Multivariate Analysis of Ecological Data

MICHAEL GREENACRE
Professor of Statistics at the Pompeu Fabra University in Barcelona, Spain

RAUL PRIMICERIO
Associate Professor of Ecology, Evolutionary Biology and Epidemiology
at the University of Tromsø, Norway

Chapter 14 Offprint

Compositional Data and Log-ratio Analysis

First published: December 2013
ISBN: 978-84-92937-50-9

Supporting websites:
www.fbbva.es
www.multivariatestatistics.org

Fundación **BBVA**

Chapter 14

Compositional Data and Log-ratio Analysis

We have already met compositional data in the form of row or column profiles in CA: these are sets of nonnegative values that add up to a constant, usually 1 or 100%. In CA the profiles are computed on data matrices of abundances or biomasses, for example by dividing by their respective row and column totals. In other contexts the original data are compositional, for example chemical or geological data where the total size of the sample, measured in units of weight or volume, is not relevant, just its decomposition into a set of components. Another example of compositional data in biology is that of fatty acid compositions in studies of marine food webs. Compositional data are special because in their original form they have the property of closure, that is the compositional values of each sample have a constant sum. There are particular methodological issues when analysing compositional data, such as subcompositional coherence and the log-ratio transformation, which we shall consider in this chapter. Although this chapter is specific to compositional data, the wider issue of rare observations is discussed and the value of the contribution biplot is again demonstrated.

Contents

Compositional data and subcompositions

To illustrate the main reason why compositional data are a special case, consider the data in Exhibit 14.1. First, there is composition consisting of four fatty acids measured in six samples, with their components adding up to 1. Second, the last component is eliminated and the composition is *closed* again, that is re-expressed as proportions that sum to 1: this is called a s*ubcomposition* of the original com-

Exhibit 14.1: *Compositional data matrix (a) and a subcomposition (b), after eliminating the last component*

(a)

	16:1(n-7)	*20:5(n-3)*	*18:4(n-3)*	*18:00*	*Sum*
B6	0.343	0.217	0.054	0.387	*1*
B7	0.240	0.196	0.050	0.515	*1*
D4	0.642	0.294	0.039	0.025	*1*
D5	0.713	0.228	0.020	0.040	*1*
H5	0.177	0.351	0.423	0.050	*1*
H6	0.209	0.221	0.511	0.059	*1*

(b)

	16:1(n-7)	*20:5(n-3)*	*18:4(n-3)*	*Sum*
B6	0.559	0.353	0.088	*1*
B7	0.494	0.403	0.103	*1*
D4	0.658	0.302	0.040	*1*
D5	0.742	0.237	0.021	*1*
H5	0.186	0.369	0.445	*1*
H6	0.222	0.235	0.543	*1*

position. If researcher A works with the data in Exhibit 14.1(a) and researcher B with the data in Exhibit 14.1(b) and they consider it interesting to compute correlations as a way of measuring association between the components, they will obtain the results in Exhibit 14.2(a) and 14.2(b) respectively. While researcher A finds that the correlations between fatty acid *18:4(n-3)* and the pair *16:1(n-7)* and *20:5(n-3)* are –0.671 and 0.357 respectively, researcher B finds that they are –0.952 and –0.139. There is clearly a paradox here – the relationship between two

Exhibit 14.2: *Correlations between the columns of the compositional data matrices in Exhibit 14.1*

(a)

	16:1(n-7)	*20:5(n-3)*	*18:4(n-3)*	*18:00*
16:1(n-7)	1	–0.038	–0.671	–0.379
20:5(n-3)	–0.038	1	0.357	–0.604
18:4(n-3)	–0.671	0.357	1	–0.407
18:00	–0.379	–0.604	–0.407	1

(b)

	16:1(n-7)	*20:5(n-3)*	*18:4(n-3)*
16:1(n-7)	1	–0.171	–0.952
20:5(n-3)	–0.171	1	–0.139
18:4(n-3)	–0.952	–0.139	1

components should be the same and not depend on whether another component (*18:00*) is present or not. We say that the correlation does not have the property of *subcompositional coherence* – it is incoherent.

The log-ratio transformation

Values that are constant in a composition and any of its subcompositions are the ratios between components. For example, consider the four-part composition $[a,b,c,d]$ with $a+b+c+d=1$, and a three-part closed subcomposition $[a,b,c]/\ (a+b+c)$. Then the ratio a/b in the composition is identical to the ratio $[a/(a+b+c)]\ /\ [b/(a+b+c)]$ in the subcomposition. Since ratios are generally compared multiplicatively rather than additively, the logarithms of the ratios provide a justifiable transformation of the compositional data and have subcompositional coherence. Exhibit 14.3(a) shows the log-ratios $\log(a/b)$ for all six pairs of components a and b in Exhibit 14.1(a), as well as their means and standard deviations. In addition, a distance d_{ab} between the two components a and b is calculated as the square root of the average sum of squares of log-ratios across the samples:

$$d_{ab} = \sqrt{\sum_{i=1}^{n} (1/n)\,[\log(a_i/b_i)]^2} = \sqrt{\sum_{i=1}^{n} (1/n)\,[\log(a_i) - \log(b_i)]^2} \qquad (14.1)$$

Exhibit 14.3: *Logarithms of ratios between all pairs of components and the root mean sum of squares of the log-ratios as a measure of proximity*

(a)

	LOG-RATIOS					
	16:1(n-7) / 20:5(n-3)	*16:1(n-7) / 18:4(n-3)*	*16:1(n-7) / 18:00*	*20:5(n-3) / 18:4(n-3)*	*20:5(n-3) / 18:00*	*18:4(n-3) / 18:00*
B6	0.458	1.849	–0.121	1.391	–0.579	–1.969
B7	0.203	1.569	–0.764	1.366	–0.966	–2.332
D4	0.781	2.801	3.246	2.020	2.465	0.445
D5	1.140	3.574	2.881	2.434	1.740	–0.693
H5	–0.685	–0.871	1.264	–0.187	1.949	2.135
H6	–0.056	–0.894	1.265	–0.838	1.321	2.159
mean	0.307	1.338	1.295	1.031	0.988	–0.043
sd	0.643	1.861	1.585	1.278	1.418	1.960
distance	0.662	2.162	1.942	1.557	1.629	1.790

(b)

	DISTANCE(LOG-RATIOS)			
	16:1(n-7)	*20:5(n-3)*	*18:4(n-3)*	*18:00*
16:1(n-7)	0	0.662	2.162	1.942
20:5(n-3)	0.662	0	1.557	1.629
18:4(n-3)	2.162	1.557	0	1.790
18:00	1.942	1.629	1.790	0

Definition (14.1) shows that this distance function is simply a Euclidean distance between the log-transformed components. In Exhibit 14.3(b) the distances have been gathered into a square matrix, which can be used in a cluster analysis or an MDS. If fatty acid *18:00* is removed and the distance function is applied to Exhibit 14.1(b), the distances between the three components of the subcomposition remain identical, hence this measure of distance between the components is subcompositionally coherent.

The “fatty acid” data set

Exhibit 14.4 shows a part of the data set “fatty acid”, compositional data on 25 fatty acids from 42 copepods of the species *Calanus glacialis*. The copepods were sampled in three different seasons and the objective is to see how the fatty acid compositions relate to these different seasons. Notice that the components with higher means also have higher standard deviations, which is typical of such data, as it is for count data. In the case of CA, the chi-square distance compensates for this disparity in variances. There is a similar issue in log-ratio analysis, which we describe now.

Log-ratio analysis

Log-ratio analysis (LRA) is the analogue of PCA that visualizes the compositional variables (also called components) transformed to log-ratios – hence it has the property of subcompositional coherence, which neither PCA nor CA have. It is a simple adaptation of PCA and has two forms: an unweighted form and a weighted form. We restrict our discussion to weighted LRA since the weighting has a number of benefits.

Notice in the last line of Exhibit 14.4 “mean(LR)^2”, the mean of the squares of the log-ratios in each column. Just as we did for each row of the mini-example

Exhibit 14.4:
Part of 42 × 25 data matrix of fatty acid compositions, expressed as percentages: each set of 25 values in the rows sums to 100%. The mean and standard deviation of each column is given, as well as the mean of the squares of log-ratios for pairs of samples in each column

	14:00	*i-15:0*	*15:00*	*16:00*	*16:1(n-7)*	···	*22:5(n-3)*	*22:6(n-3)*	*Total*
B5	14.229	1.223	0.870	12.204	6.567	···	0.543	0.446	*100*
B6	12.153	1.270	1.085	12.318	7.406	···	0.353	0.469	*100*
B7	6.640	0.790	0.529	12.272	6.804	···	0.656	0.231	*100*
B8	12.410	1.167	0.822	11.543	7.668	···	0.425	0.436	*100*
H5	6.764	0.338	0.272	8.056	6.207	···	0.298	0.464	*100*
H6	6.896	0.324	0.262	8.046	6.494	···	0.313	0.520	*100*
⋮	⋮	⋮	⋮	⋮	⋮	···	⋮	⋮	⋮
E5	5.410	0.407	0.273	12.321	6.622	···	0.273	0.257	*100*
E6	9.200	0.813	0.606	9.741	19.193	···	0.542	0.601	*100*
mean	8.366	0.678	0.546	9.196	12.818	···	0.640	9.100	*100*
sd	2.131	0.277	0.181	1.816	8.263	···	0.251	2.715	
mean(LR)^2	0.114	0.321	0.252	0.080	0.664		0.811	0.142	

in Exhibit 14.3, so we can compute log-ratios between all the 42 values in each column (there will be ½ × 42 × 41 = 861 ratios in total), which would be the basis for a distance calculation between pairs of samples. The mean square of these log-ratios has the property that it will be higher for rarer components, which can have bigger ratios than those between components at a higher level. For example, a rare component with mean 0.03% could easily have two values of 0.05% and 0.01%, which gives a ratio of 5, whereas such a large ratio would hardly ever occur for a component with values of the order of 10%, varying between 6% and 14%, say. In weighted log-ratio analysis this effect is compensated for by assigning weights to each component proportional to its mean, so that rarer components get smaller weights. This is exactly the same idea as in CA.

Technically, weighted LRA can also be thought of as a two-step procedure, performing MDS on inter-sample distances based on the log-ratios, where the components have been weighted as just described, and then adding the component variables by regression on the MDS dimensions. But more simply, it reduces to a PCA of the log-transformed data matrix which is centred row-wise, that is each row of the logged data is centred to have mean zero. This centred matrix is then subject to PCA, incorporating the column weights. Because PCA will then automatically centre the data column-wise, it follows that the log-transformed compositional data matrix is actually double-centered, row-wise and column-wise. Notice that the actual log-ratios for all pairs of components do not have to be calculated, thanks to the double-centering. LRA is thus a weighted PCA of the previously log-transformed and row-centered data, with some special features of the interpretation.

Interpretation of log-ratio analysis

Exhibit 14.5 shows the weighted LRA of the "fatty acid" data set, with samples in principal coordinates (thus approximating the log-ratio distances between them) and fatty acids in standard coordinates. Separately, we have verified that only the first two dimensions are significant. There are three clearly separated groups of samples, which we have labelled A, B and C, coinciding exactly with the three seasons in which they were sampled. As in Exhibit 13.5 we have separated the higher than average contributors to the first two dimensions from the others: these seven fatty acids are thus indicated with larger labels, and account for 90% of the variance in this biplot. A novelty of the log-ratio biplot is that it is not the vectors from the origin that define the biplot axes, but the vectors linking the component variables – these vectors are called *links*. For example, the link from *16:1(n-7)* at bottom left to *18:00* top left represents the log-ratio log(*18:00*/*16:1(n-7)*), and the direction of this link is exactly lining up with group A at the top and group C at the bottom. Similarly, the link from *18:00* to *18:4(n-3)*, as well as several others made by the group of three high-contributing fatty acids in-between, separates group A from group B. And the link from *16:1(n-7)* and *18:4(n-3)* is one that

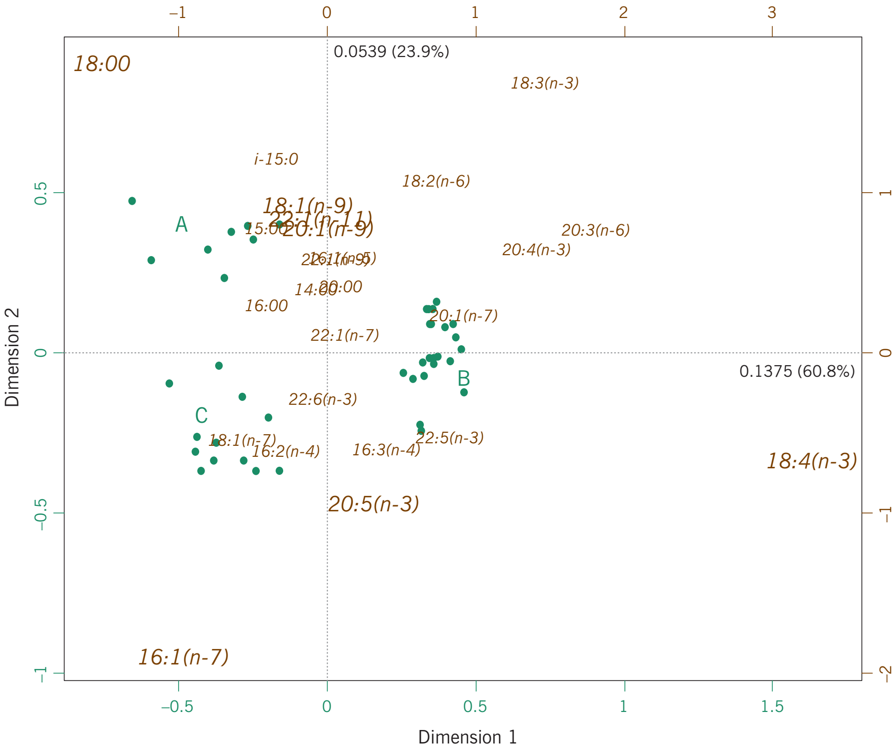

Exhibit 14.5: *Row-principal LRA biplot of "fatty acid" data set. 84.7% of the log-ratio variance is explained. The seven higher-than-average contributing fatty acids are shown in larger font. Notice the different scales for sample points and fatty acid points*

separates group C from group B. Exhibit 14.6 illustrates the group separation in a simpler scatterplot of two of these log-ratios that are suggested by these results. Since the third log-ratio log[*18:4(n-3)/18:00*] separating groups A and B is the horizontal axis of the scatterplot minus the vertical one, it can be depicted by a 45 degrees descending line, shown by the dashed arrow, perfectly coinciding with the separation of the A and B samples.

An interesting feature of the log-ratio biplot is that if components fall on straight lines (as, for example, *18:00*, *18:4(n-3)* and the group of three fatty acids inbetween, *18:1(n-9)*, *22:1(n-11)* and *20:1(n-9)* in Exhibit 14.5) then a model can be deduced between them. The Bibliographical Appendix gives a reference to this way of diagnosing models in biplots.

Relationship between CA and LRA

CA also analyses compositions, albeit compositions (i.e., profiles) computed on a matrix of counts or abundances. In fact, CA can be used to analyse purely compositional data, and likewise, LRA can be used to analyse count data or other strictly positive ratio-scale data. There is an interesting relationship between the two methods: leaving out some technical details, the main result is that if one ap-

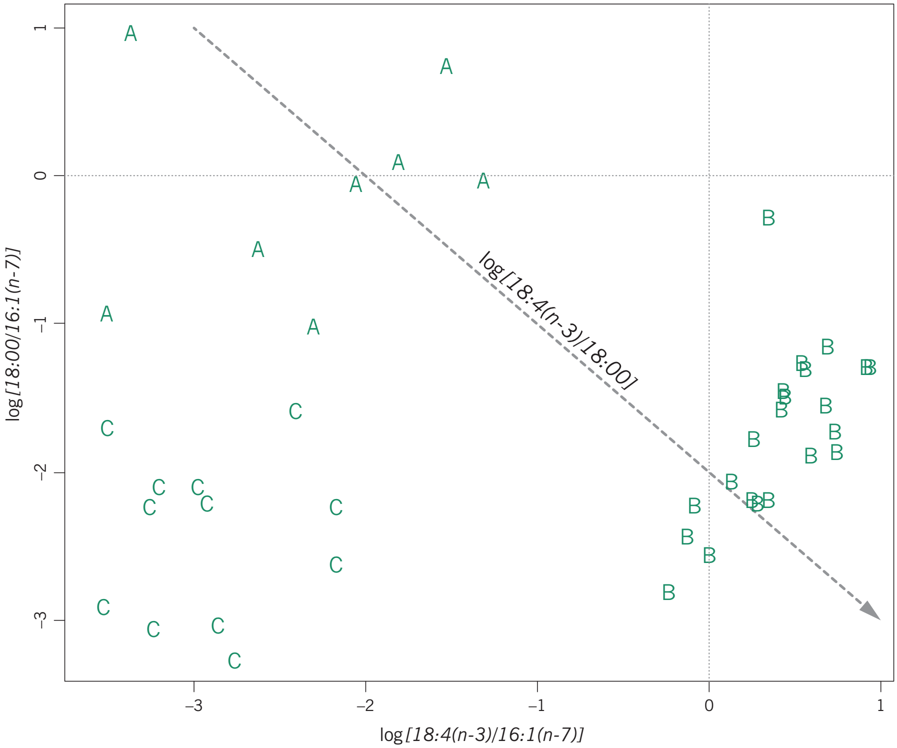

Exhibit 14.6:
Scatterplot of two log-ratios suggested by the biplot in Exhibit 14.5, perfectly separating the three groups of copepods. A third log-ratio combining the two describes a diagonal axis in the plot

plies the Box-Cox power transformation to the data (see Chapter 3 and definition (3.4)), with increasingly stronger power (for example, square root, then cube root, then fourth root, etc.), then the CA of the transformed data tends to LRA in the limit. Moreover, if the variance in the data is small, then the CA solution will be close to the LRA solution anyway. This means that CA is close to being subcompositionally coherent, and perhaps close enough for practical purposes. The CA biplot comparable to Exhibit 14.5 is given in Exhibit 14.7, and is indeed very similar. Here are some statistics comparing the two results:

	(Weighted) LRA	CA
Total variance (or inertia)	0.2260	0.1913
Variance, dimension 1	0.1375 (60.8%)	0.0882 (46.1%)
Variance, dimension 2	0.0539 (23.9%)	0.0635 (33.2%)
Percentage in two dimensions	84.7%	79.3%
Procrustes correlation between rows	0.950	
Procrustes correlation between columns	0.930	

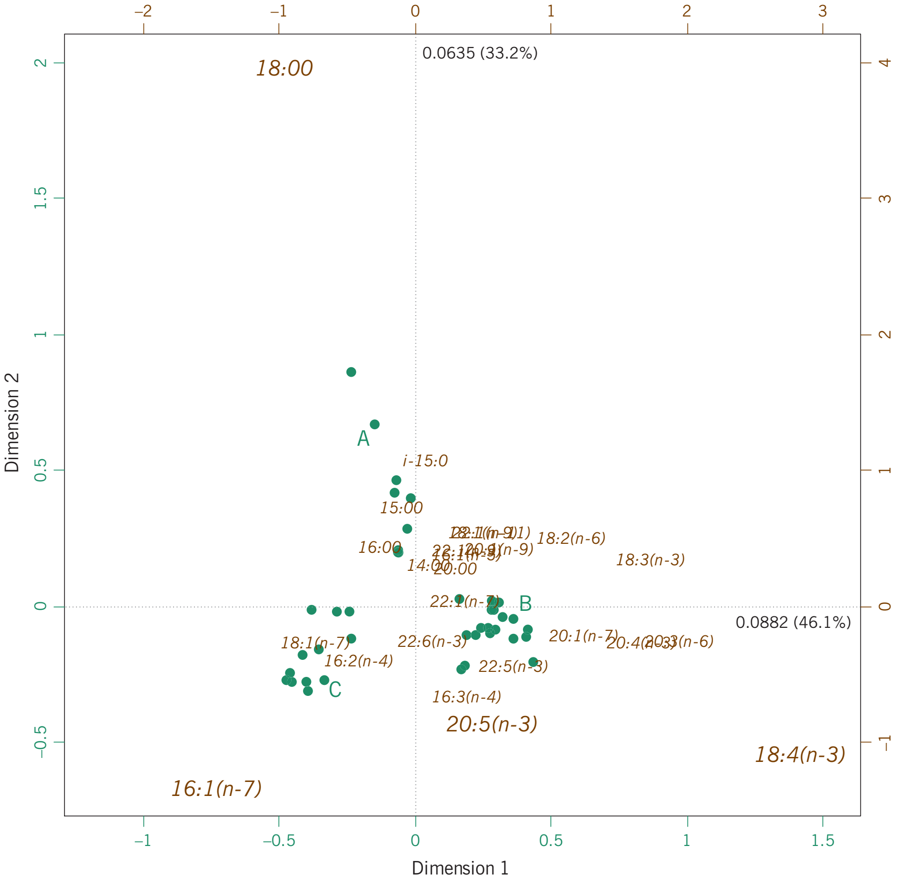

Exhibit 14.7: *Row-principal CA biplot (asymmetric map) of "fatty acid" data. Explained variance is 79.3%*

Four out of the seven fatty acids previously highlighted in the LRA are singled out as high contributors in the CA. The three groups of copepods are separated in the same way, but the interpretation of the joint plot is different. Here, as for most biplots, the biplot axes are considered through the origin to each variable point. For example, if we draw a straight line from the bottom through the origin and up to fatty acid *18:00*, then the projections of the copepods on this axis should reproduce approximately the compositional values on this fatty acid. Exhibit 14.8 verifies this and also shows how close these projections are to the actual values. In fact, the original values show some overlap between the A group and the others, whereas the estimated values perfectly separate the A group. This is due to the fact that other fatty acids are operating in the biplot to separate the groups – so group A is separated not only because it is high on *18:00* but also low on *16:1(n-7)*, which brings us right back to the idea in LRA to work with ratios.

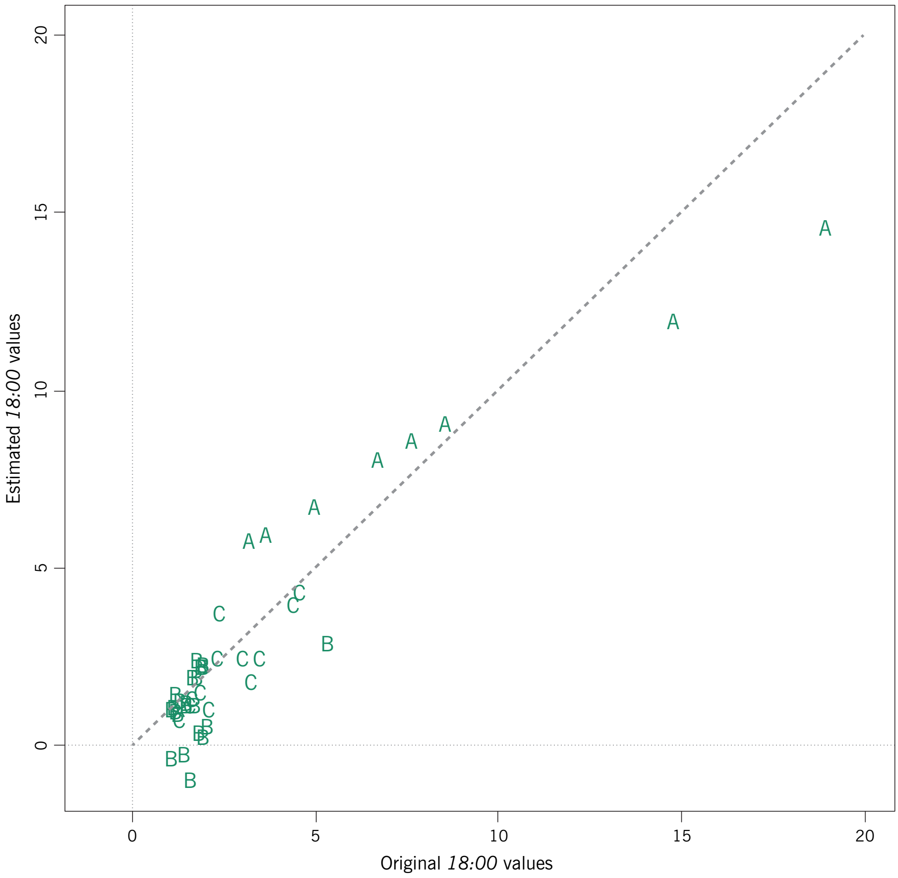

Exhibit 14.8: *Actual compositional value (as a percentage) of fatty acid 18:00 and estimated values from the CA biplot of Exhibit 14.7. The dashed line represents perfect reconstruction. The correlation is 0.928, thus the variance explained in 18:00 by the two dimensions is $0.928^2 = 0.861$, i.e. 86.1%*

Zeros in compositional data

Since LRA visualizes ratios, there should be no zero values in the data, as has been the case for the fatty acid data set used in this chapter so far. In fact, this data set, with 25 fatty acids, is a subset of a bigger one that does have an additional 15 fatty acids with some observed zeros. Collectively these 15 additional fatty acids account for between 3 and 4 percent of each sample, so they are rare fatty acids and thus sometimes observed as zeros. Let us call this data set with all 40 fatty acids the "complete fatty acid" data set, and consider how to analyse it. Zeros can arise for various reasons, one being that the presence of the fatty acid is below the detection limit of the measuring instrument. If one knows what this detection limit is, a value of half the detection limit, say, could be substituted for the zeros. This will create large log-ratios, and thus large variances, but because fatty acids are weighted in the analysis proportionally to their mean values, this will reduce the effect of these large variances in the rare fatty acids. Another option is to treat the zeros as missing values – there are ways for handling missing data by estimating values in the data table

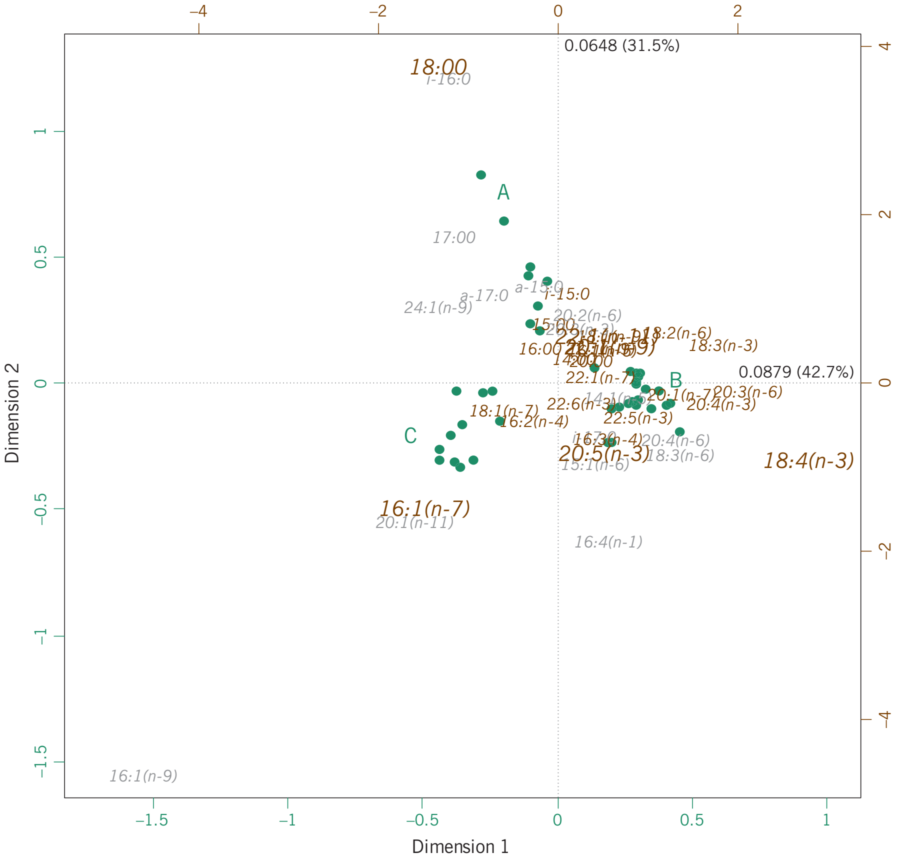

Exhibit 14.9:
CA of the "complete fatty acid" data set of 42 copepods and 40 fatty acids. The row-principal biplot is shown and the explained variance in this two-dimensional solution is 74.2%. Compared to Exhibit 14.7, the additional 15 fatty acids are coloured in gray

from the biplot. An easier solution is to recognize that CA is a good approximation to LRA and close to having subcompositional coherence, and also has no problem with zeros in the data. Exhibit 14.9 is the CA of the complete data set of 40 fatty acids, and it is clear that the extra data have not changed the results that much. The samples are in an almost identical configuration, whereas the additional fatty acids are all low contributors. This biplot illustrates what often happens with low frequency variables, such as rare species or in this case fatty acids with low proportions. Some of these are in outlying positions in the biplot, for example *i-16:0* at the top and *16:1(n-9)* at bottom left. If one does not take into account the contributions, then one might think that *16:1(n-9)*, for example, is the most important fatty acid separating out group C, whereas it is in fact *16:1(n-7)*. This can be easily corroborated by making a scatterplot of these two fatty acids, shown in Exhibit 14.10.

A better way of showing the CA results in this case is in the form of a contribution biplot (Exhibit 14.11), where the low contributing variables shrink to the centre

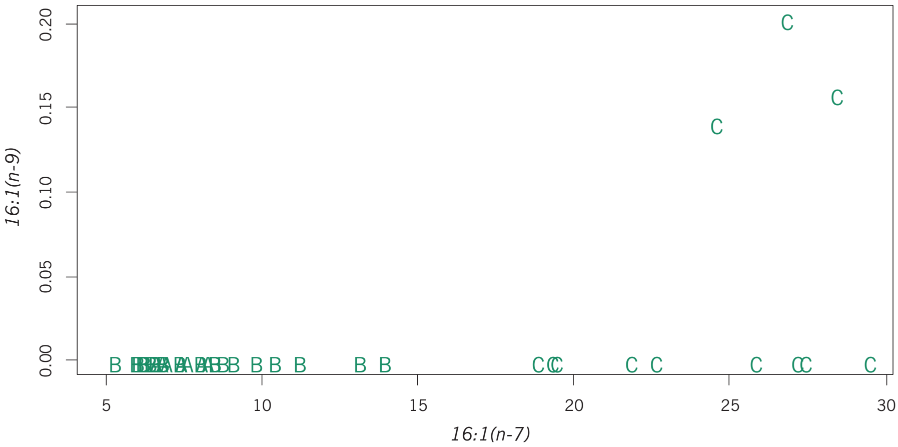

Exhibit 14.10: *Scatterplot of fatty acids 16:1(n-7) and 16:1(n-9) of the "complete fatty acids" data set, showing that 16:1(n-7) is the more important one for separating out group C of copepods. The rare fatty acid 16:1(n-9) has only three small positive percentages, coinciding with three copepods in group C*

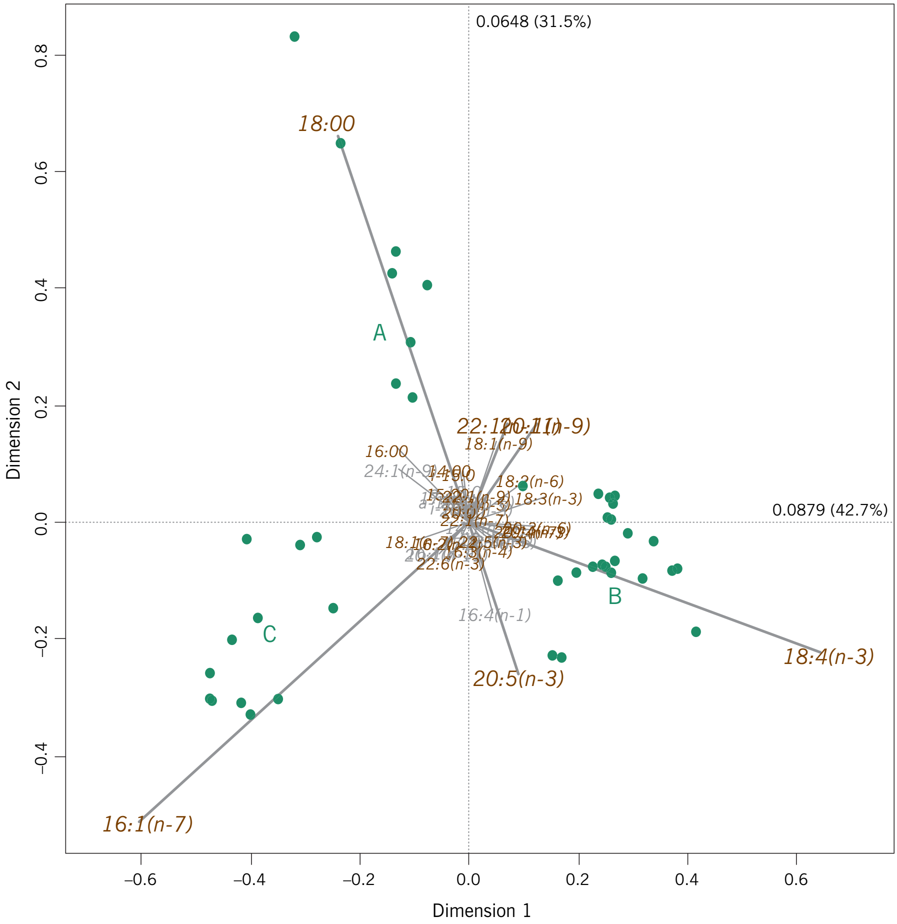

Exhibit 14.11: *CA contribution biplot of "complete fatty acid" data set. The six high contributing fatty acids stand out from the rest*

and the high contributors stand out according to their contribution. Notice too that only one scale is necessary in the contribution biplot (cf. Exhibits 14.5, 14.7 and 14.9 where there were separate scales for row and column points).

SUMMARY: Compositional data and log-ratio analysis

1. Compositional data have the property that for each sample its set of values, called *components*, sum to a constant, usually 1 (for proportions) or 100 (for percentages).

2. Because of this constant sum property, called the property of *closure*, many conventional statistics calculated on the components, such as the correlation coefficient, are inappropriate because they change when subcompositions are formed from a subset of components. Measures that do not change are said to have the property of subcompositional coherence.

3. The log-ratio transformation implies analysing all the pairwise ratios between components on a logarithmic scale. Ratios do not change in subcompositions and are thus subcompositionally coherent.

4. Log-ratio analysis (LRA) is a dimension reduction technique like PCA and CA that visualizes all the pairwise log-ratios in a biplot along with the sample points. Links between pairs of components in the biplot give directions of the log-ratio biplot axes, onto which samples can be projected to estimate the corresponding log-ratios.

5. CA turns out to have a strong theoretical link to LRA and, although not subcompositionally coherent, is close to being so. It provides a good alternative to LRA, especially when there are zero values in the data and the log-ratio approach can not be applied unless the zeros are substituted with positive values.

6. The contribution biplot is a valuable way to separate out components in the log-ratio analysis that are important for the interpretation.

List of Exhibits