


Advances in Principal Balances for Compositional Data

J. A. Martín-Fernández¹  · V. Pawlowsky-Glahn¹ ·
J. J. Egozcue² · R. Tolosona-Delgado³

Received: 23 December 2016 / Accepted: 21 October 2017 / Published online: 22 November 2017
© International Association for Mathematical Geosciences 2017

Abstract Compositional data analysis requires selecting an orthonormal basis with which to work on coordinates. In most cases this selection is based on a data driven criterion. Principal component analysis provides bases that are, in general, functions of all the original parts, each with a different weight hindering their interpretation. For interpretative purposes, it would be better to have each basis component as a ratio or balance of the geometric means of two groups of parts, leaving irrelevant parts with a zero weight. This is the role of principal balances, defined as a sequence of orthonormal balances which successively maximize the explained variance in a data set. The new algorithm to compute principal balances requires an exhaustive search along all the possible sets of orthonormal balances. To reduce computational time, the sets of possible partitions for up to 15 parts are stored. Two other suboptimal, but feasible, algorithms are also introduced: (i) a new search for balances following a constrained principal component approach and (ii) the hierarchical cluster analysis

✉ J. A. Martín-Fernández
josepantoni.martin@udg.edu

V. Pawlowsky-Glahn
vera.pawlowsky@udg.edu

J. J. Egozcue
juan.jose.egozcue@upc.edu

R. Tolosona-Delgado
r.tolosana@hzdr.de

¹ Dept. Informàtica, Matemàtica Aplicada, i Estadística, Universitat de Girona, Campus Montilivi, Edifici P-4, 17003 Girona, Spain

² Dept. d'Enginyeria Civil i Ambiental, U. Politècnica de Catalunya, Barcelona, Spain

³ Dept. Modelling and Evaluation, Helmholtz-Zentrum Dresden-Rossendorf, Helmholtz-Institut Freiberg for Resource Technology, Freiberg, Germany

of variables. The latter is a new approach based on the relation between the variation matrix and the Aitchison distance. The properties and performance of these three algorithms are illustrated using a typical data set of geochemical compositions and a simulation exercise.

Keywords Aitchison norm · Cluster analysis · Compositions · Isometric logratio coordinates · Principal component analysis · Simplex

1 Introduction

Compositional data (CoDa) convey relative information expressed in the ratios between parts. Typical examples of compositions appear in geochemistry and in environmetrics in the geosciences, but they are also found in other fields such as chemometrics, budget expenses, time-use survey data, genomics, and all omics in general. For convenience, compositions are commonly expressed in terms of proportions, percentages or parts per million (ppm) (Aitchison 1986). When analysts decide to analyze a data set \mathbf{X} ($n \times D$) using compositional methods, they are assuming that the information contained in any observation \mathbf{x} (a row of \mathbf{X}) is the same as in $k \cdot \mathbf{x}$, for any real scalar $k > 0$. Based on this property, known as scale invariance (Aitchison 1986), a composition can be defined as an equivalence class (Barceló-Vidal and Martín-Fernández 2016). According to this definition, the general expression of a scale-invariant logratio is a log-contrast (Aitchison 1986)

$$\sum_{i=1}^D a_i \ln x_i = \ln \left(\prod_{i=1}^D x_i^{a_i} \right), \quad \sum_{i=1}^D a_i = 0. \quad (1)$$

A log-contrast is, in essence, a logratio of parts because for $a_i > 0$ the corresponding part x_i appears in the numerator, but if $a_i < 0$ it appears in the denominator, while for those parts that do not contribute to the logratio $a_i = 0$ holds. Log-contrasts [Eq. (1)] have the same role as linear combinations of real variables in classical statistics. Accordingly, principal component analysis (PCA) applied to CoDa and log-ratio bases should both be based on log-contrasts. Note that ratios and logratios cannot be computed when one of the parts is zero or missing. How to deal with this difficulty, also known as the zero problem, has been described in numerous articles. For the interested reader, a general description of this topic can be found in Palarea-Albaladejo and Martín-Fernández (2015).

Using a log-contrast one can define new variables [e.g., a principal component (PC)] where the information collected in the original variables is combined. For the centered log-ratio (clr) variables (Aitchison 1986)

$$\text{clr}_k(\mathbf{x}) = \ln \frac{x_k}{(\prod x_i)^{1/D}} = \ln x_k - \overline{\ln \mathbf{x}}, \quad k = 1, \dots, D, \quad (2)$$

where $\overline{\ln \mathbf{x}}$ denotes the average of the logarithms of components in \mathbf{x} , the log-contrast expression [Eq. (1)] verifies that $a_{ki} = -1/D$ for $i \neq k$ and $a_{kk} = 1 - 1/D$. It holds

that $\sum_{k=1}^D \text{clr}_k(\mathbf{x}) = 0$, indicating that the dimension of the clr space is $D - 1$. To calculate PCs and balances, it holds that any contrast for clr coefficients is equal to a log-contrast [Eq. (1)]. That is, it holds that $\sum_{k=1}^D a_k \text{clr}_k(\mathbf{x}) = \sum_{k=1}^D a_k \ln x_k$. The inner product, distance and norm can be defined via the clr variables (Pawlowsky-Glahn and Egozcue 2001). These metric elements are used to construct orthonormal log-ratio bases (Egozcue et al. 2003).

Nowadays, there is general agreement that a statistical analysis of CoDa should be performed on coordinates with respect to a log-ratio basis (Mateu-Figueras et al. 2011). There is an infinite number of orthonormal bases in the simplex, so to perform an analysis of a data set the analyst must make a choice. Sometimes, the basis can be chosen blindly or by using a data driven criterion, such as PCA. Other times, the selection of a basis is based on expert prior knowledge, possibly using a sequential binary partition (SBP) technique (Egozcue and Pawlowsky-Glahn 2005) to construct an interpretable basis. The coordinates, called balances, may help the interpretability of the results. Here an intermediate possibility is discussed. The main goal is to identify a complete orthonormal basis of the simplex such that the coordinates are balances approaching the properties of PCA for CoDa (Aitchison 1983). The resulting procedures provide tools that improve interpretability and can also be used for an intuitive dimension reduction. These goals have been discussed within the framework of PCA of real multivariate data (Hotelling 1933; Jolliffe 2002).

PCA is an appealing technique when the PCs can be readily interpreted. However, interpretation is not always simple, because the loadings are typically nonzero and PCs are linear combinations of all the original variables. To solve this interpretation problem, different approaches have been proposed, but all of them share the practical rule of thumb: the larger the number of zero-loadings is, the easier the interpretation. Some of these approaches are based on rotating the PCs, as is done in factor analysis (Jolliffe 2002, Chapter 11), while others adapt variable selection techniques for PCA (Jolliffe et al. 2003). Chipman and Gu (2005) introduce three types of interpretable components as a result of an approximation of the PCs, while Enki et al. (2013) apply cluster analysis to the original variables as the first step in constructing interpretable PCs. The drawback is that none of these approaches retains the three properties of PCA: uncorrelation, orthogonality and optimal variance explanation (Jolliffe 2002).

This article aims to provide new algorithms for constructing principal balances (PBs), that is, a complete basis of balances based on data information showing properties similar to PCs. In Sect. 2, some basic CoDa concepts are introduced to provide a new definition for PB. Section 3 provides a new recursive algorithm to construct optimal PBs. Two different suboptimal, but faster, approaches are introduced in Sect. 4: (i) a new search for balances following a constraint PC approach and (ii) the hierarchical cluster analysis of variables. The latter consists of a new approach based on the relation between the log-ratio variances and the Aitchison distance. A geochemical example is presented in Sect. 5, where all the techniques introduced for the PBs are applied. A simple simulation exercise is also provided. Finally, in Sect. 6, some concluding remarks are presented. The new techniques discussed in this article were programmed using the open source R statistical programming language and software (R development core team 2015). The computer routines to implement the methods can be obtained from www.compositionaldata.com.

2 Balances and Principal Balances

The expression of the coordinates, known as isometric log-ratio coordinates (ilr), depends on the basis selected. When the ilr system of coordinates is defined through an SPB (Egozcue and Pawlowsky-Glahn 2005, 2006), the coordinates are proportional to a logratio of geometric means of the parts in two disjoint groups. In the first step of an SBP, the complete composition $\mathbf{x} = (x_1, \dots, x_D)$ is split into two groups of parts: one for the numerator coded as +1, and the other for the denominator coded as −1. In the steps that follow, each group is split into two groups coded with +1 and −1, respectively, while those parts not in the original group are coded as 0. That is, in step k , the r_k parts $(x_{n_1}, \dots, x_{n_{r_k}})$ in the first group are coded as +1 and placed in the numerator, and the s_k parts $(x_{d_1}, \dots, x_{d_{s_k}})$ in the second group will appear in the denominator and be coded as −1. As a result, the coordinate $\text{ilr}_k(\mathbf{x})$ or k -th balance, is

$$\text{ilr}_k(\mathbf{x}) = \sqrt{\frac{r_k \cdot s_k}{r_k + s_k}} \ln \frac{(x_{n_1} \cdots x_{n_{r_k}})^{1/r_k}}{(x_{d_1} \cdots x_{d_{s_k}})^{1/s_k}}, \quad k = 1, \dots, D-1, \quad (3)$$

where the square root term is a factor for normalizing the coordinate.

The log-contrast coefficients [Eq. (1)] for the k -th balance in a given SBP are

$$a_{ki} = \begin{cases} \frac{1}{r_k} \sqrt{\frac{r_k \cdot s_k}{r_k + s_k}} & \text{if } x_i \text{ appears in the numerator and} \\ -\frac{1}{s_k} \sqrt{\frac{r_k \cdot s_k}{r_k + s_k}} & \text{if } x_i \text{ appears in the denominator,} \\ 0 & \text{otherwise.} \end{cases} \quad (4)$$

The coefficients a_{ki} , $i = 1, 2, \dots, D$, are equal to the clr coefficients of the k -th balancing element of the selected basis. The metric elements can also be expressed in terms of ilr coordinates. For instance, the Aitchison distance satisfies $d_a(\mathbf{x}_1, \mathbf{x}_2) = d_e(\text{ilr}(\mathbf{x}_1), \text{ilr}(\mathbf{x}_2))$ regardless of the basis selected (Palarea-Albaladejo et al. 2012).

While the geometric elements of the Aitchison geometry (Pawlowsky-Glahn and Egozcue 2001) do not depend on the particular ilr coordinates used to represent compositions, an adequate choice of the basis can favor an easier interpretation of the results from a compositional analysis. A CoDa-dendrogram (Pawlowsky-Glahn and Egozcue 2011) is a descriptive tool for visualizing some univariate statistics of the ilr coordinates derived from an SBP. Table 1 shows an example of the SBP for a typical geochemical data set. The data set used contains 2097 samples of the chemical elements Ba, Ca, Fe, K, Mg, and Mn. The columns on the right show the corresponding mean, variance, and the percentage of the total variance retained by each balance. The total variance is equal to the sum of the values in the column of variances (Pawlowsky-Glahn and Egozcue 2011).

The SBP is represented by dendrogram-type links between parts, as shown in Fig. 1. The leaves of the dendrogram, represented by dotted lines, correspond to the groups of parts formed by a unique element. The location of the mean of an ilr coordinate is determined by the intersection of the horizontal segment with the vertical segment

Table 1 Example of SBP : a typical geochemical data set for the composition (Ba, Ca, Fe, K, Mg, Mn) Mean and variance correspond to the sample mean and variance of the balance. The sum of the variances is 5.62 which corresponds to the sample total variance

$\text{ilr}_k(\mathbf{x})$	Ba	Ca	Fe	K	Mg	Mn	Mean	Variance	% Variance
$\text{ilr}_1(\mathbf{x})$	+1	−1	−1	−1	−1	−1	−3.93	0.53	9.37
$\text{ilr}_2(\mathbf{x})$	0	−1	−1	+1	−1	−1	0.00	0.93	16.58
$\text{ilr}_3(\mathbf{x})$	0	−1	+1	0	−1	−1	−1.46	2.69	47.89
$\text{ilr}_4(\mathbf{x})$	0	−1	0	0	−1	+1	−3.45	1.29	23.01
$\text{ilr}_5(\mathbf{x})$	0	+1	0	0	−1	0	0.64	0.18	3.16

(variance). When these intersections are not in the middle, this indicates a major contribution from one of the groups of parts. Importantly, the variability of each balance is associated with the length of the vertical bars. When the ilr coordinate has a large variance, its vertical bar is long as, for example, in $\text{ilr}_3(\mathbf{x})$ and $\text{ilr}_4(\mathbf{x})$ that accumulate nearly 71% of the total variance. The vertical axis on the left shows the level of accumulated variance. At the top of the dendrogram this is equal to the total variance 5.62.

In some situations, very few balances can explain a large proportion of the total variability. The fact that a log-ratio basis is formed by balances defined through an SBP, and not automatically following some criterion like in PCA, motivated the definition of PBs introduced in Pawłowsky-Glahn et al. (2011). The definition of constrained PCA (Jolliffe 2002; Chipman and Gu 2005) allows for the original definition of the PBs to be redefined in the following terms

Definition 1 (Principal balances) Let $\mathbf{X} = (X_1, X_2, \dots, X_D)$ be a D -part random composition and X_i , $i = 1, 2, \dots, D$, the random parts. Principal balances (PBs) are log-linear functions $\sum_{i=1}^D a_{ki} \ln X_i$, $k = 1, 2, \dots, D - 1$, such that the vectors $\mathbf{a}_k = (a_{k1}, \dots, a_{kD})$ are constant and they maximize the variances

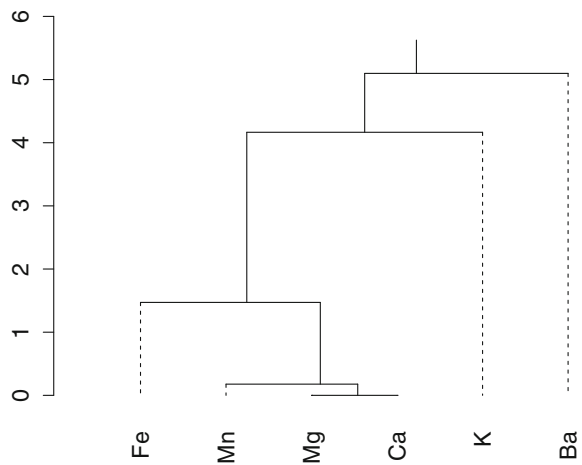
$$\text{var} \left[\sum_{i=1}^D a_{ki} \ln X_i \right],$$

subject to

- (balance condition) for $k = 1, 2, \dots, D - 1$, the coefficients a_{ki} take one of the three values $(-c_1, 0, c_2)$, for some strictly positive c_1 and c_2 ;
- (zero sum and unit norm conditions) for $k = 1, 2, \dots, D - 1$, \mathbf{a}_k satisfies $\sum_{i=1}^D a_{ki} = 0$ and $\sum_{i=1}^D a_{ki}^2 = 1$; and
- (orthogonality condition) for $k = 2, 3, \dots, D - 1$, \mathbf{a}_k is orthogonal to the previous $\mathbf{a}_{k-1}, \mathbf{a}_{k-2}, \dots, \mathbf{a}_1$, that is

$$\sum_{i=1}^D a_{ki} a_{(k-\ell)i} = 0, \quad \ell = 1, 2, \dots, k - 1.$$

Fig. 1 CoDa-dendrogram using the SBP from Table 1. Vertical axis shows the level of accumulated variance. Dotted lines represent groups of parts formed by a unique element



Importantly, the restrictions $\sum_{i=1}^D a_{ki}^2 = 1$ and $\sum_{i=1}^D a_{ki} = 0$ force the constants $-c_1$ and c_2 to take the values of the balancing elements [Eq. (4)]. That is, the log-linear functions are balances. With this definition and given an n -sample of a D -part random composition, the first PB is the balance with maximum sample variance and the k -th PB has the maximum variance conditional to its balancing element being orthogonal to the previous $(k - 1)$ balancing elements.

Remarkably, the possible solutions to the optimization problem stated in Definition 1 are not unique. At a first glance, if \mathbf{a}_k is a solution, then $-\mathbf{a}_k$ is also a solution and the variances of both log-contrasts are equal for both possibilities. It is simply a change in the orientation of the k -th element of the basis. Moreover, as is the case for real space, other uncommon cases might arise when the covariance matrix has eigenvalues with multiplicity greater than one (Jolliffe 2002, p. 27).

Because an SBP has a tree structure, the associated balances are ordered by the particular sequence of the partition. PBs, however, are ordered by the balance variances and not by the sequence of the partitions used. For instance, Table 1 and Fig. 1 correspond to the PBs of the aforementioned CoDa set. The first PB is the $\text{ilr}_3(\mathbf{x})$ that was attained in the third step of the partition. By analogy to PCA, the construction of PBs can be viewed as an orthogonal linear transformation in the simplex, restricted to transformations within the set of possible bases made of balances. It represents the data in a new coordinate system, such that the largest variance of the data comes to lie on the first coordinate (called the first PB), the second greatest variance on the second coordinate, and so on as stated in Definition 1.

3 Exhaustive Search for Principal Balances: A New Optimal Algorithm

Computing PBs requires an exhaustive search along all the possible SBPs. Given D parts, the number of different SBPs is equal to $D!(D-1)!/2^{D-1}$ (Podani 2000). This number increases dramatically with D . In fact, taking $D = 10$ parts, the number of possible SBPs is 2.57×10^9 , whereas for $D = 80$ it is 1.06×10^{212} . In this work, and to keep computational time reasonable, the feasible combinations of codes $\{-1, 0, +1\}$ for the different number of parts, from $D = 2$ to $D = 15$, were generated. However, the same procedure can be applied to generate the codes for higher dimensions. These sets of codes, representing a partition, are generated once and saved in different files so that they can be used when needed. It took, for example, 3.32 min to generate the set of all partitions for $D = 15$ with a 2.4 GHz Intel Core i5 in a Mac-OS X (version 10.9.5) and the resulting file requires 1.45 GB of disk space.

In order to present the exhaustive search algorithm, the concepts of parent, child, and top partitions are introduced. Given one partition of D parts called the current partition, a parent partition is a partition which can be followed by the current partition. For example, let $(0, -1, +1, 0, -1, -1)$ be the current partition with $D = 6$. Up to a change of signs, the possible parent partitions are $(+1, -1, -1, 0, -1, -1)$, $(0, -1, -1, +1, -1, -1)$, and $(+1, -1, -1, +1, -1, -1)$. On the other hand, a child partition is a partition which can follow the current partition. In this case, up to a change of signs, the child partitions of $(0, -1, +1, 0, -1, -1)$ are $(0, +1, 0, 0, -1, -1)$, $(0, -1, 0, 0, +1, -1)$, and $(0, -1, 0, 0, -1, +1)$. In the CoDa-dendrogram both parent and child partitions are, respectively, up and down on a consecutive level of the current partition. A top partition is a partition which does not contain zeros. In the previous example $(+1, -1, -1, +1, -1, -1)$ is a top partition, that is, the D parts are split into two groups and this would appear at the top of the CoDa-dendrogram.

Let \mathbf{X} be a CoDa (n, D) -set. The algorithm to find the PBs of \mathbf{X} is

Step 0 Initialize $d = D$

Step 1 Optimal current partition procedure

- load the file of all possible partitions of size d
- find the partition which has a balance with the maximum sample variance within all possible d -partitions and add this partition to a list of chosen partitions
- set r = number of parts marked $+1$ in the current partition and s = number of parts marked -1

Step 2 Optimal child partition procedure

- split the data set according to the parts marked $+1$ or -1 in the current partition found in Step 1
- if $r > 1$ take the parts marked with $+1$ and go to Step 1 (recursive algorithm) with $d = r$
- if $s > 1$ take the parts marked with -1 in the current partition found in Step 1 and go to Step 1 (recursive algorithm) with $d = s$

Step 3 Optimal parent partition procedure

- do-while loop: while there is a zero in the current partition, mark the non-zero parts with -1 , load the file of all possible parent partitions, find the parent partition

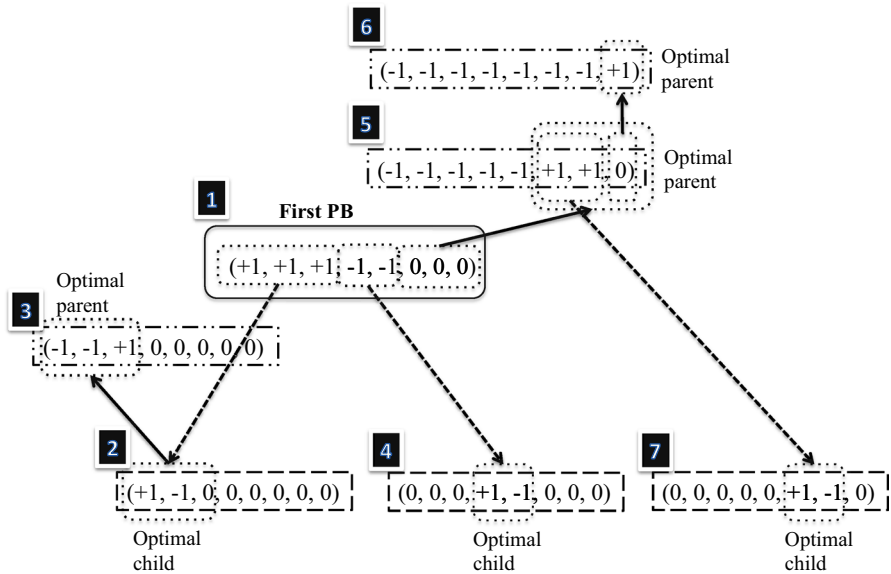


Fig. 2 Example: optimal PB algorithm. The label First PB indicates the PB that explains the maximum proportion of variance. Numbers in black rectangles indicate the sequence followed by the algorithm to construct the PBs (see text for details)

that maximizes the variance of the corresponding balance, add the selected parent partition to a list of parent partitions and chosen partitions, at the end a top partition is found

- do-for loop: for all the partitions in the list of parent partitions, take the parts marked with +1 and go to Step 1 with d equal to the number of selected parts

Step 4 Sort the chosen partitions in decreasing order of variance of their corresponding balances.

Figure 2 shows an example ($D = 8$) that illustrates the sequence followed by the algorithm to construct the PBs. First, take $d = 8$ in Step 0. In Step 1 the maximum variance associated with the partition $(+1, +1, +1, -1, -1, 0, 0, 0)$ was found and labelled as the First PB. According to the parts marked with +1 or -1, in Step 2 this partition is split into two partitions. The recursive algorithm applied to the parts marked with +1 found that the optimal child partition is $(+1, -1, 0, 0, 0, 0, 0, 0)$ (labelled 2), whose optimal parent partition is $(-1, -1, +1, 0, 0, 0, 0, 0)$ (labelled 3). The recursive algorithm applied to the parts marked with -1 in the First PB found that $(0, 0, 0, +1, -1, 0, 0, 0)$ is the optimal child partition (labelled 4). The list of consecutive optimal parent partitions of the First PB found in Step 3 is formed by the partition $(-1, -1, -1, -1, -1, +1, +1, 0)$ (labelled 5) and the top partition $(-1, -1, -1, -1, -1, -1, -1, +1)$ (labelled 6). When the parts marked with +1 in these parent partitions are sent to Step 1 the optimal child partition was found to be $(0, 0, 0, 0, 0, +1, -1, 0)$ (labelled 7). Once the SPB is completed the partitions are then sorted according to the variance of their corresponding ilr-coordinates.

Table 2 Four typical examples in geochemistry. Performance of PCA and of the PBs constructed using (O) optimal PBs, (C) constrained PCs and (W) Ward hierarchical clustering for four simple data sets (D number of parts; n number of samples). Columns in the center: computational time in seconds; columns on the right: cumulative percentage of variance explained by the first three PCs and PBs

Example	D	n	Time (sec.)				% Var. expl.			
			PCs	O	C	W	PCs	O	C	W
1	6	2097	0.013	0.041	0.028	0.003	90	88	85	88
2	8	778	0.001	0.420	0.033	0.003	68	65	64	65
3	10	87	0.006	3.061	0.026	0.002	94	85	84	82
4	13	2097	0.023	148.8	0.124	0.006	73	68	61	68

When this algorithm was applied to the geochemical data example described in Table 1, it took 0.08 s to provide the SBP in Table 1 (unsorted) and represented in Fig. 1. The PCA of CoDa, that is, the PCA applied to the clr-variables (Aitchison 1983), gives the loadings

$$\text{PC1} = (-0.06, -0.30, 0.89, -0.11, -0.28, -0.14),$$

$$\text{PC2} = (-0.08, -0.07, -0.02, -0.46, -0.23, +0.85).$$

These two PCs retain 77% of the variance. The balances

$$\text{ilr}_3(\mathbf{x}) = (0, -0.29, 0.87, 0, -0.29, -0.29) \text{ and}$$

$$\text{ilr}_4(\mathbf{x}) = (0, -0.41, 0, 0, -0.41, 0.82),$$

exhibit a similar performance in both loadings and accumulated variance (Table 1). This same harmony between the PBs and PCs was also found in the analysis of other data sets. Table 2 shows a summary for four typical examples of geochemical compositions with differing numbers of parts (D) and sample sizes (n). For the sake of simplicity, the descriptions of these examples, which are similar to the example given in Sect. 5.1, have not been provided.

4 Suboptimal (but Faster) Algorithms to Construct Principal Balances

4.1 Current Algorithms

Table 2 shows that the computation time for the optimal algorithm increases quickly with the dimension of the data set. Suboptimal strategies can then be appropriate to simplify the previous algorithm. Here the word suboptimal means that the accumulated variance explained by the first balances will be lower than or equal to the variance for the optimal algorithm. A number of approaches have already been proposed to solve this problem. Pawłowsky-Glahn et al. (2011) introduced three algorithms: the maximum explained variance hierarchical balances, the angular proximity to PCs, and the hierarchical clustering of components. The first algorithm tries to simplify the

exhaustive search by constraining it to a hierarchy of balances. It starts by searching for the full-balance (a balance that contains all the parts) that maximizes the retained variance. It consists of splitting the full composition into two sets of parts (numerator and denominator) that are recursively treated as the initial full composition. The angular proximity algorithm, in its first step, consists of finding the full-balance closest to one, but not necessarily the first, PC. After this, the algorithm recursively continues for the set of parts in the numerator and in the denominator. These two algorithms are far from being efficient (Mert et al. 2015) and so should not be considered. On the other hand, the third proposal, which uses the ward clustering method to cluster the parts of the composition, does provide better results (Mert et al. 2015). However, the SBPs constructed by this algorithm will always have the largest variance in the first balance that includes all the parts (full-balance). Other methods designed to provide directions as simple to interpret as possible have comparable drawbacks. For example, Jolliffe (2002) describes how rotating PCs produces a redistribution among the rotated PCs of the explained variance causing a loss of dominant information, while both truncating the PC loadings and the constrained PCs approach give directions that approximate the PCs, although these are not exactly orthogonal and are not uncorrelated. The SCoT (simplified component technique) and SCoTLASS (least shrinkage and selection operator) methods have tuning parameters chosen as a compromise between simplicity and variance explanation capability (Witten et al. 2011). Mert et al. (2015) introduce an algorithm to construct sparse PBs (SPB) for high-dimensional CoDa that is basically an adaptation of the algorithm implemented in Witten et al. (2011), that is, it is connected with the variable selection method introduced by Jolliffe et al. (2003). Because the algorithm is designed for high-dimensional data it focuses on dimension reduction, that is, the algorithm does not provide a complete log-ratio basis. Importantly, the authors remark that “...we want to ensure that different balances relate to different parts.” and that, “The SPB method results in balances which involve non-overlapping groups of variables...”. On the other hand, for dimension reduction purposes, the algorithm shows a remarkable reduction in the consumption of computational time. The authors acknowledge that their algorithm “...makes use of suboptimal projections ...to the clr hyperplane ...”, and that in its construction of PBs, “...the nearest balances to the resulting clr vectors are constructed [...] simply compute the arithmetic mean from the positive/negative entries ...”. These two strategies for designing interpretable dimension reduction algorithms are far from the procedure recommended in Chipman and Gu (2005).

In this work, two different suboptimal approaches for constructing PBs are presented. One is a new approach based on the information provided by the PCs and the other revisits Ward’s clustering method for parts. The latter is a new approach based on the relation between the variation matrix and the Aitchison distance. Both suboptimal approaches offer an appealing connection to the definition of balances.

4.2 Using Principal Component Loadings: New Constrained PCs Algorithm

A number of studies have been proposed to build interpretable PCs through a simplification of the loading vectors of the original PCs, both within the real space framework

(Chipman and Gu 2005; Jolliffe 2002; Jolliffe et al. 2003; Witten et al. 2011; Cox and Arnold 2016) and within the CoDa framework (Gallo et al. 2016; Mert et al. 2015). As a rule of thumb, simplifying a loadings vector means considering only the most important loadings and removing (force to zero) the rest and measuring the importance of a loading, for example by its magnitude. Consequently, the number of original variables or parts associated with each PC is reduced, making the relation more interpretable. Here, among the several proposals for simplification, the constrained PC approach introduced by Chipman and Gu (2005) was followed because it is consistent with Definition 1. Let $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \dots, \boldsymbol{\gamma}_D$ be the directions of the PCs of the clr-transformed data [Eq. (2)]. Let $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \dots, \boldsymbol{\alpha}_D$ be the corresponding simplified PCs, that is, the clr coefficients in Eq. (4). Following Chipman and Gu (2005), the components α_{ij} of the i -th vector $\boldsymbol{\alpha}_i$ only take values $-c_1, 0$, and c_2 , such that $\boldsymbol{\alpha}'_i \cdot \boldsymbol{\alpha}_i = 1$ and $\boldsymbol{\alpha}'_i \cdot \mathbf{1} = 0$. Chipman and Gu (2005) describe each vector $\boldsymbol{\alpha}_i$ as "...a difference of the average of one set of variables and the average of another set of variables, called a contrast". That is, a log-contrast when applied to log-transformed variables (Definition 1). Given the PC $\boldsymbol{\gamma}_i$, the best simplification $\boldsymbol{\alpha}_i$ minimizes the angle $\arccos(\boldsymbol{\gamma}'_i \cdot \boldsymbol{\alpha}_i)$, that is, it maximizes the inner product $\boldsymbol{\gamma}'_i \cdot \boldsymbol{\alpha}_i$. The search algorithm (Chipman and Gu 2005) starts by identifying the largest positive and negative coefficients of $\boldsymbol{\gamma}_i$ and sets, respectively, the corresponding elements of $\boldsymbol{\alpha}_i$ to $\pm \sqrt{2}/2$ [Eq. (4)]. All the other elements of $\boldsymbol{\alpha}_i$ are forced to zero. This procedure is repeated from three to D coefficients selected by absolute magnitude. Among these $D - 1$ possible elements $\boldsymbol{\alpha}_i$, the closest to $\boldsymbol{\gamma}_i$ is its best simplification. For example, for $\boldsymbol{\gamma}_1 = (-0.06, -0.30, 0.89, -0.11, -0.28, -0.14)$ the first PC of the data set in Table 1, the five candidates for simplification are the balances $(0, -1, 1, 0, 0, 0)$, $(0, -1, 1, 0, -1, 0)$, $(0, -1, 1, 0, -1, -1)$, $(0, -1, 1, -1, -1, -1)$, and $(-1, -1, 1, -1, -1, -1)$ accordingly normalized [Eq. (3)]. In this case, the closest balance, that is the balance with the smallest angle, is $\boldsymbol{\alpha}_1 = (0, -0.22, 0.89, -0.22, -0.22, -0.22)$ with an angle of 10.55° .

In essence, the constrained PCs approach is based on a restriction relaxation similar to the strategy commonly used in some operation research techniques. In fact, the calculation of the PCs directions $\boldsymbol{\gamma}_1, \boldsymbol{\gamma}_2, \dots, \boldsymbol{\gamma}_D$ is equivalent to relaxing the constraint $a_{ki} \in \{-c_1, 0, c_2\}$, $i = 1, \dots, D$, in Definition 1. Afterwards, the corresponding best simplification is constructed. Consequently, calculating the PCs and the posterior search algorithm of the balances is very straightforward. With the objective of optimizing the exhaustive searching, Steps 1 and 3 of the optimal algorithm to find approximate PBs using these techniques were replaced. That is, instead of loading the files of all possible partitions and parents of partitions, the constrained PCs were calculated respectively. As Step 1 is mainly responsible for consuming computational time, reducing it is very relevant. Indeed, the four data sets analyzed with the new algorithm (Table 2) take 0.028, 0.033, 0.026, and 0.124 s, respectively. The price paid for this time reduction is a decrease in the variance explained: the three first PBs obtained using this new algorithm for each of the four data sets in Table 2 explain, respectively, 85, 64, 84 and 61% of the total variance.

4.3 Using Hierarchical Clustering Methods for Parts: A New Approach to the Ward Method for Parts

Another different approach to constructing interpretable PCs is based on cluster techniques (Enki et al. 2013). Following Pawlowsky-Glahn et al. (2011) any hierarchical clustering of parts may be applied to create a dendrogram that can be used to construct an SBP. The hierarchical algorithms can be agglomerative or divisive, and amongst the agglomerative algorithms, perhaps the most commonly used are the linkage techniques (single, complete, average, centroid) and the Ward clustering method (Everitt et al. 2011). Each method is based on a measure of proximity or similarity between the elements and a particular rule for merging the groups. The properties and the performance of a method depend on the proximity and the rule selected. In all cases the clustering proceeds hierarchically, each being obtained by merging two clusters from the previous level.

Following Izenman (2008), a clustering method for variables generally uses a measure of proximity based on dependence measures, also known as measures of association or correlation. These measures provide some reasonable information of the closeness between two variables, that is, large correlation means that the two variables are close, and two variables for which the correlation is small are considered to be at a large distance from each other (Enki et al. 2013). Indeed, Izenman (2008, p. 439) shows that if the two variables are previously standardized (zero mean and unit variance), the squared Euclidean distance between the variables is proportional to one minus the Pearson correlation coefficient.

In CoDa analysis, variability (second order moments) can be expressed in several forms (Pawlowsky-Glahn et al. 2015). One of the more practical being the variation matrix (Aitchison 1986); which can also be used to define association measures between two parts of a composition across a sample (Lovell et al. 2015). Let \mathbf{X}_r and \mathbf{X}_s be two parts, that is, two columns of a data set \mathbf{X} ($n \times D$), with sample size n and D parts. If $\text{Var}(\ln(\mathbf{X}_r/\mathbf{X}_s))$ is exactly zero, the ratio $\mathbf{X}_r/\mathbf{X}_s$ is constant. Then, the two parts involved are proportional (Lovell et al. 2015). This work states, for the first time, that the entries of the variation matrix $\text{Var}(\ln(\mathbf{X}_r/\mathbf{X}_s))$ can also be expressed in terms of the Aitchison distance between parts. Indeed, the columns of \mathbf{X} can be considered as compositions in an n -part simplex. The i -th component of the clr transformation of \mathbf{X}_r is $\ln x_{ir} - \overline{\ln \mathbf{X}_r}$, where $\overline{\ln \mathbf{X}_r}$ is the average of the logarithms $\ln x_{ir}$ along the column \mathbf{X}_r . This is equivalent to considering \mathbf{X}^\top as a compositional data set and then taking clr transformation. This leads to

$$\begin{aligned} \text{Var} \left(\ln \left(\frac{\mathbf{X}_r}{\mathbf{X}_s} \right) \right) &= \frac{1}{n} \sum_{i=1}^n ((\ln x_{ir} - \ln x_{is}) - (\overline{\ln \mathbf{X}_r} - \overline{\ln \mathbf{X}_s}))^2 \\ &= \frac{1}{n} \sum_{i=1}^n ((\ln x_{ir} - \overline{\ln \mathbf{X}_r}) - (\ln x_{is} - \overline{\ln \mathbf{X}_s}))^2 \\ &= \frac{1}{n} d_a^2(\mathbf{X}_r, \mathbf{X}_s), \end{aligned} \quad (5)$$

where the last term is an Aitchison distance in the n -part simplex, that is, an Aitchison distance between parts. In terms of the corresponding balance [Eq. (3)], this expression is

$$\text{Var} \left(\frac{\sqrt{2}}{2} \cdot \ln \left(\frac{\mathbf{X}_r}{\mathbf{X}_s} \right) \right) = \frac{1}{2n} \cdot d_a^2(\mathbf{X}_r, \mathbf{X}_s).$$

Importantly, for a sample of size n , the matrix of the squared Aitchison distances between variables is the matrix of log-ratio variances, that is, the variation matrix (Aitchison 1986) multiplied by n .

The differences between the linkage techniques are due to the different ways of defining distance between a part and a group containing several parts, or between two groups of parts. Discussing the properties of each method is beyond the scope of this article. The Ward method is used here because of its interpretation in terms of balances. In general terms, the Ward clustering method involves merging clusters with the most similar centroids (mean vectors). The method measures this similarity defining the distance between two clusters, \mathbf{A} and \mathbf{B} , as the increase of the within cluster sum of squares when they are merged. This increase is equal to

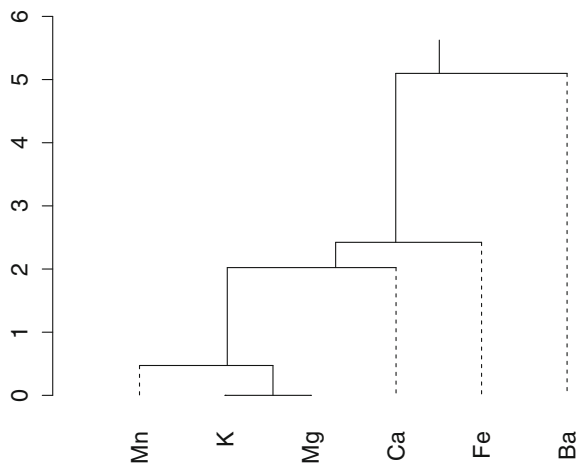
$$\frac{n_A \cdot n_B}{n_A + n_B} d_e^2(\bar{\mathbf{A}}, \bar{\mathbf{B}}),$$

where n_A, n_B are, respectively, the number of objects in each cluster, and $\bar{\mathbf{A}}, \bar{\mathbf{B}}$ stand for the centroids. In our case, the centroid of a cluster formed by the r parts $\mathbf{X}_{n1}, \mathbf{X}_{n2}, \dots, \mathbf{X}_{nr}$ is the geometric mean $\mathbf{G}_n = (\mathbf{X}_{n1} \dots \mathbf{X}_{nr})^{1/r}$. Consequently, the squared distance between the two clusters $\mathbf{X}_{n1}, \mathbf{X}_{n2}, \dots, \mathbf{X}_{nr}$ and $\mathbf{X}_{d1}, \mathbf{X}_{d2}, \dots, \mathbf{X}_{ds}$ is proportional to the variance of the balance,

$$\begin{aligned} \text{Var} \left(\sqrt{\frac{r \cdot s}{r + s}} \ln \left(\frac{\mathbf{G}_n}{\mathbf{G}_d} \right) \right) &= \frac{r \cdot s}{r + s} \text{Var} \left(\ln \left(\frac{\mathbf{G}_n}{\mathbf{G}_d} \right) \right) \\ &= \frac{r \cdot s}{r + s} \cdot \frac{1}{n} d_a^2(\mathbf{G}_n, \mathbf{G}_d) \\ &= \frac{r \cdot s}{r + s} \cdot \frac{1}{n} d_e^2(\text{clr}(\mathbf{G}_n), \text{clr}(\mathbf{G}_d)), \end{aligned}$$

where $\text{clr}(\mathbf{G}_n)$ and $\text{clr}(\mathbf{G}_d)$ are the centroids of the clr-transformed parts in each group. Given that the factor $1/n$ is common to all the entries in the distance matrix, the levels of hierarchy in the clustering are proportional to the variance of the corresponding balance. The clustering algorithm starts detecting the smallest entry in the variation matrix (matrix of log-ratio variances), and the corresponding parts are merged to form a group. The geometric mean of both columns (centroid of the group) is calculated and the variation matrix is updated. The algorithm iteratively continues merging groups of parts according to the smallest variance of the corresponding balance. The final stage consists of fusing the last two remaining groups into one, which gives the balance with the largest variance.

Fig. 3 CoDa-dendrogram using the SBP with the constrained PC algorithm



Applied to the four simple data sets in Table 2, the performance of this algorithm, with regards to the explained variance, is mostly better than the algorithm using the constrained PCs. Note too, that it is also very similar to the optimal algorithm: only in the third example is the percentage of explained variance lower, dropping from 85 to 82%. On the other hand, the decrease in computation time is now very relevant because it takes, respectively, 0.003, 0.003, 0.002 and 0.006 s.

Figure 3 shows the CoDa-dendrogram provided by the constrained PCs algorithm while Fig. 4 shows the Ward clustering method. By definition of the hierarchical clustering method, the SBP constructed will have the largest variance in the first balance (full-balance) that includes all the parts. On the other hand, the constrained PCs algorithm can find an SBP where the largest variance is in an intermediate balance (not full-balance) in harmony with the optimal algorithm (Fig. 1). However, the Ward method enhances the proportionality between groups of parts. For example, the closest parts, *Ca* and *Mg*, detected by the optimal algorithm, are also found with the clustering method, but are not detected by the constrained PCs algorithm.

5 Examples

5.1 Aar Massif Data

By means of illustration, the methods presented here are applied to characterize a data set of the geochemical composition of glacial sediment from a granodioritic-gneissic source rock (Aar Massif, CH) (Tolosana-Delgado and Eynatten 2010; Eynatten et al. 2012). This data set contains measurements of 10 major oxides and 16 trace elements from 87 samples of differing grain sizes. To keep representations simple, only the major elements converted to oxides (Al_2O_3 , CaO , Fe_2O_3 , K_2O , MgO , MnO , Na_2O , P_2O_5 , SiO_2 , TiO_2) were retained.

In practice, the analyst would start the analysis with an exploratory study using the representation of the compositional (covariance) biplot (Fig. 5). The compositional

Fig. 4 CoDa-dendrogram using the SBP with the Ward clustering method

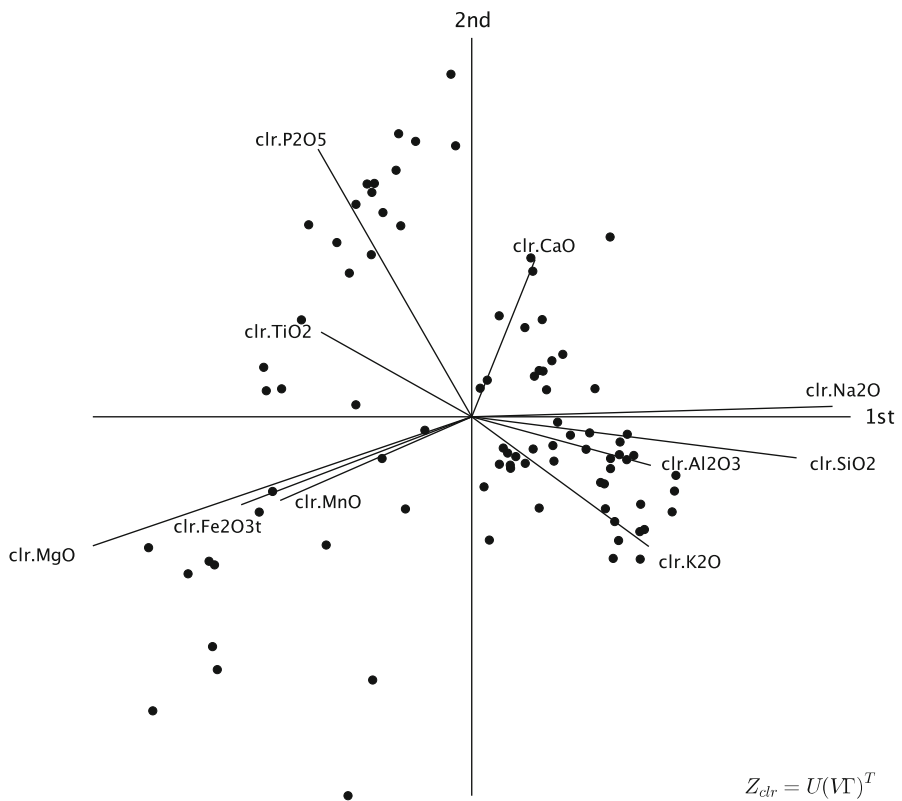
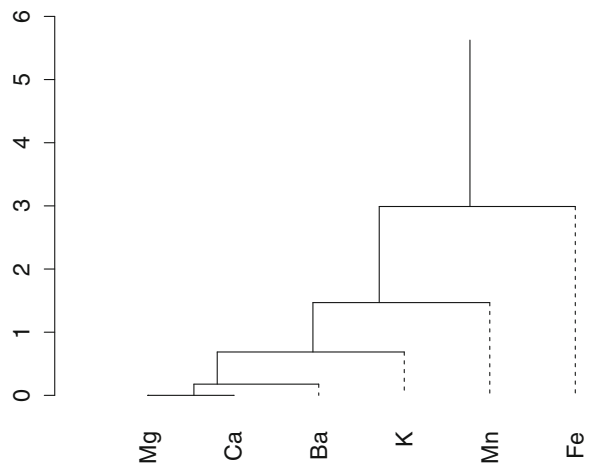


Fig. 5 Compositional biplot of Aar Massif data set

biplot is a PCA of the centered clr-transformed data [Eq. (2)] used to represent the variables and the samples simultaneously in a graph. Intuitively, an SBP approaching PBs can be constructed by examining the biplot in Fig. 5. A first partition can be identified as the variables whose rays point towards the extremes of the long link approximately parallel to the first axis. A first group is marked with + 1 (Al_2O_3 , CaO , K_2O , Na_2O , SiO_2) and the group of parts (Fe_2O_3 , MgO , MnO , P_2O_5 , TiO_2) marked with -1 . To form the second PB, one of the two groups has to be split into two. For example, the group of positive parts could be split into CaO versus (Al_2O_3 , K_2O , Na_2O , SiO_2). Afterwards, one could continue with, for example, (Al_2O_3 , K_2O) versus (Na_2O , SiO_2) or any other possibility selected by using knowledge about the problem being studied. To proceed here, the group of negative parts (Fe_2O_3 , MgO , MnO , P_2O_5 , TiO_2) is split analogously. Despite the fact that this exploratory and subjective procedure for constructing PBs might be useful, it does not guarantee a reasonable approximation to the PCs, that is, a sequential maximum explanation of variance. On the other hand, the three algorithms introduced in this article are procedures to construct PBs based on numerical criteria.

In Table 3, a PCA of the clr-transformed data [Eq. (2)] yields loadings and contributions to variance displayed (labelled “ PC_i ”). These results serve as a reference with which to compare the three methods: optimal principal balances (O), constrained principal components (C) and Ward hierarchical clustering (W). The PBs are ordered according to the percentage of variance explained. For example, PB_2 for method O was the sixth balance constructed (O_6). For a better interpretation, the values equal to zero are reported as blanks in Table 3. In terms of the computational consumption time, note that the results correspond to the (aforementioned) third example in Table 2.

The methods O and C construct the same first PB. Both PB_1 identify two main groups: felsic elements (Al_2O_3 , K_2O , Na_2O , SiO_2) versus mafic elements (Fe_2O_3 , MgO , MnO , P_2O_5 , TiO_2) and omit the element CaO , which has the lowest loading (0.08) in PC_1 . Because method W must construct a first PB using the full composition, it cannot leave out the element CaO . Method W assigns this element to the mafic elements, suggesting an association through apatite, calcium phosphate. On the other hand, method W achieves the best approximation for the second PC. Indeed, its PB_2 opposes elements CaO and P_2O_5 to the mafic parts, similar to PC_2 including the element K_2O and assigning a different sign to the element TiO_2 . Method C assigns the correct sign to the element TiO_2 , whereas it is not included by method O in the second PB. Neither methods O or C include the element CaO . The three methods construct the same third PB: (Al_2O_3 , K_2O) versus (Na_2O , SiO_2). Except for the loading of the element TiO_2 , this PB collects the largest loadings of the third PC. These types of relations between the loadings of the PCs and the PBs can be described until the last PC, where the groups of parts highlight the relation between the element Fe_2O_3 and the elements (MgO , MnO). However, sometimes the order of the PBs does not coincide with the most similar PC. For example, for methods O and W PB_6 suggests the relation between the element TiO_2 and other mafic parts that is partially shown by PC_5 . In addition to the large variance explained by the two or three first PBs, the analysis of their corresponding loadings (Table 3) corroborates that these methods provide a reasonable approximation to PCs. Figure 6 shows the dendrogram associated to each method of constructing PBs. The structure of the SBP associated

Table 3 Aar Massif data set: loadings and vector of explained variances of PCA and of the PBs constructed using (O) optimal PBs, (C) constrained PCs, and (W) Ward hierarchical clustering. PBs are ordered according to their variance. Values equal to zero are reported as blanks

	PC ₁	PC ₂	PC ₃	PC ₄	PC ₅	PC ₆	PC ₇	PC ₈	PC ₉
Al ₂ O ₃	0.23	0.12	0.32	-0.09	0.21	0.24	0.04	0.77	-0.15
CaO	0.08	-0.40	0.14	0.24	0.17	-0.68	-0.39	0.04	-0.12
Fe ₂ O ₃ t	-0.30	0.22	0.01	-0.04	-0.07	-0.29	0.20	0.16	0.78
K ₂ O	0.23	0.33	0.32	-0.56	-0.06	0.03	-0.38	-0.42	0.02
MgO	-0.50	0.33	0.19	0.55	-0.11	0.28	-0.24	-0.13	-0.21
MnO	-0.25	0.21	-0.28	-0.29	-0.08	-0.35	0.45	-0.00	-0.55
Na ₂ O	0.47	-0.03	0.19	0.37	0.17	0.07	0.56	-0.39	0.06
P ₂ O ₅	-0.20	-0.68	0.18	-0.21	-0.48	0.29	0.12	-0.02	0.02
SiO ₂	0.43	0.10	-0.63	0.18	-0.44	0.09	-0.24	0.12	0.06
TiO ₂	-0.20	-0.21	-0.44	-0.15	0.67	0.32	-0.12	-0.13	0.09
Expl. var (in %)	71.22	19.05	4.28	2.66	1.80	0.67	0.17	0.10	0.05
	O ₁	O ₆	O ₃	O ₂	O ₈	O ₇	O ₅	O ₄	O ₉
Al ₂ O ₃	0.37		0.50	-0.11				0.71	
CaO				0.95					
Fe ₂ O ₃ t	-0.30	0.29		-0.11		-0.22			0.82
K ₂ O	0.37		0.50	-0.11				-0.71	
MgO	-0.30	0.29		-0.11	0.71	-0.22			-0.41
MnO	-0.30	0.29		-0.11	-0.71	-0.22			-0.41
Na ₂ O	0.37		-0.50	-0.11			0.71		
P ₂ O ₅	-0.30	-0.87		-0.11		-0.22			
SiO ₂	0.37		-0.50	-0.11			-0.71		
TiO ₂	-0.30			-0.11		0.89			
Expl. var (in %)	64.15	14.15	6.29	4.60	3.86	3.54	2.13	0.86	0.41

Table 3 continued

	C ₁	C ₃	C ₇	C ₂	C ₄	C ₅	C ₈	C ₉	C ₆
Al ₂ O ₃	0.37		0.50	−0.11				0.71	
CaO				0.95					
Fe ₂ O ₃ t	−0.30	0.37		−0.11					0.82
K ₂ O	0.37		0.50	−0.11				−0.71	
MgO	−0.30	0.37		−0.11		0.71			−0.41
MnO	−0.30	0.37		−0.11		−0.71			−0.41
Na ₂ O	0.37		−0.50	−0.11			0.71		
P ₂ O ₅	−0.30	−0.55		−0.11	−0.71				
SiO ₂	0.37		−0.50	−0.11			−0.71		
TiO ₂	−0.30	−0.55		−0.11	0.71				
Expl. var (in %)	64.15	13.63	6.29	4.60	4.07	3.86	2.13	0.86	0.41
	W ₁	W ₂	W ₃	W ₄	W ₅	W ₆	W ₇	W ₈	W ₉
Al ₂ O ₃	0.39		0.50					0.71	
CaO	−0.26	−0.58			0.71				
Fe ₂ O ₃ t	−0.26	0.29		−0.29		−0.41		−0.71	0.71
K ₂ O	0.39		0.50						
MgO	−0.26	0.29		0.87					
MnO	−0.26	0.29		−0.29		−0.41			−0.71
Na ₂ O	0.39		−0.50				0.71		
P ₂ O ₅	−0.26	−0.58			−0.71				
SiO ₂	0.39		−0.50				−0.71		
TiO ₂	−0.26	0.29		−0.29		0.82			
Expl. var (in %)	57.63	18.28	6.29	5.96	4.61	3.83	2.13	0.86	0.41

to the three methods is very similar because all of them separate the felsic and mafic elements exhibiting the largest variability. However, methods O and C differ from method W in the treatment of the element CaO because they separate this element from the felsic and mafic elements, while method W assigns it to the mafic elements. Despite all the methods assigning the element TiO₂ to the mafic element, method O is different because it assigns the element as the last step in forming the group of mafic elements.

In Table 3 it can be appreciated that non-zero loadings corresponding to a balance (sub-tables O, C, W) admit only two values that are different, in contrast to PCs, for which the condition is that the loadings add to zero and the sum of their squares add to one. Therefore, the variance retained by the first PBs will always be less than or equal to the variance explained by the corresponding PCs. In this example this holds only for the two first balances. The first PC explains 71.22% of the variance, whereas methods O and C retain 64.15%, and with only 58.63% method W is the worst approximation. On the other hand, the 19.05% explained by PC₂ is better approximated by method W (18.28%) than by the other two methods, with, respectively, 14.15 and 13.63%.

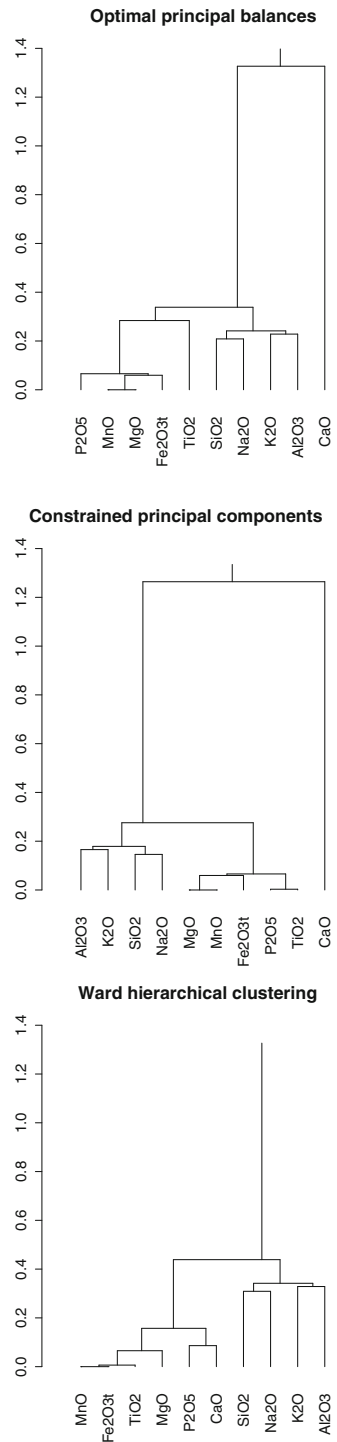
Because the set of vectors must explain the total variance of the data set, the variance explained by some of the PBs has to be greater than the variance retained by the corresponding PC. Therefore, it is necessary to define measures of effectiveness which will allow how well the PBs approach the PCs to be evaluated. Let $\mathbf{v} = (v_1, v_2, \dots, v_{D-1})$ be the vector containing the variances of estimated PCs or PBs ordered from maximum to minimum variance. The components of vector \mathbf{v} are commonly represented graphically as a scree plot (Jolliffe 2002, p. 115) with the purpose of choosing a subset of PCs for dimensionality reduction. In this work, it is preferable to represent the cumulative percentage of variance explained

$$100 \cdot \frac{\sum_{j=1}^k v_j}{\sum_{j=1}^{D-1} v_j}, \quad k = 1, \dots, D - 1.$$

Figure 7 (left) shows that, as expected, the PCs (filled circle) exhibit the best performance. Among the PB methods, method W (triangle) has the poorest performance, whereas methods O (circle) and C (cross) are nearly coincident.

The components of vector \mathbf{v} are positive and add to the total variance. This is not relevant to the analysis, as the importance the components of vector \mathbf{v} have is evaluated in terms of the proportion of the total variance (Table 3). Following Pardo et al. (2010), this suggests that vector \mathbf{v} carries relative information, that is, the compositional approach is adequate. Consequently, \mathbf{v} can be taken as a composition. Consider the sequence of subcompositions of \mathbf{v} : $\mathbf{v}_2 = (v_1, v_2)$, $\mathbf{v}_3 = (v_1, v_2, v_3)$, \dots , $\mathbf{v}_k = (v_1, v_2, \dots, v_k)$, $k \leq D - 1$. Their Aitchison norm, $\|\mathbf{v}_k\|_a$, measures the concentration of variance in the first k (ordered) components. By construction and due to its subcompositional coherence, the Aitchison norm of the subcompositions $\mathbf{v}_2, \mathbf{v}_3, \dots, \mathbf{v}_k$ provides a sequence of increasing values. Figure 7 (right) shows that these values are useful for measuring the effectiveness approaching PCs. It shows the Aitchison norm of the vector of variances for PCs (filled circle), O (circle), C (cross), and W (triangle) for different sizes of the subcomposition. Starting with vector \mathbf{v}_3 , the

Fig. 6 Aar Massif data set:
CoDa-dendrogram for PBs
obtained with each of the three
methods. Up: optimal PBs (O);
center: constrained PCs (C);
down: Ward hierarchical
clustering (W)



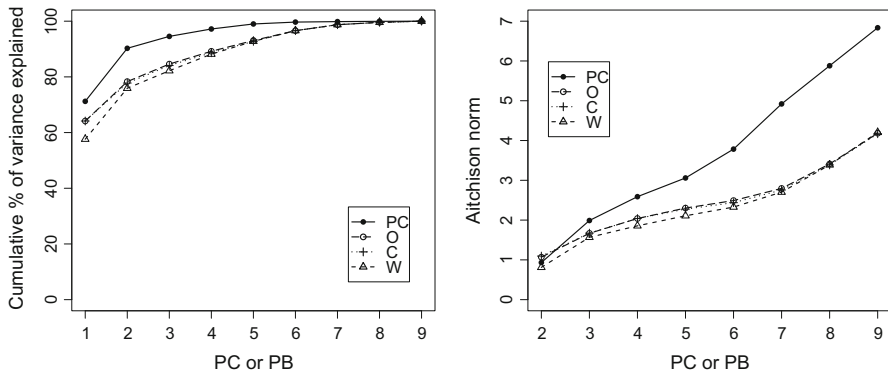


Fig. 7 Aar Massif data set. Left: cumulative percentage of explained variance for each PC or PB. Right: Aitchison norm of the vector of ordered explained variances taken as a composition

curve representing the variance explained by PC appears, as expected, above the other lines. The curve of W is always the lowest one, thus reflecting the price to be paid for estimating the first balance using only full compositions. The two recursive methods O and C produce Aitchison norm curves that almost overlap. These observations hold for the particular data set used; the shape and order of the curves will not necessarily be the same for other data and/or numbers of parts.

Further comparison of methods requires a measure of effectiveness approaching PCs. A first idea comes from the property of the PCs, which are uncorrelated random variables by construction. As the PBs approach PCs, correlations between the PBs are expected to be small. To evaluate the deviation from zero Fig. 8 shows the absolute value of the Pearson correlation coefficients between the isometric logratio coordinates [Eq. (3)] obtained by the PBs for methods O, C and W, respectively. The behavior of the three methods is very similar with regard to the minimum, maximum and median values. Method W shows the largest interquartile range, whereas the shortest is from method O. Noticeably, the PB methods do not aim to provide uncorrelated balances; on the contrary, the suggestion of both low and high correlations between ratios of parts or groups of parts may provide valuable information. For example, PB₈ and PB₉ in method O (O₄ and O₉) show a correlation coefficient equal to -0.01 , indicating that the ratio $\text{Al}_2\text{O}_3/\text{K}_2\text{O}$ can be assumed uncorrelated to the ratio $\text{Fe}_2\text{O}_3/\sqrt{(\text{MgO} \cdot \text{MnO})}$. In addition, the same method provides a correlation coefficient equal to -0.79 between PB₅ (i.e., O₈) and PB₉ (i.e., O₉) suggesting a linear relation between the parts Fe_2O_3 , MgO , and MnO .

5.2 Simulations

Although the PB algorithms presented aim to construct complete log-ratio bases, their performance as dimension reduction techniques when the number of parts is large, namely $D > 50$, is briefly explored here. For this situation, the use of the optimal algorithm would require a powerful computational system. In other words, this is not recommendable for a personal computer. Consequently, this section compares the

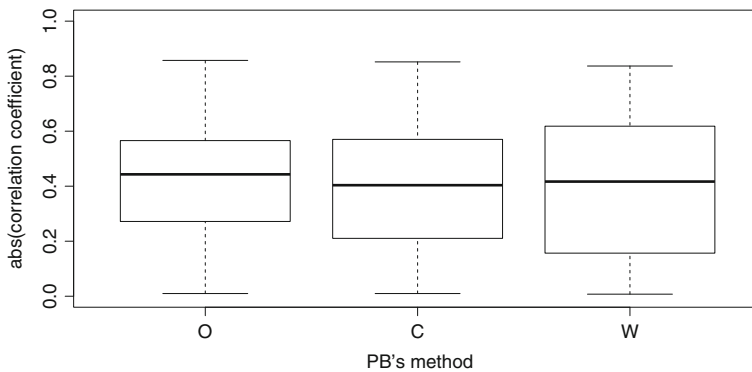


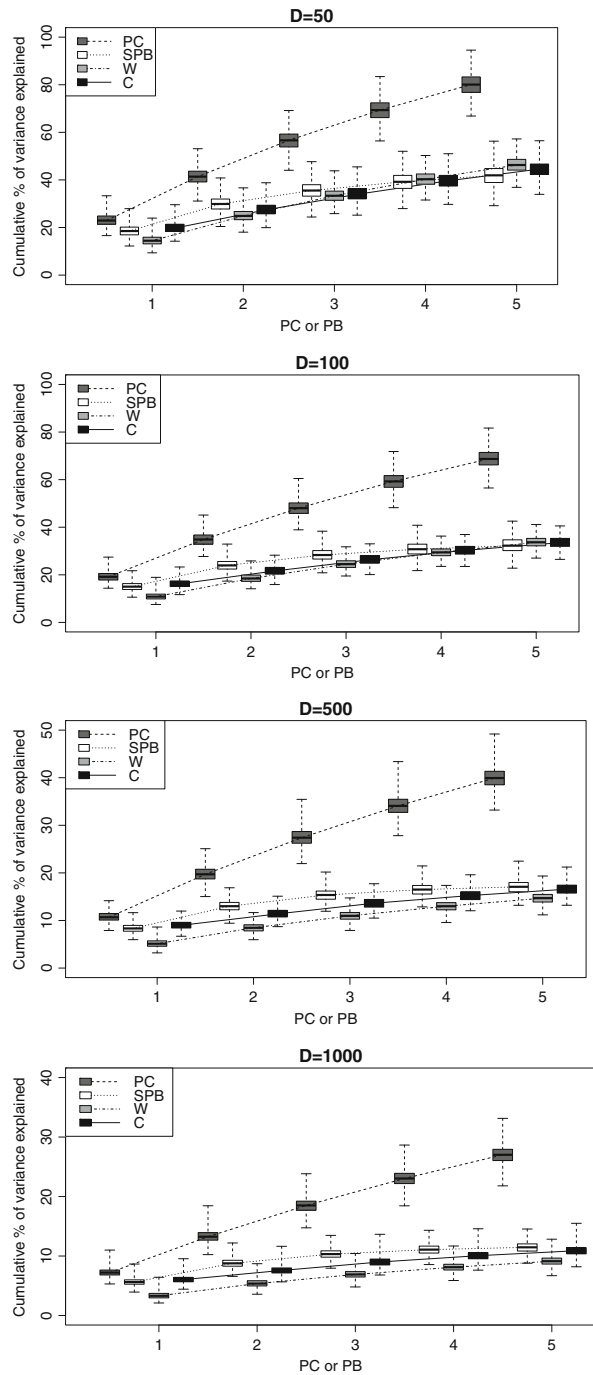
Fig. 8 Aar Massif data set: box plots of the absolute value of the Pearson correlation coefficient when the coordinates are expressed according the PBs provided by the different methods: optimal PBs (O); constrained PCs (C); and Ward hierarchical clustering (W)

constrained PCs and the Ward hierarchical clustering algorithms with the PCA and the SPB methods (Mert et al. 2015).

Following the scheme established in Mert et al. (2015), in each simulation run 1000 replications of a compositional data set with $n = 100$ observations and with the number of parts $D \in \{50; 100; 500; 1000\}$ were generated. To create one data set, a loadings matrix $(D - 1) \times (D - 1)$ was first generated in the ilr-space through a uniform distribution in the interval $[-1, 1]$. Next, a scores matrix $n \times (D - 1)$ was generated in the ilr-space using a multivariate Gaussian distribution centered at the origin of coordinates and with a diagonal covariance matrix. The $D - 1$ values of the diagonal vector of this matrix were taken as $(0.9, 0.9^2, \dots, 0.9^{10}, 0.01, \dots, 0.01)$. The form of this matrix guarantees the uncorrelation of the principal components and an exponential decrease of the first eigenvalues. Using the product of the loadings matrix and scores matrix, the matrix $n \times (D - 1)$ of ilr coordinates was obtained and back transformed to the simplex to create the matrix $n \times D$ for the compositional data set.

Because the aim is to evaluate the performance of the algorithms as a dimension reduction technique, the focus was placed on the explained variance of the resulting first balances provided by the methods. Figure 9 shows the boxplots for the cumulative percent of explained variance of the first five components for the methods: the PCA (PC), the sparse PBs (SPB) (Mert et al. 2015), and the two suboptimal algorithms proposed here: the constrained PCs (C) and the Ward hierarchical clustering (W). The four figures correspond, respectively, to the simulated data for the dimension $D \in \{50; 100; 500; 1000\}$. Each boxplot summarizes the results of the 1000 simulations. As expected, the PCA outperforms the PB methods in all the cases. As a general behavior, the SPB method seems to explain more variance for the two first PBs than the algorithms C and W do. However, the boxplot trends suggest that the suboptimal algorithms achieve, and even outperform, the same level as the SPB method for $D \in \{50, 100\}$. A similar relation is suggested between the algorithms C and W. That is, the figures suggest that the algorithm C explains more variance than the algorithm W for the

Fig. 9 Cumulative percent of explained variance of the first five components for PCA (PC), sparse PBs (SPB); constrained PCs (C); and Ward hierarchical clustering (W) for simulated data with different dimension $D \in \{50; 100; 500; 1000\}$



two first PBs, but the trend is that the algorithm W achieves the performance of the algorithm C. Another general feature is that in all the cases the differences among the three methods increase when the dimension increases.

As regards the consumption of computational time, the main disadvantage to method SPB is the previous selection of an optimal tuning parameter (Jolliffe 2002; Mert et al. 2015). This process consists of creating the PBs and calculating their cumulative explained variance for a sequence of values for the tuning parameter. In the simulation here, in a typical example with $n = 100$ and $D = 50$ and a sequence of 30 tuning values, method SPB took about 2.25 s to provide the five first PBs of the data set. For this same case, the algorithms C and W took about 0.05 and 0.01 s, respectively, to provide the complete log-ratio basis formed of 49 PBs. When the dimension was increased to $D = 1000$, the results were about 15 s for the SPB method. In this case, to provide the corresponding 999 PBs, the algorithms C and W took 2.5 and 1.5 s, respectively. However, the construction of the five first PBs without the tuning selection process for one of the replications of the data set is reasonably fast. Once the tuning parameter is fixed, the SPB method took only about 0.07 and 0.5 s, respectively, thus becoming faster than the sub-optimal algorithms for high-dimensional data sets, as stated by Mert et al. (2015).

6 Conclusions

The PCA of a set of compositions (clr-transformed and centered) has a number of appealing properties: maximum explained variance of the sequence of PCs, uncorrelated components, and orthogonal geometric axes. Because of these properties, PCA is one of the main tools for exploratory analysis and CoDa modeling. The main shortcoming of PCs is the difficulty in interpreting the resulting coordinates because a PC is a function of all the original parts.

Balances are log-contrasts resulting from a logratio of two geometric means of two groups of parts. Their interpretation is considerably simpler than that of PCs. In the present contribution, the idea of approximating the complete log-ratio basis of PCs using a set of balances, called principal balances (PBs), has been formalized. A new optimal algorithm for constructing PBs has been introduced. However, PB computation requires an exhaustive search over the possible set of orthonormal balances, which may consume a considerable amount of computational time when the number of parts increases. To avoid this, two suboptimal but faster procedures to search for PBs have been presented. A new algorithm based on the construction of constrained PCs provides similar results to the optimal algorithm. This algorithm substitutes for the exhaustive search an efficient search driven by the information provided by the PCs. The second algorithm, based on the Ward hierarchical clustering method, deals with the information provided by the variation matrix and enhances the associations between groups of parts, but forces the first PB to involve all the parts. The performance of the three algorithms has been analyzed and discussed. A new measure of effectiveness based on the Aitchison norm has been proposed. The results obtained corroborate the theoretical properties of the methods: they approximate reasonably well the complete log-ratio basis of PCs, thus improving interpretability. However, the price paid is a

smaller amount of variance explained by the first balances and the lack of uncorrelation between the coordinates.

The aim of the algorithms introduced in this work is, rather than a dimension reduction technique, to provide useful tools for data driven construction of a complete log-ratio basis. For a reduced number of parts it is feasible to apply the optimal algorithm, although its use is conditioned by the computational time, that is, by the quality and capacity of the available computational resources. If the resource available is a common personal computer, then using one of the suboptimal algorithms to increase number of parts would be preferable. To this aim, the Ward hierarchical clustering algorithm is recommended if the interest focuses on the relation of proportionality between parts or groups of parts. Despite high-dimensional data sets (hundreds or thousands of parts) perhaps not being frequent in geochemistry, they are common in other fields such as genomics or in the analysis of the microbiome. In such a scenario, one can use the SPB method (Mert et al. 2015) that was designed as a dimension reduction technique. However, when compared with the suboptimal algorithms these approaches show reasonable results with regard to the consumption of computational time and to the cumulative percentage of variance explained. For example, we found that for $D = 1000$, the first five PBs provided by the algorithms SPB and constrained PCs accumulate approximately 10 and 9.5% total variance, respectively. Importantly, because the constrained PCs approach is based on an optimization technique, its performance might be improved following new emerging algorithms (Cox and Arnold 2016). The ideas presented in this article could contribute to introducing new algorithms for high-dimensional data to improve these performances.

Acknowledgements This research has been supported by the Spanish Ministry of Economy and Competitiveness under the project CODA-RETOS (Ref: MTM2015-65016-C2-1(2)-R); and by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project COSDA (Ref: 2014SGR551). The authors gratefully acknowledge the constructive comments of the anonymous referees which have undoubtedly helped to significantly improve the quality of the paper.

References

- Aitchison J (1982) The statistical analysis of compositional data (with discussion). *J R Stat Soc B Methodol* 44:139–177
- Aitchison J (1983) Principal component analysis of compositional data. *Biometrika* 70:57–65
- Aitchison J (1986) The statistical analysis of compositional data. Monographs on statistics and applied probability. Chapman & Hall Ltd., London. (Reprinted in 2003 with additional material by The Blackburn Press)
- Aitchison J, Greenacre M (2002) Biplots for compositional data. *J R Stat Soc C Appl* 51:375–392
- Barceló-Vidal C, Martín-Fernández JA (2016) The mathematics of compositional analysis. *Austrian J Stat* 45:57–71
- Chipman HA, Gu H (2005) Interpretable dimension reduction. *J Appl Stat* 32:969–987
- Cox TF, Arnold DS (2016) Simple components. *J App Stat*. <https://doi.org/10.1080/02664763.2016.1268104>
- Enki HA, Trendafilov NT, Jolliffe IT (2013) A clustering approach to interpretable principal components. *J Appl Stat* 40:583–599
- Egozcue JJ, Pawłowsky-Glahn V (2005) Groups of parts and their balances in compositional data analysis. *Math Geol* 37:795–828
- Egozcue JJ, Pawłowsky-Glahn V (2006) Simplicial geometry for compositional data. *Geol Soc Spec Pub* 264:145–159

- Egozcue JJ, Pawłowsky-Glahn V, Mateu-Figueras G, Barceló-Vidal C (2003) Isometric logratio transformations for compositional data analysis. *Math Geol* 35:279–300
- Everitt BS, Landau S, Leese M, Stahl D (2011) *Cluster analysis*. Wiley, Chichester
- Gallo M, Trendafilov NT, Buccianti A (2016) Sparse PCA and investigation of multi-elements compositional repositories: theory and applications. *Environ Ecol Stat* 23:421–434
- Hottelling H (1933) Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 24:417–441
- Izenman AJ (2008) *Modern multivariate statistical techniques: regression, classification, and manifold learning*. Springer, New York
- Jolliffe IT (2002) *Principal component analysis*, 2nd edn. Springer series in statistics. Springer, New York
- Jolliffe IT, Trendafilov NT, Uddin M (2003) A modified principal component technique based on the LASSO. *J Comput Graph Stat* 12:531–547
- Lovell D, Pawłowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J (2015) Proportionality: a valid alternative to correlation for relative data. *PLoS Comput Biol* 11(3):e1004075. <https://doi.org/10.1371/journal.pcbi.1004075>
- Mateu-Figueras G, Pawłowsky-Glahn V, Egozcue JJ (2011) The principle of working on coordinates. In: Pawłowsky-Glahn V, Buccianti A (eds) *Compositional data analysis: theory and applications*. Wiley, Chichester, pp 31–42
- Mert MC, Filzmoser P, Hron K (2015) Sparse principal balances. *Stat Model* 15:159–174
- Palarea-Albaladejo J, Martín-Fernández JA, Soto JA (2012) Dealing with distances and transformations for fuzzy C-means clustering of compositional data. *J Classif* 29:144–169
- Palarea-Albaladejo J, Martín-Fernández JA (2015) zCompositions—R package for multivariate imputation of nondetects and zeros in compositional data sets. *Chemom Intell Lab* 143:85–96
- Pawłowsky-Glahn V, Egozcue JJ (2001) Geometric approach to statistical analysis on the simplex. *Stoch Environ Res Risk Assess* 15:384–398
- Pawłowsky-Glahn V, Egozcue JJ (2011) Exploring compositional data with the CoDa-dendrogram. *Austrian J Stat* 40:103–113
- Pawłowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2011) Principal balances. In: Egozcue JJ, Tolosana-Delgado R, Ortego M (eds) *Proceedings of the 4th international workshop on compositional data analysis*, Girona, Spain, pp 1–10
- Pawłowsky-Glahn V, Egozcue JJ, Tolosana-Delgado R (2015) *Modeling and analysis of compositional data. Statistics in practice*. Wiley, Chichester
- Podani J (2000) Simulation of random dendrograms and comparison tests: some comments. *J Classif* 17:123–142
- Prados F, Boada I, Prats A, Martín-Fernández JA, Feixas M, Blasco G, Puig J, Pedraza S (2010) Analysis of new diffusion tensor imaging anisotropy measures in the 3P-plot. *J Magn Reson Imaging* 31:1435–1444
- R development core team (2015) *R: a language and environment for statistical computing*. Vienna. <http://www.r-project.org>
- Tolosana-Delgado R, von Eynatten H (2010) Simplifying compositional multiple regression: application to grain size controls on sediment geochemistry. *Comput Geosci* 36:577–589
- von Eynatten H, Tolosana-Delgado R, Karius V (2012) Sediment generation in modern glacial settings: grain-size and source-rock control on sediment composition. *Sediment Geol* 280:80–92
- Witten D, Tibshirani R, Gross S, Narasimhan B (2011) PMA: penalized multivariate analysis. R Package Version 1:8

Reproduced with permission of copyright owner.
Further reproduction prohibited without permission.