

Correlation Between Compositional Parts Based on Symmetric Balances

Petra Kynčlová¹ · Karel Hron² · Peter Filzmoser¹

Received: 14 October 2015 / Accepted: 19 December 2016 / Published online: 4 January 2017
© The Author(s) 2017. This article is published with open access at Springerlink.com

Abstract Correlation coefficients are most popular in statistical practice for measuring pairwise variable associations. Compositional data, carrying only relative information, require a different treatment in correlation analysis. For identifying the association between two compositional parts in terms of their dominance with respect to the other parts in the composition, symmetric balances are constructed, which capture all relative information in the form of aggregated logratios of both compositional parts of interest. The resulting coordinates have the form of logratios of individual parts to a (weighted) “average representative” of the other parts, and thus, they clearly indicate how the respective parts dominate in the composition on average. The balances form orthonormal coordinates, and thus, the standard correlation measures relying on the Euclidean geometry can be used to measure the association. Simulation studies provide deeper insight into the proposed approach, and allow for comparisons with alternative measures. An application from geochemistry (Kola moss) indicates that correlations based on symmetric balances serve as a sensitive tool to reveal underlying geochemical processes.

✉ Petra Kynčlová
kynclova.petra@gmail.com

Karel Hron
hronk@seznam.cz

Peter Filzmoser
p.filzmoser@tuwien.ac.at

¹ Institute of Statistics and Mathematical Methods in Economics, Vienna University of Technology, Wiedner Hauptstrasse 8-10, 1040 Vienna, Austria

² Department of Mathematical Analysis and Applications of Mathematics, Faculty of Science, Palacký University, 17. listopadu 12, 771 46 Olomouc, Czech Republic

Keywords Correlation analysis · Compositional data · Sequential binary partitioning · Symmetric balances · Logratio transformations

1 Introduction

Compositional data are characterized by observations on compositional parts that contribute to some whole. Typical examples are the number of votes for political parties in a regional election with a given population or concentrations of chemical elements in some material with defined weight. An analysis of the associations between the compositional parts (political parties and chemical elements) based on the underlying data is often a first step to understand the multivariate data structure. However, applying correlation analysis to compositional data can lead to the so-called spurious correlations. The problem of spurious correlations dates back to the seminal paper by [Pearson \(1897\)](#), where difficulties obtained by applying the standard correlation analysis to data with a constant sum constraint are described. There was a long way with one important milestone, [Chayes \(1960\)](#) to realize that any such reasonable measure cannot be based on the original compositional parts, but rather on (log) ratios forming the only relevant information in compositions ([Aitchison 1986](#)). In the following years, it turned out that compositional data are not restricted entirely to observations with a constant sum constraint (such as proportions or percentages), but the concept covers all observations carrying relative information, with a possibility of being expressed with any prescribed sum constraint without altering the ratios between the parts ([Pawlowsky-Glahn et al. 2015](#)). The specific principles of compositional data (scale invariance, permutation invariance, and subcompositional coherence) induce the Aitchison geometry ([Pawlowsky-Glahn and Egozcue 2001](#)) with the Euclidean vector space structure that enables to express compositions in proper logratio coordinates and continue with statistical processing using the standard multivariate statistical tools.

[Aitchison \(1986\)](#) proposed to change completely the point of view on association between compositional parts by introducing the variation matrix. Accordingly, the association between two parts, expressed by the variance of the corresponding logratio, is stronger when the ratio between them tends to be constant. Although this concept turned out to be successful in a range of applications during the last 30 years ([Pawlowsky-Glahn and Buccianti 2011](#)), there are still certain limitations of the approach that inhibits its wider acceptance by the geochemical community ([Filzmoser et al. 2010](#); [Reimann et al. 2012](#)). They result mainly from the lack of possibilities of distinguishing positive and negative association, an essential feature in case of the correlation coefficient. To get an impression about such a behavior between geochemical variables, many researchers in the field tend to return back to improper preprocessing tools, such as the log transformation that violates the scale invariance principle of compositional data.

This paper proposes to measure the strength of association between compositional parts through the correlation coefficient between a particular choice of orthonormal coordinates with respect to the Aitchison geometry. The orthonormal coordinates are based on logratios, formed always by a part of interest and the remaining variables,

aggregated in terms of a weighted geometric mean. The resulting coordinates are simply logratios of individual parts to a (weighted) “average representative” of the other parts, and thus, they clearly indicate how the respective parts dominate in the composition on average. Methodologically, it follows the idea of having logratio coordinates that express all relative information about the parts of interest (Filzmoser et al. 2009). Two such coordinates need to be constructed simultaneously in a coordinate system, each corresponding to one of the parts. After a brief review of recent possibilities concerning the association between compositional parts in the next section, these coordinates are derived in Sect. 3. A detailed discussion of the new correlation measure together with some possible alternatives is provided in Sect. 4. Sections 5 and 6 employ a geochemical data set in simulations and comparisons to provide deeper insight into the properties of the proposed association measure. Section 7 concludes and provides some outlook.

2 Measures of Compositional Association

2.1 Correlation Analysis for Compositional Data

The most popular way of measuring association (relation) between variables in practice is using a correlation measure. Nevertheless, its application on compositional data is not so straightforward. Recall that a D -part composition is represented as a vector $\mathbf{x} = (x_1, \dots, x_D)'$, where all components are positive real numbers that carry only relative information (Aitchison 1986; Pawłowsky-Glahn et al. 2015). This means that only the ratios between the parts are informative and they form the basis of a reasonable (statistical) processing. Moreover, one should follow the principles of compositional data (Egozcue 2009) to have a guarantee of a reliable analysis. Particularly, the representation of a compositional vector with any sum of components (proportions, percentages, mg/kg,...) should yield the same results according to the scale invariance principle. These essential assumptions constitute the source of the problems to apply the standard correlation analysis on compositional data.

Consider compositional data with a fixed prescribed constant sum constraint (the case of proportions) that still occur sometimes in the literature. In this case, correlation analysis is also influenced by the presence of negative bias in the covariance structure. It is represented by the relations

$$\text{cov}(x_i, x_1) + \dots + \text{cov}(x_i, x_{i-1}) + \text{cov}(x_i, x_{i+1}) + \text{cov}(x_i, x_D) = -\text{var}(x_i), \quad (1)$$

for $i = 1, \dots, D$, that make the interpretation of the correlation coefficient meaningless (its value cannot freely vary between -1 and 1). Consequently, the correlation between parts of a composition with D parts can be completely in contradiction with the correlation resulting from a subcomposition containing d parts, $d \leq D$, and an illustrative example is described in Korhoňová et al. (2009). The problem is that the standard approach, when interpreting the correlation between two compositional parts, does not reflect the fact that the whole has changed when coming from the full composition to a subcomposition. On the other hand, this is intentionally recognized and

taken into account with the approach proposed in this paper. In general, correlation analysis provides an illustration of the fact that a standard statistical analysis of the original compositional data (that are driven by the Aitchison geometry) cannot be recommended in general.

The Euclidean vector space structure of the Aitchison geometry enables to get a coordinate representation of compositions in the real space, where the standard statistical methods can be applied. The resulting centered logratio (clr) coefficients (Aitchison 1986) and isometric logratio (ilr) coordinates (Egozcue et al. 2003), which seem to be recently the most popular in practice, correspond to coordinates with respect to a generating system and an orthonormal basis, respectively.

Accordingly, the clr coefficients are defined as

$$\mathbf{y} = (y_1, \dots, y_D)' = \left(\ln \frac{x_1}{\sqrt[D]{\prod_{i=1}^D x_i}}, \dots, \ln \frac{x_D}{\sqrt[D]{\prod_{i=1}^D x_i}} \right)', \quad (2)$$

imposing the zero sum constraint of the new variables, $y_1 + \dots + y_D = 0$. Although it seems to be attractive to assign each single original compositional part to a clr coefficient (and then even continue with correlation analysis), this effort has no geometrical background and should be avoided. Particularly, similar relations as those in Eq. (3)

$$\text{cov}(y_i, y_1) + \dots + \text{cov}(y_i, y_{i-1}) + \text{cov}(y_i, y_{i+1}) + \text{cov}(y_i, y_D) = -\text{var}(y_i), \quad (3)$$

for $i = 1, \dots, D$ that show a distortion of the covariance structure, support the argumentation.

Following general theoretical assumptions (Eaton 1983), correlation analysis of compositional data in the usual sense is only meaningful in logratio coordinates with respect to a basis, preferably to an orthonormal one, that guarantees isometry between the Aitchison geometry and the real space. Nevertheless, the vector of ilr coordinates has $D - 1$ elements, and it is not possible to assign a coordinate to each part in an univocal manner. Searching for interpretable orthonormal (ilr) coordinates led to the concept of balances (Egozcue and Pawłowsky-Glahn 2005) as coordinates with a specific interpretation in terms of balances between groups of compositional parts. These new coordinates are constructed using a procedure called sequential binary partitioning (SBP), where the original parts are separated sequentially into non-overlapping groups of parts (Egozcue and Pawłowsky-Glahn 2005). Although correlation analysis of balances is now possible, the interpretation is not straightforward without a deeper prior (expert) knowledge of how the SBP should be constructed. A recent discussion on the issue from the perspective of geochemical mapping can be found in McKinley et al. (2016).

Consequently, an alternative approach was introduced coming from the idea of having “automated” coordinates that would better stress the role of single compositional parts (Filzmoser et al. 2009; Fišerová and Hron 2011). A particular form of SBP leads to coordinates

$$\mathbf{z} = (z_1, \dots, z_{D-1})', \quad z_i = \sqrt{\frac{D-i}{D-i+1}} \ln \frac{x_i}{\sqrt[D-i]{\prod_{j=i+1}^D x_j}}, \tag{4}$$

$i = 1, \dots, D - 1$. It is obvious that the balance z_1 , being proportional to y_1 , contains all the relative information of the part x_1 with respect to the remaining parts of the composition, since this part is not contained in any other coordinate of Eq. (4). The variable z_1 can be interpreted in terms of dominance of x_1 to the other parts, represented by their geometric mean, and thus to their average behavior. Up to a constant, z_1 is equal to one centered logratio coordinate (Aitchison 1986), being still popular with some statistical applications (Pawlowsky-Glahn et al. 2015). Unfortunately, the same interpretation in sense of explaining all relative information cannot be assigned to z_2 and x_2 , because this balance already does not contain the first part. Nevertheless, a good candidate for the correlation between relative contributions of x_1 and x_2 in a given composition would be a symmetrical form of z_1 and z_2 because of the exclusive position of the parts of interest (x_1, x_2) in the respective coordinates. This task will be further developed in Sect. 3. Obviously, the role of x_1 and x_2 can also be interchanged, and a similar construction for different parts can be obtained by permuting the parts in Eq. (4). Without loss of generality, just the case of x_1 and x_2 will be considered in the following.

2.2 Variation Matrix as a Measure of Stability

A main tool of measuring compositional association between two compositional parts has been the variation matrix as a measure of stability (Aitchison 1986). The variation matrix of a D -part composition is a symmetric matrix of order D , defined as

$$\mathbf{T} = [t_{ij}] = \left[\text{var} \left(\ln \frac{x_i}{x_j} \right) \right], \quad i, j = 1, \dots, D, \tag{5}$$

with zero diagonal elements. When the elements of \mathbf{T} are close to 0, the ratio of x_i/x_j is nearly constant, thus the two parts x_i and x_j are almost proportional. On the contrary, high variability of the logratio indicates very different ratios of two parts among all the observations.

The logratios in Eq. (5) can also be rescaled according to Eq. (4), so that they correspond, up to orientation, to the normed coordinate of the two-part composition $(x_i, x_j)'$. The resulting normalized variation matrix (Pawlowsky-Glahn et al. 2015) is defined as

$$\mathbf{T}^* = [t_{ij}^*] = \left[\text{var} \left(\frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j} \right) \right], \quad i, j = 1, \dots, D, \tag{6}$$

where t_{ij}^* stands for the usual (sample) variance of the normalized logratio of parts i and j (balance). Subsequently, the relation between \mathbf{T} and \mathbf{T}^* is given as $\mathbf{T} = \frac{1}{2} \mathbf{T}^*$. The measure of variability could be normalized to the range (0,1] as $\tau_{ij} = \exp(-\text{var}(t_{ij}^*))$ for $1 \leq i, j \leq D, i \neq j$ (Buccianti and Pawlowsky-Glahn 2005; Filzmoser et al. 2010). The proportionality coefficient τ_{ij} tends to 0 as the variability of the logratio

increases, and conversely, smaller variabilities deliver τ_{ij} approaching 1. However, this is still just a proper scaling of the elements of the variation matrix and not a correlation measure in the common sense. Particularly, the concept of proportionality does not allow to think in terms of positive and negative association, as it is known from the correlation coefficient.

The above considerations lead to an alternative normalization of the elements of the variation matrix using the total variance measure (Pawlowsky-Glahn and Egozcue 2001); the resulting matrix is used to reveal, whether the corresponding pair of parts is less proportional than the logratio variance that would be observed in a complete non-proportional composition (Egozcue et al. 2013). For testing the elements of the variation matrix (indirectly through a regression model), it was recommended in Egozcue et al. (2013) to consider the following two balances

$$z_{1(i,j)}^v = \frac{1}{\sqrt{2}} \ln \frac{x_i}{x_j}, \quad z_{2(i,j)}^v = \frac{\sqrt{2(D-2)}}{\sqrt{D}} \ln \frac{x_i x_j}{\sqrt{\prod_{k \neq \{i,j\}} x_k}},$$

accompanied by the complementary $D - 3$ orthonormal coordinates. The first balance corresponds to an element of the normalized variation matrix, and the latter one links this logratio (capturing relative information on the subcomposition $(x_i, x_j)'$) with the remaining parts in the given composition. Nevertheless, the interpretation of the elements of the variation matrix themselves has not been further enhanced using these approaches.

3 Constructing Symmetric Balances

All the introduced approaches to measuring association between compositional parts are based, directly or indirectly, on working with orthonormal coordinates. However, constructing interpretable balances with SBP for correlation analysis needs some experience or even some prior expertise. It is also important to note that the normalized variation matrix considers only associations between two parts of a given composition through their respective logratios. Although this is relevant when the amounts (mass, matter, and volume) that gave rise to the ratios are of primary interest, one should be aware that any part in the compositional vector can be by definition dependent on ratios with all other parts in the composition. This fact should be taken into account for considering any reasonable (preferably orthonormal) coordinates that would allow for a correlation analysis between relative contributions conveyed by both parts. As mentioned in the previous section, one possible setting of coordinates would be Eq. (4). Nevertheless, it is necessary to symmetrize with respect to parts x_1 and x_2 .

Accordingly, two coordinate systems \mathbf{z} and \mathbf{z}^* resulting from the permutation of the parts in Eq. (4) are considered with a special focus on the role of x_1 and x_2 , respectively. It is obvious that the first two coordinates from each system, Eqs. (7) and (8), fully describe the subcomposition $(x_1, x_2)'$ within the given composition

$$z_1 = \sqrt{\frac{D-1}{D}} \ln \frac{x_1}{\sqrt{D-1} \sqrt{\prod_{i=2}^D x_i}}, \quad z_2 = \sqrt{\frac{D-2}{D-1}} \ln \frac{x_2}{\sqrt{D-2} \sqrt{\prod_{i=3}^D x_i}}, \tag{7}$$

$$z_1^* = \sqrt{\frac{D-1}{D}} \ln \frac{x_2}{\sqrt{D-1} \sqrt{x_1 \prod_{i=3}^D x_i}}, \quad z_2^* = \sqrt{\frac{D-2}{D-1}} \ln \frac{x_1}{\sqrt{D-2} \sqrt{\prod_{i=3}^D x_i}}. \tag{8}$$

It is now possible to build matrices of clr representations of orthonormal basis vectors corresponding to the first two balances of \mathbf{z} and \mathbf{z}^* (Egozcue et al. 2003) as

$$\mathbf{V}_{\mathbf{z}} = \begin{pmatrix} \sqrt{\frac{D-1}{D}} & 0 \\ -\frac{1}{D-1} \sqrt{\frac{D-1}{D}} & \sqrt{\frac{D-2}{D-1}} \\ -\frac{1}{D-1} \sqrt{\frac{D-1}{D}} & -\frac{1}{D-2} \sqrt{\frac{D-2}{D-1}} \\ \vdots & \vdots \\ -\frac{1}{D-1} \sqrt{\frac{D-1}{D}} & -\frac{1}{D-2} \sqrt{\frac{D-2}{D-1}} \end{pmatrix},$$

$$\mathbf{V}_{\mathbf{z}^*} = \begin{pmatrix} -\frac{1}{D-1} \sqrt{\frac{D-1}{D}} & \sqrt{\frac{D-2}{D-1}} \\ \sqrt{\frac{D-1}{D}} & 0 \\ -\frac{1}{D-1} \sqrt{\frac{D-1}{D}} & -\frac{1}{D-2} \sqrt{\frac{D-2}{D-1}} \\ \vdots & \vdots \\ -\frac{1}{D-1} \sqrt{\frac{D-1}{D}} & -\frac{1}{D-2} \sqrt{\frac{D-2}{D-1}} \end{pmatrix},$$

where $\mathbf{V}_{\mathbf{z}^*}$ results from a permutation of the first two rows of $\mathbf{V}_{\mathbf{z}}$. Consequently, the first two balances of \mathbf{z} and \mathbf{z}^* are related through an orthogonal transformation as

$$\mathbf{z}^* = \mathbf{V}_{\mathbf{z}}' \mathbf{V}_{\mathbf{z}^*} \mathbf{z}, \tag{9}$$

where the orthogonal matrix $\mathbf{V}_{\mathbf{z}}' \mathbf{V}_{\mathbf{z}^*}$ has the form:

$$\mathbf{V}_{\mathbf{z}}' \mathbf{V}_{\mathbf{z}^*} = \begin{pmatrix} -\frac{1}{D-1} & \sqrt{\frac{D-2}{D} \frac{D}{D-1}} \\ \sqrt{\frac{D-2}{D} \frac{D}{D-1}} & \frac{1}{D-1} \end{pmatrix}. \tag{10}$$

Note that both matrices $\mathbf{V}_{\mathbf{z}}$ and $\mathbf{V}_{\mathbf{z}^*}$ are closely connected to the respective coordinates. Namely, their columns $\mathbf{v}_1 = (v_{11}, \dots, v_{D1})'$, $\mathbf{v}_2 = (v_{12}, \dots, v_{D2})'$ and $\mathbf{v}_1^* = (v_{11}^*, \dots, v_{D1}^*)'$, $\mathbf{v}_2^* = (v_{12}^*, \dots, v_{D2}^*)'$ with zero sums of their elements represent logcontrast coefficients of z_1, z_2 , and z_1^*, z_2^* , respectively (Aitchison 1986)

$$z_1 = \sum_{i=1}^D v_{i1} \ln x_i, \quad z_2 = \sum_{i=1}^D v_{i2} \ln x_i, \quad z_1^* = \sum_{i=1}^D v_{i1}^* \ln x_i, \quad z_2^* = \sum_{i=1}^D v_{i2}^* \ln x_i.$$

Because of the roles of the above-mentioned coordinates with respect to the single parts x_1 and x_2 , one can construct new symmetric balances capturing their relative contributions expressed through logratios to other parts in the composition. Let x_1 be the first part of interest; the case of x_2 can be processed accordingly. Based on the basic geometry, a symmetric coordinate z_1^s capturing relative information about x_1 corresponds to an angle bisector of \mathbf{v}_1 and \mathbf{v}_2^* . Similarly, the coordinate z_2^s (that stands for x_2) would correspond to an angle bisector of \mathbf{v}_2 and \mathbf{v}_1^* . Figure 1 provides an illustration.

Particularly, the new symmetric orthonormal coordinate is computed using the respective logcontrast coefficients as

$$z_1^s = \frac{1}{\|\mathbf{v}_1 + \mathbf{v}_2^*\|} (\mathbf{v}_1 + \mathbf{v}_2^*)' \ln \mathbf{x}. \tag{11}$$

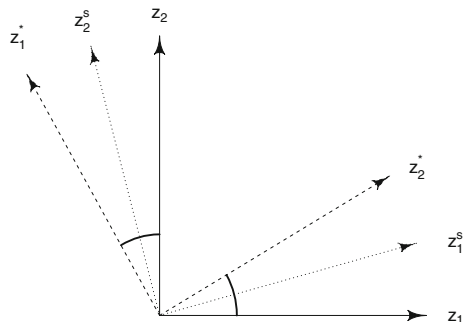
The sum of \mathbf{v}_1 and \mathbf{v}_2^* results in a vector with elements

$$\mathbf{v}_1 + \mathbf{v}_2^* = \left(\frac{D - 1 + \sqrt{D(D - 2)}}{\sqrt{D(D - 1)}}, -\frac{1}{\sqrt{D(D - 1)}}, -\frac{\sqrt{D - 2} + \sqrt{D}}{\sqrt{D(D - 1)(D - 2)}}, \dots, -\frac{\sqrt{D - 2} + \sqrt{D}}{\sqrt{D(D - 1)(D - 2)}} \right)'$$

and norm

$$\|\mathbf{v}_1 + \mathbf{v}_2^*\| = \sqrt{\frac{2 \cdot (D - 1 + \sqrt{D(D - 2)})}{D - 1}}. \tag{12}$$

Fig. 1 Graphical illustration of the symmetric balances



Subsequently, logcontrast coefficients of the symmetric coordinate z_1^s are given as

$$\frac{\mathbf{v}_1 + \mathbf{v}_2^*}{\|\mathbf{v}_1 + \mathbf{v}_2^*\|} = \left(\frac{\sqrt{D-1 + \sqrt{D(D-2)}}}{\sqrt{2D}}, -\frac{1}{\sqrt{2D(D-1 + \sqrt{D(D-2)})}}, \right. \\ \left. -\frac{\sqrt{D-2} + \sqrt{D}}{\sqrt{2D(D-2)(D-1 + \sqrt{D(D-2)})}}, \dots, \right. \\ \left. -\frac{\sqrt{D-2} + \sqrt{D}}{\sqrt{2D(D-2)(D-1 + \sqrt{D(D-2)})}} \right)'$$

followed by the resulting coordinate

$$z_1^s = \sqrt{\frac{D-1 + \sqrt{D(D-2)}}{2D}} \ln \frac{x_1}{x_2^{\frac{1}{D-1+\sqrt{D(D-2)}}} (x_3 x_4 \dots x_D)^{\frac{\sqrt{D-2} + \sqrt{D}}{\sqrt{D-2(D-1+\sqrt{D(D-2)})}}}}. \tag{13}$$

The same procedure is applied to the coordinates z_1^* and z_2 , describing information about the compositional part x_2 , in order to obtain the second symmetric coordinate z_2^s . Thus

$$z_2^s = \sqrt{\frac{D-1 + \sqrt{D(D-2)}}{2D}} \ln \frac{x_2}{x_1^{\frac{1}{D-1+\sqrt{D(D-2)}}} (x_3 x_4 \dots x_D)^{\frac{\sqrt{D-2} + \sqrt{D}}{\sqrt{D-2(D-1+\sqrt{D(D-2)})}}}}. \tag{14}$$

From the above construction, it is clear that $z_1^s, z_2^s, z_3, \dots, z_{D-1}$, or alternatively $z_1^s, z_2^s, z_3^*, \dots, z_{D-1}^*$, form orthonormal coordinates of the composition \mathbf{x} . The interpretation of the resulting symmetric balances is indeed as expected, and they both capture dominance of x_1 and x_2 , respectively, with respect to the other components in a symmetric manner. Although the coefficients in the denominator of Eqs. (13) and (14) seem to be quite complicated, one does not need to take care about them in practice, because they result just from the normalization needed to achieve orthonormality of the coordinates. More important is the powering of x_2 in z_1^s (and x_1 in z_2^s) that is different for the remaining parts, which reflects the compromise resulting from symmetrizing the input coordinates, Eqs. (7) and (8). Nevertheless, it is visible that the ratio of both powers

$$\frac{\frac{1}{D-1+\sqrt{D(D-2)}}}{\frac{\sqrt{D-2} + \sqrt{D}}{\sqrt{D-2(D-1+\sqrt{D(D-2)})}}} = \frac{\sqrt{D-2}}{\sqrt{D-2} + \sqrt{D}}$$

(Fig. 2) is stabilized quite soon with an increasing number of parts to approximately 1/2 in favor of the remaining parts. Finally, with an increasing number of parts, the effect of having one part in the denominator of the logratio with different power than

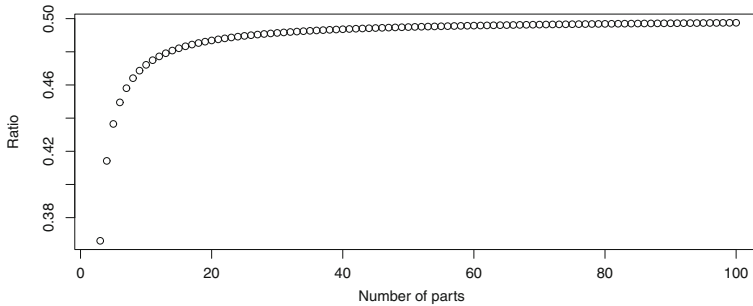


Fig. 2 Ratio between weights in symmetric balances

for the other parts is clearly suppressed. As a consequence, z_1^s and z_2^s tend to approach (up to a scaling constant) the respective clr coordinates y_1 and y_2 as $D \rightarrow \infty$.

4 Correlation Analysis with Symmetric Balances

The symmetric balances, as constructed in the previous section, allow to perform correlation analysis between coordinates which express one part of interest with respect to the other parts in the composition. For this purpose, the Pearson correlation coefficient can be taken

$$\rho(z_1^s, z_2^s) = \frac{\text{cov}(z_1^s, z_2^s)}{\sqrt{\text{var}(z_1^s)\text{var}(z_2^s)}}, \quad (15)$$

or any other alternative correlation measure. The interpretation in the sense of positive and negative association (known from the correlation coefficient) is possible and statistical inference, such as significance testing, can be performed as usual. It is just important to emphasize that it is not a correlation between the original components, but between coordinates assigned to them. They can be interpreted in terms of dominance of both parts to the average behavior of the rest as described in detail above. Hence, the remaining parts can influence the value of the correlation coefficient as well, which fully corresponds to the relative nature of compositional data. As a consequence, a positive correlation coefficient would mean that dominances of the two amounts over the respective “average representatives” of the other parts increase simultaneously and vice versa for negative correlation; a zero coefficient would mean that dominances of these two amounts are controlled by uncorrelated processes. Of course, part x_1 is contained in z_2^s and, conversely, x_2 in z_1^s . Accordingly, it is interesting to see, what happens if ratios with x_1 uniformly increase by a constant behavior of the other parts (and their ratios). From the construction of both coordinates, while z_1^s increases, z_2^s slightly decreases (x_1 is contained with reduced power in its denominator), resulting in negative correlation. This reminds to the case of correlation between two original parts, but now in a geometrically reasonable manner with orthonormal coordinates. Moreover, it is also a kind of logical result: if the dominance of one part (here x_1) increases, the dominance of another one (x_2) must necessarily decrease. Nevertheless,

the effect for the latter part cannot be the same: x_1 is just one out of $D - 1$ parts to which the dominance of x_2 is related.

Similarly, the correlation for any other pair of parts in \mathbf{x} can be calculated by permuting the parts in Eqs. (13) and (14).

By summarizing all corresponding correlation coefficients in one matrix, the compositional correlation matrix $\mathbf{R}_C(\mathbf{x})$ of dimension $D \times D$ is obtained. It is symmetric with unit diagonal as the standard correlation matrix. Moreover, any scaling and shifting in the compositional sense, which mean by perturbing \mathbf{x} with a non-random composition $\mathbf{b} = (b_1, \dots, b_D)'$ and powering with a real constant a to get a composition $a \odot \mathbf{x} \oplus \mathbf{b} = (x_1^a b_1, \dots, x_D^a b_D)$ (up to an arbitrary scaling constant), yield the same result, $\mathbf{R}_C(a \odot \mathbf{x} \oplus \mathbf{b}) = \mathbf{R}_C(\mathbf{x})$ (Pawlowsky-Glahn et al. 2015). Although by experiments with data sets, also some further interesting properties (such as positive definiteness) were indicated, it is crucial to realize that the elements of $\mathbf{R}_C(\mathbf{x})$ are formed using $D(D - 1)/2$ different coordinate systems that should be taken into account by processing it as a whole (e.g., by computing principal components).

Constructing symmetric balances seems to be the most relevant way how to perform correlation analysis between relative contributions of compositional parts. Nevertheless, the form of the coordinates \mathbf{z} and \mathbf{z}^* inspires to consider also other possibilities that will be briefly mentioned. The first option consists in taking correlation coefficients between the coordinates, Eqs. (7) and (8), respectively

$$\rho(z_1, z_2) = \frac{\text{cov}(z_1, z_2)}{\sqrt{\text{var}(z_1)\text{var}(z_2)}}, \quad \rho(z_1^*, z_2^*) = \frac{\text{cov}(z_1^*, z_2^*)}{\sqrt{\text{var}(z_1^*)\text{var}(z_2^*)}}, \quad (16)$$

and then compute their average as follows

$$\rho_{\text{ave}}(\mathbf{z}, \mathbf{z}^*) = \frac{\rho(z_1, z_2) + \rho(z_1^*, z_2^*)}{2}. \quad (17)$$

Another idea to construct a correlation coefficient with similar interpretation as for the symmetric balances follows the approach from linear discriminant analysis (Johnson and Wichern 2007) based on calculating the so-called pooled covariance matrix from

$$\Sigma_{\mathbf{z}} = \begin{pmatrix} \text{var}(z_1) & \text{cov}(z_1, z_2) \\ \text{cov}(z_2, z_1) & \text{var}(z_2) \end{pmatrix} \text{ and } \Sigma_{\mathbf{z}^*} = \begin{pmatrix} \text{var}(z_1^*) & \text{cov}(z_1^*, z_2^*) \\ \text{cov}(z_2^*, z_1^*) & \text{var}(z_2^*) \end{pmatrix}. \quad (18)$$

The pooled covariance matrix represents here an average of the covariance matrices $\Sigma_{\mathbf{z}}$ and $\Sigma_{\mathbf{z}^*}$

$$\Sigma_{\text{p}}(\mathbf{z}, \mathbf{z}^*) = \frac{\Sigma_{\mathbf{z}} + \Sigma_{\mathbf{z}^*}}{2} = \begin{pmatrix} \Sigma_{\text{p11}} & \Sigma_{\text{p12}} \\ \Sigma_{\text{p21}} & \Sigma_{\text{p22}} \end{pmatrix}, \quad (19)$$

and the elements are taken to get the resulting correlation coefficient:

$$\rho_{\text{pool}}(\mathbf{z}, \mathbf{z}^*) = \frac{\Sigma_{\text{p12}}}{\sqrt{\Sigma_{\text{p11}} \Sigma_{\text{p22}}}}. \quad (20)$$

The next section will be devoted to thorough simulation studies to investigate, whether one would benefit from employing these alternative approaches in addition to the main proposal formed by correlation analysis of symmetric balances.

5 Simulation Studies

The main aim of the following simulation studies is to investigate the properties of the different correlation coefficients as introduced in the previous section and to compare also with some other approaches that are used in the literature. In this section, randomly generated data and data obtained from the moss layer in the Kola Project (Reimann et al. 1998) are used. These data are available in the R package `mvoutlier` as data set `moss` (R Development Core Team 2015), and they contain concentrations of 31 chemical elements in more than 600 moss samples.

5.1 Simulation 1: Uniform Distribution Inside a Sphere

Data sets are generated randomly from a uniform distribution inside a unit sphere of dimension $d = D - 1$. These data are already expressed in coordinates, and by construction, all pairwise correlations are 0. The data are transformed to the original space using the inverse mapping to isometric logratio coordinates (Egozcue et al. 2003), with constant row sum 1. Without loss of generality, the interest here is in the association between the first two compositional parts. For each considered correlation measure, bootstrap confidence intervals (CIs) are constructed, using $B = 1000$ bootstrap samples of each particular data set. For each bootstrap sample, the respective correlation measure is applied. The 95% bootstrap CI is then defined as the interval given by the lower bound as the quantile 0.025 and the upper bound as the quantile 0.975 of the B correlations. For a particular dimension d and sample size n , in total $N = 1000$, samples were randomly drawn, and the averages of the lower and upper interval bounds are computed. The length of the resulting interval is reported in Fig. 3, where simulated data with different sample sizes ($n \in \{10, 50, 100, 500\}$) and dimensions ($d \in \{4(5)34\}$) have been used. Figure 4 presents the resulting coverages of the CIs. The coverage is computed as the number of intervals containing the true underlying correlation 0, divided by N . The coverage is close to 0.95 in most cases, except for the correlations based on clr coefficients and log-transformed data, and here in particular for smaller numbers of parts. The reason for the considerably smaller coverage is the negative bias for these correlations. In these cases, also the average lengths of the CIs are smaller, but still the CIs are useless. The CIs for the correlation based on pooled covariances are shorter in low dimension compared to correlations for symmetric balances and for balances describing the two parts of interest, with the drawback that the pooled covariances are not directly resulting from an orthonormal basis. Thus, from this study one would conclude that proposals Eqs. (15) and (16), and thus also Eq. (17), are performing equally and well, but symmetric balances are more adequate from an interpretation point of view.

In addition, proportionality coefficients were computed for these simulated data, but confidence intervals were not considered, since they would be meaningless.

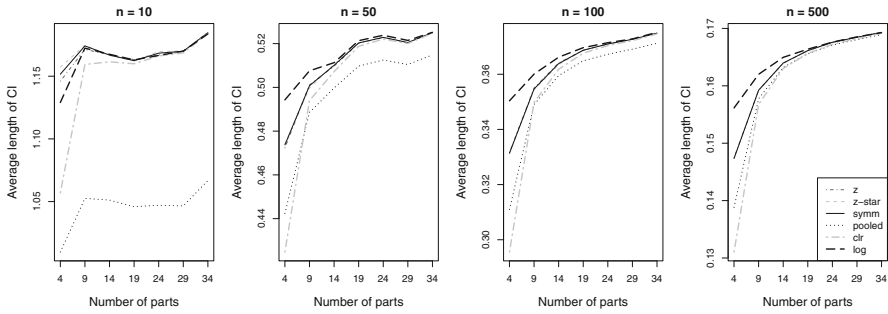


Fig. 3 Average lengths of bootstrap confidence intervals for randomly generated uniformly distributed data inside a sphere of dimension $d = D - 1$. The used correlations measures: $\rho(z_1, z_2)$ (“z”), $\rho(z_1^*, z_2^*)$ (“z-star”), $\rho(z_1^s, z_2^s)$ (“symm”), $\rho_{\text{pool}}(\mathbf{z}, \mathbf{z}^*)$ (“pooled”), $\rho(y_1, y_2)$ (“clr”) and $\rho(\log(x_1), \log(x_2))$ (“log”)

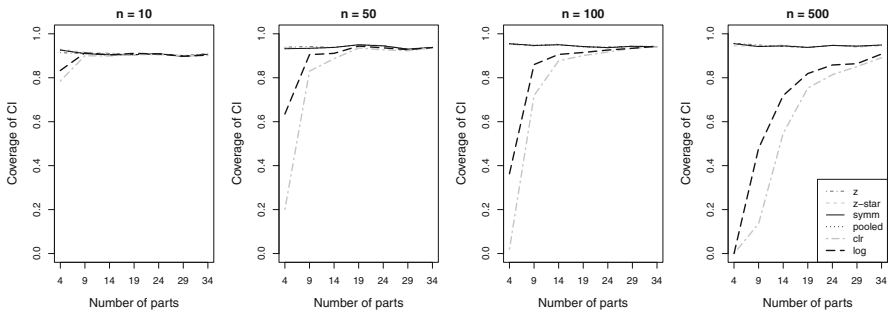


Fig. 4 Coverage of the bootstrap confidence intervals for randomly generated uniform data inside a sphere of dimension $d = D - 1$. The used correlations measures: $\rho(z_1, z_2)$ (“z”), $\rho(z_1^*, z_2^*)$ (“z-star”), $\rho(z_1^s, z_2^s)$ (“symm”), $\rho_{\text{pool}}(\mathbf{z}, \mathbf{z}^*)$ (“pooled”), $\rho(y_1, y_2)$ (“clr”) and $\rho(\log(x_1), \log(x_2))$ (“log”)

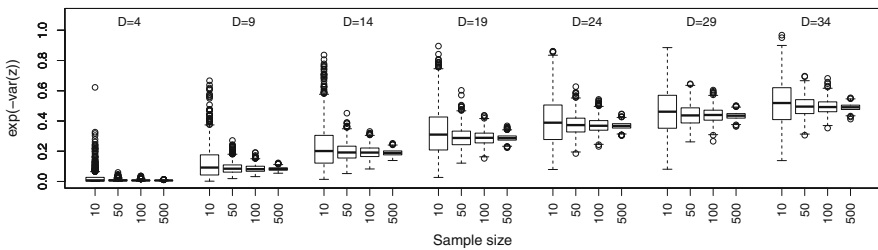


Fig. 5 Proportionality coefficients for 1000 simulated data with different sample sizes n and different numbers of independent parts D , summarized as boxplots

Figure 5 shows all 1000 results as boxplots for all different combinations of sample size and dimension. As it can be expected, sample size leads to a high variability of the proportionality coefficients. An interesting finding is, however, that the proportionality coefficients are close to 0 for a small number of parts, but they get quite high if the number of parts increases. For example, for $D = 34$ and $n = 500$, the median value for this coefficient is higher than 0.4. This raises doubts whether the proportionality coefficient as such is useful for judging the dependency between compositional

parts, even though it clearly has a different construction and interpretation than the previous (correlation) measures.

5.2 Simulation 2: Dependence on the Number of Parts

This simulation setting is based on the Kola moss data, and compares the different approaches for correlation analysis for a varying number of parts involved in the computation of the correlation coefficients. k parts ($4 \leq k \leq 30$) are randomly selected, and the correlation between the first two parts is computed (in the sense of the above proposals); the parts are always the same for the different correlation measures. For each fixed k , the random selection is done 10,000 times, resulting in 10,000 correlation values for each method. When comparing two methods, the outcomes of all results are compared for fixed k in terms of the Pearson correlation. A value close to 1 would indicate approximately the same outcome of both methods. The left panels in Fig. 6 show these pairwise comparisons of the approach based on symmetric balances with the other correlation measures, where the considered number of parts is on the horizontal axes, and the resulting correlations between the point clouds of the 10,000 outcomes on the vertical axes. The right panels show again pairwise comparisons of correlation measures, but this time, the maximum difference of the 10,000 results is computed.

It can be seen that the relation between correlations of symmetric balances and average correlations or correlations based on the pooled covariance matrix, respectively, are very close regardless of the number of parts, with some few exceptions for lower numbers of parts in the compositional data set (Fig. 6a, b). This is plausible given the way how they are constructed. In addition, the proposals given in Eq. (16) show a very similar behavior (plots not presented here).

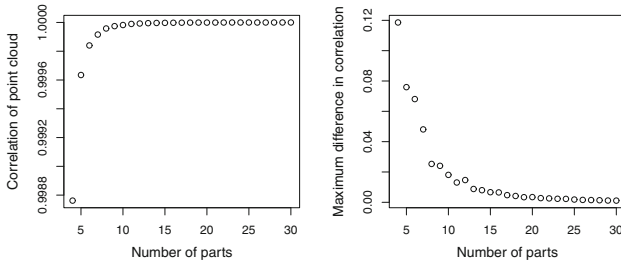
Finally, the symmetric balances approach is also compared to correlations derived from the respective clr variables (Fig. 6c) and correlations from log-transformed variables (Fig. 6d). It can be seen that with increasing number of parts, the resulting correlation structure from clr variables gets more and more similar to the correlation structure from symmetric balances. This is because also the negative bias in case of clr-based correlations gets smaller with increasing dimension. The difference in the correlation structure from log-transformed data is large in all presented cases, resulting from working in a non-appropriate geometry.

The advantage of symmetric balances is that they provide orthonormal coordinates, where reasonable statistical inference concerning their association can be performed. Even though clr coefficients can lead to similar correlations in certain cases, one does not obtain orthonormal coordinates, with possible consequences on statistical inference.

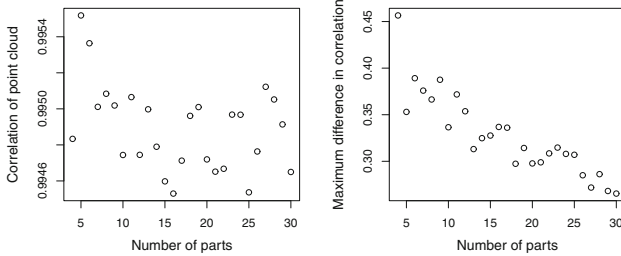
6 Example

As in the previous section, the Kola moss data set is used to compare different association measures. The resulting pairwise correlation coefficients are presented by the so-called heatmaps, Fig. 7, where the resulting correlations are simply color coded.

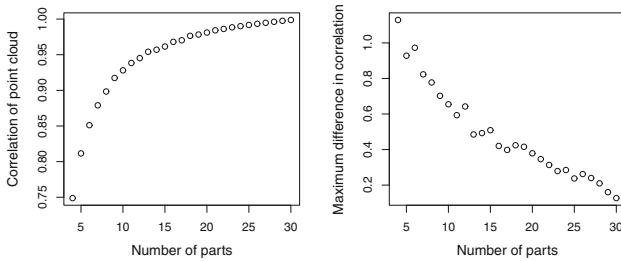
(a) Comparison of correlations between symmetric balances and average correlation coefficient



(b) Comparison of correlations between symmetric balances and pooled covariance matrix approach



(c) Comparison of correlations between symmetric balances and clr coefficients



(d) Comparison of correlations between symmetric balances and log-transformed data

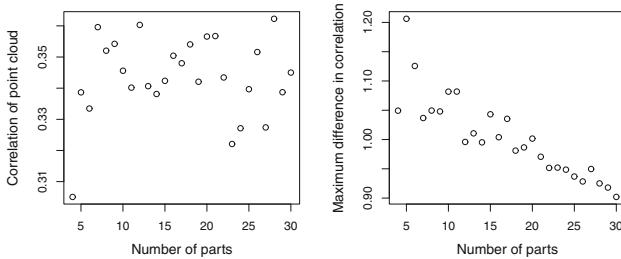


Fig. 6 Pairwise comparisons of different correlation measures based on 10,000 random selections of data sets with k parts ($4 \leq k \leq 30$); *left* Pearson correlations of the resulting point clouds; *right* maximum difference between all results

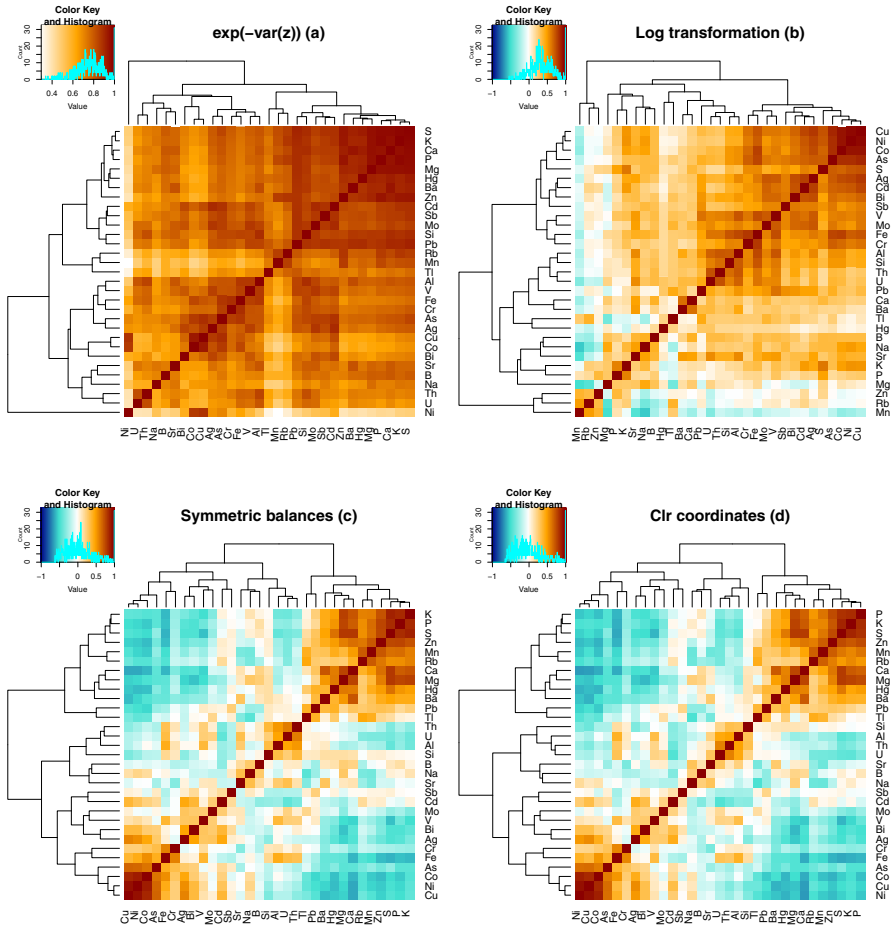


Fig. 7 Heatmaps of correlations for the moss data set based on the variation matrix coefficients (a), log-transformed data (b), symmetric balances (c), and clr coefficients (d)

In addition, the variables are grouped to identify patterns in the matrix of pairwise correlations. Figure 7 compares the heatmaps for associations based on the variation matrix coefficients (upper left), and further correlations for log-transformed data (upper right), for symmetric balances (lower left), and for clr coefficients (lower right). Due to the individual grouping in each heatmap, the order of the rows and columns within the plots changes and makes a direct comparison difficult. However, in this representation, one can clearly see the difference in patterns. The variation matrix approach leads to a very different structure due to the non-negative association measures. In addition, the heatmap for log-transformed data, still very commonly applied in geochemistry, reveals a different structure compared to that for symmetric balances. In particular, only few negative correlations, but mainly positive ones are obtained. Finally, the heatmaps for symmetric balances and for clr coefficients are very similar. This is to be expected from the simulation results, Fig. 6c, since for larger numbers of

parts, the two approaches for computing correlations get very similar. A much larger difference can be expected when investigating a subcomposition.

The heatmap for the correlations based on symmetric balances shows in the upper right plant nutrients (or, more precisely, their dominances with respect to other parts of the composition), with the major plant nutrients K, P, and S, and minor plant nutrients Zn, Mn, Rb (probably uptaken with K), Ca, and Mg. All these elements belong to the main plant nutrients. Interestingly, Hg, Ba, Pb, and Tl are in the same cluster; except of Ba, all these elements are toxic, and one can conclude that the plants play an important role in their geochemical behavior. The elements Ci, Co, and Ni, in the lower left, are the three main elements for emission. Further elements As, Fe, Cr, Ag, Bi, and V are also emitted and thus related to the the three elements. Along the diagonal, there is a cluster consisting of Th, U, Al, and Si, which may indicate dust, and the group B, Na, and Sr is related to sea spray. In the lower left corner, a block of negative correlations is identified: dominance of plant nutrients in the composition occurs with a subordination of emitting metals, and vice versa. In addition, the dust elements are negatively related to the plant nutrients.

When investigating the heatmap for the log-transformed data, some similarities to the symmetric balances outcome can be discovered: Cu, Ni, and Co are also highly correlated, and also As joins this group. However, several other elements would still be interpreted as highly correlated with this group, including Th, U, Al, and Si. It would thus not be possible to identify these elements as an own group, as in the heatmap for symmetric balances, related to a different process (dust). In addition, the plant nutrients are not as clearly separated as in the symmetric balance approach. Overall, it can be concluded that correlations based on symmetric balances are much more sensitive and useful to reveal underlying processes.

In this example, it is also obvious that the heatmap for associations based on the variation matrix coefficients (upper left) is not only different because of the lack of negative correlations, but also much less useful for identifying groups of elements. A close inspection reveals similar structures as identified in the heatmap for symmetric balances, but they are much more difficult to find.

A direct comparison of the heatmaps above is difficult, since the parts are reordered in each individual plot. Figure 8 endeavors a better comparison of the outcomes by sorting the parts according to the heatmap for the symmetric balances. Moreover, the color scheme has been modified: the median of the resulting correlations is colored in white, red tones are used for values higher than the median, and blue tones for values lower than the median. This might still somehow be inappropriate for the variation matrix coefficients, where “negative” associations cannot be expressed appropriately. However, it is interesting to see that there are still major differences in the structure of the heatmaps for symmetric balances compared to correlations from log-transformed data. As it was already observed in the simulation experiments, the distribution of the variation matrix coefficients is shifted upwards, and it is hard to judge if a value of, say, 0.5 already indicates a strong relation or not.

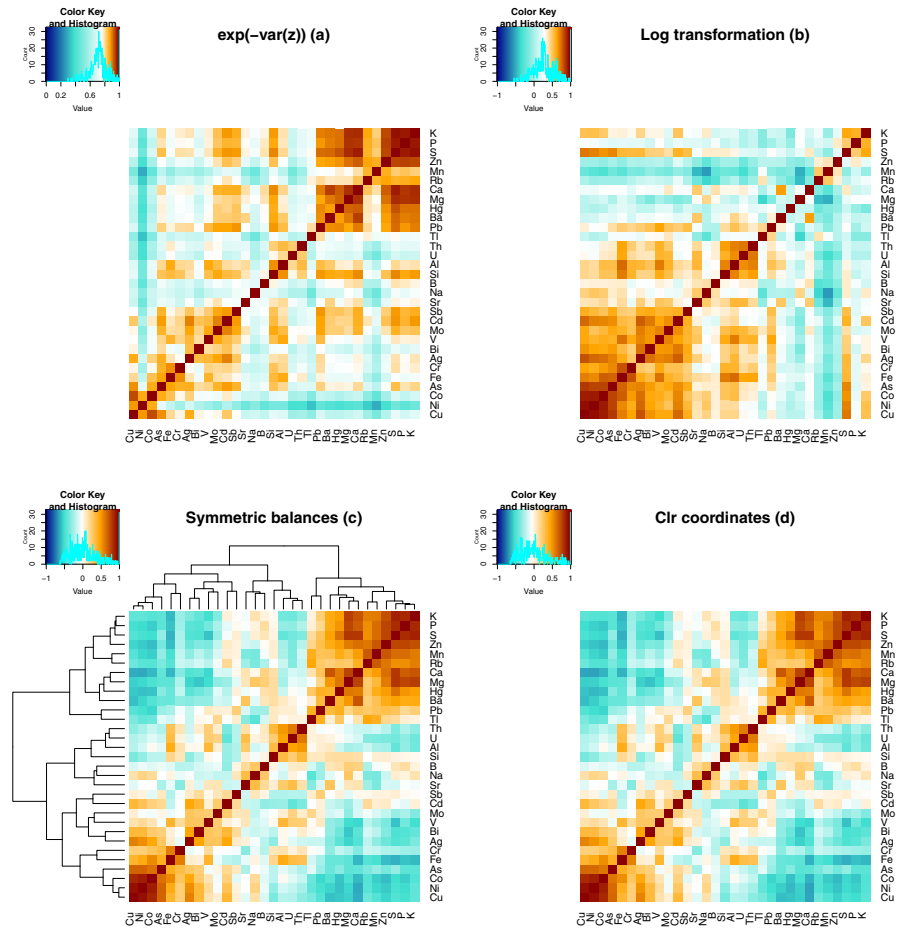


Fig. 8 Heatmaps of correlations for the moss data set based on the variation matrix coefficients (a), log-transformed data (b), symmetric balances (c), and clr coefficients (d). The parts are ordered according to the results for the symmetric balances. The *color scheme* starts with white at the median

7 Discussion

Correlation analysis of the original compositional parts fails to provide interpretable results if a fixed constant sum constraint is employed. This is due to the relative nature of compositions represented particularly by scale invariance, and it leads to a negative bias of the correlation structure. The only safe way to perform correlation analysis of compositional data is to express them in orthonormal logratio coordinates. Although sequential binary partitioning and the resulting balances can be very useful when prior knowledge about geochemical processes in the data is available, automated, and interpretable orthonormal coordinates that capture relative information, about single compositional parts can help to reveal hidden geochemical patterns when such information is not available.

For the purpose of interpretable correlation analysis in orthonormal logratio coordinates, the so-called symmetric balances were introduced using a special choice of balance coordinates. They allow to treat two compositional parts in a symmetric way in one coordinate system and to compute the correlation coefficient. Although the symmetric balances cannot be simply identified with the original compositional parts, because they capture just relative contributions of the parts within a given composition, it seems to be the first successful attempt to have correlation analysis of compositional data interpretable in terms of dominance of a pair of compositional parts. Particularly, the possibility of analyzing negative and positive associations as often required in practice (and not available using the variation matrix approach) can help to eliminate inappropriate data processing, for instance, using the popular (but scale dependent) log-transformation. Moreover, one should be aware that also other parts are naturally involved into the correlation between two given components by constructing symmetric balances. Nevertheless, it follows closely the definition of compositional data that none of the parts can be analyzed without considering relations (ratios) to the other parts. This, however, has the consequence that measurement errors in some parts may affect the resulting correlation coefficients of symmetric balances. A possible way out seems to be appropriate weighting of the parts according to their relevance, as proposed recently in Egozcue and Pawłowsky-Glahn (2015) and Filzmoser and Hron (2015). This will be further investigated in subsequent work.

Correlation coefficients can be seen as summarizing the information of the variable relations shown in scatter plots. The concept of symmetric balances allows to have an appropriate graphical representation of two compositional parts in terms of orthonormal coordinates. This can serve as a new way of investigating pairwise relations. An overview of all pairwise relations can be provided by a heatmap. In an application to the Kola moss data set, this allowed to clearly reveal processes underlying the data.

Finally, here only the Pearson correlation was used to measure association. Clearly, one can also employ alternative correlation estimators, such as the Spearman correlation for identifying non-linear relations or robust correlation estimators for downweighting the influence of outlying observations.

Acknowledgements Open access funding provided by TU Wien (TUW). The paper was supported by the Grant COST Action CRoNoS IC1408 and by the K-project DEXHELPP through COMET—Competence Centers for Excellent Technologies, supported by BMVIT, BMWFI, and the province Vienna. The COMET program is administrated by FFG. The authors are grateful to Dr. Clemens Reimann from the Geological Survey of Norway (NGU) for fruitful discussions and to the Associate Editor and two anonymous referees for valuable comments which helped to improve the quality of the paper considerably.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Aitchison J (1986) The statistical analysis of compositional data. Chapman and Hall, London
- Buccianti A, Pawłowsky-Glahn V (2005) New perspectives on water chemistry and compositional data analysis. *Math Geol* 37(7):703–727

- Chayes F (1960) On correlation between variables of constant sum. *J Geophys Res* 65(12):4185–4193
- Eaton M (1983) *Multivariate statistics. A vector space approach*. Wiley, New York
- Egozcue J (2009) Reply to “On the Harker variation diagrams; . . .” by J.A. Cortés. *Math Geosci* 41(7):829–834
- Egozcue J, Pawlowsky-Glahn V (2015) Proceedings of the 6th international workshop on compositional data analysis. In: Thió-Henestrosa S, Martín Fernández J (eds) *Changing the reference measure in the simplex and its weighting effects*, University of Girona, Girona, pp 1–10
- Egozcue JJ, Pawlowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. *Math Geol* 37:795–828
- Egozcue JJ, Pawlowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* 35:279–300
- Egozcue JJ, Lovell D, Pawlowsky-Glahn V (2013) Testing compositional association. In: Hron K, Filzmoser P, Templ M (eds) *Proceedings of the 5th International Workshop on Compositional Data Analysis*. Vorau, Austria
- Filzmoser P, Hron K (2015) Robust coordinates for compositional data using weighted balances. In: Nordhausen K, Taskinen S (eds) *Modern nonparametric. Robust and multivariate Methods*. Springer, Heidelberg
- Filzmoser P, Hron K, Reimann C (2009) Univariate statistical analysis of environmental (compositional) data: problems and possibilities. *Sci Total Environ* 407:6100–6108
- Filzmoser P, Hron K, Reimann C (2010) The bivariate statistical analysis of environmental (compositional) data. *Sci Total Environ* 408(19):4230–4238
- Fišerová E, Hron K (2011) On interpretation of orthonormal coordinates for compositional data. *Math Geosci* 43:455–468
- Johnson RA, Wichern DW (2007) *Applied multivariate statistical analysis*, 6th edn. Prentice Hall, Englewood
- Korhoňová M, Hron K, Klimčíková D, Müller L, Bednář P, Barták P (2009) Coffee aroma—statistical analysis of compositional data. *Talanta* 80(82):710–715
- McKinley J, Hron K, Grunsky E, Reimann C, de Caritat P, Filzmoser P, van den Boogaart K, Tolosana-Delgado R (2016) The single component geochemical map: fact or fiction. *J Geochem Explor* 162:16–28
- Pawlowsky-Glahn V, Buccianti A (2011) *Compositional data analysis: theory and applications*. Wiley, Chichester
- Pawlowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. *Stoch Environ Res Risk Assess (SERRA)* 15(5):384–398
- Pawlowsky-Glahn V, Egozcue J, Tolosana-Delgado R (2015) *Modeling and analysis of compositional data*. Wiley, Chichester
- Pearson K (1897) Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc R Soc Lond LX*:489–502
- R Development Core Team (2015) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna
- Reimann C, Åyräs M, VC, et al (1998) *Environmental geochemical Atlas of the Central Barents Region*. NGU-GTK-CKE Special publication, Geological Survey of Norway, Trondheim, Norway
- Reimann C, Filzmoser P, Fabian K, Hron K, Birke M, Demetriades A, Dinelli E, Ladenberger A, The GEMAS Project Team (2012) The concept of compositional data analysis in practice. Total major element concentrations in agricultural and grazing land soils of Europe. *Sci Total Environ* 426:196–210