Mathematical Geoscience manuscript No.
(will be inserted by the editor)

Analysing pairwise logratios revisited

Karel Hron · Germá Coenders · Peter Filzmoser · Javier Palarea-Albaladejo · Martin Faměra · Tomáš Matys Grygar

Received: date / Accepted: date

Abstract Even though the logratio methodology provides a range of both generic, mostly exploratory, and purpose-built coordinate representations of compositional data, simple pairwise logratios are preferred by many for multivariate analysis in the geochemical practice, principally because of their simpler interpretation. However, the logratio coordinate systems that incorporate them are predominantly oblique, resulting in both conceptual and practical problems. We propose a new approach, called backwards pivot coordinates, where each pairwise logratio is linked to one orthogonal coordinate system, and these systems are then used together to produce a concise output. In this work, principal component analysis (PCA) and regression with compositional explanatory variables are used as primary methods to demonstrate the

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 under the project CODAMET (RTI2018-095518-B-C21, 2019-2021).

K. Hron (✉)
Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University Olomouc, 17. listopadu 12, 771 46 Olomouc, Czech Republic
E-mail: hronk@seznam.cz

G. Coenders
Department of Economics, Faculty of Economics and Business, University of Girona, c. Universitat de Girona 10, 17003 Girona, Spain

M. Faměra, T. Matys Grygar
Institute of Inorganic Chemistry of the Czech Academy of Sciences, 250 68 Husinec-Řež, Czech Republic

M. Faměra
Department of Geology, Faculty of Science, Palacký University Olomouc, 17. listopadu 12, 771 46 Olomouc, Czech Republic

P. Filzmoser
Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Wiedner Hauptstrasse 8-10, 1040 Vienna, Austria

J. Palarea-Albaladejo
Biomathematics and Statistics Scotland, JCMB, The King's Buildings, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, Scotland, United Kingdom

methodological and interpretative advantages of the proposal. In the applied part of this study, sediment compositions from the Jizera River, Czech Republic, were analysed using these techniques through backwards pivot coordinates. This allowed to discuss grain size control of the element composition of sediments and clearly distinguish anthropogenically contaminated and uncontaminated strata in sediment depth profiles.

Keywords Pivot coordinates · Additive logratio coordinates · Principal component analysis · Linear regression with compositional covariates · Pairwise logratios

1 Introduction

Compositional data analysis plays a central role in the geosciences as most experimental data generated by chemical analysis consist of multivariate observations of relative nature, typically measured in units such as percentages, mg/kg (ppm) or mg/l. Although in practice they can adopt the form of closed vectors of constant sum (for example 100 if expressed as percentages) or not, they are essentially scale invariant objects (Pawlowsky-Glahn et al. 2015 Filzmoser and Hron 2019). That is, the most basic relevant information is contained in the pairwise logratios between the components or parts of the composition. This implies that the sample space of such data differs from the standard Euclidean real space for which most popular statistical methods are designed (for example principal component analysis and regression analysis as widely used in experimental studies). The sample space of compositional data is in fact formed by equivalence classes of proportional positive vectors. This means that vectors within one class are interchangeable and any representative (closed to any arbitrary sum of components) can be chosen according to the preferences in the context of application. Importantly though, the results from any relevant statistical analysis of compositional data should be equivalent regardless of the chosen representation and, amongst others, this property is guaranteed when the analysis is based on (log)ratios. Since the seminal book by Aitchison (1986), the theory for compositional data analysis has been further developed and completed, and the current state of the art is presented in several recent monographs (van den Boogaart and Tolosana-Delgado 2013 Pawlowsky-Glahn et al. 2015 Filzmoser et al. 2018 Greenacre 2018a).

The theoretical background of compositional data analysis is connected to an algebraic-geometrical structure that has become to be called the Aitchison geometry (Billheimer et al. 2001 Pawlowsky-Glahn and Egozcue 2001). This reflects the dimensionality of compositions ($D-1$ for a D-part composition) as well as their scale invariance property, and provides a solid framework for further developments of the methodology. In fact, there is no other geometrical structure known to reflect the scale invariance of compositions. One major development has been the introduction of coordinates with respect to an orthonormal basis of the Aitchison geometry, providing an isometric mapping between this and the Euclidean geometry of the real space while respecting the dimensionality of compositions. They were originally called isometric logratio (ilr) coordinates (Egozcue et al. 2003), although the naming orthonormal logratio (olr) coordinates has been recently advocated as an alternative (Martín-Fernández 2019) aiming to reflect better their distinctive feature, since other logratio representations like the centred logratio (clr) transformation (Aitchison 1983) also define an isometric mapping (although it fails to represent the actual dimensionality of the data). As with any mathematical tool, an important aspect to be respected in practice is easiness of interpretability. For this purpose, Egozcue and Pawlowsky-Glahn (2005) proposed the use of balance coordinates as a specific olr coordinate representation that is interpretable in terms of normalised balances or contrasts between groups of compositional parts. Although these are designed to reflect some natural processes for example in geochemistry, and their usefulness has been demonstrated in diverse studies (Pawlowsky-Glahn and Buccianti 2011 Buccianti 2013 Pawlowsky-Glahn et al. 2015), there are still some drawbacks mostly related to the following points:

- For the construction of balances through a sequential binary partition (SBP), as devised in Egozcue and Pawlowsky-Glahn (2005) and commonly used in applications, a well elaborated idea about a meaningful (non-overlapping) grouping of the parts is required. This is not always the case, particularly in studies with a relatively large number of compositional parts.
- Even if balances can be successfully constructed, the resulting olr coordinates usually include some which do not have a straightforward interpretation (if any at all) in the context of the particular application. In such cases their practical usability is questionable and turn out to be less appealing for geochemists.
- The details of the connection (grain-size control) with the individual components is easily lost through the grouping process resulting from the construction of balances, however this link is frequently also of scientific interest in geochemical, sedimentological or pedological research and practice.

The class of pivot logratio coordinates was proposed aiming to address the third point (Fišerová and Hron 2011 Hron et al. 2017 Filzmoser et al. 2018). They are specific balances where all the relative information about a compositional part (within the given composition) is contained in one of the coordinates, commonly the first one, which consists of a normalised balance between such part and the remaining ones summarised by their geometric mean. As such, the first pivot coordinate can be expressed as the (scaled) sum of all unique pairwise logratios involving the given part, and this can be done for each individual part of the composition sequentially by simply applying an orthogonal rotation of the coordinate system. Pivot coordinates could thus be seen as a bridge between the inappropriate statistical analysis of the original compositional parts in any fixed representation and the world of logratios. However, aggregating *all* logratios into a single coordinate mixes all pathways governing physical and chemical processes behind the observed data, particularly as the number of parts increases. This led to the proposal of a weighted counterpart to pivot coordinates (Hron et al. 2017). Still, the idea of aggregating logratios through pivot coordinates might not be welcome in areas of geochemistry where it is customary to use a component as reference to “normalise” the others with respect to it by means of the logratios. After taking logs, this results in $D-1$ logratio coordinates for a D-part composition which are not orthonormal but oblique with respect to the Aitchison geometry. This operation formally corresponds to additive logratio (alr) coordinates (Aitchison 1982). As with pivot coordinates, or centred logratio coefficients, alr coordinates cannot be simply identified with the individual original components, as they are in fact logratios, but the link with these is more clearly stated.

Having simple pairwise logratios as a coordinate representation of compositions instead of what might be perceived as involved mathematical constructs sounds appealing. A possibility to obtain logratio coordinates composed by pairwise logratios while respecting the dimensionality of compositions results from a variable selection procedure (Greenacre 2018a;b), which picks out logratios with the aim to explain as much of the total data variability as possible, and then potentially representing leading processes in the data. Pairwise logratios can also be helpful for compositions where there is an implicit ordering of parts, constructing them from adjacent parts (categories) as discussed in Vencálek et al. (2020). However, these logratio coordinates do not meet the orthonormality criterion either.

The simplicity of pairwise logratios and their interpretation in comparison to more sophisticated counterparts has recently led to discussions regarding the extent to which orthonormality of coordinates is really a fundamental property in practice, or even whether this requirement should be suppressed in favour of the simple interpretation of pairwise logratio alternatives (Greenacre 2018a; 2019). Technically, using alr coordinates or any other oblique coordinate system does not result in problems for methods whose results are invariant to affine transformations, see for example Filzmoser and Hron (2008) and Filzmoser et al. (2012). However, this is not the case for rotation invariant methods like principal component analysis. Using such class of coordinates also affects the interpretation of regression coefficients in regression with compositional explanatory variables (Coenders and Pawlowsky-Glahn 2020). Moreover, oblique coordinates violate the basic property known as subcompositional dominance (Egozcue 2009). That is, it might happen that the (Euclidean) distance between compositions expressed in such coordinates is lower than the distance between subcompositions obtained from them. This prevents from considering them as a subcompositionally coherent alternative, or even simply as a geometrically meaningful approach, although subcompositional incoherence can be minimised by a proper choice of (oblique) coordinate system (Greenacre 2018b).

The above does not imply that pairwise logratios should be necessarily avoided for meaningful statistical processing of compositional data. The goal of this paper is to show that interpretation in terms of pairwise logratios can be retained without sacrificing orthonormality and without resorting to approximations (Greenacre 2018b). 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 the 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* throughout this paper. The idea is introduced in the next section, and then developed in the context of two particular statistical techniques commonly used in geochemical studies: principal component analysis and regression analysis with compositional covariates. The methodological part is complemented with some simple motivating examples, followed by a detailed discussion of limitations. The proposed approach is demonstrated in Section 4 using sediment compositions from the Jizera River, Czech Republic. The final Section 5 then concludes with some general advice and remarks.

2 Pairwise logratios as orthonormal coordinates

Historically, the first approach to express compositional data in real space were additive logratio (alr) coordinates (Aitchison 1982), which are defined for a D-part composition $\mathbf{x} = (x_1, \dots, x_D)$ and any chosen ratioing part x_D (up to any permutation of components) as

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

These coordinates are well aligned with geochemical practice, as frequently there exists a justified normalising element to which the others are compared through the logratio. A number of geochemical studies have been performed using these coordinates, including bivariate plotting (Thomas and Aitchison 2005; 2006). However, from a geometrical perspective, their general use is not recommended as they violate the subcompositional dominance principle (Pawlowsky-Glahn et al. 2015). For example, consider two samples described by the 4-part compositions $\mathbf{x}_1 = (0.2, 0.1, 0.6, 0.1)$ and $\mathbf{x}_2 = (0.3, 0.4, 0.1, 0.2)$. Using the fourth component as ratioing part, the ordinary Euclidean distance between the respective alr-coordinate vectors is 2.6. However, when measured using alr coordinates of the subcompositions resulting from dropping the forth component (the third component is then used as ratioing part here), the Euclidean distance between samples has a greater value of 3.86 and that basic principle is not fulfilled. Moreover, distances on alr coordinates change with the permutation of the components, it means they depend on the ratioing part used. This behaviour has unacceptable consequences for any statistical method based on distances, such as the widely-used cluster analysis methods.

Recall that the relative information about a component is contained in all logratios including that specific part of the composition. From this perspective, alr coordinates could be considered as a borderline (weighted) case where only one of them is considered. All the relative information about the components is carried in clr coefficients (Aitchison 1983), defined as

$$\mathrm{clr}(\mathbf{x}) = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}}, \dots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^{D} x_i}} \right). \tag{2}$$

Note that each clr coefficient aggregates all logratios with a given component, for example for the first one we obtain

$$\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^{D} x_i}} = \frac{1}{D} \left(\ln \frac{x_1}{x_2} + \ldots + \ln \frac{x_1}{x_D} \right). \tag{3}$$

Unlike with alr coordinates, this implies that none of the parts of the composition plays any prominent role in relation to the others. Accordingly, when there is an interest in establishing a link between original components and their logratio representation, potential subjectivity related to the choice of the ratioing part is suppressed here.

The idea of aggregating all relative information about a given component (within a specified composition) led to the introduction of the concept of pivot coordinates (pc) mentioned in Section 1, where the component placed at the first position in $\mathbf{x}$ (which can be any by rotation) appears only in the first pivot coordinate. In general, this leads to D pivot coordinate systems

$$\mathrm{pc}^{(l)}(\mathbf{x})_i = \sqrt{\frac{D-i}{D-i+1}}\ln\frac{x_i^{(l)}}{\sqrt[D-i]{\prod_{j=i+1}^{D} x_j^{(l)}}},\ i=1,\ldots,D-1, \tag{4}$$

where $\mathbf{x}^{(l)} = (x_1^{(l)}, x_2^{(l)}, \ldots, x_l^{(l)}, x_{l+1}^{(l)}, \ldots, x_D^{(l)})$ stands for such a permutation of the parts $(x_1, \ldots, x_D)$ in which the l-th compositional part, $l \in \{1, \ldots, D\}$, takes the first position, $\mathbf{x}^{(l)} = (x_l, x_1, \ldots, x_{l-1}, x_{l+1}, \ldots, x_D)$. Note that pivot coordinates and clr coefficients are linked, since the collection of first pivot coordinates is nothing else than scaled clr coefficients:

$$\mathrm{pc}^{(l)}(\mathbf{x})_1 = \sqrt{\frac{D}{D-1}}\mathrm{clr}(\mathbf{x})_l. \tag{5}$$

The question we put forward here is whether a similar link could be established also with alr coordinates (or any pairwise logratios in general), which would mean to single out the corresponding pairwise logratios from a set of orthonormal coordinate systems. And, related to this, whether that link can be derived from existing coordinate systems within the pivot coordinates family.

This leads to the idea of considering each pairwise logratio as the first coordinate of an olr coordinate system and complement it with other coordinates through an appropriate SBP procedure. One such choice produces (up to permutation of parts) olr coordinates as in Egozcue et al. (2003), that are formulated in this context as

$$\mathrm{bpc}^{(l)}(\mathbf{x})_i = \sqrt{\frac{i}{i+1}}\ln\frac{\sqrt[i]{\prod_{j=1}^{i} x_j^{(l)}}}{x_{i+1}^{(l)}},\ i=1,\ldots,D-1. \tag{6}$$

The sign matrix associated to the SBP is

$$\begin{array}{lrrrrrcr}
 & x_1^{(l)} & x_2^{(l)} & x_3^{(l)} & x_4^{(l)} & x_5^{(l)} & \ldots & x_D^{(l)} \\
bpc^{(l)}(\mathbf{x})_1 & 1 & -1 & 0 & 0 & 0 & \ldots & 0 \\
bpc^{(l)}(\mathbf{x})_2 & 1 & 1 & -1 & 0 & 0 & \ldots & 0 \\
bpc^{(l)}(\mathbf{x})_3 & 1 & 1 & 1 & -1 & 0 & \ldots & 0 \\
bpc^{(l)}(\mathbf{x})_4 & 1 & 1 & 1 & 1 & -1 & \ldots & 0 \\
\ldots & & & & & & \ldots & 0 \\
bpc^{(l)}(\mathbf{x})_{D-1} & 1 & 1 & 1 & 1 & 1 & \ldots & -1
\end{array}$$

where the element (i, j) of the matrix is 1 if the j-th part is in the numerator in the i-th step of the partition, or -1 if it is in the denominator, or it is 0 if it is not involved in that step.

It must be noted that the ratioing part is placed at the second position in $\mathbf{x}^{(l)}$, so that the pairwise logratio of interest is

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

and the resulting coordinate systems $\mathrm{bpc}^{(1)}(\mathbf{x}), \ldots, \mathrm{bpc}^{(D-1)}(\mathbf{x})$ with the specific interpretation of the first coordinate could be considered as a counterpart to alr coordinates (1) using the ratioing part x_D. Here the superscript l obviously varies only between 1 and $D-1$ for a given ratioing part, but ratioing parts of interest for the practitioner could be selected and up to $D(D-1)$ coordinate systems could be built to represent any possible pairwise logratio. The construction of coordinates using (6) as a reversed run of pivot coordinates (4) motivated the name *backwards pivot coordinates* (bpc). This way we establish a direct relationship between the first backwards pivot coordinate and the respective alr coordinate, namely

$$\mathrm{bpc}^{(l)}(\mathbf{x})_1 = \frac{1}{\sqrt{2}} \mathrm{alr}(\mathbf{x})_l.$$

Here $\mathrm{alr}(\mathbf{x})_l$ stands for the lth alr coordinate and in both cases we use the same ratioing element. Note that backwards pivot coordinates can be used also to represent pairwise logratios resulting from the variable selection procedure described in Greenacre (2018b), adjacent logratio coordinates (Vencálek et al. 2020), or any other set of pairwise logratios which are meaningful for the practitioner, as many as $D(D-1)$.

For instance, with a three-part composition, of components A, B and C, three pairwise logratios are possible, leading to at most three coordinate systems, in which the first coordinate is the backwards pivot. The backward pivots in the first two systems correspond to the alr coordinates with B as ratioing part:

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

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

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

Importantly, as most multivariate statistical methods rely on orthogonal coordinates, pairwise logratios that are determined through olr coordinates using (6) are more appropriate than those which originate from oblique coordinate systems containing the pairwise logratios such as alr coordinates. This is further elaborated in the next section.

3 Consequences for statistical processing

For multivariate statistical methods which focus on analysing samples rather than variables and rely on affine equivariant estimators of location and scale, it is possible to use any of the logratio coordinate systems introduced in the previous section. The results in terms of decisions from statistical hypothesis testing, allocation of objects to specific classes, etc. are always the same. This is for instance the case for outlier detection procedures based on the Mahalanobis distance (Filzmoser and Hron 2008) and linear, quadratic or Fisher discriminant analysis (Filzmoser et al. 2012). Nevertheless, the isometric nature of the coordinates matters for distance-based methods like clustering (Palarea-Albaladejo et al. 2012 Filzmoser et al. 2018) or multidimensional scaling. This is also relevant for regression analysis with compositional response (Egozcue et al. 2012). Here an oblique coordinate representation does not allow to decompose a multivariate regression into univariate regressions as ordinarily without violating the respective decomposition of the residual sum of squares. This would consequently affect statistics based on this decomposition, like the coefficient of determination R^2. Similarly, partial correlations between two orthonormal (orthogonal) coordinates (Erb 2020), for example non-overlapping pairwise log-ratios, would depend on the choice of the other coordinates if they do not complement those two to form an olr coordinate system.

When the logratio coordinates are actually meant to be interpreted as part of the data analysis, then a proper choice is even more important. The problem here is not only the fact that angles, norms and distances are not preserved with oblique coordinate systems (Pawlowsky-Glahn et al. 2015). A lack of orthonormality prevents from using any method that is invariant to orthogonal rotations, like the popular principal component analysis. Additionally, the covariance structure becomes biased when using oblique coordinates, as these do not preserve the total variance of the data set (Greenacre 2018b).

In the following sections we focus on two particular methods that are commonly used in geochemistry, and likewise popular across many other fields: principal component analysis and regression with compositional covariates. We choose these methods to illustrate how using backwards pivot coordinates as introduced in this work provides a more relevant (interpretable) picture of the underlying processes in the data as contained in pairwise logratios, while respecting the structure of the Aitchison geometry and the basic properties of compositional data analysis. Of course these ideas can be extended to other multivariate methods facing an analogous challenge, for example to partial least squares regression (Kalivodová et al. 2015) in a high-dimensional context.

3.1 Principal component analysis

Principal component analysis (PCA) is a well-known dimension reduction method which is also popular in compositional data analysis (Aitchison 1983 Filzmoser et al. 2009). In practical geochemical studies, PCA is sensitive to outliers and its results are very influenced by how the input compositional data are preprocessed (Reid and

Spencer 2009). PCA is commonly used along with the compositional biplot as a two-dimensional graphical representation of the information in the data through the resulting loadings (arrows in the biplot referring to the variables) and scores (points in the biplot referring to the samples) (Aitchison and Greenacre 2002). 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 log-ratios (Greenacre 2018a;b). Focusing on the second option, the PCA method is invariant only to orthogonal rotations, and changes in oblique coordinate systems (like alr or adjacent coordinates) lead to mutually different PCA scores and hence to mutually different loadings of the pairwise log-ratios. These undesirable effects are however avoided using olr coordinates. Although it is possible to minimise those differences by an appropriate selection of pairwise logratios (Greenacre 2018b), there is no guarantee that the differences will be negligible enough so that an acceptable result is obtained regardless of the chosen oblique coordinates.

Instead of using clr coefficients for the construction of compositional biplots as ordinarily done, Kynčlová et al. (2016) introduced the so-called *composed compositional biplot*, where the PCA scores are computed from any pivot coordinate system (4) and the set of PCA loadings results from putting together the first coordinates from each of D pivot coordinate systems isolating in turn all relative information about each component. Given the relationship (5) between olr coordinates and clr coefficients, these loadings differ only by a scaling factor, thus the resulting compositional biplot is visually the same. The main difference, as the naming used for this class of biplot indicates, is that the respective loadings are composed of D pivot coordinate systems. Accordingly, as for the ordinary compositional biplot, it is fully meaningful to interpret the length (approximating the standard deviation of the coordinates) and direction of the arrows. However, when interpreting the links between arrowheads, which indicate proportionality between components in the ordinary clr-based biplot, or the angles between arrows, one should be aware that the loadings come from different coordinate systems. Although for the composed compositional biplot constructed using pivot coordinates (4) this has not practical implications, due to its relation to the compositional biplot with a specific interpretation of the loadings (Aitchison and Greenacre 2002), using the same strategy for any other olr coordinate system could lead to biased conclusions. This is because the loadings depicted in the biplot can correspond to coordinates (pairwise logratios) which are not necessary orthogonal. Accordingly, considering relationships between such variables (through the respective loadings) in terms of correlations should be avoided. The relevance of the composed compositional biplot for the current study is in the idea of composing loadings from first pivot coordinates, that can be extended to *any* coordinates from an olr coordinate system.

We then propose an analogous strategy to develop biplots based on pairwise logratios, in contrast to biplots displaying loadings and scores of PCA performed in alr coordinates (or any other oblique system). Note that in this case not just scores but also loadings, corresponding to pairwise logratios obtained from backwards pivot coordinates, would in general differ from those produced by an oblique coordinate system. Similarly to the case of pivot coordinates, we can interpret direction and length of the respective arrows in the composed compositional biplot. Although we cannot interpret angles between rays in terms of correlations for the same reason as above, we can discuss the results in terms of *proximity* (if both rays point towards the same direction), or *anti-proximity* (if they are orientated in opposite directions). A remarkable advantage of such biplot is that the effect of pairwise logratios on the multivariate data structure can be observed in a realistic manner with respect to the Aitchison geometry. The interpretation of a specific pairwise logratio does not depend on the choice of backwards pivot coordinate system (or any olr coordinate system in general) containing such a logratio. The cost is the limitations of this biplot to investigate the relationships between variables through the loadings.

Another supporting argument for using bpc for PCA of pairwise logratios is the form of the total variance of a random composition $\mathbf{x}$ (Pawlowsky-Glahn et al. 2015),

$$\mathrm{totvar}(\mathbf{x}) = \frac{1}{D}\sum_{i=1}^{D}\sum_{j=i+1}^{D}\mathrm{var}\left(\ln\left(\frac{x_i}{x_j}\right)\right), \tag{7}$$

which can be alternatively expressed using bpc (but in general using *any* orthonormal coordinates) as

$$\mathrm{totvar}(\mathbf{x}) = \sum_{i=1}^{D-1}\mathrm{var}\left(\mathrm{bpc}^{(l)}(\mathbf{x})_i\right)$$

for any $l \in \{1,\ldots,D\}$. However, by using alr coordinates, only part of the total variance is represented, and consequently the corresponding PCA cannot explain the total variance. On the contrary, PCA using bpc corresponds to PCA considering all $D(D-1)/2$ pairwise logratios, as a consequence of the fact that the resulting scores differ only by the scaling constant $1/\sqrt{D}$ (the sum of variances of pairwise logratios in (7) multiplied by $1/D$) and the loadings corresponding to the pairwise logratios have the same direction. Note, however, that using all parwise logratios has its limitations for methods requiring regular data sets. Because the dimensionality of D-part compositions is just $D-1$, the covariance matrix of all pairwise logratios is clearly singular. This would inhibit, for example, estimation of parameters in regression analysis as introduced in the next section, but also computation of robust counterparts to classical estimates in multivariate procedures (see below).

Finally note that, as real-world geochemical data often contain outlying observations, it might be desirable to use robust methods to estimate the parameters involved in compositional PCA (Filzmoser et al. 2009). When computing robust principal components based on a decomposition of the covariance matrix, it is thus necessary to robustly estimate the covariance matrix (as well as the measure of location or central position). In this case, when PCA is based on olr coordinates, the robust covariance estimator needs to be orthogonal equivariant. A popular estimator which is even affine equivariant is the minimum covariance determinant (MCD) estimator (Maronna et al. 2006), which will also be considered below. Since the dimensionality of the compositional data is only $D-1$, it is unavoidable to only use the relevant $D-1$ pairwise logratios as an input, because otherwise (for example when using all pairs) the estimator would not be computable due to a singularity issue. Here, using the bpc approach is thus inevitable and can be recommended on a general basis on the grounds of its generalisability for any estimation method.

3.2 Regression with compositional explanatory variables

A proper coordinate representation of the explanatory composition in a regression model with real response variable (Tolosana-Delgado and van den Boogaart 2011) is key to a meaningful interpretation of regression coefficients. As clearly demonstrated in Coenders and Pawlowsky-Glahn (2020), when an oblique logratio coordinate system is used, the interpretation of the regression coefficient does not correspond to the coordinate to which the parameter is assigned, but its values are interpreted in terms of the other coordinates, following the usual rule when interpreting a regression model: “keeping all other regressors constant”. The inconsistency of the behaviour of alr coordinates in contrast to olr coordinates is also discussed in McGregor et al. (2020) in relation to the magnitude and sign of the regression coefficients in the context of Cox regression with compositional covariates. In brief, pairwise logratios as explanatory variables cannot be interpreted as intended. That is, as the effect of a pairwise trade-off between two parts or, in other words, as the effect of increasing just one part at the expense of decreasing just the other. The regression coefficient of a pairwise logratio must be interpreted as how much the response variable is expected to increase when that logratio coordinate increases while all the remaining logratio coordinates in the equation are kept constant. Thus it depends on the manner in which the remaining logratio coordinates are defined (Coenders and Pawlowsky-Glahn 2020). These caveats can be easily illustrated numerically using a simple example.

We use one of the classical data sets provided by Aitchison (1986), called *Coxite*, which is freely available in the R package ‘compositions’ (van den Boogaart and Tolosana-Delgado 2013). Note that the data set is actually simulated and even the mineral names are fictional. Thus, no meaningful interpretation can be performed from a mineralogical point of view. The data set consists of mineral compositions of 25 rock specimens of coxite type. We consider the subcomposition of three minerals formed by albite, blandite, and cornite to explain porosity through regression analysis. Let us assume that the effect on porosity of the pairwise logratio between albite and cornite is of especial interest to the researcher . With three components, two linear regression models containing $\ln\left(\frac{\text{albite}}{\text{cornite}}\right)$ are possible whose coefficients (estimated by the least-squares method) are shown with standard errors within parentheses in the following:

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

and

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

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. If the ratio $\frac{\text{blandite}}{\text{cornite}}$ remains constant, increasing

$\frac{\text{albite}}{\text{cornite}}$ implies increasing $\frac{\text{albite}}{\text{blandite}}$ by the same factor (Coenders and Pawlowsky-Glahn 2020). Thus, the regression coefficient of $\ln\left(\frac{\text{albite}}{\text{cornite}}\right)$ refers to the effect of increasing albite while decreasing blandite and cornite by the same factor. Conversely, in the second model, the interpretation of the effect of increasing $\ln\left(\frac{\text{albite}}{\text{cornite}}\right)$ must assume that $\ln\left(\frac{\text{albite}}{\text{blandite}}\right)$ is kept constant and, hence, as increasing albite and blandite by a common factor at the expense of decreasing cornite. These interpretations however do not correspond to what the practitioner commonly intends when using pairwise logratios, which is to estimate the effect of increasing just one part at the expense of decreasing just another one, in our case, according to the researcher's interest, the effect of increasing just albite at the expense of reducing just cornite. Note that not only the interpretation changes, also the coefficient estimates and standard errors do so.

The first model above corresponds to using alr coordinates with cornite as ratioing part. Focusing on the alr case, we further illustrate the issues with oblique coordinates in comparison with orthonormal coordinates. The following equation shows the estimates from choosing albite as alr-ratioing part instead:

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

Both the first and the third models constructed from alr coordinates include the trade-off between cornite and albite, but simply expressed by the reciprocal logratios $\ln\left(\frac{\text{albite}}{\text{cornite}}\right)$ and $\ln\left(\frac{\text{cornite}}{\text{albite}}\right)$. Hence, it would be reasonably expected that the associated regression coefficients were the same in magnitude but opposite in sign. However, although the predictions of porosity from the models are the same, the coefficients in fact differ (coefficient $+16.2949$ versus -9.5564), implying a different measure of the influence of that trade-off on porosity and inconsistent interpretations. This issue casts uncertainty on how to breakdown the variance of the response variable amongst the different ratios on the explanatory side. In contrast, the following equation shows estimates from a regression model in which the trade-off between albite and cornite is constructed as a backwards pivot coordinate:

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

When interpreting the effect of $\ln\left(\frac{\text{albite}}{\text{cornite}}\right)$ as a backwards pivot coordinate, blandite does not increase together with either albite nor 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. In other words, the ratio $\frac{\text{albite}}{\text{blandite}}$ increases by a given factor while the ratio $\frac{\text{cornite}}{\text{blandite}}$ gets reduced by the inverse factor. It is thus a matter of the trade-off between albite and cornite only.

3.3 Regression on pairwise logratio analysis through backwards pivot coordinates

When a common interpretation of pairwise logratios, including alr coordinates, is of interest for regression analysis, we propose to employ backwards pivot coordinates (6) by following on the strategy by Hron et al. (2012). Accordingly, being Y a real response variable, up to $D(D-1)$ regression models of the form

$$Y = \beta_0 + \beta_1^{(l)} \mathrm{bpc}^{(l)}(\mathbf{x})_1 + \ldots + \beta_{D-1}^{(l)} \mathrm{bpc}^{(l)}(\mathbf{x})_{D-1} + \varepsilon \tag{8}$$

can be considered, where only the regression coefficient corresponding to $\mathrm{bpc}^{(l)}(\mathbf{x})_1$ is retained. The intercept term β_0 is the same for all models in (8) as well as the overall model statistics, it means the overall F-statistic used for model significance testing, the coefficient of determination R^2, etc. (Johnson and Wichern 2007). As already argued for PCA, it is often useful to consider robust estimators, because outliers could have undesirable effects on the traditional least-squares estimates. Many robust regression estimators have been proposed in the literature, and MM-regression (Maronna et al. 2006) is nowadays a popular choice. This estimator has excellent robustness properties, and also robust statistical inference in terms of hypothesis tests for the regression coefficients is feasible.

Regression coefficient point and variability estimates, associated significance test statistics, and so on, are commonly summarised in a table of results. A regression coefficient $\beta_1^{(l)}$ can be interpreted in the usual manner: it informs about how much the response variable is expected to vary when that explanatory variable, the scaled pairwise logratio $\frac{1}{\sqrt{2}} \ln \frac{x_1^{(l)}}{x_2^{(l)}}$, varies while keeping the remaining coordinates constant. This means increasing the ratio of $x_1^{(l)}$ to $x_2^{(l)}$ while keeping constant all possible pairwise logratios among the parts $x_3^{(l)}$ to $x_D^{(l)}$, and also the ratio between $x_1^{(l)} x_2^{(l)}$ and $\prod_{j=3}^{D} x_j^{(l)}$ (Coenders and Pawlowsky-Glahn 2020).

Unlike with ordinary pairwise logratios, or with any oblique logratio coordinates in general, using backwards pivot coordinates as explanatory terms in regression analysis allows the interpretation of the first regression coefficient $\beta_1^{(l)}$ to correspond to the pairwise trade-off between the two involved parts $x_1^{(l)}$ and $x_2^{(l)}$ only. Moreover, it is not necessary to specify the remaining coordinates, as any other collection of coordinates resulting from an orthogonal rotation of the basis (retaining $\mathrm{bpc}^{(l)}(\mathbf{x})_1$ as the first coordinate) would produce the same result for the term of interest $\beta_1^{(l)}$. Nevertheless, similarly as with pivot coordinates in general, one should be aware when interpreting pairwise logratio regression coefficients that each of them is coming from a different coordinate system, although this has no direct practical implications for analyzing the effect of a pairwise logratio on the response.

Finally, the interpretation of regression coefficients can be further simplified by using *orthogonal* rather than orthonormal coordinates (Müller et al. 2018). These may rely on an alternative base of the logarithm (like the binary logarithm) and suppress the normalising constant in $\mathrm{bpc}^{(l)}(\mathbf{x})_1$, while preserving all the desirable properties of olr coordinates in the regression context. When using the binary logarithm without the normalising constant $\frac{1}{\sqrt{2}}$, the coefficient $\beta_1^{(l)}$ is interpreted as the effect of doubling the ratio between $x_1^{(l)}$ and $x_2^{(l)}$. It is important to note however that ignoring the normalising constant would not be feasible in PCA.

4 Application to sediment compositions

In the following we demonstrate the usefulness of backwards pivot coordinates in practice by an application to analyse sediment compositions from the Jizera River (Czech Republic).

The catchment of the Jizera River is formed by crystalline rocks in headwaters and sedimentary rocks (Matys Grygar et al. 2013) and occasional outcrops of basalts in middle river reach (Faměra et al. 2018). A previous geochemistry study (Matys Grygar et al. 2013) showed relatively simple grain-size control of sediment composition; several lithogenic elements show similar grain-size control as risk elements Cu, Pb, and Zn. This is a fact that justifies the use of the local enrichment factors for quantification of sediment contamination (Bábek et al. 2015 Matys Grygar and Popelka 2016). The sampling site in Horky nad Jizerou is a downstream city of Mladá Boleslav, with traditional production of car batteries, motorcycles and cars (Škoda), which generated weak to moderate contamination in the upper strata of the floodplain sediments (Matys Grygar et al. 2013).

The Jizera River sediments were obtained from 4 drill cores by manual soil corer (grove corer, 3 cm internal diameter, Eijkelkamp, The Netherlands). The sediments were dried, powdered with a planetary micromill, and analysed by X-ray fluorescence spectrometry (XRF) as in preceding studies (Álvarez-Vázquez et al. 2020). The cores were situated in a meander scar, it means abandoned channel (MFJ1 and MFJ2), a former point bar (MFJ3), and a former floodplain outside the erosion bank (MFJ4), see Fig. 1. The samples thus included all major sedimentary facies, it means active channel sediments (MFJ1 to MFJ3, deeper strata), abandoned channel sediments (middle part of MFJ1 and MFJ2), point bar deposits (middle part of MFJ3), overbank fines (upper part of MFJ3 and MFJ4), and levee deposits (uppermost part of MFJ4). The complete information about granulometry is contained in particle size distributions which were analysed recently, in the context of another case study on sediment geochemistry (Matys Grygar et al. 2018). However, for the purpose of this study, we resort to size of particles (in μm) at the 50th percentile in cumulative size distribution function (D50), which characterises major grain size of sediment particles. This was further log-transformed to honour its relative scale.

Elements Al and Si are preferred as ratioing parts in logratios because they are major elements with strongest relation to major mineral constituents of studied sediments and they are immobile in pedogenesis. Concentrations of both these elements are controlled by grain size. The Al concentration is frequently used as a proxy for all sediment constituents finer than quartz sand and element correlating with majority of other elements (except for, say Ca, Si, Zr). The Al and Si concentrations are indirectly proportional as their grain-size control is usually opposite, however, logratios to both Al and Si correct other element concentrations for variable organic matter content. Most other elements than Al and Si, including other major elements, have more complicated grain-size controls (Ti, K) or are not immobile (Fe). Logratios to other elements than Al and Si would bring further sources of variability and produce geochemically more tricky patterns.

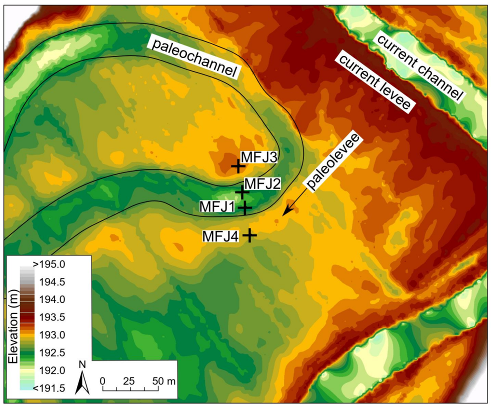

Fig. 1: Digital terrain model of the valley of the Jizera River around sampling sites MFJ1 to MFJ4.

4.1 Regression analysis

In the first part of the analysis we investigate how the granulometry in terms of D50 can be explained by the lithogenic elements Al, Si, K, Ti, Rb, Zr by using pairwise logratios. Content of all these elements is known to be mainly controlled by sediment grain size and not by post-depositional migration or anthropogenic impacts. In the study area, as well as in other temperate regions of Central Europe, also K and Rb behave as other lithogenic elements, being mostly present in primary minerals inherited from parent rocks and not, for example, in soluble salts.

As we know from previous sections, these logratios can be considered, for example, as elements of alr and bpc coordinate systems. The alr approach corresponds to the usual way of treating pairwise logratios and the bpc approach benefits from the properties of orthogonal coordinates. We used orthogonal rather than orthonormal coordinates (Müller et al. 2018) for the regression modelling with bpc both as a simplification and to enhance comparability with the alr case. This means we suppressed the $\frac{1}{\sqrt{2}}$ normalising constant in $\mathrm{bpc}^{(l)}(\mathbf{x})_1$.

Log-transforming D50 led, in addition to express it in the interval scale, to reductions in skewness and kurtosis from 3.85 and 16.73 to 1.71 and 2.95, respectively. After log-transformation, some outliers were evident (maximum Cook's distance equal to 1.06). Given this, following van den Boogaart et al. (2020) we relied on MM robust regression estimation using the lmrob function available in the *robustbase* R package (Maechler et al. 2020).

As argued above, we consider pairwise logratios with Al and Si as ratioing parts both in the bpc and alr approaches. For the alr case, the model was thus 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 (Coenders and Pawlowsky-Glahn 2020). An adjusted $R^2 = 0.853$ confirmed a good fit to the data. The results are summarised in Table 1.

The first significant bpc in the top panel of Table 1 is interpreted as follows. According to the model, increasing the ratio of Al over the remaining parts (K, Ti, Rb, Zr), while decreasing the ratio of Si over these parts by a common factor, leads to a reduction in D50. This means that 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 D50. This is tantamount to saying that there is an effect of increasing Al at the expense of reducing Si on D50, and this corresponds to the natural interpretation of the pairwise logratio between Al and Si. In the same vein, increasing Ti at the expense of reducing Si, and increasing Zr at the expense of reducing Si leads to a statistically significant decrease in D50. The estimates for the ratio Si/Al are redundant with those obtained considering Al/Si, as they are identical in all respects but sign.

These results have a natural geological interpretation. The Al/Si ratio is a widely used proxy for sediment grain size, best performing in mixtures of fine particles of clay minerals (Al and Si are their main element components) and coarser particles of quartz (SiO_2) sand (Bouchez et al. 2011 von Eynatten et al. 2016 Matys Grygar and Popelka 2016 Matys Grygar et al. 2018; 2019 Álvarez-Vázquez et al. 2020). The elements Ti and Zr show concentration maxima in the medium size fraction, fine silt (Ti) and coarse silt/very fine sand (Zr) due to particle size of their carriers and hydrodynamic sorting before sediment deposition (von Eynatten et al. 2016 Matys Grygar and Popelka 2016 Guo et al. 2018 Matys Grygar et al. 2019 Álvarez-Vázquez et al. 2020). Actually, the content of clay and silt fractions relative to fine sand are crucial for D50 in the Jizera River floodplain deposits, which is consistent with the significance of Al/Si, Ti/Si, and Zr/Si when using bpc-based regression.

This straightforward interpretation is not apparent in the bottom panels of Table 1, where we have the corresponding results from the alr-based regression coefficients referring to the same $\ln(\mathrm{Al}/\mathrm{Si})$, $\ln(\mathrm{Ti}/\mathrm{Si})$ and $\ln(\mathrm{Zr}/\mathrm{Si})$ logratios. For example, the p-value for $\ln(\mathrm{Al}/\mathrm{Si})$ is $p = 0.012$ from bpc-based regression, which is in stark contrast to the $p = 0.928$ obtained using alr-based regression. Recall that, as stressed in the previous section, the regression coefficients of alr coordinates do not strictly correspond in terms of interpretation with the trade-off between the components represented in the numerator and denominator of the corresponding pairwise logratio. The fact that the coefficient of the $\ln(\mathrm{Si}/\mathrm{Al})$ ratio is not equal to that of $\ln(\mathrm{Al}/\mathrm{Si})$

Table 1: Robust MM regression of $\ln(\mathrm{D50})$ on the sediment composition using bpc-based models and alr-based models with Si and Al as the ratioing parts.

	Estimate	Std. Error	t-value	p-value
bpc:				
ln(Al/Si)	-1.279	0.494	-2.589	0.012 *
ln(K/Si)	-2.575	1.732	-1.487	0.143
ln(Ti/Si)	-1.944	0.755	-2.574	0.013 *
ln(Rb/Si)	-0.353	1.308	-0.270	0.788
ln(Zr/Si)	-1.150	0.545	-2.112	0.039 *
ln(Si/Al)	1.279	0.494	2.589	0.012 *
ln(K/Al)	-1.296	2.047	-0.633	0.529
ln(Ti/Al)	-0.665	1.065	-0.624	0.535
ln(Rb/Al)	0.926	1.065	0.869	0.389
ln(Zr/Al)	0.128	0.543	0.236	0.814
alr (with Si):				
ln(Al/Si)	-0.124	1.359	-0.091	0.928
ln(K/Si)	-2.716	2.905	-0.935	0.354
ln(Ti/Si)	-1.454	0.843	-1.724	0.091
ln(Rb/Si)	1.728	2.139	0.808	0.423
ln(Zr/Si)	0.133	0.520	0.255	0.800
alr (with Al):				
ln(Si/Al)	2.433	1.031	2.361	0.022 *
ln(K/Al)	-2.716	2.905	-0.935	0.354
ln(Ti/Al)	-1.454	0.843	-1.724	0.091
ln(Rb/Al)	1.728	2.139	0.808	0.423
ln(Zr/Al)	0.133	0.520	0.255	0.800

with opposite sign constitutes a further illustration of the problems to interpret pairwise trade-offs between parts from alr coordinates or any oblique coordinates.

4.2 Principal component analysis

Here we resort to principal component analysis (PCA) where the set of lithogenic elements is complemented by the above mentioned anthropogenic elements Cu, Pb and Zn. In Fig. 2 it is easy to see that for different alr coordinate systems, with Al and Si as denominator of logratios respectively, both the scores and loadings change dramatically. This raises doubts about the reliability of using either alr version for further considerations. On the other hand, the scores are by construction the same for both versions of the composed compositional biplot based on bpc for the respective alr-like pairwise logratios. Green-brown colouring confirms good separation of samples taken from bigger depth and closer to the surface. There are, logically, some differences between loadings in both composed compositional biplots due to different effects of logratios with Al and Si, respectively. Nevertheless, there is an interesting role of the logratio between the reference elements Al and Si. While the arrows have, as expected, opposite directions in the composed compositional biplots, they are somewhat unrelated in their alr alternatives.

From the geochemical perspective, PCA in alr and bpc representations of Al, Si, K, Ti, Rb, Zr, Cu, Zn, and Pb concentrations separates lithogenic elements (Al,

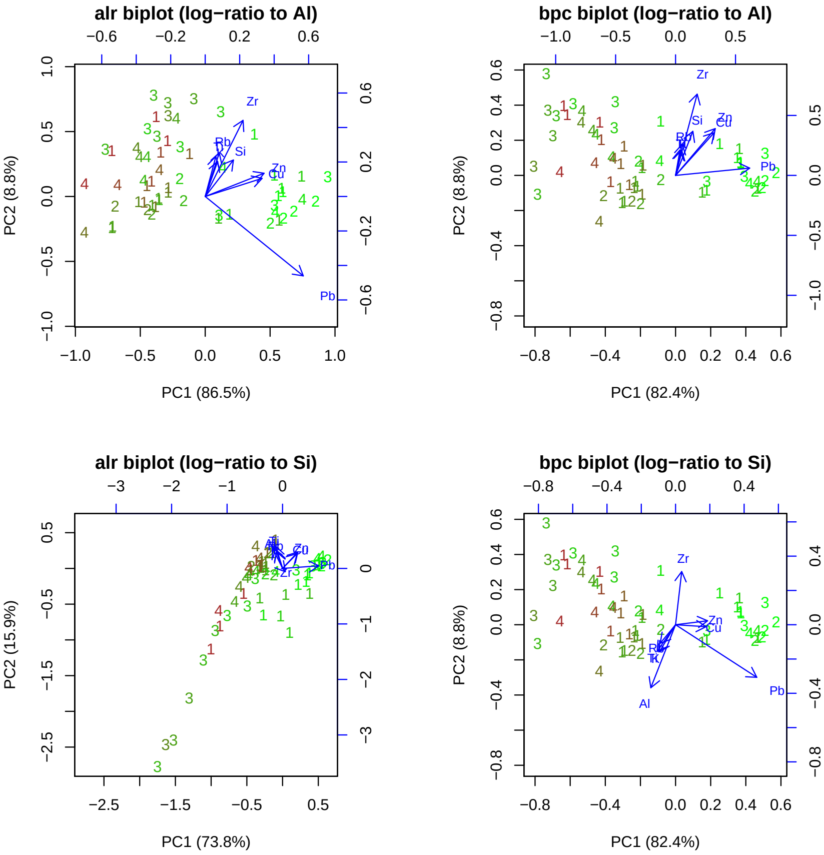

Fig. 2: Biplots of principal component analysis using alr coordinates (left column) and backwards pivot coordinates (right column) where Al (upper row) and Si (lower row) are used as denominator of pairwise logratios. Numbers correspond to drill cores and colour to depths (from brown – deeply drilled samples – to green ones, close to the surface).

Si, K, Ti, Rb and Zr), mostly associated to PC2, and risk elements Cu, Zn and Pb, mostly associated to PC1 (Fig. 2), evidencing a desirable separation of geochemical signals. The plots in Fig. 2 also display correctly the presence of risk-element enriched sediments in the top strata of the depth profiles. The choice of Al or Si as ratioing elements mostly impacted the sign of the loadings for Zr, resulting from distinct grain-size control of Al, Zr, and Si. Zirconium was most abundant (relatively to other elements) in the medium fraction, Al in the finest fraction, and Si in the coarsest fractions of the studied sediments. This is, however, clearly shown only in the composed compositional (bpc) biplots (right column). The performance of PCA can additionally be judged by unequivocal discrimination of uncontaminated and contaminated sediments in their depth profiles. In contaminated floodplains, the top stratum is typically enriched in the risk elements relative to the deeper strata (Matys Grygar et al. 2013; 2014). Risk element concentrations of Cu, Pb, and Zn are controlled by two major factors: grain size (lithology) and the anthropogenic contamination. Both factors also act in the Jizera River floodplain (Fig. 3). The contamination is conventionally characterised by enrichment factor (EF), an empirical way to correct the risk element concentrations in sediments for lithological variations using concentration (log-)ratios. The most common way to compute EF is the so called "double normalisation" (Grosbois et al. 2012 Matys Grygar and Popelka 2016)

$$EF = (M/M_{ref})/(M/M_{ref})_{UCC},$$

where M is the risk element concentration, M_{ref} is the selected reference (lithogenic) element concentration and UCC is the mean upper continental crust as a global geochemical reference. Most authors use Al as M_{ref} (Grosbois et al. 2012 Chen et al. 2014 Matys Grygar and Popelka 2016), however this is not always applicable (Matys Grygar and Popelka 2016 Matys Grygar et al. 2018 Álvarez-Vázquez et al. 2020). The EF works with concentration ratios, which correct for dilution by components not-containing risk and reference elements. If M and M_{ref} have similar grain-size control, their ratio partly corrects M also for the sediment grain size; and EF is a relative measure because the actual element ratio is divided by the element ratio expected for the natural background (for example UCC). The EF is thus dimensionless and concentration scale invariant. In Fig. 3, the logarithm of EF was taken which enabled to express the EF as a difference between the concentration logratios computed for each sample and the UCC logratio. Accordingly, departures from 0 show global enrichment ($\ln(EF) > 0$) or global depletion ($\ln(EF) < 0$). Note that only a single M_{ref} is included in the EF formula, although grain size cannot be fully characterised by a single number as shown in the discussion about the regression analysis for D50. Thus, it is clear that EF is only a coarse approximation.

Fig. 3 shows depth profiles of $\ln(EF)$ and PC1 scores in sediment core MFJ1. A typical property of correctly calculated relative enrichment EF in depth profile is to have stable values near 1 ($\ln(EF)$ near 0) for uncontaminated sediment strata, irrespective of lithology (Matys Grygar et al. 2013; 2014), and stable larger values in the top strata. This is because contamination was homogenised in the Jizera River floodplain by bioturbation and ploughing and the topmost sediments have received persistent secondary contamination (Matys Grygar et al. 2013). The enrichment factor validity is preconditioned by several assumptions, for example applicability of the global reference (UCC) and shape of background function for M versus M_{ref} (Matys Grygar and Popelka 2016). Fig. 3 illustrates how sediment lithology causes scatter in log-transformed Pb concentrations (thus capturing also their original scale) both in the lowermost uncontaminated strata and in the middle of the contaminated strata. Hereby, major features of the sediment lithology are expressed as D50 and also D90 (90th percentile in cumulative grain size curve, thus it shows grain size in the coarse end of the grain size curve). Lithological (grain size) effect is partly eliminated in EF, and that is also well eliminated in the first principal component (PC1) of the bpc representation. This produced satisfactory step-like changes from uncontam-

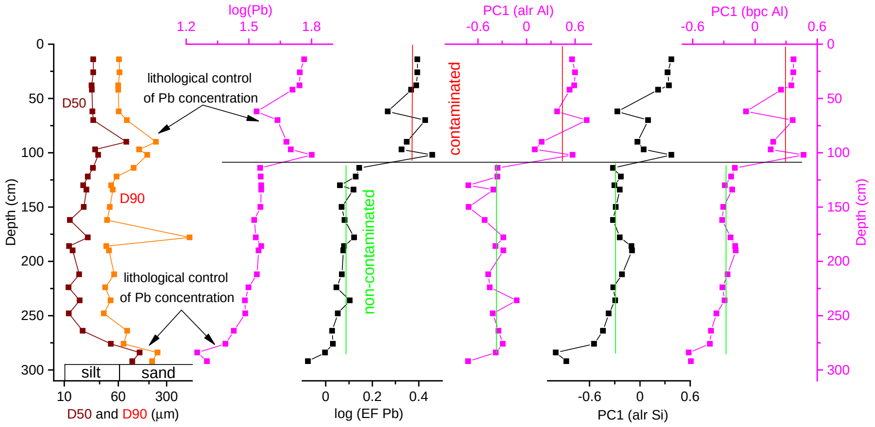

Fig. 3: Sediment lithology compared to EF and PCA outputs.

inated to contaminated sediments, although here the whole multivariate information was used as the input.

It is important to emphasise once again, that the bpc-based PCA scores were produced irrespective of the particular (and subjective) choice of Al, Si, or any other denominator in the logratios in alr coordinates. Because grain size cannot be fully characterised by a single element concentration or a single element concentration ratio (D50 involved at least Al, Si, Ti, and Zr logratios in regression, see Table 1), the results using backwards pivot coordinates should be more reliable also from this perspective. On the contrary, Figures 2 and 3 show different PC1 scores depending on whether either Al or Si is used as ratioing parts in alr-based PCA. This is also relevant for the EF methodology, where the selection of M_{ref} is more critical than it used to be recognised by the researchers employing EF. The EF values really depends on M_{ref} (Álvarez-Vázquez et al. 2020), and a wrong choice of M_{ref} can distort EF by introducing secondary grain-size effect (Matys Grygar and Popelka 2016). Accordingly, the respective bpc loadings show reasonable directions with respect to the multivariate structure of compositional data. This can be observed also in Fig. 2 (upper right), with loadings corresponding to log-ratios with Al in the denominator. Here PC1, capturing the vast majority of the total variation, is dominated by the Pb contamination logratio. The loadings of the two other contaminants (Zn and Cu) are placed partially in the direction of PC2, overriding the effect of lithogenic element logratios (except of Zr). In other words, from the bpc-based biplot, it follows that the contaminant logratios tend to play a dominant role in the multivariate data structure. From the perspective of separating contaminated and uncontaminated samples, the corresponding alr-based biplot (upper left), where PC1 is influenced by all contaminant logratios, seems to produce more convincing results, see Fig. 3, although the scores in the non-contaminated part are still fairly scattered. However, the price to pay using alr coordinates is that the scores can quite heavily depend on the chosen reference element. This can be seen when using scores of alr-based PCA based on Si as denominator, which result in the least convincing contamination indicator amongst the alternatives discussed here.

5 Conclusions

In geological practice, simple pairwise logratios are a common data representation for statistical analysis. Although advances in the logratio methodology over the past few decades have brought some sophisticated alternatives, which might be useful in specific contexts, recent developments in the field suggest that the Occam's Razor rule should be considered. Although there is always an element of subjective judgement, some formulations might sometimes add an unnecessary burden to interpretability in empirical terms. The tools we use for multivariate statistical analysis, here based on logratios, should be as simple as possible, but not simpler. The orthonormality of logratio coordinates is from this perspective a necessary requirement for sound statistical analysis in our view, at least for principal component analysis and regression analysis with compositional explanatory variables as demonstrated in this work. The approach based on backwards pivot coordinates introduced here thus represents a flexible and reliable alternative to using pairwise logratios derived from either oblique coordinate systems or the generating system of all pairwise logratios (or any subset of them).

There are still a number of multivariate statistical methods for which relaxing the orthonormality assumption of coordinates does not matter, as discussed at the beginning of Section 3. Nevertheless, using olr coordinates is a kind of guarantee that things cannot go wrong inadvertently. This does not mean that using sets of pairwise logratios, in the form of alr coordinates or any other representation, will necessarily lead to flawed analysis and results, especially when a careful variable selection is performed (Greenacre 2018b), but one definitely needs to be careful. Accordingly, further research can be carried out with other multivariate statistical methods to show when using conventional pairwise logratios or bpc matters, including, for instance, the machine learning methods which have recently increased their usage in the compositional data community (Tolosana-Delgado et al. 2019) and mediation analysis (Sohn and Li 2019).

A further, challenging point from the empirical perspective is the choice of the reference element for geochemical normalisation in the risk-element enrichment factor method, which is popular in environmental geochemistry. That is, choosing the ratioing element in alr (bpc) coordinates. More reference elements should be preferably used for an "unbiased" enrichment factor in order to reflect that even the median sediment grain size depends substantially on more than one lithogenic element, as demonstrated and rationalised in this work. To gain a first insight, one can still resort to less focused alternatives, like (weighted) pivot coordinates (Filzmoser et al. 2018 Hron et al. 2017).

Finally, note that the general concept of pivot-*like* coordinates can be useful in any context where some logratios, which are not necessarily orthogonal, have to be analysed within one model, as they can always be assigned a pivotal role within an olr coordinate system. After considering the limitations of this approach, which result primarily from the fact that each such logratio corresponds to a *different* olr coordinate system, both theoretical and practical advantages show a strong potential for pivot-like coordinates within the logratio methodology for compositional data analysis.

Acknowledgements The authors thank to Štěpánka Tůmová (IIC Řež) for providing the digital terrain model of the sampling site.

Conflict of interest

The authors declare that they have no conflict of interest.

Availability of data and material

Data are available from the authors on request.

Authors' contributions

KH, GC and TMG conceived this research and designed the experiments; TMG provided the geochemical dataset and interpretations; GC and PF performed the experiments and analysis; KH and GC wrote the first draft of the paper and KH, GC, TMG, PF and JPA all participated in the revisions of it; MF planned and performed sediment sampling, which represented the studied floodplain, and supervised sediment analyses.

References

Aitchison J (1982) The statistical analysis of compositional data (with discussion). Journal of the Royal Statistical Society, Series B (Statistical Methodology) 44(2):139–177

Aitchison J (1983) Principal component analysis of compositional data. Biometrika 70(1):57–65

Aitchison J (1986) The Statistical Analysis of Compositional Data. Chapman & Hall, London. (Reprinted in 2003 with additional material by The Blackburn Press)

Aitchison J, Greenacre M (2002) Biplots for compositional data. Journal of the Royal Statistical Society, Series C (Applied Statistics) 51(4):375–392

Álvarez-Vázquez M Á, Hošek M, Elznicová J, Pacina J, Hron K, Fačevicová K, Talská R, Bábek O, Matys Grygar T (2020) Separation of geochemical signals in fluvial sediments: new approaches to grain-1 size control and anthropogenic contamination. Applied Geochemistry

Bábek O, Matys Grygar T, Faměra M, Hron K, Nováková T, Sedláček J (2015) How to separate the effects of sediment provenance and grain size with statistical rigour? Catena 135:240–253

Billheimer D, Guttorp P, Fagan W (2001) Statistical interpretation of species composition. Journal of the American Statistical Association 96(456):1205–1214

Bouchez J, Lupker M, Gaillardet J, France-Lanord C, Maurice L (2011) How important is it to integrate riverine suspended sediment chemical composition with depth? Clues from Amazon river depth-profiles. Geochimica et Cosmochimica Acta 75(22):6955–6970

Buccianti A (2013) Is compositional data analysis a way to see beyond the illusion? Computers & Geosciences 50:165–173

Chen J B, Gaillardet J, Bouchez J, Louvat P, Wang Y N (2014) Anthropophile elements in river sediments: Overview from the Seine River, France. Geochemistry, Geophysics, Geosystems 15:4526–4546

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 J J (2009) Reply to "On the Harker variation diagrams; . . . " by J.A. Cortés. Mathematical Geosciences 41(7):829–834

Egozcue J J, Daunis-i Estadella J, Pawlowsky-Glahn V, Hron K, Filzmoser P (2012) Simplicial regression. The normal model. Journal of Applied Probability and Statistics 6(1-2):87–108

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

Egozcue J J, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. Mathematical Geology 35(3):279–300

Erb I (2020) Partial correlations in compositional data analysis. Applied Computing and Geosciences 6:100026

Faměra M, Matys Grygar T, Elznicová J, Grison H (2018) Geochemical normalization of magnetic susceptibility for investigation of floodplain sediments. Environmental Earth Sciences 77:189

Filzmoser P, Hron K (2008) Outlier detection for compositional data using robust methods. Mathematical Geosciences 40(3):233–248

Filzmoser P, Hron K (2019) Comments on: Compositional data: the sample space and its structure. TEST 28(3):639–643

Filzmoser P, Hron K, Reimann C (2009) Principal component analysis for compositional data with outliers. Environmetrics 20:621–632

Filzmoser P, Hron K, Templ M (2012) Discriminant analysis for compositional data and robust parameter estimation. Computational Statistics 27(4):585–604

Filzmoser P, Hron K, Templ M (2018) Applied Compositional Data Analysis. Springer Series in Statistics. Springer, Cham

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

Greenacre M (2018a) Compositional Data in Practice. CRC Press, Boca Raton

Greenacre M (2018b) Variable selection in compositional data analysis using pairwise logratios. Mathematical Geosciences 51:649–682

Greenacre M (2019) Comments on: Compositional data: the sample space and its structure. TEST 28(3):644–652

Grosbois C, Meybeck M, Lestel L, Lefévre I, Moatar F (2012) Severe and contrasted polymetallic contamination patterns (1900–2009) in the Loire River sediments (France). Science of the Total Environment 435:290–305

Guo Y L, Yang S Y, Su N, Li C, Yin P, Wang Z B (2018) Revisiting the effects of hydrodynamic sorting and sedimentary recycling on chemical weathering indices. Geochimica et Cosmochimica Acta 227:48–63

Hron K, Filzmoser P, Caritat P d, Fišerová E, Gardlo A (2017) Weighted pivot coordinates for compositional data and their application to geochemical mapping. Mathematical Geosciences 49(6):797–814

Hron K, Filzmoser P, Thompson K (2012) Linear regression with compositional explanatory variables. Journal of Applied Statistics 39(5):1115–1128

Johnson R, Wichern D (2007) Applied Multivariate Statistical Analysis. Upper Saddle River: Prentice Hall, 6th edition

Kalivodová A, Hron K, Filzmoser P, Najdekr L, Janečková H, Adam T (2015) PLS-DA for compositional data with application to metabolomics. Journal of Chemometrics 29(1):21–28

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

Maechler M, Rousseeuw P, Croux C, Todorov V, Ruckstuhl A, Salibian-Barrera M, Verbeke T, Koller M, Conceicao E, di Palma M (2020) robustbase: Basic Robust Statistics. R package version 0.93-6

Maronna R, Martin D, Yohai V (2006) Robust Statistics: Theory and Methods. Chichester: John Wiley & Sons

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

Matys Grygar T, Elznicová J, Bábek O, Hošek M, Engel Z, Kiss T (2014) Obtaining isochrones from pollution signals in a fluvial sediment record: A case study in a uranium-polluted floodplain of the Ploučnice River, Czech Republic. Applied Geochemistry 48:1–15

Matys Grygar T, Hošek M, Pacina J, Štojdl J, Bábek O, Sedláček J, Hron K, Talská R, Kříženecká S, Tolaszová J (2018) Changes in the geochemistry of fluvial sediments after dam construction (the Chrudimka River, the Czech Republic). Applied Geochemistry 98:94–108

Matys Grygar T, Mach K, Martinez M (2019) Checklist for the use of potassium concentrations in siliciclastic sediments as paleoenvironmental archives. Sedimentary Geology 382:75–84

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. Science of the Total Environment 452:233–245

Matys Grygar T, Popelka J (2016) Revisiting geochemical methods of distinguishing natural concentrations and pollution by risk elements in fluvial sediments. Journal of Geochemical Exploration 170:39–57

McGregor D E, Palarea-Albaladejo J, Dall P M, Hron K, Chastin S F (2020) Cox regression survival analysis with compositional covariates: Application to modelling mortality risk from 24-h physical activity patterns. Statistical Methods in Medical Research 29(5):1447–1465

Müller I, Hron K, Fišerová E, Šmahaj J, Cakirpaloglu P, Vančáková J (2018) Interpretation of compositional regression with application to time budget analysis. Austrian Journal of Statistics 47(2):3–19

Palarea-Albaladejo J, Martín-Fernández J A, Soto J A (2012) Dealing with distances and transformations for fuzzy c-means clustering of compositional data. Journal of Classification 29(2):144–169

Pawlowsky-Glahn V, Buccianti A, editors (2011) Compositional Data Analysis: Theory and Applications. Wiley, Chichester

Pawlowsky-Glahn V, Egozcue J J (2001) Geometric approach to statistical analysis on the simplex. Stochastic Environmental Research and Risk Assessment (SERRA) 15(5):384–398

Pawlowsky-Glahn V, Egozcue J J, Tolosana-Delgado R (2015) Modeling and Analysis of Compositional Data. Wiley, Chichester

Reid M, Spencer K L (2009) Use of principal components analysis (PCA) on estuarine sediment datasets: The effect of data pre-treatment. Environmental Pollution 157:2281–2275

Sohn M, Li H (2019) Compositional mediation analysis for microbiome studies. The Annals of Applied Statistics 13(1):661–681

Thomas C, Aitchison J (2006) Log-ratios and geochemical discrimination of Scottish Dalradian limestones: A case study. In Buccianti A, Mateu-Figueras G, Pawlowsky-Glahn V, editors, Compositional Data Analysis in the geosciences: From Theory to Practice, Geological Society, London, 25–41

Thomas C W, Aitchison J (2005) Compositional data analysis of geological variability and process: a case study. Mathematical Geology 37(7):753–772

Tolosana-Delgado R, Talebi H, Khodadadzadeh M, Van den Boogaart K (2019) On machine learning algorithms and compositional data. In Ortego M, editor, Proceedings of the 8th International Workshop on Compositional Data Analysis (CoDaWork2019): Terrassa, 3-8 June, 2019, Universitat Politècnica de Catalunya-BarcelonaTECH, 172–175

Tolosana-Delgado R, van den Boogaart K (2011) Linear models with compositions in R. In Pawlowsky-Glahn V, Buccianti A, editors, Compositional Data Analysis: Theory and Applications, Wiley, Chichester, 356–371

van den Boogaart K, Filzmoser P, Hron K, Templ M, Tolosana-Delgado R (2020) Classical and robust regression analysis with compositional data. Mathematical Geosciences :1–36

van den Boogaart K, Tolosana-Delgado R (2013) Analyzing Compositional Data with R. Springer, Heidelberg

Vencálek O, Hron L, Filzmoser P (2020) A comparison of generalised linear models and compositional models for ordered categorical data. Statistical Modelling 20(3):249–273

von Eynatten H, Tolosana-Delgado R, Karius V, Bachmann K, Caracciolo L (2016) Sediment generation in humid mediterranean setting: grain-size and source-rock control on sediment geochemistry and mineralogy (Sila Massif, Calabria). Sedimentary Geology 336:68–80