Analysing Pairwise Logratios Revisited

Germà Coenders

GRECS
GRUP DE RECERCA EN ESTADÍSTICA, ECONOMETRIA I SALUT

Universitat de Girona
Departament d'Economia

Presentation based on an E-pub. ahead of print in *Mathematical Geosciences*, with the same title (K. Hron, G. Coenders, P. Filzmoser, J. Palarea-Albaladejo, M Faměra, T. Matys Grygar).

KH, PF, MF and TMG gratefully acknowledge the support of the Czech Science Foundation (GA19-01768S). GC and JPA were supported by the Spanish Ministry of Science, Innovation and Universities (RTI2018-095518-B-C21).

Introduction

The backbone of compositional data analysis is an algebraic-geometrical structure that has become to be called the Aitchison geometry. This reflects the dimensionality D-1 for a D-part composition, as well as the scale invariance property.

One major development has been the coordinates with respect to an orthonormal basis, providing an isometric mapping with the Euclidean geometry of the real space while respecting the dimensionality.

They were originally called isometric logratio (ilr) coordinates , although the naming orthonormal logratio (olr) coordinates has been recently advocated (Martín-Fernández, 2019) to reflect their distinctive feature, since the centred logratio (clr) transformation is also isometric but does not respect the dimensionality.

An important aspect is interpretability. For this purpose, Egozcue and Pawlowsky-Glahn (2005) proposed balance coordinates as a specific olr interpretable in terms of balances between groups of parts. There are still some drawbacks:

- A meaningful non-overlapping grouping of the parts is required.
- The resulting olr coordinates usually include some which do not have a straightforward interpretation.
- The connection with the individual parts is easily lost.

Pivot logratio coordinates were proposed to address the third point (Fišerová and Hron, 2011).

They are specific balances where the relative information about a part within the composition is contained in one of the coordinates, which consists of a balance between that part and the remaining ones.

Pivot coordinates might not be welcome in areas where it is customary to use a component as reference to “normalise” the others. This results in D-1 logratio alr coordinates which are not orthonormal but oblique with respect to the Aitchison geometry.

Having simple pairwise logratios instead of what might be perceived as over complex mathematical constructs sounds appealing.

Pairwise logratios are not limited to alr but can be computed for any of the D(D-1) possible pairs of components.

A possibility to obtain pairwise logratios while respecting the dimensionality of compositions results from a variable selection procedure, which picks out logratios which explain as much of the total data variability as possible, and hopefully represent leading processes in the data (Greenacre, 2018).

However, these logratio coordinates do not meet the orthonormality criterion either.

The simplicity of pairwise logratios has led to discussions regarding whether orthonormality is really a fundamental assumption, or should be dispensed with in favour of the simpler interpretation.

Using oblique coordinates does not result in problems for methods which are invariant to affine transformations. However:

- This is not the case in methods like principal component analysis or cluster analysis.
- Oblique coordinates affect the interpretation of regression coefficients with compositional explanatory variables.
- Oblique coordinates violate the subcompositional dominance property.
- Distances change with the oblique coordinate choice.

The above does not imply that pairwise logratios should be necessarily avoided.

The goal of this paper is to show that interpretation in terms of pairwise logratios can be retained without sacrificing orthonormality.

Once a set of interpretable pairwise logratios is determined, a collection of orthonormal coordinate systems containing these pairwise logratios as one of the coordinates is built.

Accordingly, the pairwise logratios are used for statistical analysis underpinned by their respective orthonormal coordinate systems.

Because the idea of compiling results from several olr coordinate systems is borrowed from the concept of pivot coordinates, just in a kind of "reverse order", we will refer to this strategy as *backwards pivot coordinates* (bpc).

The beaten track: alr, clr, and pivots

The first approach to express compositional data were additive logratio (alr) coordinates, which are defined for a D-part composition $\mathbf{x} = (x_1, \ldots, x_D)$:

$$\operatorname{alr}(\mathbf{x}) = \left(\ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D} \right).$$

- These coordinates are well aligned with geochemical practice, as frequently there is a justified normalising element.
- They are a particular case of pairwise logratios. Permutation of components leads to all possible pairwise logratios.
- They are oblique.

An alternative way of obtaining easy-to-interpret logratios, rather than relating parts one-to-one is one-to-all. The clr coefficients are:

$$\operatorname{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \ldots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right).$$

Original components and their logratio representation are linked.

This led to pivot coordinates, where the component placed at the first position (which can be any by permutation) appears only in the first pivot coordinate. In general, this leads to D olr coordinate systems where $\mathbf{x}^{(l)} = (x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)$ stands for such a permutation in which the *l*th part takes the first position.

The sign matrix of the sequential binary partition (SBP), of which the first balance is the pivot coordinate, is:

$$\begin{array}{lcccccc}
 & x_1^{(l)} & x_2^{(l)} & x_3^{(l)} & x_4^{(l)} & \ldots & x_D^{(l)} \\
pc^{(l)}(\mathbf{x})_1 & 1 & -1 & -1 & -1 & \ldots & -1 \\
pc^{(l)}(\mathbf{x})_2 & 0 & 1 & -1 & -1 & \ldots & -1 \\
pc^{(l)}(\mathbf{x})_3 & 0 & 0 & 1 & -1 & \ldots & -1 \\
\ldots & & & & & \ldots & -1 \\
pc^{(l)}(\mathbf{x})_{D-1} & 0 & 0 & 0 & 0 & \ldots & -1
\end{array}$$

Pivot coordinates and clr coefficients are linked, since the first pivot coordinate is nothing else than a scaled clr coefficient:

$$\mathrm{pc}^{(l)}(\mathbf{x})_1 = \sqrt{\frac{D-1}{D}} \ln \frac{x_1^{(l)}}{\sqrt[D-1]{\prod_{j=2}^{D} x_j^{(l)}}} = \sqrt{\frac{D}{D-1}} \mathrm{clr}(\mathbf{x})_l .$$

Pairwise logratios as orthonormal coordinates

Here we put forward that a similar link can be established with pairwise logratios.

This leads to considering each pairwise logratio as the first coordinate of an olr coordinate system and complement it with other coordinates through an appropriate SBP:

$$\begin{array}{cccccccc}
 & x_1^{(l)} & x_2^{(l)} & x_3^{(l)} & x_4^{(l)} & \dots & x_D^{(l)} \\
bpc^{(l)}(\mathbf{x})_1 & 1 & -1 & 0 & 0 & \dots & 0 \\
bpc^{(l)}(\mathbf{x})_2 & 1 & 1 & -1 & 0 & \dots & 0 \\
bpc^{(l)}(\mathbf{x})_3 & 1 & 1 & 1 & -1 & \dots & 0 \\
\dots & & & & & \dots & 0 \\
bpc^{(l)}(\mathbf{x})_{D-1} & 1 & 1 & 1 & 1 & \dots & -1
\end{array}$$

The normalising part is placed at the second position, so that the pairwise logratio of interest is:

$$\mathrm{bpc}^{(l)}(\mathbf{x})_1 = \frac{1}{\sqrt{2}} \ln \frac{x_1^{(l)}}{x_2^{(l)}}$$

and the resulting D-1 coordinate systems could be considered as a counterpart to alr coordinates using the normalising part x_2.

The superscript (*l*) varies between 1 and D-1 for a given normalising part, but more normalising parts of interest could be selected and up to D(D-1) coordinate systems could be built to represent any possible pairwise logratio.

The construction of coordinates as a reversed run of pivot coordinates motivated the name *backwards pivot coordinates* (bpc).

With a three-part composition, three pairwise logratios are possible, leading to at most three coordinate systems, in which the first coordinate is the bpc.

The bpc in the first two systems correspond to the alr coordinates with B as normalising part:

$$\sqrt{\frac{1}{2}}\ln\left(\frac{A}{B}\right),\sqrt{\frac{2}{3}}\ln\left(\frac{\sqrt{A\cdot B}}{C}\right)$$

$$\sqrt{\frac{1}{2}}\ln\left(\frac{C}{B}\right),\sqrt{\frac{2}{3}}\ln\left(\frac{\sqrt{C\cdot B}}{A}\right)$$

$$\sqrt{\frac{1}{2}}\ln\left(\frac{A}{C}\right),\sqrt{\frac{2}{3}}\ln\left(\frac{\sqrt{A\cdot C}}{B}\right).$$

As some statistical methods rely on orthogonal coordinates, pairwise logratios that are determined through olr coordinates are more appropriate than oblique coordinates.

The analysis has to be rerun as many times as pairwise logratios are of interest to the researcher, and results of the first coordinates (the pivots) are brought together.

We focus on two particular methods: principal component analysis and regression with compositional covariates.

Regression with compositional explanatory variables

When an oblique logratio coordinate system is used, the interpretation of the regression coefficient does not correspond to the construction of the coordinate (Coenders and Pawlowsky-Glahn, 2020) but to the rule “keeping all other regressors constant”.

Thus, it depends on the manner in which the remaining coordinates are defined.

Pairwise logratios as explanatory variables cannot be interpreted as intended, that is, as the effect of a trade-off between two parts: the effect of increasing just one part at the expense of decreasing just another.

We use one of the classical simulated data sets provided by Aitchison, called *Coxite*. We consider the subcomposition of three minerals (albite, blandite, and cornite) to explain porosity through regression analysis.

Let us assume that the pairwise logratio between albite and cornite is of especial interest to the researcher. Two linear regression models containing $\ln\left(\dfrac{\text{albite}}{\text{cornite}}\right)$ are possible whose OLS estimates are:

$$\text{porosity} = \underset{(1.6855)}{-0.2427} + \underset{(1.0656)}{16.2949}\ \ln\left(\frac{\text{albite}}{\text{cornite}}\right) - \underset{(1.0580)}{6.7385}\ \ln\left(\frac{\text{blandite}}{\text{cornite}}\right)$$

$$\text{porosity} = \underset{(1.6855)}{-0.2427} + \underset{(0.9033)}{9.5564}\ \ln\left(\frac{\text{albite}}{\text{cornite}}\right) + \underset{(1.0580)}{6.7385}\ \ln\left(\frac{\text{albite}}{\text{blandite}}\right).$$

The effect of increasing $\ln\left(\frac{\text{albite}}{\text{cornite}}\right)$ in the first model must be interpreted assuming that $\ln\left(\frac{\text{blandite}}{\text{cornite}}\right)$ is kept constant. Thus, increasing $\frac{\text{albite}}{\text{cornite}}$ implies increasing $\frac{\text{albite}}{\text{blandite}}$ by the same factor, and blandite and cornite both decrease.

This interpretation does not correspond to the effect of increasing just albite at the expense of decreasing just cornite.

The first model corresponds to using alr coordinates with cornite as normalising part. The estimates with albite are:

$$\text{porosity} = \underset{(1.6855)}{-0.2427} \underset{(0.9033)}{-9.5564} \ln\left(\frac{\text{cornite}}{\text{albite}}\right) \underset{(1.0580)}{-6.7385} \ln\left(\frac{\text{blandite}}{\text{albite}}\right).$$

Both the first and the third models include the trade-off between cornite and albite, but simply expressed by the reciprocal logratios.

It would be expected that the regression coefficients were the same in magnitude but opposite in sign.

The estimates with the trade-off between albite and cornite constructed as a bpc are:

$$\text{porosity} = \underset{(1.6855)}{-0.2427} + \underset{(1.1798)}{18.2796}\sqrt{\frac{1}{2}}\ln\left(\frac{\text{albite}}{\text{cornite}}\right) - \underset{(1.2958)}{8.2530}\sqrt{\frac{2}{3}}\ln\left(\frac{\sqrt{\text{albite}\cdot\text{cornite}}}{\text{blandite}}\right)$$

When interpreting the effect of $\frac{\text{albite}}{\text{cornite}}$ as a bpc, blandite does not increase together with either albite or cornite, but keeping the second coordinate constant ensures that blandite remains proportional to the geometric mean of the two components whose trade-off is of interest to the researcher.

It is thus a matter of the trade-off between albite and cornite only.

Accordingly, being Y a real response variable, up to D(D-1) regression models of the form

$$Y = \beta_0 + \beta_1^{(l)} \mathrm{bpc}^{(l)}(\mathbf{x})_1 + \ldots + \beta_{D-1}^{(l)} \mathrm{bpc}^{(l)}(\mathbf{x})_{D-1} + \varepsilon$$

can be considered, where only the regression coefficient corresponding to $\mathrm{bpc}^{(l)}(\mathbf{x})_1$ is interpreted.

$\beta_1^{(l)}$ corresponds only to the pairwise trade-off between the two involved parts $x_1^{(l)}$ and $x_2^{(l)}$.

The intercept term is the same for all models as well as the overall model statistics (F, R², etc.).

Principal component analysis and compositional biplot

The arrows in the compositional biplot may represent the relative importance of the parts within the given composition (through their D clr coefficients) or trade-offs between pairs of parts from a selection of pairwise logratios (Greenacre, 2018).

Focusing on the second option, the PCA method is invariant only to orthogonal rotations, and changes in oblique coordinate systems lead to different PCA scores and hence to different loadings. These undesirable effects are avoided using olr coordinates.

Kyncˇlová et al. (2016) introduced the so-called *composed compositional biplot*, where the PCA scores are computed from any pivot coordinate system and the set of PCA loadings results from putting together the first coordinates from each pivot coordinate system.

We propose an analogous strategy to develop biplots based on pairwise logratios, by using bpc. Scores and loadings of a specific pairwise logratio do not depend on the choice of bpc system.

The only instance in which pairwise logratios are equivalent to bpc is when all possible D(D-1) pairwise logratios are simultaneously included in the analysis. This may result in a cramped biplot and makes robust principal component extraction methods unfeasible.

Application to sediment compositions

We demonstrate the usefulness of bpc with sediment compositions from the Jizera River (Czech Republic).

The sampling site in Horky nad Jizerou is downstream the city of Mladá Boleslav, with production of batteries, motorcycles, and cars, which generated contamination in the upper sediment strata (Matys Grygar et al., 2013).

We resort to size of particles at the 50th percentile which is log-transformed.

Al and Si are used as normalising parts because they are major elements, have strong relation to major mineral constituents, and are related to grain size.

Regression analysis

In the first part of the analysis we investigate how the granulometry can be explained by the lithogenic elements Al, Si, K, Ti, Rb, Zr by using pairwise logratios.

We employ MM robust regression estimation (`lmrob` function in the `robustbase` R package).

We consider pairwise logratios with Al and Si as normalising parts both in the bpc and alr approaches.

For the alr case, the model was run twice, while it had to be run 10 times for the bpc approach, and only the first coefficient in each run is shown.

All 12 model runs yield the same predictions and goodness of fit indicators (Adjusted R^2=0.853).

The first significant bpc (Al/Si) is interpreted as follows. Increasing the ratio between Al/Si (while keeping constant both the mutual ratios among K, Ti, Rb, Zr and their ratios with respect to the geometric mean of Al and Si) has the effect of reducing particle size.

The effect of increasing Al at the expense of reducing Si corresponds to the natural interpretation of the pairwise logratio.

This interpretation is not apparent in the bottom panels of the table whith the alr regression coefficients referring to the same logratios.

The fact that the coefficient of the Si/Al ratio is not equal to that of Al/Si with opposite sign constitutes a further illustration of the problems with oblique coordinates.

		Estimate	s.e.	p value
bpc	ln(Al/Si)	– 1.279	0.494	0.012*
	ln(K/Si)	– 2.575	1.732	0.143
	ln(Ti/Si)	– 1.944	0.755	0.013*
	ln(Rb/Si)	– 0.353	1.308	0.788
	ln(Zr/Si)	– 1.150	0.545	0.039*
	ln(Si/Al)	1.279	0.494	0.012*
	ln(K/Al)	– 1.296	2.047	0.529
	ln(Ti/Al)	– 0.665	1.065	0.535
	ln(Rb/Al)	0.926	1.065	0.389
	ln(Zr/Al)	0.128	0.543	0.814
alr (with Si)	ln(Al/Si)	– 0.124	1.359	0.928
	ln(K/Si)	– 2.716	2.905	0.354
	ln(Ti/Si)	– 1.454	0.843	0.091
	ln(Rb/Si)	1.728	2.139	0.423
	ln(Zr/Si)	0.133	0.520	0.800
alr (with Al)	ln(Si/Al)	2.433	1.031	0.022*
	ln(K/Al)	– 2.716	2.905	0.354
	ln(Ti/Al)	– 1.454	0.843	0.091
	ln(Rb/Al)	1.728	2.139	0.423
	ln(Zr/Al)	0.133	0.520	0.800

Principal component analysis

We resort to principal component analysis (PCA) where the set of lithogenic elements is complemented by the anthropogenic elements Cu, Pb and Zn.

For different alr coordinate systems (left), with Al and Si as denominator respectively, both the scores and loadings change dramatically. This raises doubts about the reliability of using either alr version.

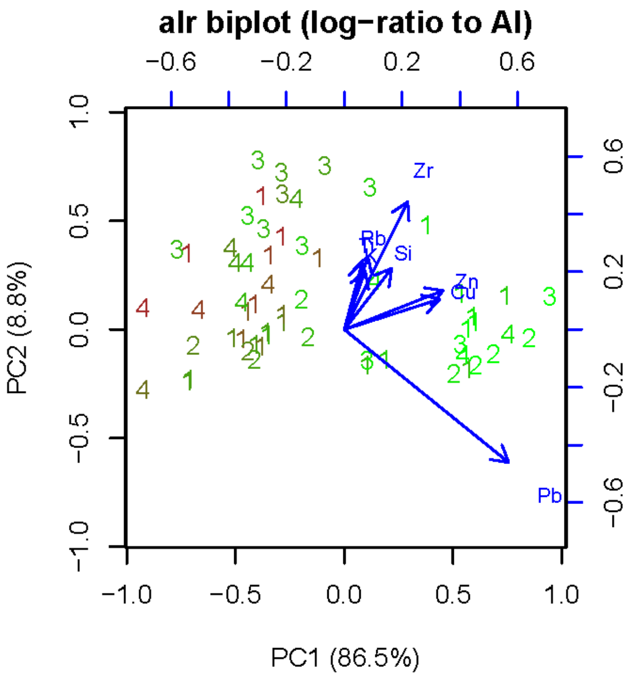
On the other hand, the scores are the same for both versions of the composed compositional biplot based on bpc (right).

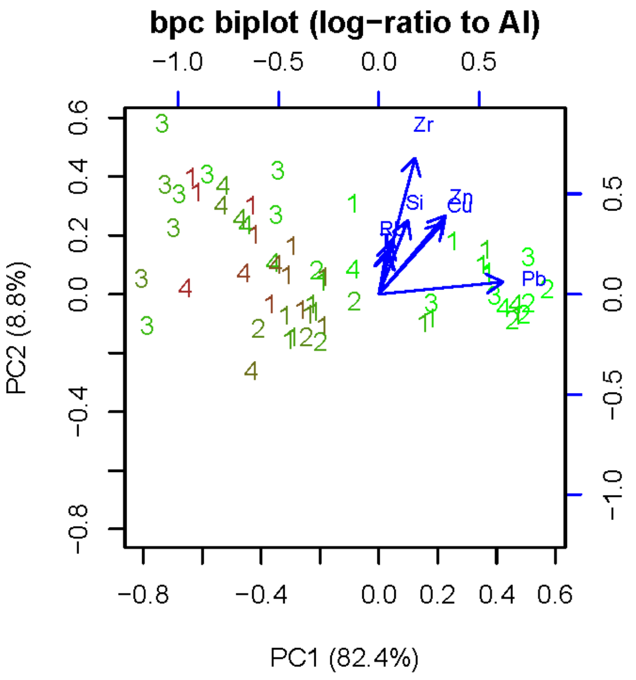
Green (surface)-brown (deep) colouring confirms good separation of samples taken deeper or closer to the surface. The two logratios between the reference elements Al and Si have, as expected, opposite directions.

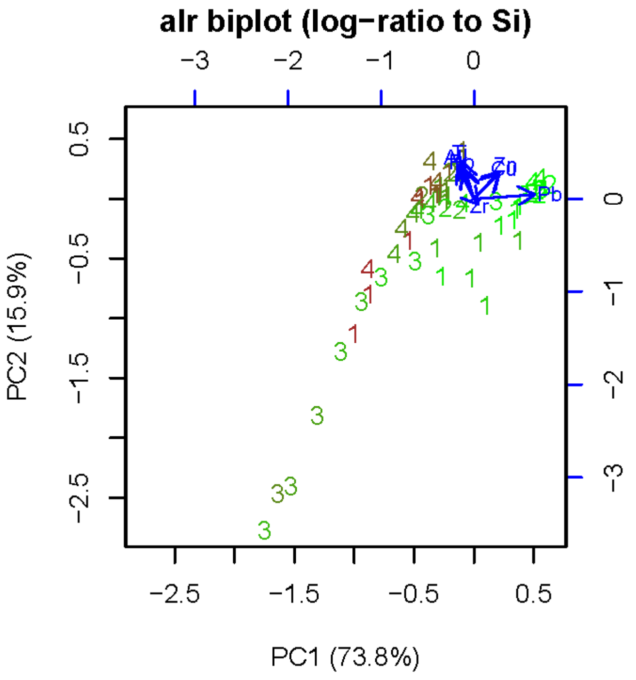
Lithogenic elements (Al, Si, K, Ti, Rb and Zr) are mostly associated to PC2, and risk elements Cu, Zn and Pb, mostly associated to PC1.

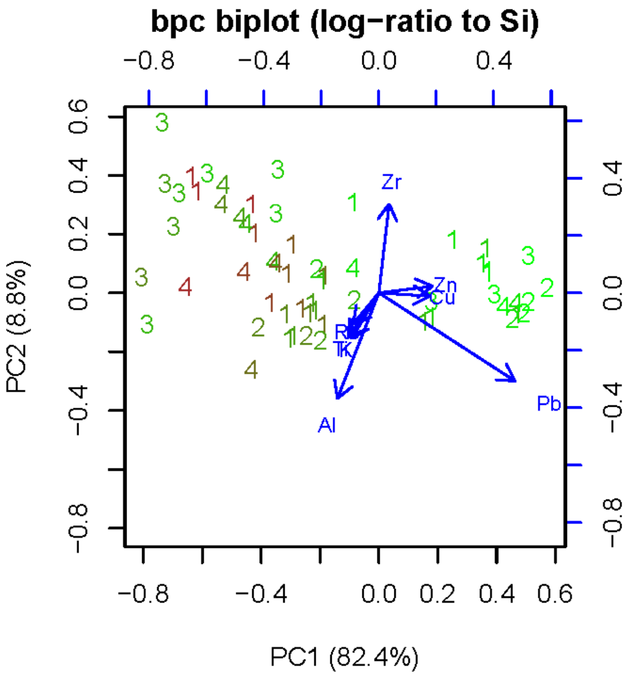
The corresponding alr-based biplot (upper left), also produces convincing results. However, the price of using alr coordinates is that the scores can heavily depend on the chosen reference element.

This can be seen when using scores of alr-based PCA based on Si as denominator (lower left), which result in the least convincing contamination indicator amongst the alternatives discussed here.


alr biplot (log−ratio to Al)
−0.6
−0.2
0.2
0.6
1.0
0.5
0.0
−0.5
−1.0
PC2 (8.8%)
0.6
0.2
−0.2
−0.6
−1.0
−0.5
0.0
0.5
1.0
PC1 (86.5%)
Zr
Rb
Si
Zn
Cu
Pb


bpc biplot (log−ratio to Al)
−1.0
−0.5
0.0
0.5
0.6
0.4
0.2
0.0
−0.4
−0.8
PC2 (8.8%)
0.5
0.0
−0.5
−1.0
−0.8
−0.4
0.0
0.2
0.4
0.6
PC1 (82.4%)
Zr
Si
Zn
Cu
Pb


alr biplot (log−ratio to Si)
−3
−2
−1
0
0.5
−0.5
−1.5
−2.5
PC2 (15.9%)
0
−1
−2
−3
−2.5
−1.5
−0.5
0.5
PC1 (73.8%)
Al
Zn
Cu
Zr
Pb


bpc biplot (log−ratio to Si)
−0.8
−0.4
0.0
0.4
0.6
0.4
0.2
0.0
−0.4
−0.8
PC2 (8.8%)
0.4
0.0
−0.4
−0.8
−0.8
−0.4
0.0
0.2
0.4
0.6
PC1 (82.4%)
Zr
Zn
Cu
Pb
Al

Conclusions

In geological practice, simple pairwise logratios are common. Although advances in the logratio methodology have brought some sophisticated alternatives, recent developments in the field suggest that the Occam's Razor rule should be considered.

The tools we use should be as simple as possible, but not simpler. The orthonormality of logratio coordinates is a requirement for sound statistical analysis, at least for principal component analysis and regression analysis with compositional predictors.

There are multivariate statistical methods for which orthonormality does not matter. Nevertheless, using olr coordinates is a guarantee that things cannot go wrong.

References

Coenders G, Pawlowsky-Glahn V (2020) On interpretations of tests and effect sizes in regression models with a compositional predictor. SORT 44(1):201–220

Egozcue JJ, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. Math Geol 37(7):795–828

Fišerová E, Hron K (2011) On interpretation of orthonormal coordinates for compositional data. Math Geosci 43(4):455–468

Greenacre M (2018) Compositional data in practice. CRC Press

Kyncˇlová P, Filzmoser P, Hron K (2016) Compositional biplots including external non-compositional variables. Statistics 50(5):1132–1148

Martín-Fernández J (2019) Comments on: Compositional data: the sample space and its structure. TEST 28(3):653–657

Matys Grygar T, Nováková T, Bábek O, Elznicová J, Vadinová N (2013) Robust assessment of moderate heavy metal contamination levels in floodplain sediments: A case study on the Jizera River, Czech Republic. Sci Total Environ 452:233–245

Thanks for your attention

Germà Coenders
Department of Economics, University of Girona
C/ Universitat 10, 17003 Girona, Spain
tel ++34972418736
www3.udg.edu/fcee/professors/gcoenders
http://orcid.org/0000-0002-5204-6882
germa.coenders@udg.edu