17th Applied Statistics Conference, Slovenia, 20–22 September 2021

Compositional data analysis of high-dimensional biological datasets: a revalidation of the additive logratio transformation

Michael Greenacre

Universitat Pompeu Fabra, Barcelona

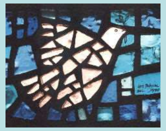
michael.greenacre@upf.edu

Marina Martínez–Álvaro

Scotland's Rural College, Edinburgh

Agustín Blasco

Universitat Politècnica de València



www.econ.upf.edu/~michael
www.globalsong.net
www.multivariatestatistics.org
youtube.com/StatisticalSongs
youtube.com/CARMEnetwork
Course website:
www.econ.upf.edu/~michael/cda

includes a video of this talk

Limerick summary

Compositional data can be fun,
Their values always add up to one.
Drop a category and re-express,
Their values change: it’s a mess!
Rather use **ratios**, then you’re done!

Compositional data

- Compositional data are sets of non-negative data that have totals which are not of interest – rather, the relative values are of interest. These datasets are then conveniently expressed as ***compositions***, i.e. proportions (adding up to 1), or percentages (adding up to 100%). This "relativization" of the data is referred to as ***normalization*** or ***closure***.
- The components of a composition are called its ***parts***.
- If a subset of the parts are considered and the data are re-expressed with respect to the respective subtotals, i.e. renormalized (or reclosed), the data now consist of ***subcompositions*** of the original compositions. The fact that the compositional values of the parts change, after renormalization, makes compositional data special, and needing special approaches.
- In a research study, any given compositional dataset inevitably consists of subcompositions of a much larger one that is not observed or not even observable.

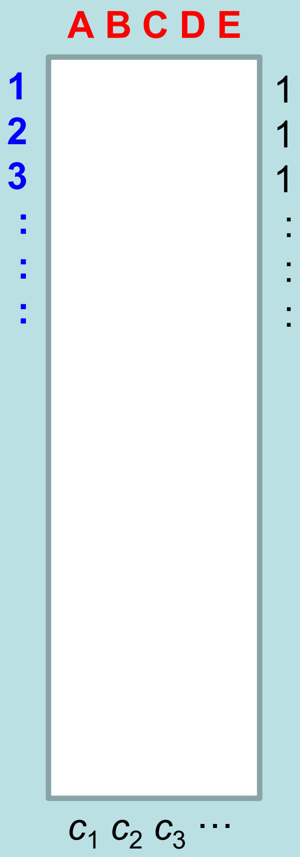
Compositional data – peculiarities

- Compositional data should not be treated like regular statistical variables
- For example, take some random nonnegative data, and create a fictitious matrix of samples-by-variables, the variables will have near-zero correlation, by construction.
- Now normalize (normalize) the data, i.e. express each sample’s data as a proportion of its sum. Then there will be some high correlations between the parts and there will necessarily be some negative correlations (if some proportions go up, others have to come down, since they add up to 1)
- So the values for any particular part depend on the choice of the other parts. This is not the case with “regular” statistical variables. For example, a temperature value remains the same irrespective of what other variables are in the data set such as salinity, depth, pressure, etc…

Compositional data – logratios

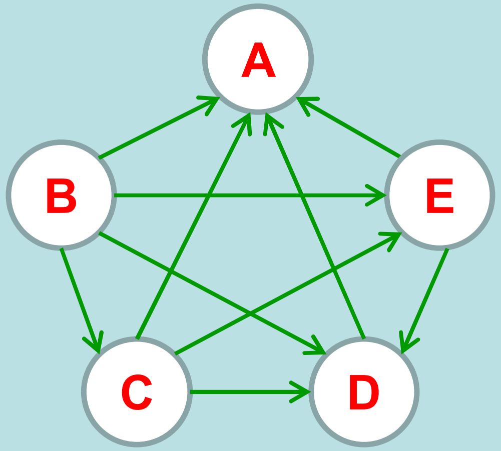
- Since the pioneering work of John Aitchison in the 1980s, it has been recognized that the way to make a statistical approach to compositional data “coherent” is to use ratios of parts, specifically logarithms of these ratios, or ***logratios***. Ratios remain constant even when parts are added or removed and the data renormalized.
- For a data set with D parts, $X_1, X_2, \ldots, X_D$, there are $\frac{1}{2}D(D-1)$ possible pairwise logratios.
- A D–part compositional dataset is known to have rank (or dimensionality) $K = D-1$, assuming the number N of sampling units, or cases, is bigger than D. Two of the three data sets dealt with here have many more parts than samples: $D > N$, then the dimensionality is determined by the cases: $K = N-1$.
- Any linearly independent set of K pariwise logratios will explain 100% of the logratio variance (variances of all pairwise logratios).

Graph representation of logratios



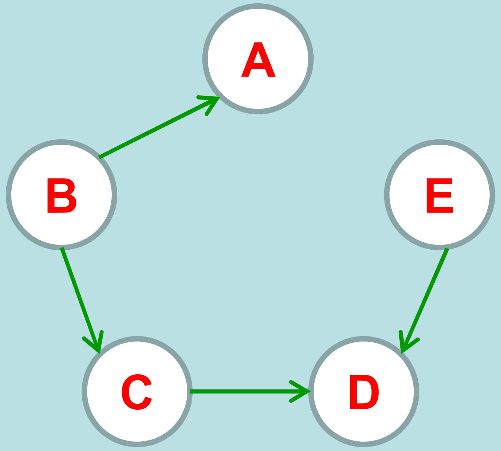
N cases
D parts

- e.g. for $D = 5$



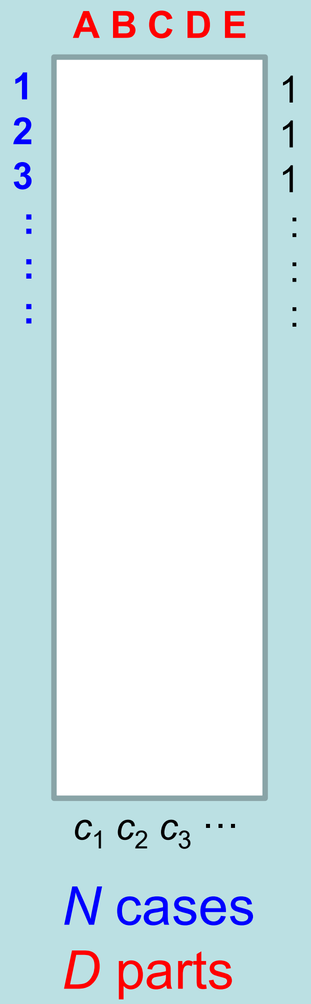
All 10 pairwise logratios are ***edges*** of the graph, directed or undirected

- Dimensionality of the data set is D–1 = 4: only 4 edges are required to completely account for the logratio variance, but these need to correspond to independent ratios, i.e. they should form an ***acyclic graph***

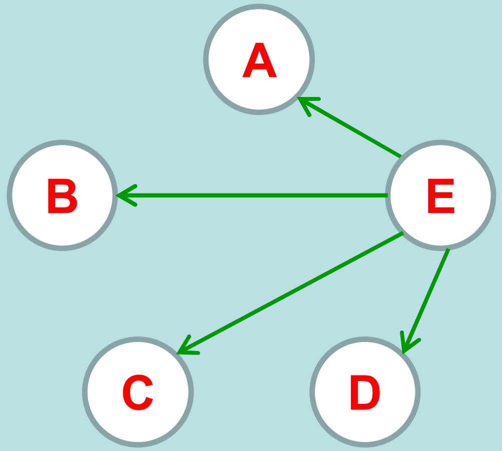


For general D there are D^{D-2} possible acyclic graphs
(Cayley's formula)

Additive logratios (ALRs)



- e.g. for a 5-part composition



4 logratios
with respect to a
chosen ***reference***
part, e.g. E:
log(A/E), log(B/E),
log(C/E), log(D/E)

- ALRs are the simplest set of pairwise logratios, since they merely involve choosing the "best" reference part, hence there are only D possibilities. ALRs have been generally rejected by those who favour the so-called "isometric" logratios, involving ratios of geometric means and consequently complicated and difficult to interpret.
- We will show, however, that especially for datasets with a high number of parts, the ALR transformation is the one of choice.

Simplicity for the practitioner is high on our priorities

- John Aitchison himself was in favour of simple solutions and specifically spoke out against the development of the more complex isometric logratio alternatives. In fact, he said (quote from his keynote address at the 2008 CoDA Workshop):

  > *"The ALR transformation methodology has, in my view, withstood all attacks on its validity as a statistical modeling tool. Indeed, it is an approach to practical compositional data analysis which I recommend particularly for non-mathematicians. The advantage of its logratios involving only two components, in contrast to CLR and ILR (centered and isometric logratio transformations ...), which use logratios involving more than two and often many components, makes for simple interpretation and far outweighs any criticism, more imagined than real, that the transformation is not isometric."*

- The issue here is that the ALRs are not mathematically exactly ***isometric***.

Isometry

- This property is highly prized by many proponents of the logratio approach, to the extent that logratio transformations that are not isometric are ruled out of consideration.

- The exact logratio geometry can be thought of as the K–dimensional configuration of the cases coded by all $\frac{1}{2}D(D-1)$ pairwise logratios, equivalently as all the Euclidean distances between the cases in the multidimensional logratio space.

- Logratio transformations that produce K–dimensional configurations that reproduce exactly the logratio geometry are called isometric. The problem is that isometric transformations are more complicated (as Aitchison said), involving ratios of geometric means of groups of parts, and presenting difficulties in their specification and interpretation.

- But we will show that, thanks to an optimal way of choosing the reference part, the simpler ALR–transformed compositions can be made measurably very close to being isometric; so close that – for all practical purposes – they are isometric. The degree of isometry can be measured using ***Procrustes analysis.***

Datasets considered here

• The Rabbits data	89×3937 dataset of 3937 microbial genes
• The Mice data	28×3147 dataset of mRNA transcripts
• The Cows data	211×127 dataset of NMR intensities

- Two datasets are “wide” and one “narrow”.
- In all the datasets the sample totals are of no relevance, so each set of sample values is expressed relative to the respective total.

Methodological approach: summary

For each dataset:

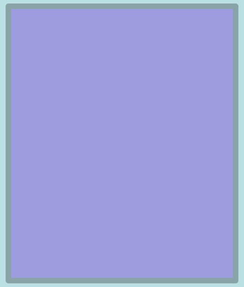
- Consider each part as a reference in an ALR-transformation, as many transformations as there are parts
- For each ALR transformation, compute the Procrustes correlation between the exact logratio geometry and the ALR geometry
- Identify which ALR transformation has maximum Procrustes correlation, i.e. which is the closest to being isometric
- Identify the references with the lowest log-transformed variances

ALR transformation

$$\log(X_j / X_{ref}) = \log(X_j) - \log(X_{ref}) \ , \qquad j=1,\ldots,D \quad j \neq ref$$

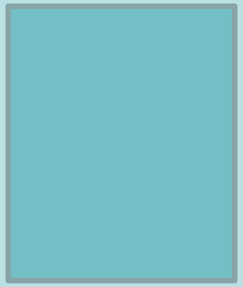
Primary criterion: Maximizing Procrustes correlation

K–dimensional exact geometry



LRA

For each reference part *ref*, K-dimensional ALR geometry



PCA

- Procrustes analysis: rescale and rotate one of the geometries to fit the other.
- Procrustes correlation: correlation between the two sets of geometrical coordinates (one set is Procrustified).

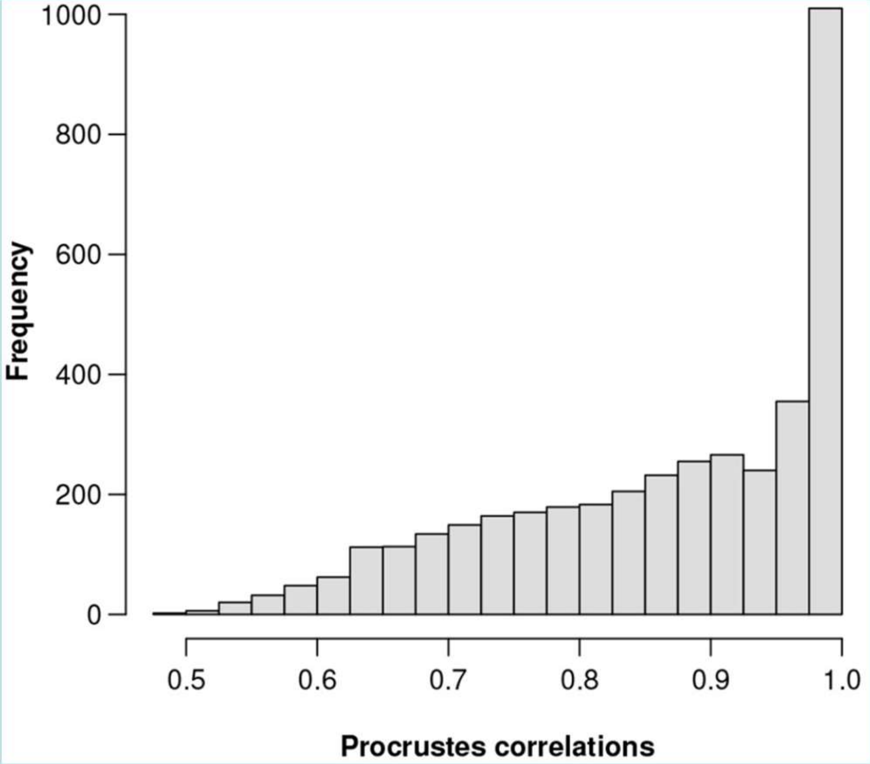
Secondary criterion: Minimizing variance of $\log(X_{ref})$

For each reference part *ref*, compute the variance of $\log(X_{ref})$. For near constant $\log(X_{ref})$ the ALR can be identified with the numerator part, for easier interpretation. Near constant $\log(X_{ref})$ is equivalent to near constant proportionality of X_{ref} with original sample totals.

Rabbits data

89 × 3937 dataset of 3937 microbial genes

Histogram of 3937 Procrustes correlations



Maximum $r = 0.9991$ for gene #856

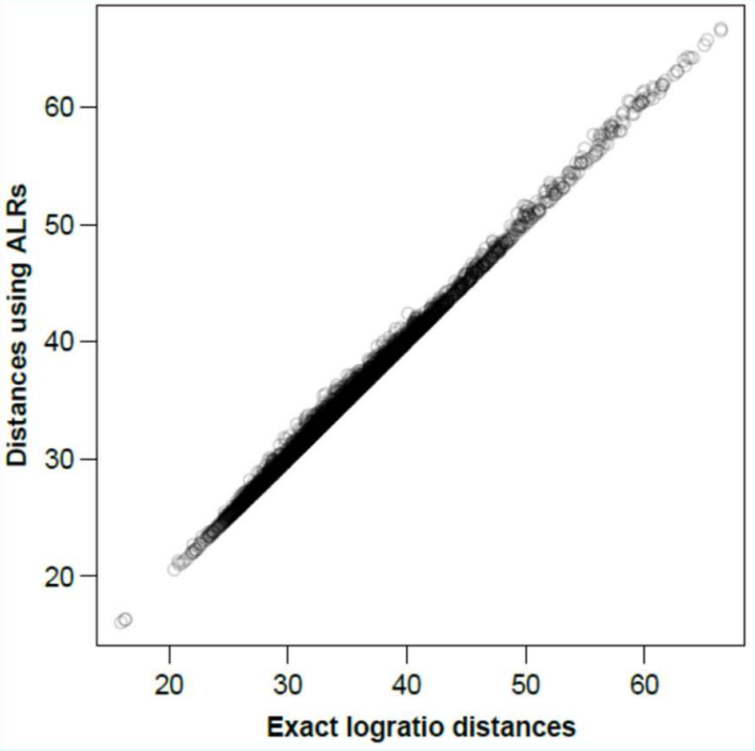
This gene also minimizes the variance of log(component) :

$\mathrm{var}(\log(X_{856}))= 0.00117$

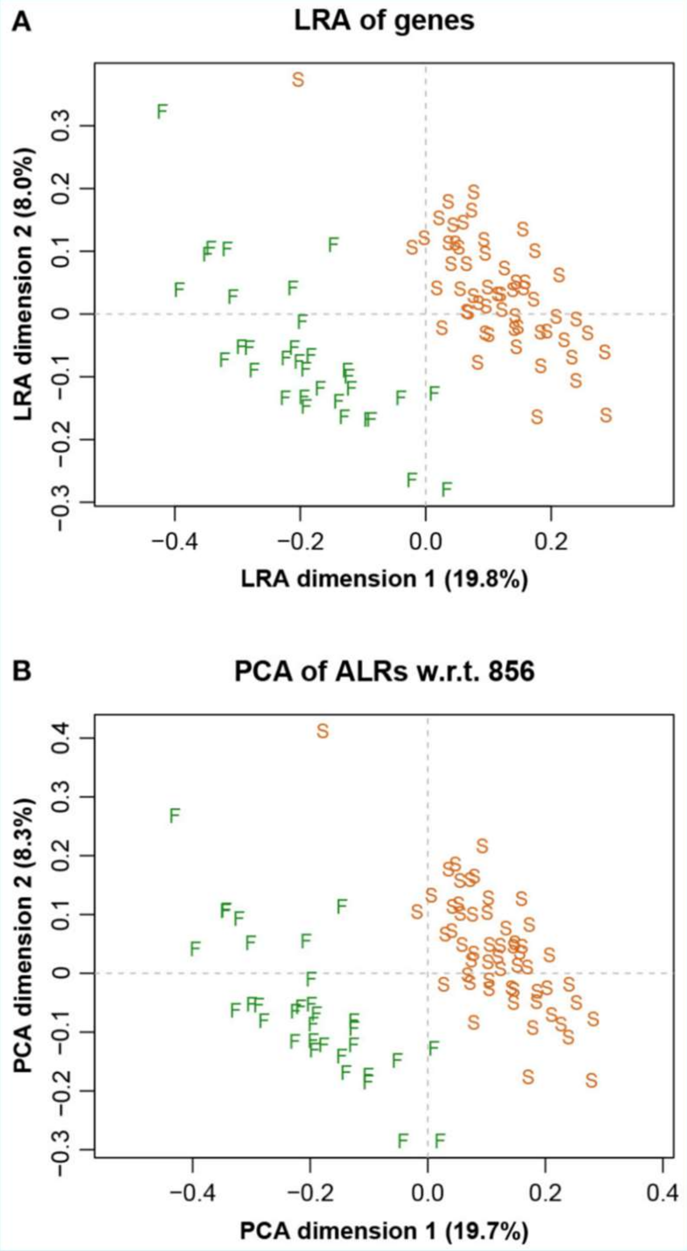
minimum = −6.97 first quartile = −6.89 median = −6.87 third quartile = −6.84 maximum = −6.76

Rabbits data

ALRs quasi–isometric:



½×89×88 = 3916 distances plotted

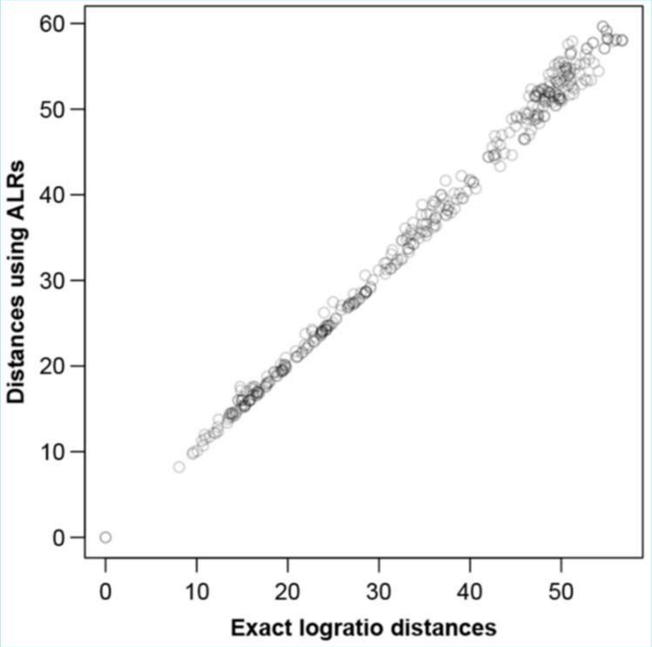


The logratio analysis (LRA) is based on the exact distances, the PCA on the ALR distances. The two groups of points are due to two sequencing laboratories, indicated here by F and S. This effect was later eliminated in the analysis

Mice data

28 × 3147 dataset of 3147 mRNA transcripts

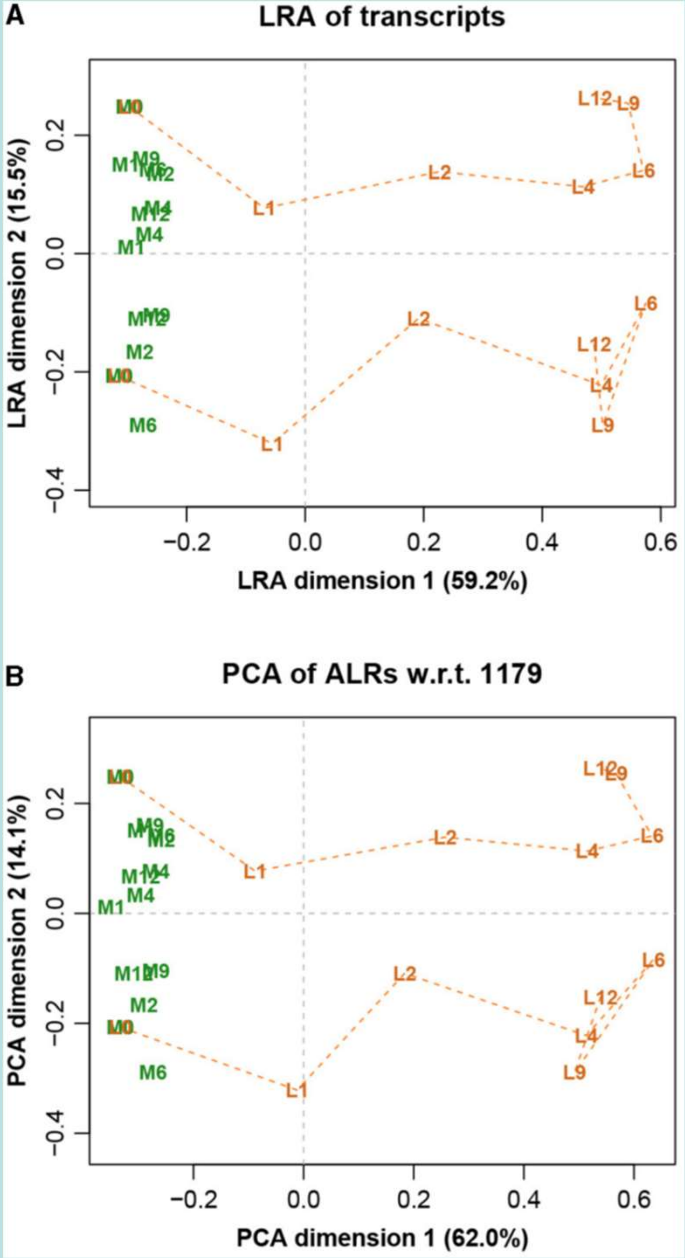
ALRs quasi-isometric:



½×28×27 = 378 distances plotted

Maximum Procrustes correlation = 0.9977 (reference #1318) but we used the second highest correlation = 0.9974 (reference #1179) because it had much lower log-variance

M: control group, L: treatment group. Numbers are times in hours of the experiment. Two replicates (upper and lower groups of points)



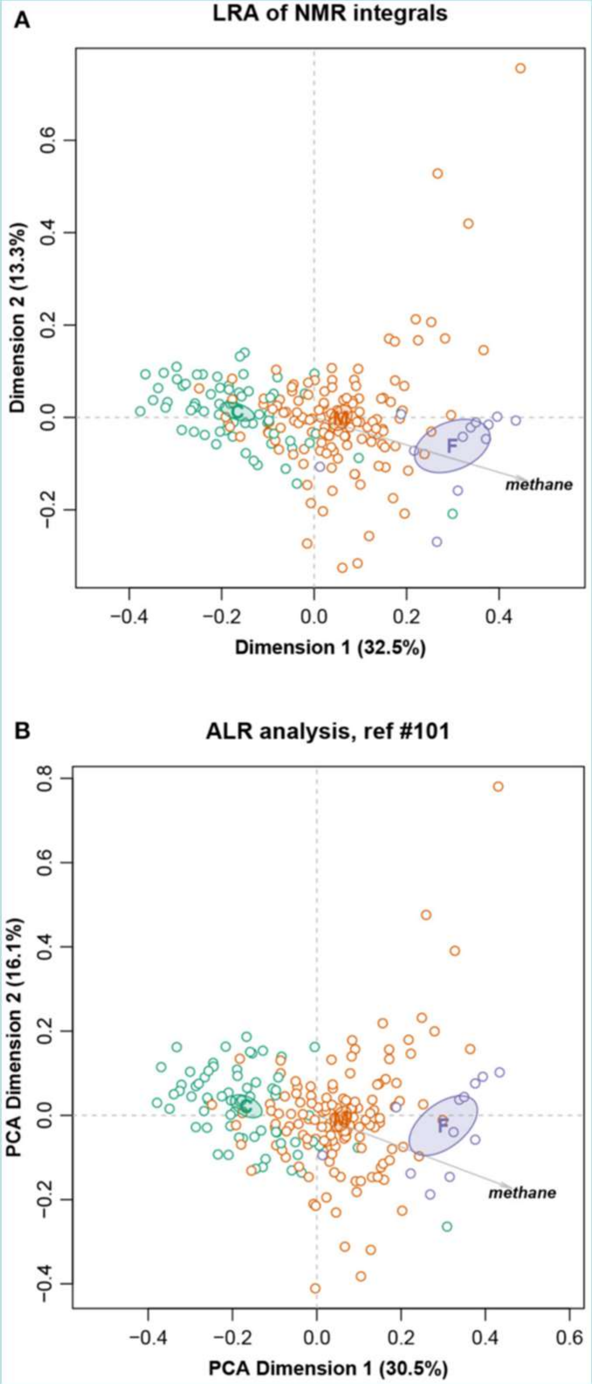
Cows data

211 × 127 dataset of 127 nuclear magnetic resonance (NMR) integrals, in a study of methane emisión from cattle

Even though there are much fewer parts, the maximum Procrustes correlation = 0.9902 (for reference #101).

But the reference part has a fairly high variance of 0.0115 and thus a wider range than the previous examples, so the ALRs have to be interpreted as actual logratios, not as approximate log-transformed parts in the numerators.

Once more the dimensión-reduced logratio geometry is faithfully reproduced by the ALRs (see opposite).



95% confidence ellipses shown for the three groups of samples according to diet: C=concentrate, M=mixed, F=forage. The supplementary variable methane is also shown as an arrow in each analysis

Summary and conclusions

- For three datasets with high numbers of parts it was possible to identify a suitable ALR transformation which:

  (a) explains 100% of the variance (this is a property of any set of pairwise logratios that contains each part at least once)

  (b) comes very close to being isometric

  (c) in two datasets had a reference part which was almost constant, which facilitated the interpretation of the ALRs

- The resistance to the ALR transformation by various authors is difficult to understand. Why adhere to the strict requirement of exact isometry, using complex transformations, when for a tiny sacrifice of isometry, a much simpler transformation can be used?

- Another advantage of the ALR transformation is that it is easy to identify an optimal reference part. It can take significant computational time when the number of parts is high, but programming is simple (and will be part of the easyCODA package soon).

- Remembering John Aitchison's words, the ALR is revalidated!

Selected References

- Egozcue JJ et al. (2003). Isometric logratio transformations for compositional data analysis. *Math. Geol.* 35:279-300
- Hron K et al. (2017). Weighted pivot coordinates for compositional data and their application to geochemical mapping. *Math. Geosci.* 49: 777-796.
- van den Boogaart, KG and Tolosana-Delgado R (2013). *Analyzing Compositional Data with R*. Springer-Verlag
- Filzmoser P et al. (2018). *Applied Compositional Data Analysis*. Oxford University Press

Favour isometric transformations, reject ALRs because they are “oblique” and not exactly isometric

- 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. `https://core.ac.uk/download/pdf/132548276.pdf`
- Greenacre, M (2018). *Compositional Data Analysis in Practice*. Chapman & Hall / CRC Press.
- Greenacre, M et al. (2021). Compositional data analysis of microbiome and any-omics datasets: a validation of the additive logratio transformation. In press at *Frontiers in Microbiology*. With supplementary material; data & R scripts.

Prefer pairwise logratios and ALRs, which provide simple solutions. We have shown they work well in practice and can* satisfy isometry almost exactly

*true for datasets with very many parts - needs to be checked for each case