Principal balances

V. PAWLOWSKY-GLAHN[1], J. J. EGOZCUE[2] and R. TOLOSANA-DELGADO[3]

[1]Dept. Informàtica i Matemàtica Aplicada, U. de Girona, Spain (vera.pawlowsky@udg.edu)
[2]Dept. Matemàtica Aplicada III, U. Politècnica de Catalunya, Barcelona, Spain (juan.jose.egozcue@upc.edu)
[3]Dept. Eng. Hidraulica, Marìtima i Ambiental, U. Politècnica de Catalunya, Barcelona, Spain
(raimon.tolosana@upc.edu)

Abstract

Principal balances are defined as a sequence of orthonormal balances which maximize successively the explained variance in a data set. Apparently, computing principal balances requires an exhaustive search along all possible sets of orthogonal balances. This is unaffordable for even a small number of parts. Three suboptimal, but feasible, alternatives are explored. The approach is illustrated using a data-set of geochemical composition of glacial sediments.

Introduction

Principal component analysis (PCA), based on statistical criteria, is well known in statistics (Hotelling, 1933). The procedure linearly transforms a number of centered variables into a new set of uncorrelated variables called principal components (PC). PCs are selected so that (a) they are uncorrelated; (b) the first PC is the linear combination of the original variables which attains the largest sample variance; subsequent PCs maximize explained variance conditional to be uncorrelated to preceding PCs. Geometrically, each PC is associated with a direction represented by a unitary vector (also called Principal Direction (PD)). PDs constitute an orthonormal basis of the space. The sample values of PCs, called scores, are expressed as coordinates with respect to the PDs. When PCA is applied to centred-log-ratio (clr) transformed compositional data, it provides isometric–log-ratio (ilr) coordinates as scores of each PC, and an orthonormal basis of the simplex given by the PDs. Consequently, PCA for compositional data (CoDa) is a powerful tool in exploratory analysis.

However, the obtained ilr-coordinates can be difficult to interpret as they are log-contrasts generally involving all the parts of the composition with irregular coefficients. Although the CoDa-biplot (Aitchison and Greenacre, 2002; Greenacre, 2011) may help to simplify interpretation of PCs, the interpretation problem is still a difficult one, specially when the number of parts of the composition is large. To overcome this difficulty, balances were introduced by Egozcue and Pawlowsky-Glahn (2005). *Balances* are log-contrasts which are log-ratios of geometric means of two non-overlapping groups of parts. They are then normalised so that they are the coordinate of the composition with respect to a unitary vector called *balancing element* (Egozcue et al., 2003). The general expression of a balance is

$$b = \sqrt{\frac{rs}{r+s}} \ln \frac{\mathrm{g_m}(\mathbf{x}_+)}{\mathrm{g_m}(\mathbf{x}_-)} \ , \tag{1}$$

where $\mathbf{x}_+$, $\mathbf{x}_-$ are two non-overlapping groups of parts of a complete composition $\mathbf{x}$ of D parts; r and s, $r+s \leq D$, are the number of parts in $\mathbf{x}_+$, $\mathbf{x}_-$ respectively, and $\mathrm{g_m}(\cdot)$ denotes the geometric mean of the arguments. The corresponding balancing element is

$$\mathbf{e} = \mathcal{C} \exp(v_1, v_2, \ldots, v_D) \ ,$$

where $\mathcal{C}$ denotes closure, exp applies componentwise and $(v_1, v_2, \ldots, v_D) = \mathrm{clr}(\mathbf{e})$. The clr components v_i have the following values: $v_i = 0$ if the i-th part is neither in $\mathbf{x}_+$ nor in $\mathbf{x}_-$; $v_i = (s/(r(r+s)))^{1/2}$ if the i-th part is in $\mathbf{x}_+$; and $v_i = -(r/(s(r+s)))^{1/2}$ if the i-th part is in $\mathbf{x}_-$.

A set of orthonormal balances is easily defined using a sequential binary partition (SBP), resulting in ilr-coordinates. The corresponding orthonormal basis of the simplex is made of the corresponding balancing elements (Egozcue and Pawlowsky-Glahn, 2005). It is frequently straightforward to interpret, specially when based on expert knowledge.

Egozcue, J.J., Tolosana-Delgado, R. and Ortego, M.I. (eds.)
ISBN: 978-84-87867-76-7

A set of orthonormal balances, with properties similar to those of CoDa-PCs, appears thus to be an exploratory tool more intuitive than CoDa-PCA, and at the same time simpler than the subjective selection of a SBP. Given a compositional centered sample, we define the first principal balance as the balance which maximizes the explained sample variance. Subsequent principal balances, being orthogonal to the preceding ones, also maximize the explained remaining variance.

It is important to note that principal balances not necessarily will coincide in order with a SBP, but we assume that a basis made of principal balances will be associated with a SBP, possibly in a different order. This question requires further study and remains open.

Up to our knowledge, computing principal balances requires an exhaustive search along all possible sets of orthogonal balances. This computation appears as unaffordable when the number of parts is large. This number increases dramatically with D, as shown in Table 1. Suboptimal approaches are

D	3	4	5	6	7	10	12
nr.	3	18	180	2700	56700	2.57×10^9	9.34×10^{12}

Table 1: Number of different orthonormal basis made of balancing elements for different values of D. See appendix A for a proof.

then appropriate to simplify search algorithms. There are several criteria to approach the properties of the CoDa-PCs. For instance, CoDa-PCs can be taken as a starting point and balancing elements for principal balances are then selected minimizing the geometric angle to one PD, or maximizing the sample correlation with one PC (statistical angle). Other possibilities try to simplify the exhaustive search constraining it to a hierarchy of balances. The ideas of cluster analysis of some set of log-ratios also provide efficient but suboptimal algorithms to approach principal balances.

Here we explore and compare three strategies. (1) Minimize the geometric angle of a balancing element corresponding to a first order partition of the composition to one PD. The same strategy is applied to the subcompositions obtained in previous steps until a full SBP is obtained. (2) Use clustering algorithms based on the variation matrix. This choice for hierarchical clustering of components has the advantage of being subcompositionally coherent, a property not shared by other classical choices (e.g. cosine metric). The Ward clustering method uses as distance between two groups of parts the variance of their balance, thus offering an appealing connection with principal balances. (3) Look for the first order partition which balancing element maximises the explained sample variance, taking as a starting point the first PD in a PCA and then using the signs of the loadings to define the initial partition.

1 Theory

1.1 Basic concepts

Balances (Egozcue and Pawlowsky-Glahn, 2005) based on a SBP are a tool to build orthonormal basis, enhancing interpretation when based on expert knowledge. Nevertheless, frequently the question for a blind construction of such a basis optimizing some criterion has been risen, question motivated by the easy way of constructing PCs. The following approach is based on it as a motivating rule.

Definition 1.1 (Principal balances) *Given an n-sample of a D-part random composition, the set of Principal Balances (PB) is a set of $D-1$ balances satisfying the following conditions:*

- *Each sample PB is obtained as the projection of a sample composition on a unitary composition or balancing element associated to the PB;*
- *The first PB is the balance with maximum sample variance;*
- *The i-th PB has maximum variance conditional to its balancing element being orthogonal to the previous $1st$, $2nd$, ..., $(i-1)th$ balancing elements.*

Therefore, PBs are orthonormal coordinates with respect to a basis of balancing elements selected so that they maximize the explained variance of a data set in decreasing order. The total variance of a compositional sample is decomposed into variances of orthonormal balances. Therefore, the sample variance of a balance can be interpreted as the part of the total variance explained by the balance.

By analogy to PCA, Principal Balances Analysis (PBA) can be defined as an orthogonal linear transformation restricted to transformations within the set of possible basis made up of balances that transforms the data into a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate (called the first principal balance), the second greatest variance on the second coordinate, and so on.

1.2 Construction of principal balances

By definition, PBs reproduce the properties of principal components as close as possible. The construction calls for an exhaustive search among all possible bases of balances, i.e. an exhaustive search of all possible sequential binary partitions of a set of compositional parts into groups of parts, a problem which is analogous to the problem of optimization in *cluster analysis*. As shown in the appendix A, such an exhaustive search is unaffordable for even a relatively small number of parts in a composition. Therefore, the approach presented looks for suboptimal strategies.

Criteria to approach PCA using balances can be based on different properties. From definition of PB, orthogonality and maximization of variance explained by balances seems to be the main criterion. Another criterion would be proximity between PCs and uncorrelated balances. Also geometrical angles (in the simplex) of PD and balancing elements can be used as a criterion. As a first approach, three suboptimal procedures for construction of PBs are analysed below.

1.2.1 Angular proximity to principal components (AP)

The definition of PBs is inspired in CoDa-PCs, and therefore it seems reasonable to use the information obtained from PCA. A possible strategy consists thus in computing CoDa-PCs for a given realisation of a D-part random composition $\mathbf{X}$ and iterating the following steps from 1 to $D-1$:

1. Look for a binary partition of the whole composition $\mathbf{X}$ such that the associated balancing element minimises the geometric angle with one (not necessarily the first) of the PDs associated with the CoDa-PCs.

2. Eliminate the approximated PD from the set of directions.

3. Take each of the groups defined in the previous steps separately and look for a binary partition of each such that the associated balancing element minimizes the angle with one of the remaining PDs associated with the CoDa-PCs.

4. Repeat steps 2 and 3 until a complete SBP is obtained.

1.2.2 Hierarchical clustering of components (HC)

Another way of looking for a SBP is as the result of a hierarchical cluster analysis of components of the random composition $\mathbf{X}$. The agglomeration criterion to be used is the variance of the balance between the two groups (Eq. 1). This criterion can be used even when the groups have a single component each, and can be interpreted as a *measure of proportionality* between components. This offers an appealing connection with principal balances. It can be shown that it is equivalent to standard clustering techniques with the entry in the variation matrix, i.e. $d^2(i,j) = Var[\log(x_i/x_j)] = t_{ij}$, as distance criterion between individual components, and the Ward clustering method (Everitt, 1993) as agglomeration criterion. The choice is justified because the variation matrix does actually behave as one intuitively expects for a dissimilarity matrix between components. First, its elements $t_{ij} > 0$, and $t_{ij} = 0$ if $i = j$ or components i and j are perfectly proportional. Second, the larger t_{ij} the more unreliable a relationship between the two variables appears and they likely belong to different groups. It is worth mentioning that t_{ij} between components i and j depends exclusively on these two

components. This is due to the subcompositional coherence of the variation matrix. Therefore, this approach appears as a natural choice for hierarchical clustering of components, given that none of the classical choices (e.g. cosine metric) is subcompositionally coherent.

1.2.3 Maximum explained variance hierarchical balances (MV)

The approach assumes that the first principal balance corresponds to a binary partition of the whole composition into two groups of parts. This assumption can fail because a simpler log-contrast can explain larger variances. Computation of the optimum partition is heuristically based on the result of CoDa-PCA: parts with positive loadings are initially assigned to the + group and parts with negative loading are initially assigned to the − group to define a balance between these two groups of parts. Optimality of the explained variance of the corresponding balance is then checked moving one part from the + group to the − group. This procedure is computationally quite efficient because the number of checks for this first balance is only of the order of involved parts, D. Once the first balance has been determined, data are projected into the subcomposition formed by the larger number of components and a new PCA is performed. Again the method assumes that the maximum explained variance is attained for a balance corresponding to a partition of the selected subcomposition and the procedure is applied again in the subsequent steps.

This approach generates a hierarchy of balances following a SBP by construction. Despite the hierarchical character of the SBP generated, explained variances for the sequence of balances can be non decreasing although this seldom occurs.

2 Application

The three methods presented (AP, HC, and MV) have been applied to the characterization of a data set of geochemical composition of glacial sediment from a granodioritic-gneissic source rock (Aar Massif, Switzerland; von Eynatten and Tolosana-Delgado, 2008). This data set contains measurements of 10 major oxides and 16 trace elements of 87 samples of different grain sizes. From these elements, only 21 elements were kept (those without zeros). A CoDa-PCA yields loadings and contributions to variance displayed respectively in table 2 (labeled "comp") and figure 1 (left: circled black solid line). These results serve as a reference to compare the approximative methods proposed (AP, HC,

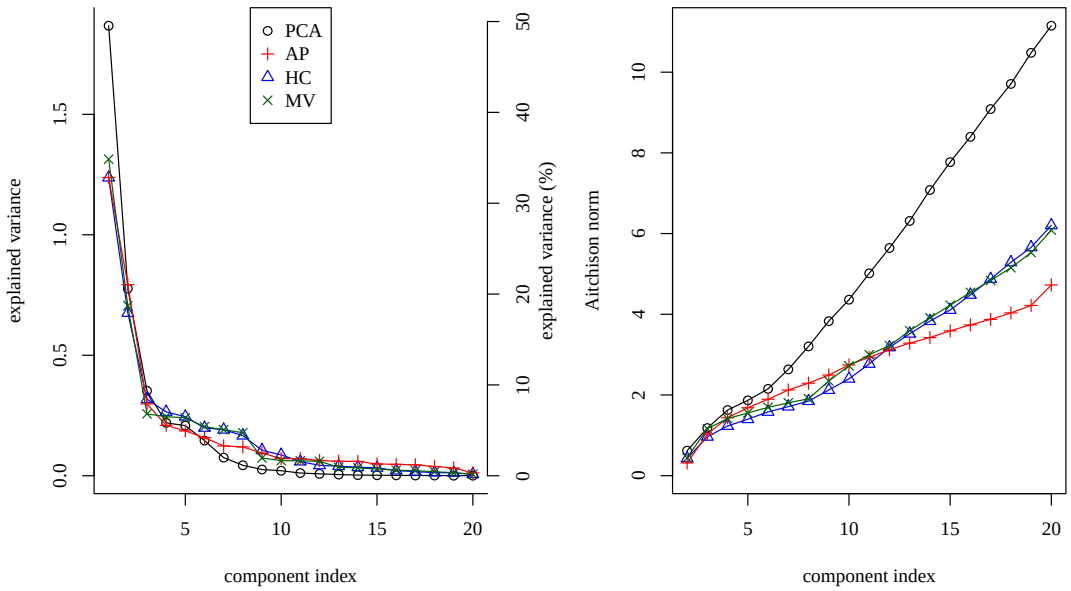

Figure 1: Left: explained variance for each PC/PB ordered in decreasing order. Right: Aitchison norm of the vector of ordered explained variances taken as a composition

and MV). In Table 2 it can be realized that loadings corresponding to a balance (sub-tables AP, HC, MV) only admit two different values as loadings different from zero, in contrast with PCs, in which

the only condition is that the loadings add to zero and the sum of their squares add to one. Therefore, depending on the data set, the approach of PCs based on PBs can be poor. The angular proximity method (AP) produces principal balances approaching the principal components by minimal angle. Compare AP.8 with Comp.1 (large positive Zr and Nd vs large negative loadings in mafic elements), or AP.9 with Comp.5 (a balance between Nd and Zr), or AP.1 with Comp.2 (essentially mafic vs. felsic elements) in table 2. Note that the variances of these principal balances are not necessarily ordered hierarchically, as can be seen in Fig. 2. Hierarchical order means that, starting from a simple balance

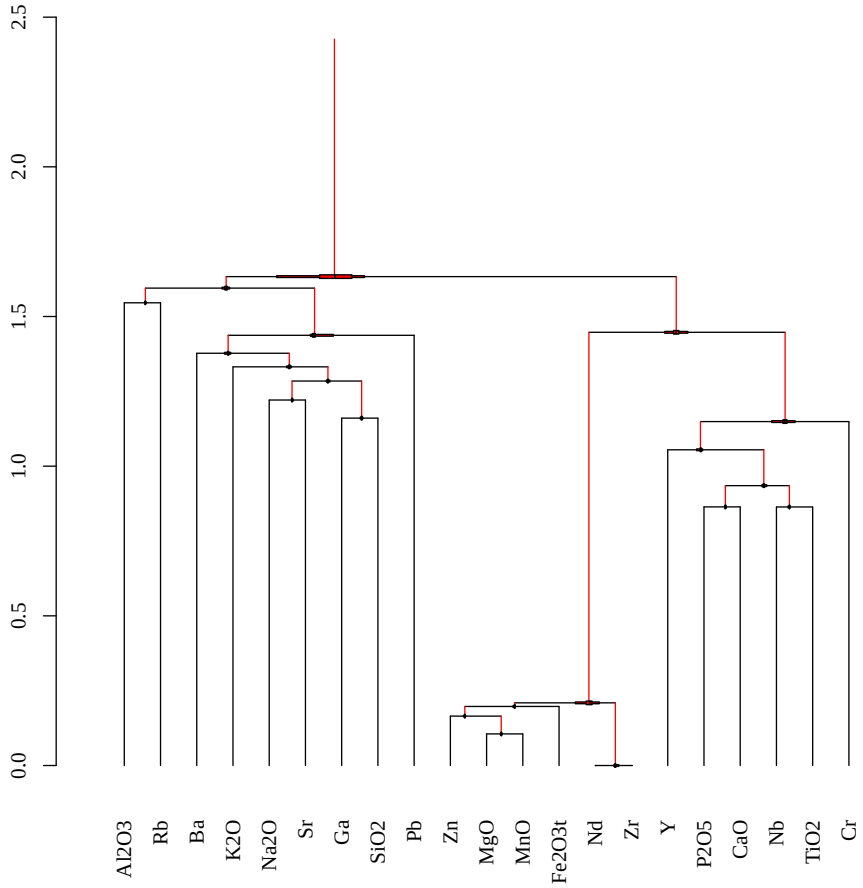

Figure 2: CoDa-dendrogram for principal balances approached by AP method.

(few parts included in the two compared groups) and plotted in the lower part of a CoDa-dendrogram, and moving upwards to more complex balances, an increasing variance is expected (longer vertical bars). This is not the case in Fig. 2. The hierarchical order can be observed in the CoDa-dendrograms of Figures 3 and 4. The hierarchical clustering of components (HC) yields, by construction, a series of balances with increasing variance (upwards in the dendrogram, Fig. 3). Hence, the largest variance balances tend to be the last ones. It is not surprising that HC.20 and Comp.1 have large correlations (mafic vs. felsic major oxides, but a poor structure in trace elements), as do HC.19 and Comp.2 (traces in heavy minerals vs. felsic components), or Comp.4 with HC.15 (ultramafic Cr vs. mafic elements).

Comparing the three CoDa-dendrograms, Figures 2, 3 and 4, the only clear difference is that hierarchical order is broken using AP, whereas HC and MV produce well organized CoDa-dendrograms from the hierachical point of view. While the aspect of hierarchically ordered CoDa-dendrograms seems to be more comfortable to the user, they do not allow to detect a large variance in balances involving only a few parts which may facilitate a straightforward interpretation. A detailed inspection of the three CoDa-dendrograms reveals that the differences between the three proposed methods are not as dramatic as they appear at the first glance. For instance, the first estimated principal balances using AP, HC, MV, separate in different groups some common elements. For instance AP.8 (Table 2, Fig. 2), the first PB estimated using AP, assigns MgO, MnO, Fe2O3t, Zn to one group that is compared to Zr and Nd. The first principal balances estimated using HC and MV (HC.20, MV.1) situate the

Table 2: Loadings of the 6 PCs ("Comp") and of the 6 PBs obtained using the AP, HC, and MV methods, with the largest variance. Explained variance and % contribution to the total variance are also reported.

	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5	Comp.6	AP.8	AP.1	AP.3	AP.9	AP.2	AP.14
SiO2	0.288	-0.189	0.043	-0.201	0.132	0.166	0	-0.252	0	0	0	-0.154
TiO2	-0.124	0.206	-0.094	0.021	-0.145	0.026	0	0.189	-0.183	0	0.289	0
Al2O3	0.103	-0.258	-0.03	0.074	-0.02	-0.006	0	-0.252	0	0	0	0
MnO	-0.206	0.055	0.078	-0.266	-0.005	0.039	-0.289	0.189	0	0	-0.289	0
MgO	-0.417	0.025	-0.116	-0.332	-0.05	0.194	-0.289	0.189	0	0	-0.289	0
CaO	0.069	0.031	-0.388	0.121	-0.118	-0.041	0	0.189	-0.183	0	0.289	0
Na2O	0.292	-0.303	-0.155	0.094	-0.033	0.106	0	-0.252	0	0	0	-0.154
K2O	0.085	-0.316	0.154	-0.022	0.035	-0.12	0	-0.252	0	0	0	-0.154
P2O5	-0.099	0.295	-0.453	0.25	-0.016	-0.343	0	0.189	-0.183	0	0.289	0
Fe2O3t	-0.259	0.013	-0.025	-0.248	-0.033	0.049	-0.289	0.189	0	0	-0.289	0
Ba	0.004	-0.245	-0.256	-0.037	-0.018	0.075	0	-0.252	0	0	0	-0.154
Cr	-0.263	0.042	0.193	0.654	0.308	0.546	0	0.189	0.913	0	0.289	0
Ga	-0.017	-0.187	0.053	0.108	-0.044	-0.171	0	-0.252	0	0	0	-0.154
Nb	0.071	0.251	0.268	0.126	0.01	-0.213	0	0.189	-0.183	0	0.289	0
Pb	-0.069	-0.106	0.272	-0.019	0.075	-0.146	0	-0.252	0	0	0	0.926
Rb	-0.027	-0.243	0.308	-0.047	0.056	-0.262	0	-0.252	0	0	0	0
Sr	0.097	-0.112	-0.326	0.059	-0.054	0.05	0	-0.252	0	0	0	-0.154
Y	0.133	0.259	0.201	0.19	-0.029	-0.397	0	0.189	-0.183	0	0.289	0
Zn	-0.371	0.074	0.095	-0.153	-0.011	-0.053	-0.289	0.189	0	0	-0.289	0
Zr	0.383	0.409	-0.065	-0.316	0.625	0.129	0.577	0.189	0	0.707	-0.289	0
Nd	0.33	0.299	0.243	-0.055	-0.664	0.371	0.577	0.189	0	-0.707	-0.289	0
var.	1.868	0.777	0.353	0.22	0.208	0.146	1.238	0.793	0.298	0.209	0.186	0.158
% var.	49.54	20.603	9.353	5.826	5.517	3.871	32.835	21.042	7.912	5.554	4.935	4.182
	HC.20	HC.19	HC.18	HC.17	HC.16	HC.15	MV.1	MV.8	MV.9	MV.18	MV.2	MV.3
SiO2	0.138	0.156	0.228	-0.598	0	0	-0.154	0.169	-0.258	0	0	0
TiO2	-0.345	0	0	0	0	0.183	0.309	0	0	0	-0.598	0
Al2O3	0.138	0.156	0.228	0.239	0	0	-0.154	0.169	-0.258	0	0	0
MnO	-0.345	0	0	0	0	0.183	0.309	0	0	0	0.239	0.224
MgO	-0.345	0	0	0	0	0.183	0.309	0	0	0	0.239	0.224
CaO	0.138	0.156	-0.399	0	0	0	-0.154	0.169	-0.258	0	0	0
Na2O	0.138	0.156	0.228	-0.598	0	0	-0.154	0.169	-0.258	0	0	0
K2O	0.138	0.156	0.228	0.239	0	0	-0.154	0.169	0.387	0	0	0
P2O5	0.138	0.156	-0.399	0	0	0	0.309	0	0	0	-0.598	0
Fe2O3t	-0.345	0	0	0	0	0.183	0.309	0	0	0	0.239	0.224
Ba	0.138	0.156	-0.399	0	0	0	-0.154	0.169	-0.258	0	0	0
Cr	-0.345	0	0	0	0	-0.913	0.309	0	0	0	0.239	-0.894
Ga	0.138	0.156	0.228	0.239	0	0	-0.154	0.169	0.387	0	0	0
Nb	0.138	-0.428	0	0	0.289	0	-0.154	-0.423	0	0.289	0	0
Pb	0.138	0.156	0.228	0.239	0	0	-0.154	0.169	0.387	0	0	0
Rb	0.138	0.156	0.228	0.239	0	0	-0.154	0.169	0.387	0	0	0
Sr	0.138	0.156	-0.399	0	0	0	-0.154	0.169	-0.258	0	0	0
Y	0.138	-0.428	0	0	0.289	0	-0.154	-0.423	0	0.289	0	0
Zn	-0.345	0	0	0	0	0.183	0.309	0	0	0	0.239	0.224
Zr	0.138	-0.428	0	0	-0.866	0	-0.154	-0.423	0	-0.866	0	0
Nd	0.138	-0.428	0	0	0.289	0	-0.154	-0.423	0	0.289	0	0
var.	1.237	0.675	0.318	0.265	0.245	0.199	1.314	0.706	0.257	0.245	0.24	0.204
% var. %	32.818	17.918	8.428	7.029	6.505	5.271	34.844	18.722	6.812	6.505	6.379	5.403

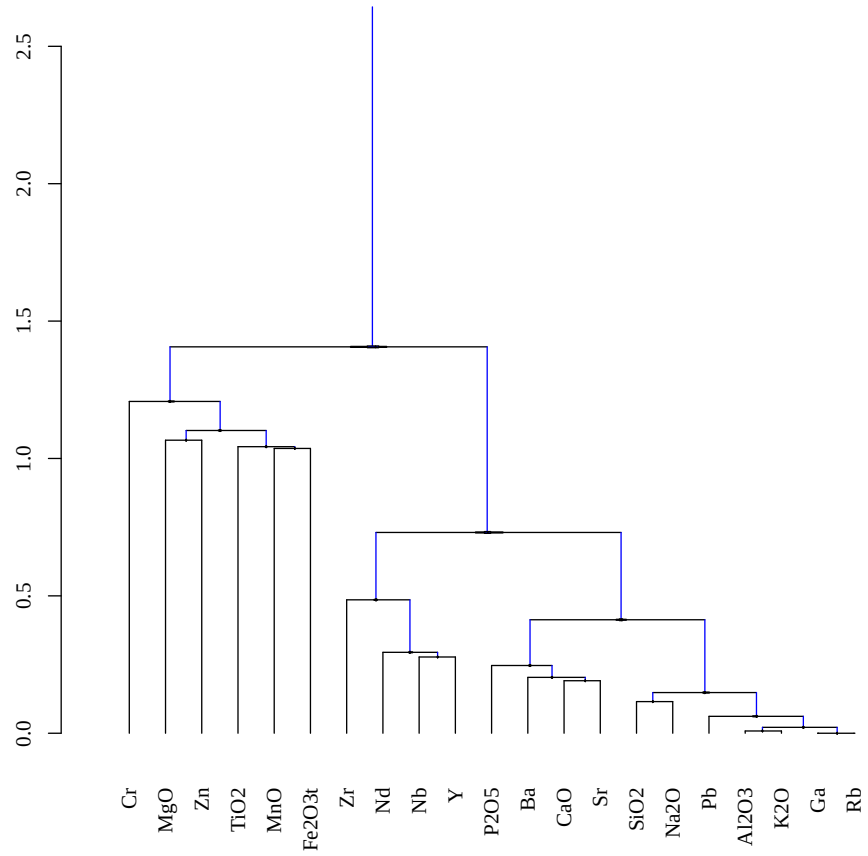

Figure 3: CoDa-dendrogram for principal balances approached by HC method.

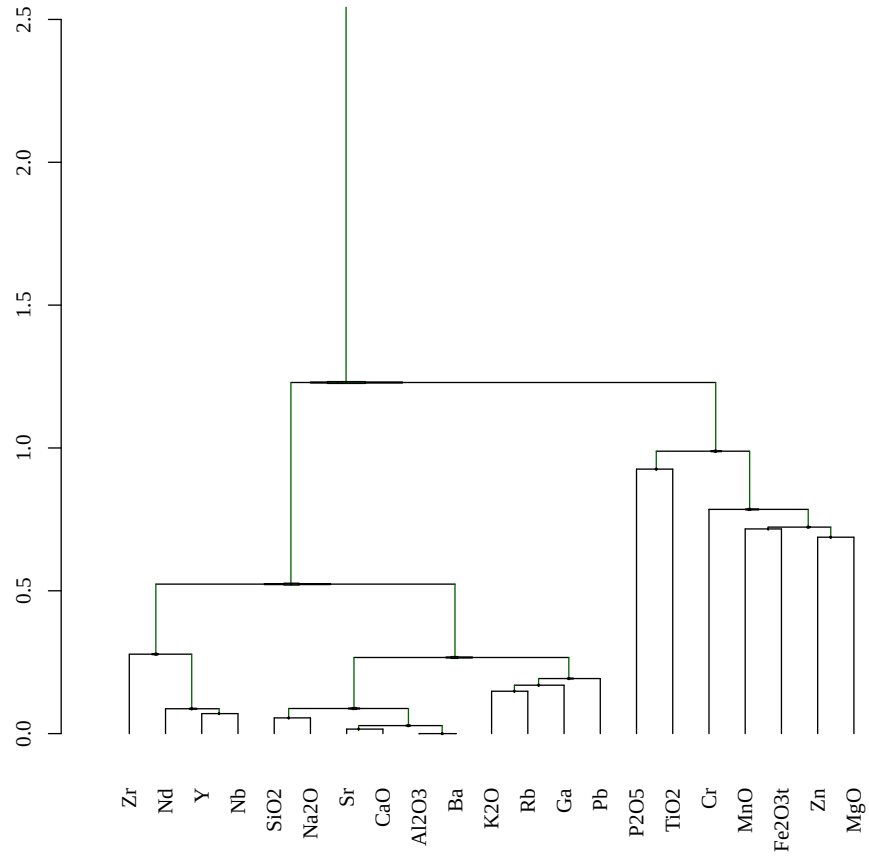

Figure 4: CoDa-dendrogram for principal balances approached by MV method.

Table 3: Sample correlation matrix for the 6 first AP-principal balances (PB). Numbering in labels AP means the order in which they have been identified.

	AP.8	AP.1	AP.3	AP.9	AP.2	AP.14
AP.8	1	-0.241	-0.595	0.120	0.301	-0.479
AP.1	-0.241	1	0.121	0.042	-0.059	0.473
AP.3	-0.595	0.121	1	-0.079	-0.067	0.411
AP.9	0.120	0.042	-0.079	1	0.032	-0.083
AP.2	0.301	-0.059	-0.067	0.032	1	-0.239
AP.14	-0.479	0.473	0.411	-0.083	-0.239	1

Table 4: Sample correlation matrix HC-principal balances (PB). Numbering in labels HC means the order in which they have been identified.

	HC.20	HC.19	HC.18	HC.17	HC.16	HC.15
HC.20	1	-0.302	0.334	-0.816	-0.413	-0.035
HC.19	-0.302	1	0.099	0.302	0.321	0.004
HC.18	0.334	0.099	1	-0.106	0.106	-0.153
HC.17	-0.816	0.302	-0.106	1	0.548	-0.012
HC.16	-0.413	0.321	0.106	0.548	1	-0.056
HC.15	-0.035	0.004	-0.153	-0.012	-0.056	1

same elements in different groups although mixed with other parts. Although one can follow these similarities through most of the estimated principal balances for the three methods, the comparison procedure is tedious and difficult. Therefore, comparison of methods require a measure of effectiveness approaching principal balances. A first idea comes from the property of PCs which are uncorrelated random variables by construction. As PBs approach PCs, correlations between PBs are expected to be small. Tables 3, 4, and 5 show the obtained correlations between the 6 first estimated principal balances for methods AP, HC, and MV respectively. They show that off-diagonal correlations are not high, thus behaving as expected, but the difficulty of comparing these three tables is still present. Another possibility of comparison is to compute the geometrical angles between the principal directions PD and the balancing elements computed. But the resulting angle-tables, not shown here, are also difficult to compare.

A measure of effectiveness approaching PCs follows. Consider the vector containing the variances of estimated PCs or PBs ordered from maximum to minimum variance, i.e. the first component corresponds to the first PC or PB, the second component to the second PC or PB, etc. These vectors can be considered compositional because the total variance is not relevant to this analysis. Therefore, the Aitchison norm (Pawlowsky-Glahn and Egozcue, 2001) of the variance vector measures the concentration of variance in the first (ordered) components. On the other hand, the Aitchison norm of sub-vectors (or subcompositions) including the first variance up to a given (increasing) number of parts provides a sequence of increasing Aitchison norms. This measures the effectiveness approaching the first PCs. Figure 1 (right), shows the Aitchison norm of the vector of variances for PCs (black), AP (red), HC (blue), MV(green) for different sizes of the subcomposition. PC appear, as expected, above the other curves. The curve of AP is the higher one for subcompositions containing up to 10 variances; for larger subcompositions it becomes the lowest one, thus reflecting the price to be payed for an easy interpretation of the first estimated principal balance. The two hierarchical methods HC and MV produce Aitchison norm curves that almost overlap. These are observations for the particular data set used and the shape and order of the curves will not necessarily be the same for other data and/or number of components. In this particular example the use of AP-principal balances would be recommended to a user interested in a straightforward interpretation of the very first principal balances. If interest is centered in how major and trace elements are associated in the groups of the principal balances, HC and MV are more appropriate. If a blind selection of ilr-coordinates is required, either for exploratory analysis or for modelling, then PCA maintains all its virtues.

Table 5: Sample correlation matrix MV-principal balances (PB). Numbering in labels MV means the order in which they have been identified.

	MV.1	MV.8	MV.9	MV.18	MV.2	MV.3
MV.1	1	0.181	0.499	0.377	0.489	0.130
MV.8	0.181	1	0.180	0.309	0.535	0.080
MV.9	0.499	0.180	1	0.500	0.690	0.022
MV.18	0.377	0.309	0.500	1	0.344	-0.018
MV.2	0.489	0.535	0.690	0.344	1	0.034
MV.3	0.130	0.080	0.022	-0.018	0.034	1

3 Conclusion

Principal component analysis (or singular value decomposition) of a compositional data set (clr-transformed and centered) has a number of appealing properties: maximum explained variance of the sequence of principal components, uncorrelated components, orthogonal geometric axes. Due to these properties CoDa-PCA is one of the main tools for exploratory analysis and modelling of compositional data. The main shortcoming of CoDa-PCs is the difficulty in interpreting the resulting coordinates. Biplots provide a helpful tool for a reduced number of significative PCs but may fail for an increasing number of them.

Balances are log-contrasts resulting from a log-ratio of two geometric means of two groups of parts and its interpretation may be considerably simpler than the interpretation of a PC. In the present contribution, the idea of approaching CoDa-PCs using a set of balances, called principal balances (PB), has been formalized. Computation of PBs may require an exhaustive, unaffordable search over the possible set of orthonormal balances for a moderate number of parts. To avoid it, suboptimal but feasible procedures to search for principal balances are required. Three methods, based on different principles, are presented. A measure of effectiveness based on the Aitchison norm is proposed.

Acknowledgements

This research has been supported by the Spanish Ministry of Science and Innovation under project "CODA-RSS" Ref. MTM2009-13272; and by the *Agència de Gestió d'Ajuts Universitaris i de Recerca* of the *Generalitat de Catalunya* under project Ref: 2009SGR424.

References

Aitchison, J. and M. Greenacre (2002). Biplots for compositional data. *Applied Statistics 51*(4), 375–392.

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

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

Everitt, B. S. (1993). *Cluster Analysis*. Edward Arnold, Cambridge (UK). 170 p.

Greenacre, M. (2011). Compositional data and correspondence analysis. In V. Pawlowsky-Glahn and A. Buccianti (Eds.), *Compositional Data Analysis: Theory and Applications.*, John Wiley & Sons. (in press).

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal Educ. Psychology 24*, 417–441 + 498–520.

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

von Eynatten, H. and R. Tolosana-Delgado (2008). A log-linear model of grain size influence on the geochemistry of sediments. In J. Martín-Fernández and J. Daunis-i Estadella (Eds.), *Compositional Data Analysis Workshop – CoDaWork'08, Proceedings*. Universitat de Girona, http://ima.udg.es/Activitats/CoDaWork08/.

A Number of orthonormal basis made of balancing elements

In a D-part simplex an orthonormal basis made of balancing elements is obtained performing a sequential binary partition (SBP) of the compositional vector (Egozcue and Pawlowsky-Glahn, 2005). The number of possible SBP for a fixed number of parts D can be computed as follows. Consider the D groups of parts made of a single element. In a first step, join two of the available groups. The number of ways of doing this, is combining two groups at a time from the total of D. In subsequent steps, with k available groups of one or more parts, two groups are joined. The number of possible unions is combining two elements from a total of k. We have $D-2$ steps to get a single group of D parts. Therefore, the number of possibilities is

$$N = \left(\begin{array}{c} D \\ 2 \end{array}\right)\left(\begin{array}{c} D-1 \\ 2 \end{array}\right)\left(\begin{array}{c} D-2 \\ 2 \end{array}\right)\cdots\left(\begin{array}{c} 2 \\ 2 \end{array}\right) = \frac{D!(D-1)!}{2^{D-1}} \ .$$

This number increases dramatically with D. For example, for $D = 3, 7, 10, 12$, $N = 3, 56700, 2.6\times10^9$, 9.3×10^{12}, respectively.