots, x_J\}$, there are $\frac{1}{2}J(J-1)$ unique LRs of the form:

$$\mathrm{LR}(j,j') = \log\left(\frac{x_j}{x_{j'}}\right) \quad j,j' = 1,\ldots,J, \;\; j < j'. \qquad 1.$$

Any subset of $J-1$ linearly independent LRs that includes all the compositional parts forms a basis that can generate all the other LRs by linear combinations. The simplest such subset is that of the additive logratios (ALRs), where a specific reference part is contrasted with all the other parts (here, the last part is chosen as reference in the denominator):

$$\mathrm{ALR}(j|J) = \log\left(\frac{x_j}{x_J}\right) \quad j = 1,\ldots,J-1. \qquad 2.$$

Table 1 Logratio transformations of a composition consisting of J parts

Abbreviation	Name	Description
LR	Pairwise logratio	The log of the ratio of two parts
ALR	Additive logratio	A pairwise logratio (LR) that is one of a set of $J-1$ ALRs having the same denominator (or numerator)
SLR	Summated (amalgamated) logratio	The log of the ratio of the sums (amalgamations) of two subsets of parts
CLR	Centered logratio	The log of the ratio of a part and the geometric mean of all the parts; usually one of a set of J CLRs, each with one of the J parts in the numerator
ILR	Isometric logratio[a]	The log of the geometric means of two subsets of parts
PLR	Pivot logratio[a]	The log of the ratio of a single part and the geometric mean of a subset of the parts; usually one of a set of $J-1$ PLRs

SLRs, ILRs and PLRs are often qualified as balances. SLRs are straightforward balances of amalgamated parts, whereas ILRs and PLRs are balances of geometric means.

[a]Both ILRs and PLRs have a scale factor; a PLR is a simpler form and special case of an ILR.

Parts can be amalgamated by adding them together (Aitchison 1982), and these amalgamations can be used in logratios. An amalgamation (summated) logratio (SLR) balance is defined as follows, for two subsets J_1 and J_2 of parts:

$$\mathrm{SLR}(J_1, J_2) = \log\left(\frac{\sum_{j\in J_1} x_j}{\sum_{j\in J_2} x_j}\right). \quad 3.$$

SLRs are usually defined on substantive grounds: In biochemistry, for example, saturated FAs (SFA) and monounsaturated FAs (MUFA) are routinely amalgamated, and their ratio, SFA/MUFA, is computed for all samples (see the footnote at the end of Section 3 for an explanation of these groupings). Another example is in water chemistry, where all dissolved solids are amalgamated to form total dissolved solids, and ratios are investigated with respect to other water components. Since an SLR contrasts two groups of parts, it is a balance between the groups, where the SLR is 0 if the numerator and denominator amalgamations are equal (thus balanced) and is positive or negative depending on whether the numerator is greater than the denominator or vice versa (Greenacre 2020, Greenacre et al. 2020, Quinn & Erb 2020).

There are several logratio transformations that rely on geometric means to combine parts, the most important being the centered logratio (CLR) transformation (Aitchison 1986). A CLR is the logratio between a part and the geometric mean of all J parts in the composition. There are thus J CLRs, defined as:

$$\mathrm{CLR}(j) = \log\left(\frac{x_j}{\left(\prod_{j'} x_{j'}\right)^{1/J}}\right) = \log(x_j) - \frac{1}{J}\sum_{j'}\log(x_{j'}) \quad j = 1, \ldots, J. \quad 4.$$

Thus, a CLR is the logarithm of a part, centered with respect to the mean of the logarithms of all the parts in the sample. The CLRs serve very useful computational purposes as a substitute for analyzing all the LRs. For example, the total logratio variance of the CLRs is the same as that of all the LRs (see Section 3), the differences between two CLRs is the same as the LR of the two numerator parts, and the CLRs can be analyzed in a reduced-dimensional component analysis to represent the equivalent analysis of all the LRs (Aitchison & Greenacre 2002), illustrated in Section 5.

The isometric logratio (ILR) transformation (Egozcue et al. 2003) is related to a log-contrast (Aitchison 1983, Aitchison & Bacon-Shone 1984), which is a linear combination of the log-transformed parts where the sum of the coefficients of the combination sum to zero: $\sum_j a_j \log(x_j)$, $\sum_j a_j = 0$. If the positive coefficients $a_j > 0$ define the index subset J_1 and the negative ones $a_j < 0$ the subset J_2, the log-contrast can be written as a logratio:

$$\sum_{j\in J_1} |a_j| \log(x_j) - \sum_{j\in J_2} |a_j| \log(x_j) = \log\left(\frac{\prod_{j\in J_1} x_j^{|a_j|}}{\prod_{j\in J_2} x_j^{|a_j|}}\right), \quad 5.$$

where $\sum_{j\in J_1} |a_j| = \sum_{j\in J_2} |a_j|$.

Suppose that $|J_1|$ and $|J_2|$ denote the numbers of positive and negative coefficients, respectively. Then an ILR balance is a special case when the $|J_1|$ positive coefficients are $a_j = 1/|J_1|$ and the $|J_2|$ negative coefficients are $a_j = -1/|J_2|$. The log-contrast then forms a balance between the geometric means of the two respective groups of parts. There is an additional scale factor (explained below) that depends on the numbers of parts $|J_1|$ and $|J_2|$ (Egozcue & Pawlowsky-Glahn 2005):

$$\begin{aligned}\mathrm{ILR}(J_1, J_2) &= \sqrt{\frac{|J_1||J_2|}{|J_1| + |J_2|}} \log\left(\frac{\left(\prod_{j\in J_1} x_j\right)^{1/|J_1|}}{\left(\prod_{j\in J_2} x_j\right)^{1/|J_2|}}\right) \quad 6.\\ &= \sqrt{\frac{|J_1||J_2|}{|J_1| + |J_2|}} \left(\frac{1}{|J_1|}\sum_{j\in J_1}\log(x_j) - \frac{1}{|J_2|}\sum_{j\in J_2}\log(x_j)\right).\end{aligned}$$

Amalgamation (summated) logratio (SLR) balance: log of ratio of two sums of parts

Centered logratios (CLRs): log of each part relative to the geometric mean of all parts

Log-contrast: linear combination of logs of parts, with coefficients summing to 0

Isometric logratio (ILR) balance: log of ratio of geometric means of two subsets of parts, with scale factor

The factor rescales the set of coefficients $1/|J_1|$ and $-1/|J_2|$ that define the ILR balance to have unit length, since $\sum_{j\in J_1}(1/|J_1|)^2 + \sum_{j\in J_2}(1/|J_2|)^2 = (|J_1|+|J_2|)/(|J_1|\,|J_2|)$. Thus, when there are $J-1$ sets of such coefficients orthogonal to one another, the resultant $J-1$ ILR balances are defined with respect to an orthonormal basis. Apart from the scale factor, Equation 6 is equal to the average of the $|J_1|\times|J_2|$ LRs $\log(x_j/x_{j'})$ where $j\in J_1, j'\in J_2$.

Pivot logratio (PLR): ILR balance of single part relative to the geometric mean of a subset of parts

A simpler case of ILR balances is a set of pivot logratios (PLRs), which are a succession of $J-1$ ILR balances where the numerator in the ratio is a single part and the denominator is all those parts to the right in the particular ordered list of parts:

$$\mathrm{PLR}(j) = \sqrt{\frac{|J_2|}{1+|J_2|}}\log\left(\frac{x_j}{\left(\prod_{j'\in J_2}x_j\right)^{1/|J_2|}}\right) = \sqrt{\frac{|J_2|}{1+|J_2|}}\left(\log(x_j) - \frac{1}{|J_2|}\left(\sum_{j'\in J_2}\log(x_{j'})\right)\right), \quad 7.$$

where $j=1,\ldots,J-1$ and J_2 is the set of parts $J_2=\{j+1,j+2,\ldots,J\}$ (Hron et al. 2017, Filzmoser et al. 2018). A PLR, with its single part in the numerator, has the interpretational advantage of being equal to the average of a set of LRs where the numerator is the same for each LR. For example, the first PLR is, apart from the scale factor, equal to the average of $J-1$ logratios $(\log(x_1/x_2)+\log(x_1/x_3)+\cdots+\log(x_1/x_J))/(J-1)$.

With the exception of the CLRs, which involve all parts, the abovementioned logratios are coherent. Since LRs are coherent, so are the special case of ALRs, as are ILRs and PLRs. Like ILR balances, SLR balances are coherent with respect to adding parts to the composition or removing parts that are not in the numerator or denominator groups (Greenacre 2020).

Ratios can be represented as graphs in the form of a network, where each edge represents a ratio of two parts (**Figure 2*a–c***). Each of the graphs in **Figure 2*b*** and **Figure 2*c*** is an example of an acyclic connected graph: All parts are connected and there are no cycles. Such graphs necessarily represent a set of linearly independent LRs, consisting of one less LR than the number of parts. **Figure 2*d*** shows how the parts can be represented as a dendrogram, where each node splits into two subgroups of parts defining a ratio in an SLR or ILR balance.

A set of $J-1$ linearly independent logratios, of any of the types described above, involving all the parts at least once, can be inverted back (i.e., back-transformed) to the original J closed part values. Similarly, the complete set of J CLRs can be back-transformed to the closed parts. The values of the $J-1$ SLR or ILR balances defined by the nodes of **Figure 1*d*** can also be back-transformed. These inversion formulae, based on simple matrix computations, are summarized in the **Supplemental Appendix**.

Supplemental Material >

3. LOGRATIO VARIANCE

The transformations in Section 2 refer to a single general composition. Since LRs are the central concept in CoDA, the definition of total variability in a compositional data set consisting of several observed compositions is made in terms of them. Henceforth, a data matrix of compositions, $\mathbf{X}$ ($I\times J$), is considered: I rows (with index $i=1,\ldots,I$) observed on J compositional parts, as before (with $j=1,\ldots,J$). A double subindex notation $_{jj'}$ is introduced to denote a pair of parts. The mean of the (j,j')th LR is thus denoted by $\mathsf{Mean}_{jj'} = (1/I)\sum_i \log(x_{ij}/x_{ij'})$.

The variance of the (j,j')-th LR, denoted by $\mathsf{Var}_{jj'}$, is defined as

$$\mathsf{Var}_{jj'} = \frac{1}{I}\sum_{i=1}^{I}\left(\log(x_{ij}/x_{ij'}) - \mathsf{Mean}_{jj'}\right)^2, \quad 8.$$

where the sum of squared deviations from the mean is divided by I rather than $I-1$ (thereby assigning an equal weight of $1/I$ to each sample, weights that can be varied if necessary). The total

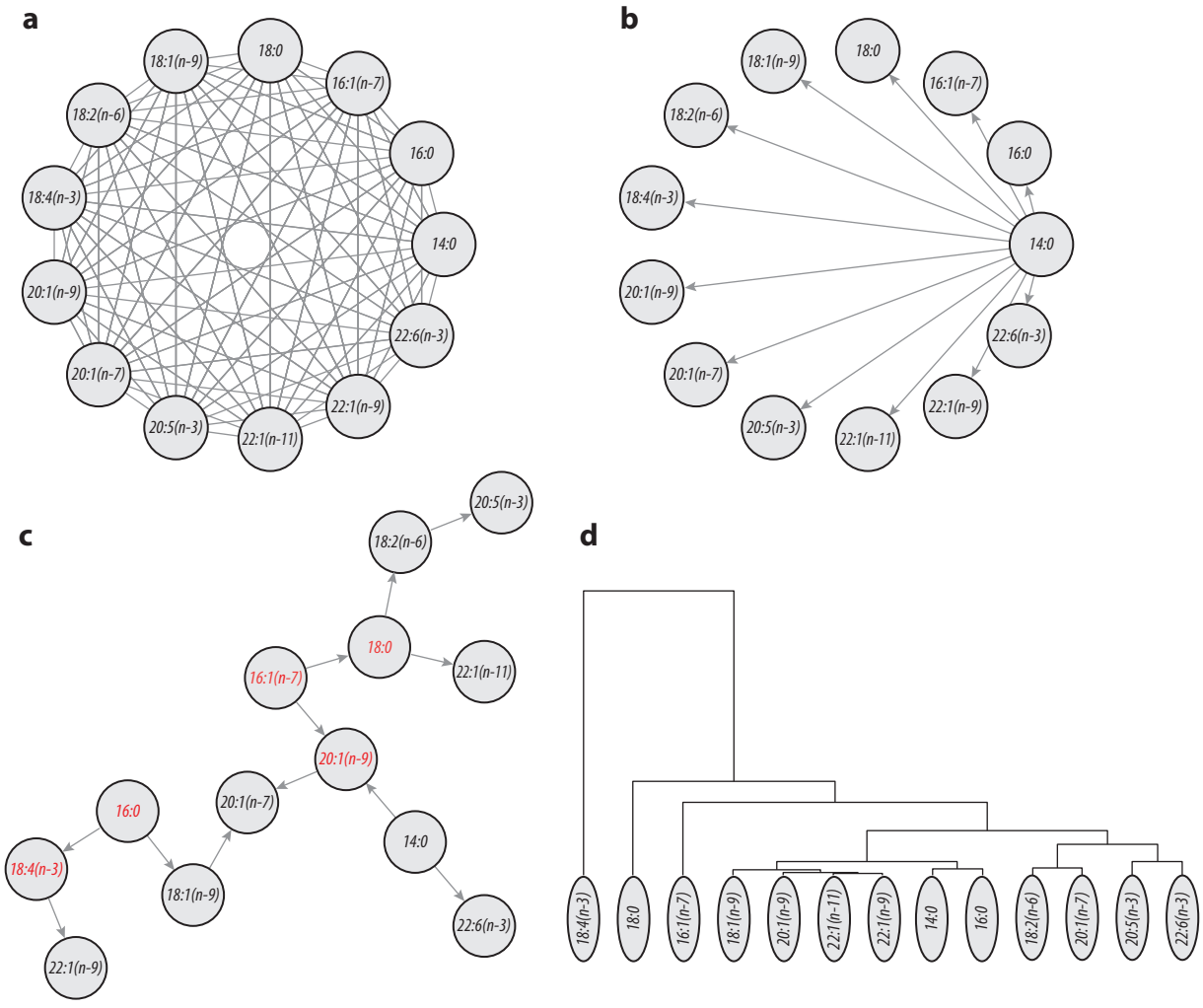

Figure 2

Graphs of logratios of 13 parts (fatty acids) as the edges of a network: (*a*) all $13 \times 12/6 = 78$ pairwise logratios (LRs), (*b*) the 12 additive logratios (ALRs) with *14:0* as the denominator part, (*c*) an acyclic connected graph of 12 independent LRs (the five parts in red are explained in Section 5), and (*d*) a dendrogram defining balances of two subsets of parts at each node. The examples in panels *b* and *c* are shown as directed graphs where each arrow points to the numerator part, often referred to by the acronym DAG (directed acyclic graph).

logratio variance, denoted by TotVar, is then defined as:

$$\mathsf{TotVar} = \frac{1}{J^2} \sum \sum_{j<j'} \mathsf{Var}_{jj'}. \qquad 9.$$

Total logratio variance: the sum of all LR variances divided by J^2; equivalently, the average of the CLR variances

Notice that there are $\frac{1}{2}J(J-1)$ terms in the above sum, but the division is by J^2, explained below.

A shortcut to obtain TotVar is to average the variances of the CLRs, denoted by Var_j ($j = 1, \ldots, J$), with a single subindex, which gives an identical solution:

$$\mathsf{TotVar} = \frac{1}{J} \sum_j \mathsf{Var}_j, \qquad 10.$$

where the variances are similarly computed by dividing by I, not by $I - 1$. Notice that the equivalent definitions in Equations 9 and 10 are equal to what Aitchison (1986) defined as the total variance, namely $\sum_j \mathsf{Var}_j$, divided by J. The division by J, the number of parts, is for good reason, because then each part has an equal weight of $1/J$, with weights summing to 1, and these can be varied if differential weights are required (see Section 8).

Contained variance: the contribution of a logratio to the total logratio variance

Explained variance: the contribution of a logratio to explaining total variance in the sense of regression

FA compositions of marine copepods: a 42×40 compositional data matrix

Logratio analysis (LRA): PCA of all pairwise LRs, equivalent to PCA of all CLRs

When quantifying the importance of each LR to TotVar, the distinction needs to be made between contained variance and explained variance. Each LR contains a part of variance of the total, which is simply its (weighted) contribution $\mathsf{Var}_{jj'}/J^2$ to the sum in Equation 9, where the weight of an LR is the product of the weights of the jth and j'th parts, each being $1/J$ in this equally weighted definition. However, a particular LR can explain a much larger part of the total variance than its own variance since it is correlated with many other LRs. In fact, as stated at the end of Section 2, only $J - 1$ independent LRs are needed to explain 100% of the total variance of all $\frac{1}{2}J(J-1)$ LRs, whereas their cumulated contained variances can be a relatively small percentage of the total.

To illustrate these concepts and the analyses to follow, a real data set from biochemistry is now introduced.

Data set (FA compositions of marine copepods). Calanoid copepods are small marine organisms that are a very important food source in the Arctic. They were collected during an extensive field study in the Rijpfjorden fjord, Svalbard, a high Arctic sea-ice-dominated ecosystem, during the International Polar Year 2007/2008, in three different seasons: spring, summer, and winter. The objective was to investigate the seasonal development of the key marine species *Calanus glacialis* (Søreide et al. 2010). This data set is composed of 42 copepods and 40 FAs and is available online (see the **Supplemental Appendix**).

Supplemental Material >

Initially, attention is restricted to the FAs that have an average occurrence of at least 0.01 (i.e., 1%). This reduces the number of FAs from 40 to a subcomposition of 13, exactly the subset that was used in **Figure 2**. This subset of the data matrix is shown as percentages in **Table 2**, closed to sum to 100%. These 13 FAs have no data zeros, so the logratio transformations are possible—later, the complete data set is analyzed, after the issue of data zeros is dealt with.

The total logratio variance of this 13-part data set is computed, using either of the equivalent forms of Equations 9 and 10, as $\mathsf{TotVar} = 0.2462$. The (contained) contributions by the individual FAs (using Equation 10) range from 1.1% for FA *22:1(n-9)*[2] to 45.7% for *18:4(n-3)*.

The R package `easyCODA` (Greenacre 2018, R Core Team 2020) is used for most of the computations.

4. LOGRATIO DISTANCES: COMPONENT AND CLUSTER ANALYSIS

Once compositional data have been logratio-transformed, multivariate analysis can essentially be conducted as before for regular interval-scale data, with appropriate adaptation of the interpretation to the fact that the variables are now logratios of varying complexities. The principal component analysis (PCA) of all $\frac{1}{2}J(J-1)$ LRs is called logratio analysis (LRA) (Greenacre & Lewi 2009; Greenacre 2018, 2019). LRA is equivalent to the PCA of the J CLRs (Aitchison 1990; Aitchison & Greenacre 2002, appendix), the latter being much more efficient computationally.

[2]FAs, affectionately called the fats of life, are long chains of hydrocarbons coded according to their chemical structure in the format *XX:Y(n-ZZ)*, where *XX* = the number of carbon atoms, *Y* = the number of double carbon bonds, and *ZZ* = number of carbons from the last double bond to the methyl, or omega, end. Saturated FAs have no double bonds, monounsaturated FAs have one double bond, and polyunsaturated FAs have two or more double bonds.

Table 2 Fatty acid percentages* in a 13-part subcomposition of a 40-part data set

Season	14:0	16:0	16:1(n-7)	18:0	18:1(n-9)	18:2(n-6)	18:4(n-3)	20:1(n-9)	20:1(n-7)	20:5(n-3)	22:1(n-11)	22:1(n-9)	22:6(n-3)
Winter	15.78	13.54	7.28	7.25	8.29	2.26	1.88	17.8	0.96	4.92	11.28	1.44	7.34
Winter	13.44	13.63	8.19	9.24	8.73	2.10	1.29	16.83	0.78	5.18	11.92	1.51	7.16
Winter	7.11	13.14	7.29	15.64	5.91	2.89	1.51	15.34	1.16	5.96	13.43	1.74	8.89
Winter	13.49	12.55	8.34	8.08	8.61	1.94	1.03	19.44	0.96	5.40	10.87	1.46	7.84
Winter	16.54	12.28	8.74	3.25	8.89	1.47	0.84	14.12	0.97	8.08	11.02	1.42	12.37
Winter	8.42	12.53	7.50	20.15	7.08	1.75	0.25	17.37	0.71	4.84	10.71	1.39	7.31
Winter	5.01	11.17	8.15	5.13	9.66	2.16	0.57	20.87	0.85	8.33	14.79	1.80	11.52
Winter	8.19	11.81	9.16	3.71	8.11	1.57	0.26	19.64	0.81	9.39	13.05	2.08	12.21
Spring	9.93	10.51	20.71	2.28	6.56	0.99	2.26	11.30	0.76	13.07	7.92	1.06	12.65
Spring	8.79	10.39	27.45	2.06	4.46	0.71	2.99	10.80	0.67	15.58	6.17	0.93	9.01
Spring	8.56	10.09	30.21	1.50	5.52	0.94	1.65	10.81	0.79	11.90	6.79	1.01	10.21
Spring	9.47	10.30	31.49	1.51	4.54	0.80	1.19	10.36	0.82	11.15	7.24	1.03	10.10
Spring	10.32	10.92	24.97	4.69	5.83	0.96	0.72	13.30	0.78	7.99	8.74	1.22	9.56
Spring	9.00	12.68	21.73	4.52	5.47	0.92	1.88	14.23	0.68	11.06	7.56	1.29	8.99
Spring	7.36	9.89	30.49	1.19	5.02	1.18	1.85	11.32	0.83	13.98	6.82	1.01	9.07
Spring	9.78	10.2	32.26	1.80	4.98	1.23	0.91	9.28	0.68	10.30	7.10	1.10	10.37
Spring	7.45	11.53	23.83	2.98	4.42	0.79	0.93	10.21	0.70	13.33	7.20	1.14	15.50
Spring	9.31	10.72	21.17	2.37	6.28	1.19	1.09	18.02	0.93	8.40	11.44	1.74	7.35
Spring	7.02	10.93	28.00	3.52	5.10	0.80	1.37	10.58	0.76	12.05	7.40	1.09	11.39
Spring	6.60	11.16	29.73	3.27	4.67	0.80	1.10	8.04	0.79	13.53	5.56	1.06	13.68
Summer	7.34	8.74	6.73	1.89	5.23	2.54	16.11	13.34	2.16	13.37	9.15	1.01	12.38
Summer	7.51	8.76	7.07	2.00	5.08	2.64	17.31	12.66	2.38	7.47	9.67	0.98	16.47
Summer	7.05	8.67	8.01	1.74	4.73	2.49	15.12	13.94	2.28	12.59	9.57	1.09	12.72
Summer	5.83	13.27	7.13	5.50	4.69	4.33	9.66	5.67	3.08	14.01	5.52	0.88	20.41
Summer	10.15	9.91	7.04	1.68	9.48	1.89	10.34	17.20	1.47	11.69	10.19	1.44	7.50
Summer	10.18	8.89	7.53	1.60	8.44	1.90	10.90	16.77	1.69	13.06	10.83	1.45	6.75
Summer	10.57	9.49	6.70	1.88	9.04	1.69	11.14	17.09	1.31	12.17	9.89	1.22	7.82
Summer	10.81	8.68	12.5	1.38	6.58	1.56	10.92	16.25	1.34	13.48	9.26	1.32	5.94
Summer	8.22	7.58	9.29	1.61	5.55	2.03	11.56	18.14	1.18	14.55	10.45	1.50	8.34
Summer	9.20	8.16	10.75	1.39	5.91	1.94	11.75	18.35	1.35	11.34	9.35	1.36	9.13
Summer	8.53	9.34	6.54	1.88	8.79	1.67	10.71	18.44	1.28	12.87	10.52	1.38	8.07
Summer	9.38	8.82	7.45	1.71	8.08	1.79	11.03	17.68	1.27	12.67	10.03	1.40	8.69
Summer	9.34	9.12	5.73	1.85	8.19	1.63	10.88	17.09	1.22	13.48	11.48	1.45	8.52
Summer	6.13	6.06	9.46	1.50	4.00	3.09	18.96	14.27	1.58	13.74	8.90	1.45	10.86
Summer	7.61	7.36	14.46	1.30	3.94	1.95	12.10	14.49	1.42	15.78	8.56	1.22	9.82
Summer	6.21	7.17	15.44	0.95	4.01	1.73	11.67	15.78	1.39	15.70	9.42	1.32	9.20
Summer	9.39	8.22	8.89	0.99	7.03	1.65	11.18	16.19	1.22	14.69	10.97	1.48	8.09
Summer	9.91	8.10	6.77	1.05	7.36	2.04	11.67	17.59	1.32	13.32	11.23	1.50	8.16
Summer	8.78	7.43	6.52	1.18	7.41	2.99	12.94	16.82	1.23	12.39	11.30	1.52	9.48
Summer	9.51	8.56	11.61	0.92	7.53	1.75	11.06	16.96	1.26	13.39	9.46	1.26	6.73
Summer	9.47	8.57	9.38	1.08	7.78	1.67	12.66	16.22	1.26	13.13	9.68	1.32	7.79
Summer	9.07	8.02	9.99	1.15	7.39	1.83	12.19	16.24	1.48	13.11	9.45	1.37	8.70

*The percentages in each row sum to 100%, due to closure of the subcomposition.

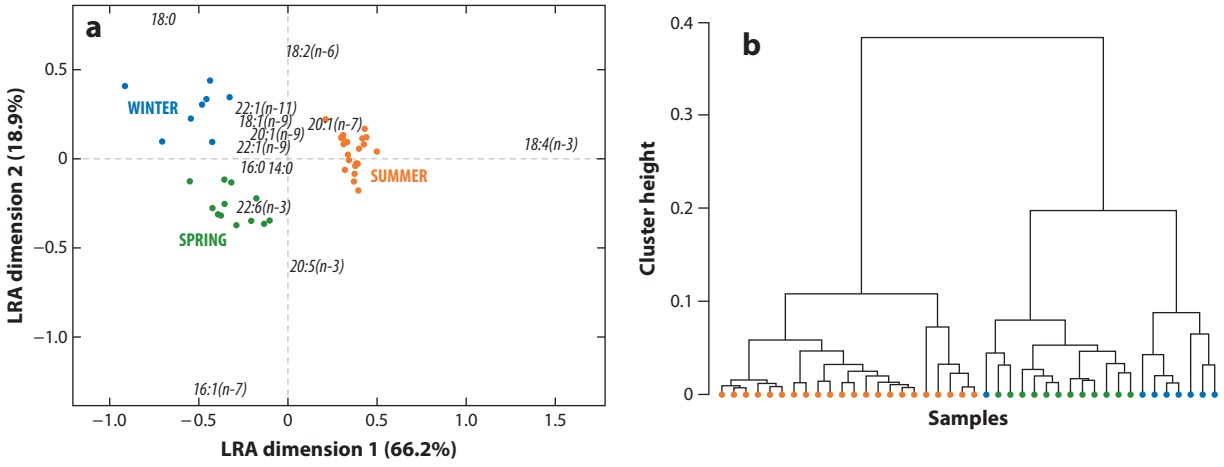

Figure 3

(*a*) Logratio analysis (LRA) form biplot (Greenacre 2010a) of the 13-part fatty acid (FA) subcomposition, where the intersample distances approximate logratio distances, and the directions between pairs of FAs represent logratio biplot axes. (*b*) Ward clustering of the logratio distances between the 42 samples.

Both the matrix of all LRs and the matrix of the CLRs have the same rank (i.e., dimensionality), equal to $J - 1$. The principal components of the CLR matrix have the property that their coefficients sum to 0, so that they reduce to log-contrasts of the parts (Equation 5). In fact, thanks to their orthogonality, principal components are variables where the contained and explained variances are the same and so make an identical decomposition of TotVar.

Figure 3*a* shows the LRA of the FA data set, using function `LRA()` of the `easyCODA` package for the computations, with added color coding of the samples according to season. While it is tempting to interpret the positions of the FAs as variables in a regular PCA, they only have meaning in their pairwise positions. For example, the logratio log(*16:1(n-7)/18:0*) would be in the direction of the connection between these two FAs, pointing almost vertically downward, since *16:1(n-7)* is in the numerator. Hence, it should always be remembered that the PCA of the CLRs is just a shortcut to analyzing all the pairwise LRs, the latter being the variables of primary interest.

As the number of parts increases, biplots such as **Figure 3*a*** soon become crowded. This problem can be alleviated greatly by either using contribution biplots where only the highly contributing parts are shown (Greenacre 2013) (see Section 9.1 for an example) or by performing variable selection (see Section 5).

For clustering, the usual hierarchical or nonhierarchical clustering algorithms can be performed on the samples, using Euclidean distances defined on all the LRs, but again more efficiently computed using the CLRs. If the matrix $\mathbf{Y} = [y_{ij}]$ denotes the CLR-transformed data set, and $\mathbf{Z} = [z_{i,jj'}]$ denotes the matrix of LRs $\log(x_{ij}/x_{ij'})$, then $d_{ii'}$, the logratio distance between samples i and i', can be defined in two equivalent forms, analogous to the definitions in Equations 9 and 10, as:

Logratio distance between samples: can be computed between LRs or between CLRs

$$d_{ii'} = \sqrt{\frac{1}{J^2}\sum\sum_{j<j'}(z_{i,jj'} - z_{i',jj'})^2} = \sqrt{\frac{1}{J}\sum_{j}(y_{ij} - y_{ij'})^2}. \qquad 11.$$

Figure 3*b* shows the Ward clustering of these distances between samples, where three clear clusters are identified, with only one winter sample misclassified among the spring ones. Using

the function `WARD()` from the `easyCODA` package, the vertical height scale is such that the sum of squared heights of all 41 nodes is equal to TotVar, the total logratio variance.

Notice that $z_{i,jj'} - z_{i',jj'} = \log(x_{ij}/x_{ij'}) - \log(x_{i'j}/x_{i'j'}) = \log(x_{ij}/x_{i'j}) - \log(x_{ij'}/x_{i'j'})$, i.e., the difference between two logratios row-wise is identical to a difference computed on the same four values column-wise. Both are equal to $\log\left((x_{ij}x_{i'j'})/(x_{ij'}x_{i'j})\right)$, which is denoted by $s_{ii',jj'}$, i.e., the logarithm of the cross-product ratio $(x_{ij}x_{i'j'})/(x_{ij'}x_{i'j})$. The first definition of intersample distance in Equation 11 can thus be written as $d_{ii'} = \sqrt{\frac{1}{J^2}\sum\sum_{j<j'}(s_{ii',jj'})^2}$. In a symmetric fashion, the corresponding logratio distance between parts j and j' would involve the sum of the same squared values, $(s_{ii',jj'})^2$, across all unique pairs of rows: $d_{jj'} = \sqrt{\frac{1}{I^2}\sum\sum_{i<i'}(s_{ii',jj'})^2}$. To get the same between-part distances using CLRs, the CLRs have to be first recomputed columnwise—that is, each column of the matrix of log-transformed compositional data has to be centered with respect to the column mean, and then a formula similar to the right-hand side of Equation 11 is applied, averaging over the I rows.

The total logratio variance can be equivalently obtained from the sum of either the intersample or interpart squared logratio distances:

$$\mathsf{TotVar} = \frac{1}{I^2}\sum\sum_{i<i'} d_{ii'}^2 = \frac{1}{J^2}\sum\sum_{j<j'} d_{jj'}^2. \qquad 12.$$

Yet another way of obtaining the total logratio variance and the logratio distances, between rows or between columns, is to compute the double-centered matrix of the log-transformed data matrix $\log(\mathbf{X})$:

$$\mathbf{H} = \left(\mathbf{I} - \frac{1}{I}\mathbf{1}_I\mathbf{1}_I^\mathsf{T}\right)\log(\mathbf{X})\left(\mathbf{I} - \frac{1}{J}\mathbf{1}_J\mathbf{1}_J^\mathsf{T}\right), \qquad 13.$$

where $\mathbf{1}_I$ and $\mathbf{1}_J$ are vectors of I and J ones, respectively. Then the total variance is the average of the squared elements of $\mathbf{H}$: $\mathsf{TotVar} = \frac{1}{IJ}\sum_i\sum_j h_{ij}^2$; the squared interrow distances are the averages of the squared differences between rows (samples): $d_{ii'}^2 = \frac{1}{J}\sum_j(h_{ij} - h_{i'j})^2$; and the squared intercolumn distances are the averages of the squared differences between columns (parts): $d_{jj'}^2 = \frac{1}{I}\sum_i(h_{ij} - h_{ij'})^2$. Moreover, the LRA (**Figures 3*a*** and **4*a***) is equivalent to performing the singular value decomposition (SVD) of $\mathbf{H}$. This is because the CLR transformation removes the row means of the log-transformed data, $\log(\mathbf{X})$, after which PCA removes the column means; hence, $\log(\mathbf{X})$ is double-centered in the LRA.

It turns out that, with the present definitions of variance and logratio distance, the squared logratio distance between two parts j and j' is identical to the variance of the LR of the respective parts, denoted in Section 3 by $\mathsf{Var}_{jj'}$: $d_{jj'}^2 = \mathsf{Var}_{jj'}$. Gathered in a square symmetric matrix, the quantities $\mathsf{Var}_{jj'}$ form the variation matrix [Aitchison (1986), who denotes them by $\tau_{jj'}$]. Hence, a multidimensional scaling and cluster analysis of the parts are effectively working on distances equal to $\sqrt{\tau_{jj'}}$, i.e., the standard deviations of the LRs. **Figure 4*a*** shows the same LRA as in **Figure 3*a***, but now the parts (FAs) have a distance interpretation, and **Figure 4*b*** shows the corresponding Ward clustering of the interpart distances.

The squared distances between parts (i.e., variances of the LRs) are related to the concept of proportionality between parts (Lovell et al. 2015, Erb & Notredame 2016, Quinn et al. 2017). If two parts were perfectly proportional, related by a constant factor across all the samples, then the corresponding logratio would be a constant and have a variance of zero, i.e., zero distance apart, and thus be perfectly correlated. Attempts have been made to turn this measure into a correlation-type coefficient between parts, the challenge being to define an upper bound on the LR variance in order to have a coefficient lying between 1 (proportionality 0, i.e., perfect correlation) and 0

Logratio distance between parts: can also be computed between LRs or between CLRs centered columnwise

Total logratio variance: can be computed from the squared distances, between-samples or between-parts

Double-centered matrix of log-transformed data: gives total variance, all distances as well as the biplots

Proportionality between parts: equals zero if the two parts are constant multiples of each other

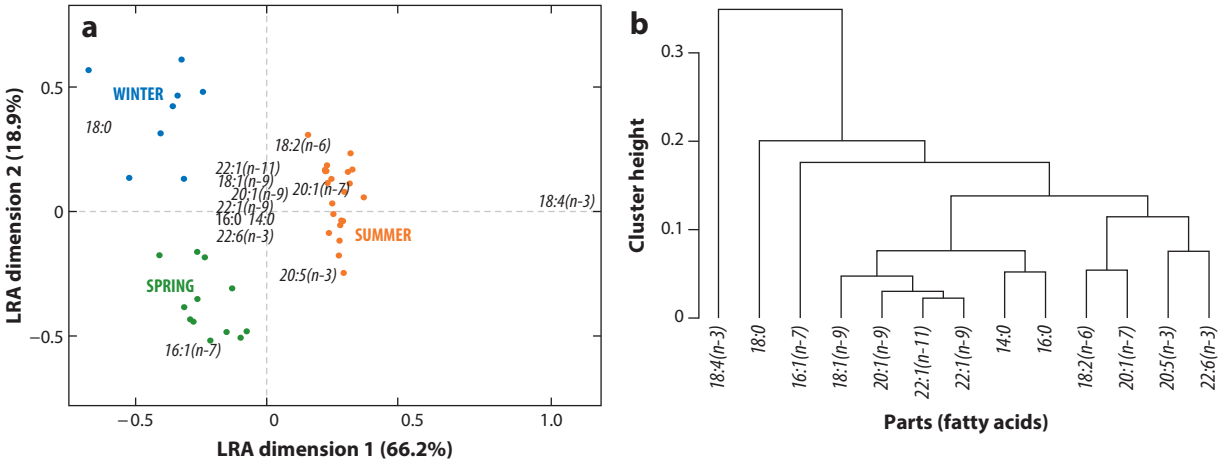

Figure 4

(*a*) Same analysis as **Figure 3*a***, but the logratio analysis (LRA) covariance biplot version, where logratio distances between fatty acids (FAs) are now approximately shown, and where these displayed distances between FAs are approximately equal to the standard deviations of the respective logratios. (*b*) Ward clustering of the logratio distances between FAs.

(no correlation). Aitchison (1997) proposed the transformation $e^{-\tau_{jj'}}$—Quinn et al. (2017) provide further discussion and alternatives.

5. LOGRATIO VARIABLE SELECTION

With so many LRs available in a compositional data set, the question arises as to which are the important ones to focus on and which ones can be ignored in order to make our understanding of the structure more parsimonious. This question can be similarly posed as identifying the significant subcompositions, or alternatively, SLR, ILR, or PLR balances. There are many ways to answer this question, but the essential idea behind a solution is inspired by the work of Krzanowski (1987), who in the context of PCA considered how to select variables that were responsible for the main structure of the data set.

The first consideration in the present case is whether the context is one of unsupervised or supervised learning. In the former context, we want to identify the LRs, or the subcomposition of parts, that are essentially defining the logratio structure of the data set. In the latter context, we want to identify the LRs, or subcomposition of parts, that are explaining some observed response variable or known grouping structure in the samples. For example, in the FA data set, samples are from three seasons, so we would be interested to identify the logratios that are separating the three groups of points. In this particular example, looking at the clear division between the seasons already observed in **Figure 3** in both the PCA and the cluster analysis, both approaches should arrive at essentially the same results.

Redundancy analysis (RDA): multiresponse linear regression

Adopting an unsupervised approach, the data have a total logratio variance, TotVar, equal to 0.2462, so the question is: Which LR explains a maximum part of this variance? Any LR explains its own contained part of variance, but it also explains parts of variances contained in many other LRs with which it is correlated. It turns out that, on the one hand, the LR log(*16:0/18:4(n-3)*) explains 65.6% of TotVar, more than any other LR. The contained variance of this LR is, on the other hand, only 5.0% of TotVar, but it is the explained variance which is relevant for variable selection. The analysis used to determine the explained variance is called redundancy analysis (RDA), a

generalization of regression analysis to multiresponse data (van den Wollenberg 1977, Gittins 1985, Zuur et al. 2007)—additional details are provided in the **Supplemental Appendix**. Once again, the smaller number of CLRs conveniently forms the set of response variables, equivalent to using all the LRs.

Supplemental Material >

This first LR given above is retained, and then the next step is to find which other LR adds a maximum additional explained variance—the stepwise process is explained in detail by Greenacre (2019). The next LR is identified as log(*16:1(n-7)/18:0*), explaining an additional 18.2%, bringing the variance explained by these two LRs to 83.8%. Adding more LRs in this way would bring the variance explained to 100% when $J - 1 = 12$ LRs have entered—in fact, **Figure 2*c*** shows the 12 chosen ratios as an acyclic connected graph. However, just the first three selected LRs bring the variance explained to over 90%, which could be considered satisfactory, effectively replacing the whole 13-part data set with only three LRs. These three LRs are part of the graph in **Figure 2*c***, connecting the five FAs labeled in red.

The three LRs are given in **Figure 5**, which also shows how much of the logratio variance of individual FAs (specifically, the variance of their CLRs) each of the three LRs explains. The FAs are ordered in descending order of logratio variance (the bar chart in the right panel displays these as percentages of total variance). The bars in the left panel show that the high logratio variances have most of their variance explained by the LRs, with parts of unexplained variance concentrated mostly in the FAs with low variances.

The PCA of these three LRs alone, involving only five parts (FAs), and the cluster analysis of the samples based on just these three are shown in **Figure 6**. The PCA and the cluster analysis both show an improved separation of the seasons, even though knowledge of this grouping has

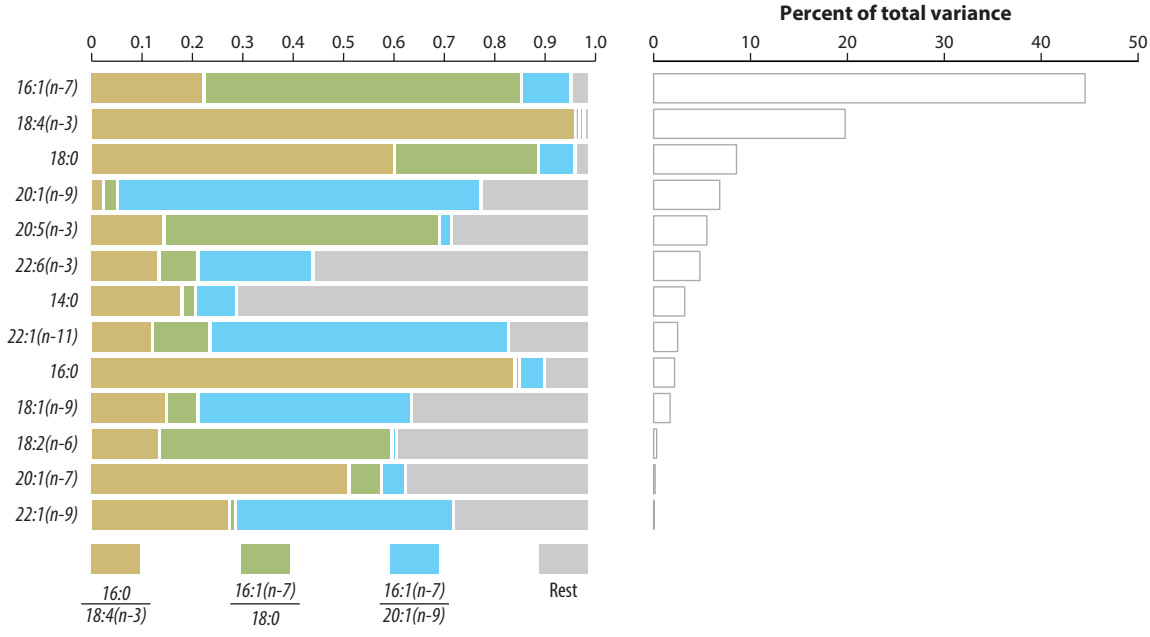

Figure 5

Explanation of logratio variance by the three LRs (these three LRs connect the five red FAs in **Figure 2*c***) chosen stepwise. The horizontal bar charts on the left show proportionally how much the three LRs explain the logratio variance of each of the parts (FAs), shown as rows, as well as the unexplained variance in each case (*gray*). The FAs are listed in descending order of their logratio variances, shown as percentages of the total logratio variance in the bar chart on the right. Abbreviations: FA, fatty acid; LR, pairwise logratio.

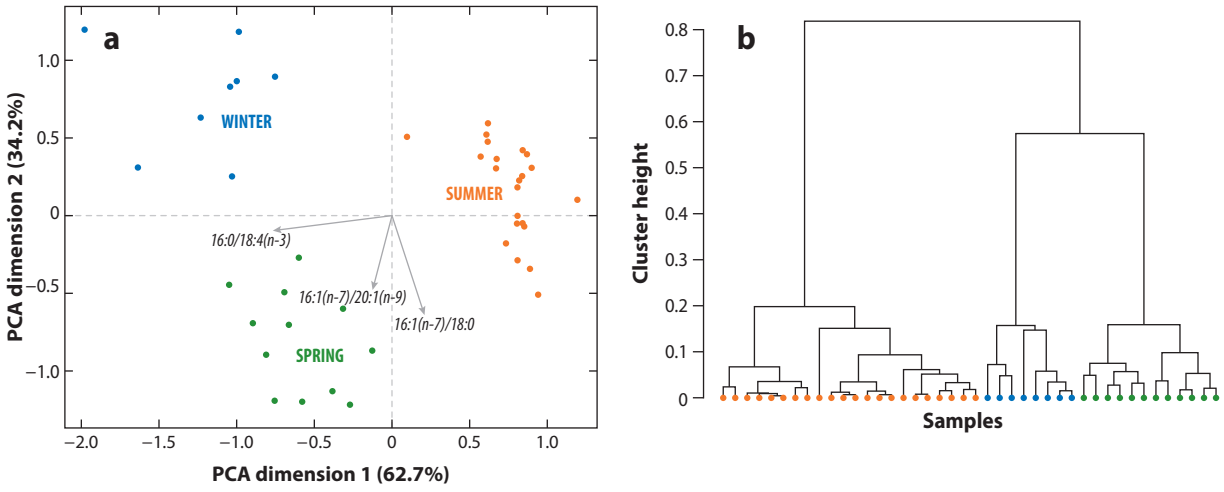

Figure 6

(*a*) PCA of the three pairwise logratios that explain 90.9% of the total logratio variance of the 13-part fatty acid data set; 96.9% of the variance of these three pairwise logratios (LRs) is explained by the two-dimensional solution, i.e., 0.969 × 0.909 = 0.881, or 88.1%, of the total logratio variance. (*b*) Ward clustering of the samples based on these three LRs only, showing perfect clustering of the seasons.

not been taken into account in the analysis. The cluster analysis now perfectly coincides with the three seasonal groups.

The variance explained by these three LRs is 90.9% of the total logratio variance. They involve five parts, *16:0*, *16:1(n-7)*, *18:0*, *18:4(n-3)*, and *20:1(n-9)*, which could be used as a 5-part subcomposition. Using the five CLRs of this subcomposition would explain 92.7% of the total variance, 1.8 percentage points more. This is because the three LRs do not connect all five parts of the subcomposition, shown in red in **Figure 2*c***—one additional LR would be needed to make an acyclic connected graph of the five parts. Any one connection linking them, e.g., the logratio of *16:0* relative to *16:1(n-7)*, will make the connection, resulting in the four LRs explaining 92.7% of the logratio variance, the same as the five CLRs of the subcomposition.

Splitting the parts into two subcompositions, the 5-part one given above and the complementary 8-part one, gives variance explained by these two subcompositions as 94.7% and 86.1% respectively (**Figure 7*a***)—each of these subcompositions explains a substantial amount of variance in common. This common part can be identified by again doing an RDA but first partialling out the variance of either subcomposition and then seeing how much of the residual variance is explained—it turns out that 81.1% of the variance is common to these two subcompositions. In **Figure 7*b***, the parts of total variance contained in the LRs of the two subcompositions are shown, as well the parts of total variance in all the LRs that connect the two.

As a final comment to this section, the set of LRs chosen stepwise might not always be the ones that an expert, in this case a biochemist, deems the most suitable for representing the compositional data. In fact, at each step there are several candidate LRs for entering that are very close to explaining the maximum variance, and one of these might be regarded as substantively more interesting. In a collaboration between a biochemist and a statistician, Graeve & Greenacre (2020) detail such an exercise on two FA data sets where the LR chosen at each step is either the optimal one or a slightly suboptimal one that has more substantive meaning biochemically. This idea is echoed in Section 6.2 when LRs are used as predictors of a response variable.

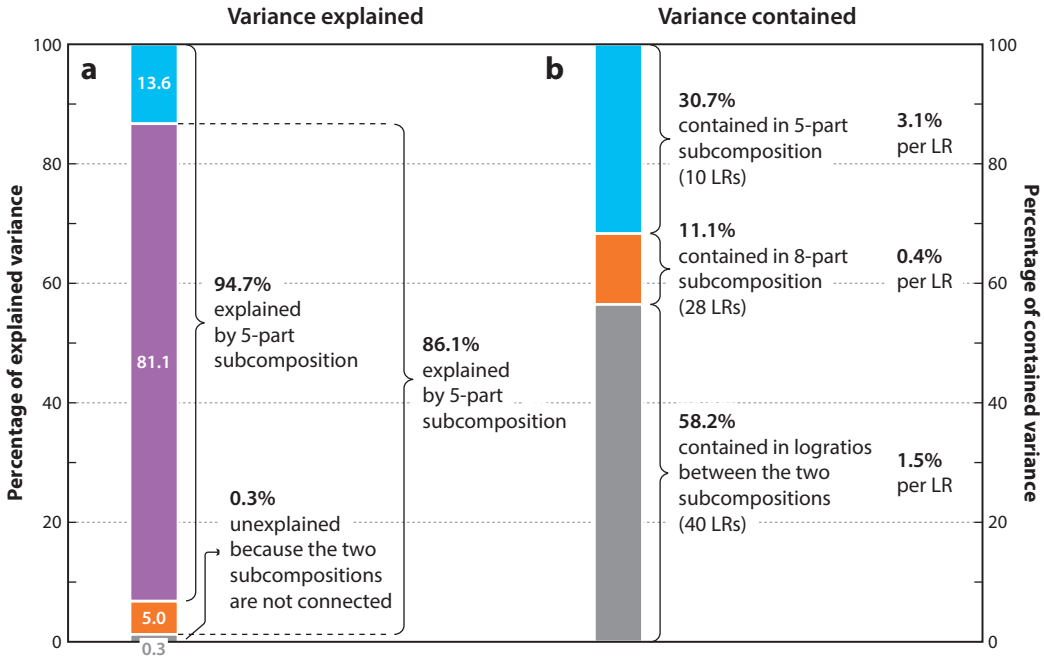

Figure 7

(*a*) Percentages of variance explained by the two disjoint subcompositions, including the common part explained by both. The residual variance of 0.3% would be explained by any pairwise logratio (LR) that connects the two subcompositions. (*b*) Percentages of variance contained in the LRs of the two disjoint subcompositions, and the LRs between them. The percentage contained per LR is highest in the subcomposition that optimally explains the total logratio variance.

6. MODELING WITH LOGRATIOS

In the previous section, individual LRs themselves were investigated as variables explaining the set of all LRs, but there was no issue of interpretation of the effects of the chosen LRs—interest was only in replacing the original data set with a reduced set of LRs that best represented the original one. When it comes to using LRs of any type in actual models, some care is needed in their interpretation, especially when the LRs are used as explanatory variables. Two cases are distinguished here: compositional data as response variables and compositional data as explanatory variables.

6.1. Logratios as Response Variables

A single LR, or several selected ones, or all of them, can form a set of response variables, modeled in terms of some explanatory variables, which could be continuous or categorical. A simple application is given by Faes et al. (2011) of modeling a single LR response. In the case of several, or all, of the logratios as responses, an approach similar to that of Section 5 can be adopted: that is, investigating which of the explanatory variables explain large and significant parts of the total logratio variance, using RDA.

For example, in the FA data, the seasons could be considered a categorical variable that explains logratio variance. An RDA of the complete set of CLRs, with the three dummy variables for the seasons as explanatory variables, yields the result that a large part, 76.3%, of the total

logratio variance is explained by the seasons. This is identical to the percentage of between-season variance, so that the within-season variance is 23.7%. Restricted to the 5-part subcomposition identified by the best three LRs (see Section 5), the seasonal variable explains an even higher percentage of variance, 84.3%. By contrast, only 46.7% of the variance of the complementary 8-part subcomposition is explained by the seasons. Section 3 in the **Supplemental Appendix**, titled Redundancy Analysis (RDA), provides more details about this result and an additional graphical representation of the explained variances.

Supplemental Material >

6.2. Logratios as Explanatory Variables

If LRs are considered as explanatory variables, then more care needs to be taken in the model interpretation. Usually, LRs will be combined additively as explanatory variables in the model, whether it be regression or any other generalized linear model. In regular regression modeling, the effect size of an explanatory variable is judged by considering a unit increase while all other explanatory variables are fixed. This approach can present difficulties in the special case of compositional data, with its unit sum constraint, since changing the values of any parts necessarily means changes in some of the other parts.

Aitchison & Shen (1984) first proposed a regression relationship in terms of a log-contrast of compositional parts, for example, for a response variable y and J parts:

$$y = a_0 + \sum_{j=1}^{J} a_j \log(x_j) + e, \quad \text{where} \sum_{j=1}^{J} a_j = 0. \tag{14.}$$

Coenders & Pawlowsky-Glahn (2020) show how different types of explanatory logratios can reduce to the log-contrast form above. For example, Equation 14 can be reparameterized as a model with any set of ALRs, for example, the ALRs with respect to the last part x_J:

$$y = b_0 + \sum_{j=1}^{J-1} b_j \log(x_j/x_J) + e, \tag{15.}$$

where $a_j = b_j$, $j = 0, 1, \ldots, J-1$ and $a_J = -(b_1 + b_2 + \cdots + b_{J-1})$.

In the **Supplemental Appendix**, further relationships between linear combinations of $(J-1)$ independent logratios and the log-contrast form are given, using the concept of a logratio pattern matrix, which is also useful in the back-transformation of logratios to the original compositional parts.

The more important issue in practice is that of the interpretation of the regression coefficients of a set of explanatory LRs in a model. To facilitate the expression of the effect sizes, Müller et al. (2018) proposed using logarithms to the base 2 so that a unit increase in the logratio corresponds to a doubling of the ratio. Since a change in an LR affects the values of other LRs, depending on the particular mix of explanatory LRs that are included, this can present difficulties in interpreting the effect sizes.

As an illustration, the same FA data set described up to now includes another variable, total lipids, in grams, denoted by y. This variable is regarded as a response variable and, after log-transformation, is modeled as a function of selected LRs of the 13 FAs $x_1, x_2, \ldots, x_{13}$, choosing the logratios in a stepwise fashion. The first LR selected, out of the pool of available $\frac{1}{2} \times 13 \times 12 = 78$ logratios, is $\log(x_8/x_4)$, the log of *20:1(n-9)/18:0*, explaining 74.3% of the variance of y. A second LR $\log(x_7/x_9)$ is then selected, the log of *18:4(n-3)/20:1(n-7)*, with an additional 9.0% variance explained, bringing the variance explained by the two logratios up to 83.3%. The

regression equation, using logs to the base 2 throughout, and showing p-values, is:

$$\mathrm{E}\left(\log_2(y)\right) = -0.349 \;+\; \underset{(p<10^{-5})}{0.495\,\log_2(x_8/x_4)} \;+\; \underset{(p<10^{-4})}{0.323\,\log_2(x_7/x_9)}. \qquad 16.$$

The two LRs involve different compositional parts, so the interpretation is fairly straightforward. A unit increase in the LR $\log_2(x_8/x_4)$, i.e., doubling the ratio, keeping $\log_2(x_7/x_9)$ fixed, implies an additive change in $\mathrm{E}\left(\log_2(y)\right)$ of 0.496—that is, a multiplicative change in the response of $2^{0.495} = 1.410$, or a 41.0% increase. The doubling of (x_8/x_4) would have an effect on the other parts (e.g., doubling x_8 with respect to the same value of x_4 would imply decreasing some of the other parts), so keeping (x_7/x_9) fixed can mean either that x_7 and x_9 have not changed or that x_7 is changing by the same factor as x_9.

Notice that negative regression coefficients on the explanatory variables can always be made positive, if preferred, by inversion of the ratios. For further details about the interpretation of logratios as explanatory variables, readers are directed to Coenders & Pawlowsky-Glahn (2020).

If an additional LR is added to the model to link $\{x_8, x_4\}$ with $\{x_7, x_9\}$ and thus form an acyclic connected graph, for example $\log(x_8/x_7)$, the estimated model is:

$$\mathrm{E}\left(\log_2(y)\right) = -1.192 + 0.415\,\log_2(x_8/x_4) + 0.656\,\log_2(x_7/x_9) + 0.241\,\log_2(x_8/x_7). \qquad 17.$$

The log-contrasts implied by Equations 16 and 17 are, respectively:

$$-0.495\,\log_2(x_4) + 0.323\,\log_2(x_7) + 0.495\,\log_2(x_8) - 0.323\,\log_2(x_9)$$

$$-0.415\,\log_2(x_4) + 0.416\,\log_2(x_7) + 0.655\,\log_2(x_8) - 0.656\,\log_2(x_9).$$

The latter log-contrast, corresponding to the model in Equation 17, would be the same if any set of three independent LRs of the subset of variables $\{x_4, x_7, x_8, x_9\}$ were used—also any set of three independent ILR balances, or any of the 4! sets of PLRs, or any three of the four CLRs.

The log-contrasts above both suggest that the effect is concentrated in raising parts x_7 and x_8 while lowering x_4 and x_9, suggesting the ratio of the amalgamation of the former pair of parts relative to the amalgamation of the latter pair. This simpler model, with just one logratio, is easier to interpret but does reduce the explanatory power compared with the models represented by Equations 16 and 17:

$$\mathrm{E}\left(\log_2(y)\right) = -0.704 + 0.909\,\log_2\left(\frac{x_7 + x_8}{x_4 + x_9}\right). \qquad 18.$$

Alternatively, the average of the two logratios as a single explanatory variable could be used:

$$\mathrm{E}\left(\log_2(y)\right) = -0.225 + 0.794\left[\left(\log_2\left(\frac{x_8}{x_4}\right) + \log_2\left(\frac{x_7}{x_9}\right)\right)/2\right]. \qquad 19.$$

The explanatory variable in Equation 19 is identical to the ILR balance of the parts $\{x_8, x_7\}$ in the numerator and $\{x_4, x_9\}$ in the denominator—see Equation 6, which has a scaling factor of 1 in this special case where $|J_1| = |J_2| = 2$.

In the same vein as the comment at the end of Section 5, a biochemist (M. Graeve, personal communication) was asked to intervene in the stepwise procedure, as an expert with domain knowledge of this data set. This exercise resulted in the second LR $\log_2(x_7/x_9)$ being chosen in the first step and a different LR chosen in the second step, $\log_2(x_{12}/x_2)$, the log of the ratio *16:0/22:1(n-9)*. Even though this entailed a loss of 3.1% in explained variance, it gave a model with a clearer biochemical interpretation. In this way domain knowledge can be combined with statistical criteria to arrive at a meaningful final model.

For large compositional data sets, when the above stepwise procedure is not viable, variable selection can be achieved using penalized regression methods (see, for example, Shi et al. 2016, Combettes & Müller 2019).

Procrustes correlation: measure of similarity of multivariate structure of two configurations of the same points

7. THE PROBLEM OF ZEROS

Up to now, the 13-part subcomposition of the original 40-part composition of FAs was used, where each part had an average occurrence at least 0.01 (1%). In this subcomposition there were no data zeros and hence no problem with the various logratio transformations. To analyze the complete 40-part data set, including mainly rarer FAs, a decision has to be made about the 187 zeros in this 42 × 40 data set, which are about 11% of the data. In this case, the zeros are due to values being below the detection limit and thus recorded as zero (Palarea-Albaladejo et al. 2007). A pragmatic choice is to base the decision on an assumption of the probability distribution of values near zero, for example, a triangular distribution from 0 to the smallest positive value $x_{\min}$, which gives an expected value equal to $\frac{2}{3}x_{\min}$ (Martín-Fernández et al. 2003). There are many iterative algorithms designed to substitute zeros, of various levels of sophistication (e.g., Martín-Fernández et al. 2012), summarized by Filzmoser et al. (2018, chapter 13).

The question is whether the particular method chosen makes any substantive difference to the eventual results of the data analysis. Since the substituted values will be small numbers, engendering large negative or positive logratios, the total logratio variance of the imputed matrices can be assessed across the alternatives. More specifically, the change induced in the multivariate structure of the data can be measured using the Procrustes correlation, since the distance structure of the samples is fundamental to all the results obtained subsequently. The function `protest` in the R package `vegan` (Oksanen et al. 2019) was used (Peres-Neto & Jackson 2001).

Table 3 shows some results for the simple $\frac{2}{3}x_{\min}$ method as well as three alternative iterative methods in the two R packages `zCompositions` (Palarea-Albaladejo & Martín-Fernández 2015) and `robCompositions` (Templ et al. 2011). Some of the methods broke down owing to the two FAs that had 40 and 39 zeros, respectively, out of the 42 samples, so these two parts were eliminated for this exercise, leaving a 42 × 38 matrix with 108 zeros (6.8% of the data). The total logratio variances are given down the diagonal, and off-diagonal are the Procrustes correlations between the logratio configurations of the samples in multivariate space. The methods lrDA (logratio data augmentation) and BDLs (below detection limits) engender large total logratio variances due to the high number of very small imputed values close to zero (see **Figure 8**) creating large logratios

Table 3 Logratio variances (on-diagonal) of zero-substituted tables and Procrustes correlations (off-diagonal) between their resultant multivariate structures

	(2/3)min[a]	lrDA[b]	lrEM[c]	BDLs[d]
(2/3)min	*0.351*	0.825	0.941	0.834
lrDA	0.825	*0.799*	0.851	0.798
lrEM	0.941	0.851	*0.455*	0.829
BDLs	0.834	0.798	0.829	*0.819*

[a] $\frac{2}{3}$ times minimum positive values of the respective parts.
[b] Logratio data augmentation algorithm, function `lrDA` in R package `zCompositions`.
[c] Logratio expectation-maximization algorithm, function `lrEM` in R package `zCompositions`.
[d] Stands for "below detection limits"; robust model-based procedure, function `BDLs`in R package `robCompositions`.
Two parts with 40 and 39 zeros (out of 42 samples) were removed for this exercise, as the three iterative methods failed to impute them. A total of 108 zeros were substituted, plotted in **Figure 8**.

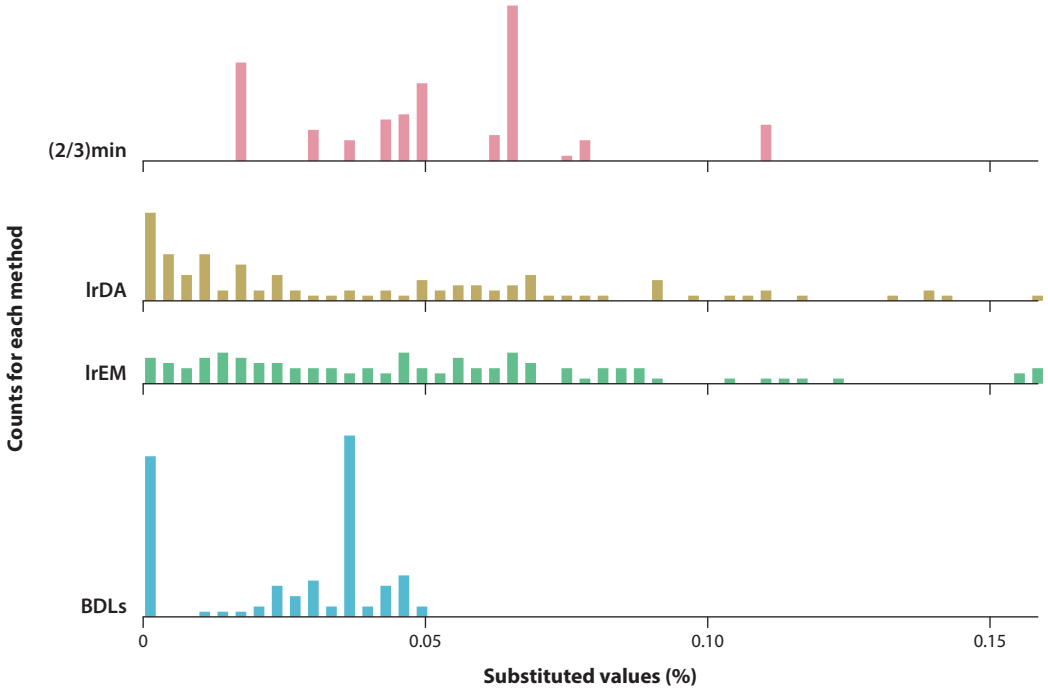

Figure 8

Histograms of the 108 zero substitutions made by four different algorithms. Abbreviations: (2/3)min, $\frac{2}{3}$ times minimum positive values of the respective parts; BDLs, below detection limits; lrDA, logratio data augmentation; lrEM, logratio expectation-minimization.

in absolute value (for example, 29 values less than 0.001% are substituted by the BDLs method). The highest concordance between the multivariate structures is between the simplest substitution method, (2/3)min, and the logratio expectation-minimization algorithm lrEM (Procrustes correlation =0.941). Some correlations are quite low for this type of comparison, where the configurations should be matched at a correlation of at least 0.9. Admittedly, these results apply only to this particular example, but this nevertheless shows that the method of zero substitution can have a strong effect on the structure of the compositional data set and influence its subsequent analysis. Hence, a sensitivity analysis of the results of statistical analyses using more than one replacement method is desirable.

8. WEIGHTING THE PARTS

When it comes to the joint analysis of compositional parts, a problem can arise that some parts excessively dominate the solution. A case in point is given by Greenacre (2018, 2019) in the context of an archaeometric data set of oxide compositions of Roman glass cups. Manganese oxide, the rarest oxide in the data, has the highest component of logratio variance, due to large ratios created between its small values. The relative error in the values of this oxide is extremely high and can distort any multivariate analysis performed on these data, which was pointed out in the original publication by Baxter et al. (1990).

Lewi (1976, 1986) realized the importance of weighting both the rows and the columns of a table of activity spectra of drug compounds, and defined what he called spectral mapping. In his

Weighted logratio distance between samples: can be computed between LRs or between CLRs (which have been computed with weights)

own words, Lewi (2005, p. 215) advocated weighting "to reduce the leverage of less important row and column items." Spectral mapping is the SVD of a double-centered matrix of log-transformed data, just like Equation 13, but using weighted centering on both the rows and columns and using a weighted SVD in the least-squares matrix approximation (Greenacre 2016a, 2018). The proposed weights were the marginal sums of the table, divided by their totals, which are exactly the weights used in correspondence analysis (CA) (Benzécri 1973, Greenacre 2016a) (see Section 9). Spectral mapping can thus be called weighted LRA and is a useful alternative when one wants to reduce the influence of rare parts that have high relative error (Mert et al. 2016). A compositional data set, once closed, will have equal row margins, so only the parts would have differential weighting. If different weights are required to be applied to the samples, for whatever reason, these can be easily introduced as well.

All the definitions given before of total logratio variance and of logratio distance can be very easily adapted to include this part weighting. For example, the logratio distance between samples given in Equation 11 has a weighted form, the weighted logratio distance between samples, for part weights $c_1, c_2, \ldots, c_J$, all positive and with $\sum_j c_j = 1$, as:

$$d_{ii'} = \sqrt{\sum\sum_{j<j'} c_j c_{j'} (z_{i,jj'} - z_{i',jj'})^2} = \sqrt{\sum_j c_j (y_{ij} - y_{ij'})^2}. \qquad 20.$$

The unweighted version in Equation 11 is then just a special case of this weighted definition, with weights $c_j = 1/J$ for all j.

Notice that the CLRs y_{ij} also have to be computed with weights on the parts: That is, in Equation 20, $y_{ij} = \log(x_{ij}) - \sum_{j'} c_{j'} \log(x_{ij'})$. The weights can be based on the means of the part values, as in spectral mapping, or on any prior knowledge about the relative errors of the parts. Also, parts originally with many zeros can be downweighted to reduce the sensitivity of the analysis to zero substitution. As the weights tend to zero for some parts, the influence of these parts diminishes and they become so-called supplementary, or passive, parts, using the concept from CA. That is, parts with zero weight play no role in the dimension reduction but can still be related to the solution afterward and visualized.

The issue of part weighting is similar to the problem of variable standardization. The Euclidean distance is particularly sensitive to the range of the component parts, and it is a good idea to investigate, for example, the variance–mean relationship in a data set before proceeding with multivariate analysis. Here expert knowledge is key to making a decision whether weighting should be introduced or not. When it comes to modeling, the issue is only important when the logratios are regarded as responses, not when they are explanatory variables.

9. SUBCOMPOSITIONAL INCOHERENCE, CORRESPONDENCE ANALYSIS, AND THE OMICS CHALLENGE

As a final section, alternative options that do not follow the ideal requirements of the logratio approach are considered, since these requirements can become restrictive when it comes to analyzing very large data sets, especially those emanating from microbiome and genetic research (Li 2015). These data sets are almost always compositional (Gloor & Reid 2016, Gloor et al. 2017) and typically involve hundreds or thousands of parts, usually with 50–80% of the data being zeros.

The underpinning principle of CoDA is that of subcompositional coherence, which recognizes that any compositional data set is actually a subcomposition and could be extended by extra parts, or certain parts could be excluded because of missing values or substantive considerations—in these cases, the compositional data change, but the ratios between the parts do not. The question

is whether this principle can be sacrificed to a measurably small extent in order to allow other methods to be used—for example, methods that do not have the zero problem.

Stress: a measure of difference between two distance matrices

Microbiome data set: a 490 × 335 matrix of counts, treated compositionally

9.1. Subcompositional Incoherence in Assessing Data Structure

Seeing that the multivariate structure of the parts is defined by the distances between them, Greenacre (2011b) considered several distance functions alternative to the logratio distance and defined a measure of subcompositional incoherence, in the same spirit as one might assess deviation from the ideal of normality, for example, or lack of model fit. This measure was based on comparing the matrix of interpart distances $\mathbf{D}$ based on the full composition at hand, with the matrix of interpart distances $\mathbf{D}_{[2]}$ based on 2-part subcompositions after closure. A measure of stress $S(\mathbf{D}, \mathbf{D}_{[2]})$, which is used in multidimensional scaling, was proposed, specifically the measure called stress-1 (Borg & Groenen 2010):

$$S(\mathbf{D}, \mathbf{D}_{[2]}) = \sqrt{\frac{\sum\sum_{j<j'}(d_{jj'} - d_{[2]jj'})^2}{\sum\sum_{j<j'}(d_{jj'})^2}}. \qquad 21.$$

For the logratio distance, these two matrices are identical, coherence is perfect, and incoherence, as measured by stress, is zero. For any other alternative distance measure between parts, however, the 2-part subcomposition, after dropping all other parts and reclosing, gives different distances. Clearly, the closed 2-part subcompositions are the worst case that an alternative distance measure can be subjected to. Using the above approach, and using a 11-part data set, the Euclidean distance used in PCA turned out to fare the worst, while the chi-square distance in CA (Greenacre 2016a, Nenadić & Greenacre 2007) fared the best (Greenacre 2011b; see also Jackson 1997, where CA is proposed as an alternative to the logratio approach).

This is no guarantee that this will be the same for other data sets, but there is a more compelling reason why CA might be an acceptable alternative approach. It has been demonstrated that, for strictly positive data, CA of power-transformed compositional data converges to LRA as the power decreases and tends to 0. The result is due, first, to the fact that both CA and LRA are SVDs of double-centered matrices, and second, to the Box-Cox transformation (for more details, see Greenacre 2009, 2010b; see also Stewart 2017). There are two ways that the power transform can be applied: First, it can be applied to the original data, in which case the convergence is to unweighted LRA (notice that the margins of the power-transformed table tend to constants at the limit). Second, it can be applied to the contingency ratios, which are the positive elements of the data matrix divided by their expected values using the row and column margins, in which case the original row and column weights in CA are conserved. In this latter case, weighted double-centering is used, and convergence is to Lewi's spectral map. Thus, for a compositional data set with constant row margins, the convergence is to weighted LRA, remembering once again that this is only true for strictly positive data.

For a compositional data set with zeros, the convergence of the unweighted or weighted form will start to break down as the power parameter descends to 0, but there could be a value of the power when incoherence is minimized. For example, a typical sparse microbiome data set is now considered, with hundreds of parts (bacteria) and more than 50% of the data values zero. The question is: For such a data set, how can the logratio approach and its ideal of subcompositional coherence be maintained?

Data set (bacterial counts in stool samples in a study on colorectal cancer). This data set (Baxter et al. 2016) results from 16S rRNA gene sequencing of stool samples of 490 patients in a study of colorectal cancer. A total of 335 bacterial operational taxonomic units (OTUs) were identified and

counted, and 58.7% of the data counts in the 490 × 335 matrix are zeros. The patients were classified into three groups: adenoma (a benign form of the tumor), cancer, and normal. The data should be regarded as compositional, since the total counts in each sample are not relevant. Several covariates are available, but here only the disease classification (three categories as dummy variables) and age (a continuous variable) are considered.

Canonical correspondence analysis (CCA): correspondence analysis, using chi-square distances, with explanatory variable constraints on the solution

After applying a fourth-root transformation of the data to bring the analysis closer to an analysis of logratios, and then closing the data, CA is applied, with a constraint on the solution that the dimensions be linearly related to the three dummy variables for disease and the variable age. This version of CA, called canonical correspondence analysis (CCA), is one of the most popular methods for analyzing sparse matrices of abundance data in ecology (ter Braak 1986).

The resultant ordination, shown in **Figure 9**, uses the contribution biplot scaling (Greenacre 2013), where the bacteria furthest out from the center have the highest contributions to the dimensions—only the most important contributors are shown, omitting those that have minor contributions. The horizontal dimension clearly separates the cancer group on the right, and the top 10 contributors to this dimension can be identified as OTUs {*260*, *310*, *105*, *281*, *264*, *297*, *057*,

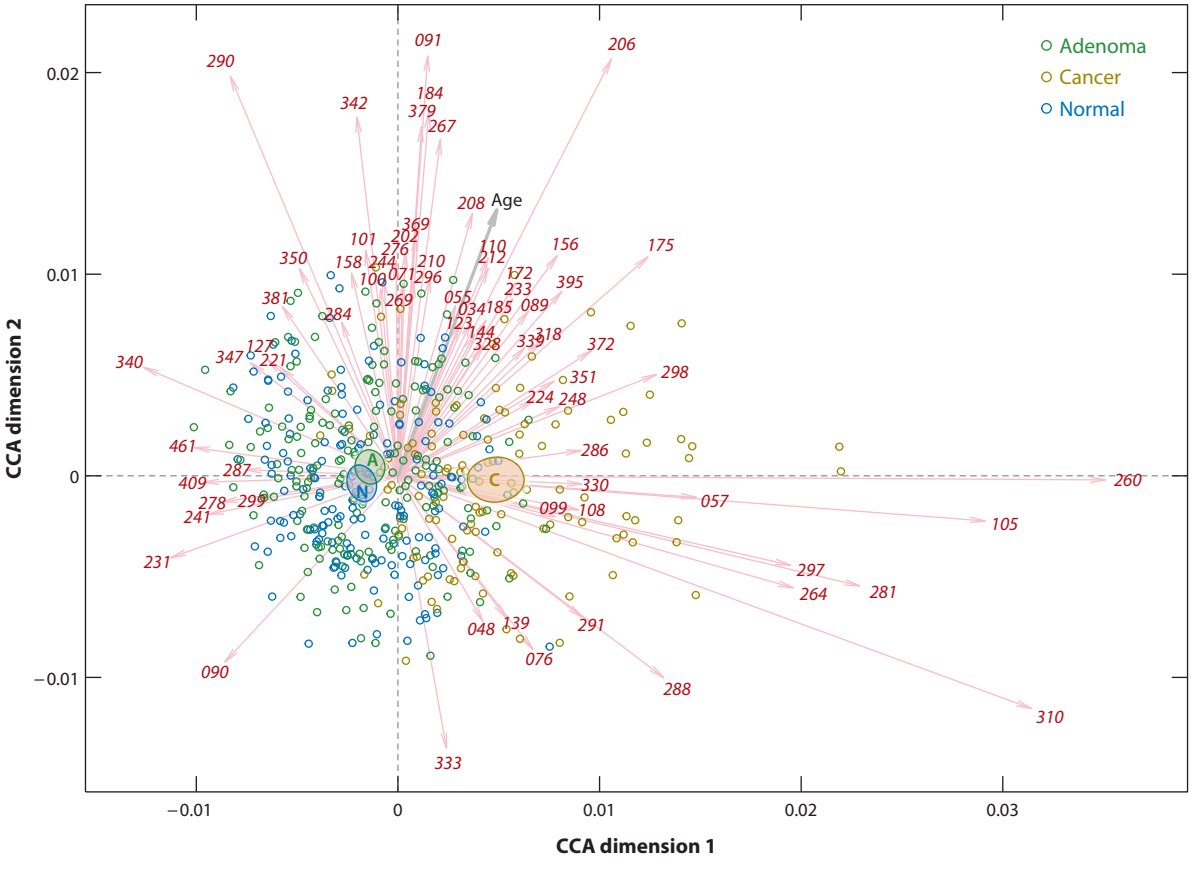

Figure 9

Canonical correspondence analysis (CCA) of relative bacterial counts constrained by three disease categories and age. Three ellipses labeled as A, N, and C are 99% confidence regions for the means of the three disease groups, adenoma, normal, and cancer.

288, *298*, *340*}, with OTU *340* being the only one with negative value, i.e., it is low in the cancer group. This list contains almost all of the OTUs identified as important by Baxter et al. (2016), who used random forests to predict cancer. The 99% confidence ellipse (Greenacre 2016b) for cancer is well separated from those for the adenoma and normal groups, which are overlapping. The solution is actually three-dimensional, and the **Supplemental Video** shows that the confidence ellipsoids of adenoma and normal in the 3D solution space do separate. The biplot axis of the variable Age lines up the three groups with cancer projected onto it as the oldest and normal as the youngest. This concords with the highly significant difference in age between the groups, with the normal group on average 10 years younger than the cancer group ($p < 10^{-12}$).

Supplemental Material >

All the above interpretation makes substantive sense, but the problem from the CoDA viewpoint is that CA, and its variant CCA, are not subcompositionally coherent. To study how incoherent the methods are, CA and CCA can be repeatedly performed for various subcompositions of the OTUs of different sizes, comparing the chi-square distances between the OTUs in each subcomposition with those in the full composition, using the stress measure in Equation 21. The comparison is made in the full 334-dimensional space of the data (i.e., the CA space). Notice that it makes no sense here to use 2-part subcompositions for such a large data set, where a worst-case scenario might rather be something like a subcomposition of 10% of the OTUs, not just two of them. Neither is it realistic to exclude OTUs with high frequencies, since in practice it is the rare OTUs that might be excluded from or added to a given composition. Hence, the extraction of a subcomposition should give a higher probability to the frequent parts being included.

The results of this exercise are given in **Figure 10**. For each percentage of the 335 parts, 100 subcompositions are selected at random, closed in each case, and then the interpart chi-square distances are compared with their corresponding ones in the full composition. The compositional data are not subject to any root transformation for this exercise. The stress-1 measure, often expressed as a percentage, shows a low value (median stress = 6.0%) for subcompositions of 10% of the parts (i.e., omitting 90% of the parts), going down to near zero (median stress = 0.3%) for subcompositions of 90% of the parts (i.e., omitting 10% of the parts). It seems that, at least in this example and using CA, the lack of subcompositional coherence is not a critical issue.

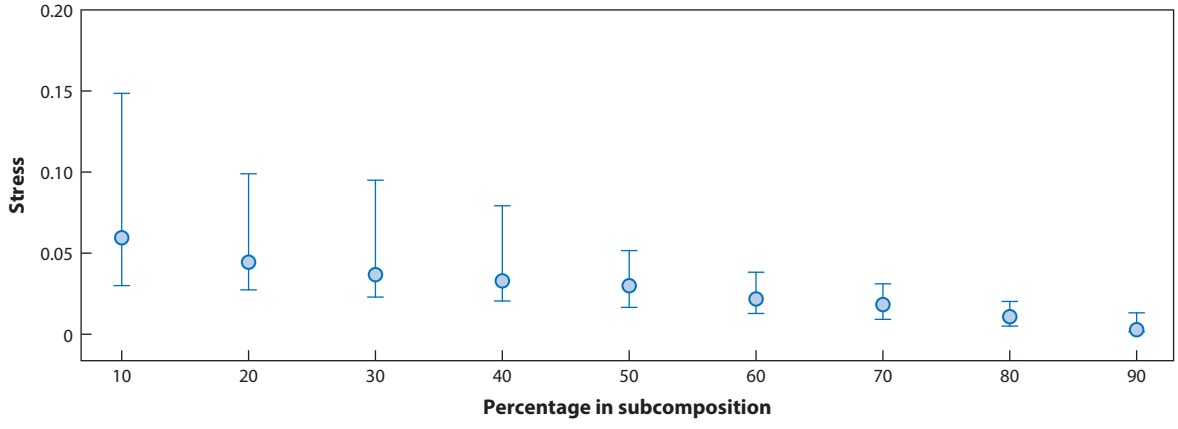

Figure 10

Subcompositional incoherence in the chi-square distances measured by stress: For each percentage of parts in the subcomposition, the 2.5% to 97.5% estimated quantiles are plotted, and the median is represented by the circle.

9.2. Subcompositional Incoherence in Modeling

In a similar exercise, the probability of being a cancer patient was modeled in a logistic regression as a function of the 10 highly contributing OTUs identified in the CA. The relative abundances were square-root transformed, which is a regular transformation of proportional data. Two of these OTUs turned out to be significant, *OTU310* and *OTU105*, with the following linear regression equation for the logit of probability p of cancer:

$$\text{logit}(p) = -1.67 + 6.92\,OTU310^{0.5} + 19.16\,OTU105^{0.5}. \qquad 22.$$
$$(p = 0.0002) \qquad (p < 10^{-4})$$

Then 99 random subcompositions of 50% of the OTUs were generated, discarding 167 OTUs each time and reclosing the subcomposition, with the condition that the above two OTUs were always included. The logistic regression equation was re-estimated for each subcomposition of the data, and **Figure 11** shows how the regression coefficients varied as well as the associated p-values. The variation of the regression coefficients as well as the p-values is relatively small. It is clear that if any 50% subcomposition of the data were analyzed, the results would be essentially the same as analyzing the full composition, with the regression coefficients varying well within the confidence interval of the coefficients estimated in the full composition. As before, it seems that the incoherence is low and does not affect the overall results of the modeling.

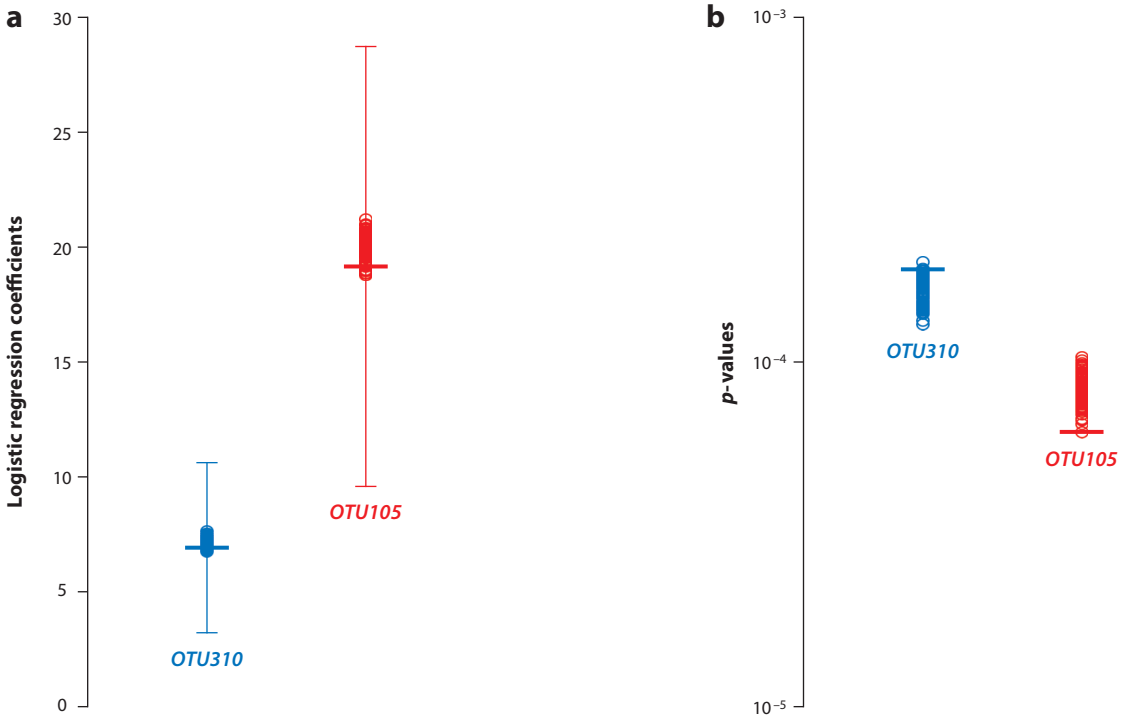

Figure 11

Subcompositional incoherence in estimation of coefficients in logistic regression in random 50% subcompositions. (*a*) The regression coefficients (*overlapping dots*), also showing the original estimated coefficients (*wider horizontal lines*) and the extent of the 95% confidence interval for the original estimates in the full 335-part composition. (*b*) The p-values (*overlapping dots*), also showing the original p-values (wider horizontal lines) when the two variables were in the full composition.

SUMMARY POINTS

1. Compositional data are multivariate data for which the relative values are important, not their absolute values. The components, or parts, constituting a compositional sample are usually expressed as proportions summing to 1 or, equivalently, as percentages summing to 100%.
2. The fundamental principle underlying compositional data analysis is that of subcompositional coherence: Relationships between parts should remain constant irrespective of the other parts present in the composition.
3. Ratios of parts are subcompositionally coherent, and the basic transformation of compositional data is the logarithm of part ratios, or logratios. Once transformed, statistical analysis, both univariate and multivariate, as well as statistical modeling, can proceed very much as before, while taking into account the relationship between the logratios and the original compositional values.
4. Important logratios, which effectively drive the multivariate structure of a compositional data set, can be identified and interpreted, preferably in collaboration with a researcher who has domain knowledge. Variable selection helps to simplify the interpretation of the data structure.
5. Zeros in compositional data are a major issue: Various zero substitution strategies have been proposed, and the impact of any one of these on the eventual results should be investigated.
6. Since some parts can excessively dominate a data set due to high variance of the logratios they engender, a differential weighting of the parts can be considered.
7. Other approaches that do not use logratios can be investigated, for example, in the analysis of large sparse data sets where the problem of zeros is acute. The deviation from subcompositional coherence, i.e., incoherence, can be measured for the data set as a whole by comparing interpart distances in the original compositional data set with their distances in subcompositions of varying sizes.
8. Likewise, if the part values themselves are used as explanatory variables in a statistical model, rather than their logratios, then the effects on the model can be quantified by repeating the estimation procedure using various random subcompositions.

FUTURE ISSUES

1. Compositional data are ubiquitous, and there will be increasingly more applications in the future, especially in the area of microbiome and genetic research. Practitioners should be aware that these data constitute a special case in statistics and need careful treatment.
2. For data sets with very many compositional parts, the issue of variable selection is of the highest importance. Signal has to be separated from the high level of noise in these data because of the high variability in such data as well as high measurement error.
3. Zeros in compositional data are the Achilles heel of the logratio approach. Present methods of zero substitution need to be critically examined and compared. When many zeros

are substituted in an application, a sensitivity analysis should be obligatory. Other ways of dealing with zeros that arise due to measuring instruments not being able to detect very low values can be investigated, for example, using measurement error models.

4. Alternative approaches to using logratios can be investigated more widely, especially those that admit data zeros, using the general idea of subcompositional incoherence to measure deviation from the ideal approach using logratios.
5. The cases of ordered parts and parts that form natural hierarchies need to be investigated. The use of amalgamated parts can be especially useful when groupings are predetermined based on domain knowledge and the research question.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

I express my sincere gratitude to many colleagues who have communicated with me about issues in CoDA: Germà Coenders and Ionas Erb (Catalonia), Eric Grunsky (Canada), John Bacon-Shone (Hong Kong), and Thom Quinn (Australia), as well as various teams that I am working with at the moment on CoDA projects: in Spain, headed by Jaume Bertranpetit (Barcelona) and Agustín Blasco (Valencia); in Germany, headed by Martin Graeve; and in South Korea, headed by Taesung Park (Seoul). A special thanks to Antonio Garrido of Sevilla, who has been of immense help in giving me feedback about my software package `easyCODA`. All these contacts have enriched my knowledge and experience.

LITERATURE CITED

Aitchison J. 1981. A new approach to null correlations of proportions. *Math. Geol.* 13:175–89

Aitchison J. 1982. The statistical analysis of compositional data (with discussion). *J. R. Stat. Soc. Ser. B* 44:139–77

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

Aitchison J. 1986. *The Statistical Analysis of Compositional Data*. London: Chapman and Hall

Aitchison J. 1990. Relative variation diagrams for describing patterns of variability of compositional data. *Math. Geol.* 22:487–512

Aitchison J. 1997. The one-hour course in compositional data analysis, or compositional data analysis is simple. In *Proceedings of IAMG'97, the Third Annual Conference of the International Association for Mathematical Geology*, ed. V Pawlowsky-Glahn, pp. 3–35. Barcelona: CIMNE

Aitchison J. 2005. A concise guide to compositional data analysis. In *Proceedings of CoDaWork05*. **http://ima.udg.edu/Activitats/CoDaWork05/A_concise_guide_to_compositional_data_analysis.pdf**

Aitchison J. 2008. *The single principle of compositional data analysis, continuing fallacies, confusions and misunderstandings and some suggested remedies*. Keynote address presented at CoDaWork08, Girona, Spain, May 27–30. **https://core.ac.uk/download/pdf/132548276.pdf**

Aitchison J, Bacon-Shone J. 1984. Log contrast models for experiments with mixtures. *Biometrika* 71:323–30

Aitchison J, Greenacre M. 2002. Biplots for compositional data. *J. R. Stat. Soc. Ser. A* 51:375–92

Aitchison J, Shen SM. 1984. Measurement error in compositional data. *Math. Geol.* 16:637–50

Bacon-Shone J. 2011. A short history of compositional data analysis. In *Compositional Data Analysis: Theory and Applications*, ed. V Pawlowsky-Glahn, A Buccianti, pp. 3–11. New York: Wiley

Baxter MJ, Cool HEM, Heyworth MP. 1990. Principal component and correspondence analysis of compositional data: some similarities. *J. Appl. Stat.* 17:229–35
Baxter N, Ruffin M, Rogers M, Schloss P. 2016. Microbiota-based model improves the sensitivity of fecal immunochemical test for detecting colonic lesions. *Genom. Med.* 8:37
Benzécri J-P. 1973. *Analyse des Données*, Tôme 1: *L'Analyse des Correspondances*. Paris: Dunod
Borg I, Groenen PJF. 2010. *Modern Multidimensional Scaling: Theory and Applications*. New York: Springer. 2nd ed.
Coenders G, Pawlowsky-Glahn V. 2020. On interpretations of tests and effect sizes in regression models with a compositional predictor. *SORT* 20:201–20
Combettes PL, Müller CL. 2019. Regression models for compositional data: general log-contrast formulations, proximal optimization, and microbiome data applications. arXiv:1903.01050v1 [math.ST]
Egozcue JJ, Pawlowsky-Glahn V. 2005. Groups of parts and their balances in compositional data analysis. *Math. Geol.* 37:795–828
Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C. 2003. Isometric logratio transformations for compositional data analysis. *Math. Geol.* 35:279–300
Erb I, Notredame C. 2016. How should we measure proportionality on relative gene expression data? *Theory Biosci.* 135:21–36
Faes C, Molenberghs G, Hens N, Muller A, Goossens H, Coenen S. 2011. Analysing the composition of outpatient antibiotic use: a tutorial on compositional data analysis. *J. Antimicrob. Chemother.* 66:vi89–94
Filzmoser P, Hron K, Templ M. 2018. *Applied Compositional Data Analysis*. Oxford, UK: Oxford Univ. Press
Gittins R. 1985. *Canonical Analysis: A Review with Applications in Ecology*. Berlin: Springer-Verlag
Gloor GB, MacKlaim JM, Pawlowsky-Glahn V, Egozcue JJ. 2017. Microbiome datasets are compositional: and this is not optional. *Front. Microbiol.* 8:2224
Gloor GB, Reid G. 2016. Compositional analysis: a valid approach to analyze microbiome high-throughput sequencing data. *Can. J. Microbiol.* 62:692–703
Gower J, Dijksterhuis GB. 2004. *Procrustes Problems*. New York: Springer
Graeve M, Greenacre M. 2020. The selection and analysis of fatty acid ratios: a new approach for the univariate and multivariate analysis of fatty acid trophic markers in marine organisms. *Limnol. Oceanogr. Methods* 18:196–210
Greenacre M. 2009. Power transformations in correspondence analysis. *Comput. Stat. Data Anal.* 53:3107–16
Greenacre M. 2010a. *Biplots in Practice*. Bilbao, Spain: BBVA Found. **https://www.multivariatestatistics.org**
Greenacre M. 2010b. Log-ratio analysis is a limiting case of correspondence analysis. *Math. Geosci.* 42:129–34
Greenacre M. 2011a. Compositional data and correspondence analysis. In *Compositional Data Analysis: Theory and Applications*, ed. V Pawlowsky-Glahn, A Buccianti, pp. 104–13. New York: Wiley
Greenacre M. 2011b. Measuring subcompositional incoherence. *Math. Geosci.* 43:681–93
Greenacre M. 2013. Contribution biplots. *J. Comput. Graph. Stat.* 22:107–22
Greenacre M. 2016a. *Correspondence Analysis in Practice*. Boca Raton, FL: Chapman and Hall/CRC Press
Greenacre M. 2016b. Data reporting and visualization in ecology. *Polar Biol.* 39:2189–205
Greenacre M. 2017. 'Size' and 'shape' in the measurement of multivariate proximity. *Methods Ecol. Evol.* 8:1415–24
Greenacre M. 2018. *Compositional Data Analysis in Practice*. Boca Raton: Chapman and Hall/CRC
Greenacre M. 2019. Variable selection in compositional data analysis using pairwise logratios. *Math. Geosci.* 51:649–82
Greenacre M. 2020. Amalgamations are valid in compositional data analysis, can be used in agglomerative clustering, and their logratios have an inverse transformation. *Appl. Comput. Geosci.* 5:100017
Greenacre M, Grunsky E, Bacon-Shone J. 2020. A comparison of amalgamation and isometric logratios in compositional data analysis. *Comput. Geosci.* In press. **https://doi.org/10.1016/j.cageo.2020.104621**
Greenacre M, Lewi PJ. 2009. Distributional equivalence and subcompositional coherence in the analysis of compositional data, contingency tables and ratio-scale measurements. *J. Classif.* 26:29–54
Hron K, Filzmoser P, de Caritat P, Fiserova E, Gardlo A. 2017. Model-based replacement of rounded zeros in compositional data: classical and robust approaches. *Math. Geosci.* 49:797–814
Jackson DA. 1997. Compositional data in community ecology: the paradigm or peril of proportions? *Ecology* 78:928–40

Krzanowski WJ. 1987. Selection of variables to preserve multivariate data structure, using principal components. *J. R. Stat. Soc. Ser. A* 36:22–33
Lewi PJ. 1976. Spectral mapping, a technique for classifying biological activity profiles of chemical compounds. *Arz. Forsch.* 26:1295–300
Lewi PJ. 1986. Analysis of biological activity profiles by Spectramap. *Eur. J. Med. Chem.* 21:155–62
Lewi PJ. 2005. Spectral mapping, a personal and historical account of an adventure in multivariate data analysis. *Chem. Intell. Lab. Syst.* 77:215–23
Li H. 2015. Microbiome, metagenomics and high-dimensional compositional data analysis. *Annu. Rev. Stat. Appl.* 2:73–94
Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J. 2015. Proportionality: a valid alternative to correlation for relative data. *PLOS Comput. Biol.* 11(3):e1004075
Martín-Fernández JA, Barceló-Vidal C, Pawlowsky-Glahn V. 2003. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. *Math. Geol.* 35:253–78
Martín-Fernández JA, Hron K, Templ M, Filzmoser P, Palarea-Albaladejo J. 2012. Model-based replacement of rounded zeros in compositional data: classical and robust approaches. *Comput. Stat. Data. Anal.* 56:2688–704
Mert C, Filzmoser P, Hron K. 2016. Error propagation in compositional data analysis: theoretical and practical considerations. *Math. Geosci.* 48:941–61
Mosimann JE. 1962. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika* 49:65–82
Müller I, Hron K, Fišerová E, Šmahaj J, Cakirpaloglu P, Vančaková J. 2018. Interpretation of compositional regression with application to time budget analysis. *Austrian J. Stat.* 47:3–19
Nenadić O, Greenacre M. 2007. Correspondence analysis in R, with two- and three-dimensional graphics: the ca package. *J. Stat. Softw.* 20. **http://dx.doi.org/10.18637/jss.v020.i03**
Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, et al. 2019. vegan: community ecology package. *R package*, version 2.5-6. **https://CRAN.R-project.org/package=vegan**
Palarea-Albaladejo J, Martín-Fernández JA. 2015. zCompositions—R package for multivariate imputation of left-censored data under a compositional approach. *Chemometr. Intell. Lab.* 143:85–96
Palarea-Albaladejo J, Martín-Fernández JA, Gómez-García J. 2007. A parametric approach for dealing with compositional rounded zeros. *Math. Geol.* 39:625–45
Pawlowsky-Glahn V, Buccianti A. 2011. *Compositional Data Analysis: Theory and Applications*. New York: Wiley
Pearson K. 1897. Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurements of organs. *Proc. R. Soc.* 60:489–98
Peres-Neto PR, Jackson DA. 2001. How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia* 129:169–78
Quinn TP, Erb I. 2020. Amalgams: data-driven amalgamation for the reference-free dimensionality reduction of zero-laden compositional data. bioRxiv 968677. **https://www.biorxiv.org/content/10.1101/2020.02.27.968677v1**
Quinn TP, Erb I, Richardson MF, Crowley TM. 2018. Understanding sequencing data as compositions: an outlook and overview. *Bioinformatics* 34:2870–78
Quinn TP, Richardson MF, Lovell D, Crowley TM. 2017. propr: an R-package for identifying proportionally abundant features using compositional data analysis. *Sci. Rep.* 7:16252
R Core Team. 2020. R: a language and environment for statistical computing. *Statistical Software*, R Found. Stat. Comput., Vienna
Shi P, Zhang A, Li H. 2016. Regression analysis for microbiome compositional data. *Ann. Appl. Stat.* 10:1019–40
Søreide JE, Leu E, Berge J, Graeve M, Falk-Petersen S. 2010. Timing of blooms, algal food quality and *Calanus glacialis* reproduction and growth in a changing Arctic. *Glob. Change Biol.* 16:3154–63
Stewart C. 2017. An approach to measure distance between compositional diet estimates containing essential zeros. *J. Appl. Stat.* 44:1137–52
Templ M, Hron K, Filzmoser P. 2011. robCompositions: an R-package for robust statistical analysis of compositional data. In *Compositional Data Analysis: Theory and Applications*, ed. V Pawlowsky-Glahn, A Buccianti, pp. 341–55. New York: Wiley

ter Braak C. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* 67:1167–79

Tolosana-Delgado R, van den Boogaart KG. 2011. Linear models with compositions in R. In *Compositional Data Analysis: Theory and Applications*, ed. V Pawlowsky-Glahn, A Buccianti, pp. 356–71. New York: Wiley

Tsilimigras MCB, Fodor AA. 2016. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann. Epidemiol.* 26:330–35

van den Boogaart KG, Tolosana-Delgado R. 2013. *Analyzing Compositional Data with R*. Berlin: Springer-Verlag

van den Wollenberg AL. 1977. Redundancy analysis, an alternative for canonical analysis. *Psychometrika* 42:207–19

Zuur AF, Ieno EN, Smith GM. 2007. *Analysing Ecological Data*. New York: Springer