Mathematical Geology, Vol. 37, No. 7, October 2005 (© 2005)
DOI: 10.1007/s11004-005-7381-9

Groups of Parts and Their Balances in Compositional Data Analysis[1]

J. J. Egozcue[2] and V. Pawlowsky-Glahn[3]

Amalgamation of parts of a composition has been extensively used as a technique of analysis to achieve reduced dimension, as was discussed during the CoDaWork'03 meeting (Girona, Spain, 2003). It was shown to be a non-linear operation in the simplex that does not preserve distances under perturbation. The discussion motivated the introduction in the present paper of concepts such as group of parts, balance between groups, and sequential binary partition, which are intended to provide tools of compositional data analysis for dimension reduction. Key concepts underlying this development are the established tools of subcomposition, coordinates in an orthogonal basis of the simplex, balancing element and, in general, the Aitchison geometry in the simplex. Main new results are: a method to analyze grouped parts of a compositional vector through the adequate coordinates in an ad hoc *orthonormal basis; and the study of balances of groups of parts (inter-group analysis) as an orthogonal projection similar to that used in standard subcompositional analysis (intra-group analysis). A simulated example compares results when testing equal centers of two populations using amalgamated parts and balances; it shows that, in certain circumstances, results from both analysis can disagree.*

KEY WORDS: simplex, Euclidean geometry, log-ratio analysis, orthogonal projection, subcomposition, amalgamation.

INTRODUCTION

By convention, n-part compositional data are vectors of n strictly positive real components,$[x_1, x_2, \ldots, x_n]$, such that $x_1 + x_2 + \cdots + x_n = \kappa > 0$. κ is generally 100 (percentages) or 1 (proportions). Compositional data analysis, as introduced by Aitchison (1982, 2003a), is based on ratios of parts, and this is the essence of a deeper understanding of the nature of these type of data. A mathematical justification for the central role of ratios can be found in Barceló-Vidal (2000) and Barceló-Vidal, Martín-Fernández, and Pawlowsky-Glahn (2001). Once the

[1]Received 14 March 2004; accepted 28 February 2005.

[2]Departamento de Matemàtica Aplicada III, Universitat Politècnica de Catalunya (Campus Nord), Jordi Girona Salgado 1-3, Mod. C2, Barcelona, Spain; e-mail: juan.jose.egozcue@upc.edu

[3]Departamento d'Informàtica i Matemàtica Aplicada, Universitat de Girona, Girona, Spain; e-mail: vera.pawlowsky@udg.es

0882-8121/05/1000-0795/1 © 2005 International Association for Mathematical Geology

relevance of ratios is accepted, the Aitchison geometry of the simplex is a natural framework for statistical analysis of compositional data (Pawlowsky-Glahn and Egozcue, 2001). It implies, e.g. that statistical analysis of a single part is meaningless, and that the value of the closure constant κ is irrelevant, as already stated in Aitchison (1986).

Aitchison geometry in the simplex is based on specific operations such as perturbation and powering—which take the place of the ordinary sum of vectors and multiplication by a real constant—and definitions of Aitchison distance, norm and inner product. They induce a finite-dimensional Hilbert-space structure in the simplex. The appendix gives a summary of the essential definitions, as well as of the basic properties of the Aitchison geometry of the simplex. For further details see Aitchison (2003a), Aitchison and others (2002), Billheimer, Guttorp, and Fagan (2001), Egozcue and others (2003), and Pawlowsky-Glahn and Egozcue (2001).

Statistical analysis of compositional data occasionally requires an interpretation of results in terms of ratios and log-ratios, which are more difficult to interpret than real vectors in standard multivariate analysis. In order to simplify the analysis, parts can be ordered in such a way that they can be grouped into two or more subsets, which are interpretable in some way. For instance, chemical compositions may include a group of anions and another group of cations; rocks may be described by proportions of minerals grouped into silicates and other components or, alternatively, grouped into trace-elements and major-elements; in political surveying, parties may be grouped as left-wing, right-wing and other parties; in analysis of abundance of animal species, one may consider mammals and *other species* and, afterwards, subdivide mammals into carnivores and herbivores. In these and other similar situations, the analyst may be interested in studying two features of the sample compositions: (a) the relationship or *balance* between these groups of parts or *inter-group* analysis; and (b) the behavior of parts within a group or *intra-group* analysis.

Groups of parts can be viewed either as a subcomposition or as a group inside the whole composition. Subcompositional analysis is intended to deal with parts within the group and relations with respect to other groups or parts are obviated. The concept of subcomposition was established right from the beginning of compositional data analysis. To study such relationships between subcompositions, Aitchison introduced the concept of amalgamation or addition of several parts of a composition and the concept of partition, which is a set of mutually exclusive and exhaustive subcompositions together with the amalgamations of each one of them (Aitchison, 2003a, p. 36–42). Once the partition is obtained, the behavior of the ratios of amalgamations can be studied. At the *CoDaWork'03* meeting (Girona, Spain, 2003) several contributors, including the present authors, used amalgamation as a way of reducing the dimension of their respective problems. Subsequent discussion at the meeting pointed out that the non-linear character

of amalgamation with respect to the Aitchison geometry in the simplex led to problems and prompted the present paper.

Amalgamation of parts inside each group or subcomposition may be justified when the amalgamation has a clear, well-defined sense and interest is centered in studying the variability of, let us say, $(x_1 + \cdots + x_r)/(x_{r+1} + \cdots + x_n)$, or its logarithm. Now, when the analysis of amalgamated parts is performed simultaneously with some compositional analysis of the original—non-amalgamated—parts, one would expect both analyses to be compatible and their interpretation to be coherent. Unfortunately, these two kinds of analysis (original parts and amalgamated parts) can be—and frequently are—incompatible.

For instance, let be $\mathbf{x}_j = [x_{j1}, x_{j2}, x_{j3}]$, $j = 1, 2, \ldots, J$, a sample of a three-part random composition. The center of this sample is the geometric mean

$$\operatorname{cen}(\mathbf{x}) = \mathcal{C}\left[\left(\prod_{j=1}^{J} x_{j1}\right)^{1/J}, \left(\prod_{j=1}^{J} x_{j2}\right)^{1/J}, \left(\prod_{j=1}^{J} x_{j3}\right)^{1/J}\right],$$

where $\mathcal{C}[\cdot]$ stands for closure to κ. Let it be amalgamated into two parts,

$$\mathcal{C}\left[\left(\prod_{j=1}^{J} x_{j1}\right)^{1/J} + \left(\prod_{j=1}^{J} x_{j2}\right)^{1/J}, \left(\prod_{j=1}^{J} x_{j3}\right)^{1/J}\right]. \tag{1}$$

One would expect that the same center would be obtained by amalgamating the sample, i.e. $\mathcal{C}[x_{j1} + x_{j2}, x_{j3}]$, $j = 1, 2, \ldots, J$, and then, finding the center. However, the two-part center obtained is

$$\mathcal{C}\left[\left(\prod_{j=1}^{J} (x_{j1} + x_{j2})\right)^{1/J}, \left(\prod_{j=1}^{J} x_{j3}\right)^{1/J}\right],$$

which obviously differs from Equation (1). This disappointing behavior of amalgamation with respect to the center of the sample is only a partial aspect of a more general situation: amalgamation does not preserve Aitchison distances in the simplex and distances of amalgamated compositions have a complicated, non-monotonic behavior with respect to original distances.

For illustration, we consider three-part compositions and we select as a reference composition the neutral composition $\mathbf{n} = [1/3, 1/3, 1/3]$. Figure 1(a) shows the Aitchison distance from each composition, $\mathbf{x} = [x_1, x_2, x_3]$, to the reference. Compositions are represented in the plane of two orthogonal coordinates (Egozcue

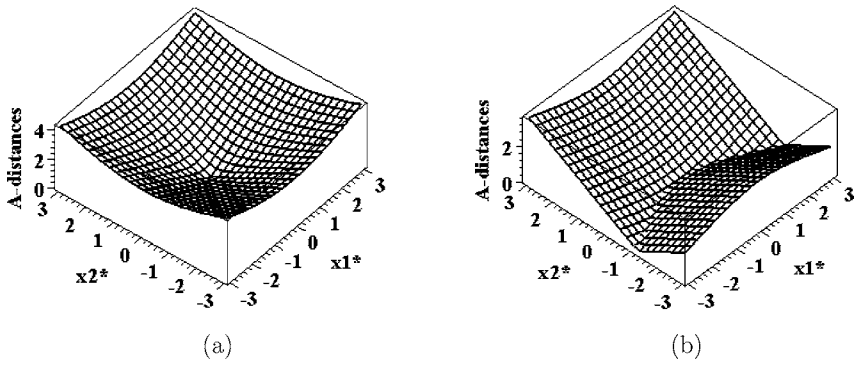

Figure 1. (a) Aitchison distance in $\mathcal{S}^3$, from $[1/3, 1/3, 1/3]$ to compositions $[x_1, x_2, x_3]$ whose coordinates are (x_1^*, x_2^*). It is a conic surface. (b) Aitchison distance in $\mathcal{S}^2$, from $[2/3, 1/3]$ to $[x_1 + x_2, x_3]$. The surface is not monotonic.

and others, 2003),

$$x_1^* = \frac{1}{\sqrt{2}} \ln \frac{x_1}{x_2}, \quad x_2^* = \frac{1}{\sqrt{6}} \ln \frac{x_1 x_2}{x_3^2}.$$

In this representation, the Aitchison distance $d_a(\mathbf{x}, \mathbf{n})$ is the ordinary Euclidean distance; it appears as an inverted, circular cone, which has its vertex at $(x_1^*, x_2^*) = (0, 0)$. Amalgamation of parts x_1 and x_2 of each composition $\mathbf{x}$ leads to the two-part composition $[x_1 + x_2, x_3]$. The Aitchison distance from the amalgamated composition to the amalgamated reference $[2/3, 1/3]$ is shown in Figure 1(b). One would expect that, when following lines in the coordinate plane for which the distance in $\mathcal{S}^3$ increases, increasing, or at least non-decreasing, distances for amalgamated compositions would be obtained. However, there are lines in that plane for which this does not hold. For instance, Figure 2 represents two cross-sections of surfaces of Figure 1(a) and (b) for constant coordinate x_2^*. One corresponds to $x_2^* = +1$, the other to $x_2^* = -1$. Distances in the three-part simplex (full-line parabola) are equal for both cross-sections and are always larger than distances obtained for amalgamated compositions. For $x_2^* = +1$ (circles) the expected behavior is obtained: when distance in the three-part simplex increases, so does distance of amalgamated compositions. But for $x_2^* = -1$ (triangles) the distance of amalgamated compositions is not monotonic. This behavior equally corresponds to constant x_1^*, i.e. constant ratio x_1/x_2, which means there is no influence of the subcomposition constituted by parts 1 and 2.

The change of monotony of distances when amalgamating parts affects multivariate analysis performed on the amalgamated sample. For instance, the result of a cluster analysis based on Aitchison distances might be completely different

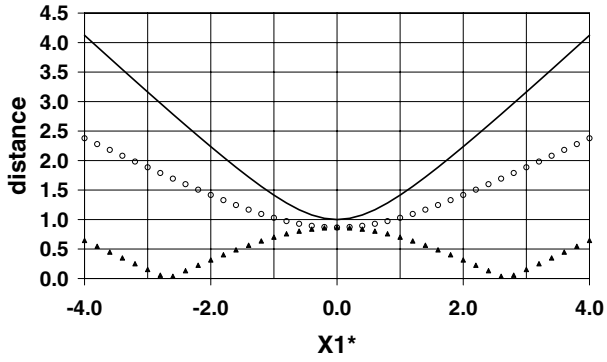

Figure 2. Aitchison distance in $\mathcal{S}^3$, from [1/3, 1/3, 1/3] to compositions $[x_1, x_2, x_3]$ whose coordinates are x_1^* in the horizontal axis and $x_2^* = \pm 1$ (full line), and Aitchison distance in $\mathcal{S}^2$, from [2/3, 1/3] to $[x_1 + x_2, x_3]$ for $x_2^* = +1$ (circles) and for $x_2^* = -1$ (triangles).

when using original parts or amalgamated parts. Also, measures of dispersion are strongly affected, and Aitchison distances between sample points are not invariant even for such simple operations as centering and standardization to unit total variance in the simplex.

An example of non-invariant behavior under perturbation is shown in Table 1. Two three-part compositions at an Aitchison distance of 1.035 are shown. After perturbation by [0.2, 0.7, 0.1] they maintain, as expected, the distance. However, when amalgamation into two parts is carried out, the distance in the two-part simplex is different before and after perturbation.

Despite this undesirable behavior, amalgamation techniques are very frequent because they are an easy and apparently intuitive way of grouping parts, especially to obtain a reduction of dimensionality of compositional data. Therefore, when interest lies in analyzing both the whole composition and lower-dimensional representations, an alternative and coherent way of analyzing grouped parts inside a composition is needed. The main requirement of such an alternative technique is

Table 1. Effect of Perturbation by [0.2, 0.7, 0.1] on Aitchison Distances, d_a, Before (left) and After (right) Amalgamation

	x_1	x_2	x_3	d_a in $\mathcal{S}^3$	$x_1 + x_2$	x_3	d_a in $\mathcal{S}^2$
Unperturbed	0.1	0.8	0.1		0.9	0.1	
	0.3	0.6	0.1	1.035	0.9	0.1	0.000
Perturbed	0.034	0.949	0.017		0.983	0.017	
	0.123	0.857	0.020	1.035	0.980	0.020	0.134

that it should, if possible, be easily interpretable and that it should be compatible with the Aitchison geometry of the simplex, e.g. invariance of distances under perturbation (before and after grouping parts). This is our goal.

The next section introduces the concept of orthonormal basis of a sequential binary partition of a composition and illustrates the procedure with some examples. Linked to this orthonormal basis, the concept of balance between groups of parts is introduced and connected to the subcomposition of a group of parts. A separate section is devoted to study formally some properties of projections, bases, subspaces and balances necessary for a detailed understanding of balances and subcompositions. Finally, the last section describes how to analyze grouped parts in practice, and provides guides to the use of these new techniques. An example on testing centers of two populations for equality compares techniques based on amalgamations and those based on balances.

ORTHONORMAL BASIS OF A SEQUENTIAL BINARY PARTITION

Sequential Binary Partitions

As mentioned in the "Introduction" section, compositional vectors of n parts are frequently partitioned into groups of parts presenting a certain affinity. Any grouping of parts can be viewed as an intermediate state of a sequential binary partition. Initially, we have a compositional vector $[x_1, x_2, \ldots, x_n]$ in the simplex of n parts, $\mathcal{S}^n$. A first-order binary partition consists of making two groups of parts. A second-order partition is obtained by subdividing one of the first-order groups into two groups; the procedure is iterated until all groups contain only a single part. The number of binary divisions of a group to attain the end of the process is $n - 1$. We take into account the order in which the binary partitions have been done and, therefore, call them *sequential binary partitions*. We point out that the concept of partition used by Aitchison (2003a, p. 40), although similar, differs from the present one. Here we use partition in the usual sense: separation of a whole—the parts of a composition—into non-overlapping sets or groups of parts. Aitchison attached the amalgamations of each group of parts to this ordinary sense partition of a vector of parts.

In order to denote a sequential binary partition, we separate the grouped parts by one or more vertical bars. The number of separators between contiguous parts points out the order in which the separation was done: if ν is the number of vertical bars between two parts, the sequential order of the separation is $n - \nu$; the larger ν, the more important the separation and the lower the sequential order. For instance, $[x_1||x_2|x_3]$ means that we first divide $[x_1|x_2, x_3]$ and the first-order partition is made of two groups. Then, we subdivide group $\{x_2, x_3\}$ into two single-part groups. Then, the second-order partition is made of three groups of a

Table 2. Sequential Binary Partition of a Seven-Part Composition

Order													
0	x_1	,	x_2	,	x_3	,	x_4	,	x_5	,	x_6	,	x_7
1	x_1	,	x_2	,	x_3	|	x_4	,	x_5	,	x_6	,	x_7
2	x_1	,	x_2	,	x_3	||	x_4	,	x_5	|	x_6	,	x_7
3	x_1	|	x_2	,	x_3	|||	x_4	,	x_5	||	x_6	,	x_7
4	x_1	||	x_2	,	x_3	||||	x_4	,	x_5	|||	x_6	|	x_7
5	x_1	|||	x_2	|	x_3	|||||	x_4	,	x_5	||||	x_6	||	x_7
6	x_1	||||	x_2	||	x_3	||||||	x_4	|	x_5	|||||	x_6	|||	x_7

single part. The most important division is the first one, which separates part 1 from the other two parts. The sequential order of the separation is then $3 - 1 = 2$.

The following example is used to illustrate the ideas in this section. Let $n = 7$ be the total number of parts and assume they have been ordered conveniently. We start with a partition into two sets, e.g. $[x_1, x_2, x_3|x_4, x_5, x_6, x_7]$. We now proceed to again subdivide the vector of parts by adding a new separator. The original partition is now denoted by two vertical bars, recalling that it was the first separation in the sequential binary partition. An example of a complete sequential binary partition is shown in Table 2.

As the order of parts in compositional data analysis is arbitrary, at least in the mathematical sense, any grouping of parts into $\ell + 1$ sets of parts can be obtained as an ℓ-order partition. In practice, this process requires an intuitive ordering of parts so that the sequential partitions maintain the affinity between contiguous parts and inside the desired groups. Also, in practice, interest may be centered on a limited number of groups, and they can be attained as a partition at a sequential order less than $n - 1$. In those cases, for computational purposes, the sequential binary partition should be arbitrarily completed up to order $n - 1$. Note that the last row in Table 2 , corresponding to order 6, contains all the coded information necessary to reconstruct the whole process of the sequence of binary partitions.

Orthonormal Basis of a Partition

The idea underlying the next development is to associate an orthonormal basis of the n-part simplex, $\mathcal{S}^n$, with a sequential binary partition. The corresponding coordinates are the balances between the groups of parts separated in each step of a binary partition, and they allow us both subcompositional analysis, i.e. intra-group ratios, and grouped parts analysis, i.e. inter-group ratios. The main results on orthonormal bases in the simplex and how they are associated with a single partition were developed by Egozcue and others (2003). Coordinates of a composition with respect to a given orthonormal basis were

called ilr-coordinates by those authors, who noted the isometric character of such a representation by coordinates. Here we drop the qualifier ilr because the isometric character of the representation holds whenever the reference basis is orthonormal.

We associate one unit norm compositional vector with each order of a binary partition. The $n-1$ unitary compositional vectors so associated with the whole sequential binary partition constitute an orthonormal basis of $\mathcal{S}^n$. Assume that, in the ℓ-order binary partition, we separate parts $x_{k+1}, x_{k+2}, \ldots, x_{k+r}$ (r parts) from $x_{k+r+1}, x_{k+r+2}, \ldots, x_{k+r+s}$ (s parts). Furthermore, assume the remaining parts, if any, were separated in previous sequential order partitions. They are represented by $x_1, \ldots, x_k$ (k parts), respectively $x_{k+r+s+1}, \ldots, x_n$ (j parts). This means that $n = k + r + s + j$, $\ell \leq n - r - s + 1$ and that k and j can be null. The unitary vector associated with the ℓ-order binary partition, called the *balancing element*, is defined as

$$\mathbf{e}_\ell = \mathcal{C}\left[\exp\left(\underbrace{0, 0, \ldots, 0}_{k \text{ elements}}, \underbrace{a, a, \ldots, a}_{r \text{ elements}}, \underbrace{b, b, \ldots, b}_{s \text{ elements}}, \underbrace{0, 0, \ldots, 0}_{j \text{ elements}}\right)\right], \qquad (2)$$

where

$$a = \frac{\sqrt{s}}{\sqrt{r(r+s)}} \quad \text{and} \quad b = \frac{-\sqrt{r}}{\sqrt{s(r+s)}}.$$

The unit norm and orthogonality of these vectors is easily checked using Equation (A.3) (see the appendix).

In order to build up the basis associated with the sequential binary partition of Table 2, we proceed from order 1 to 6. For each order, the balancing element of the last partition is included in the basis. In our example, six basis elements are built up in this manner, leading to the following associated orthonormal basis:

$$\begin{aligned}
\mathbf{e}_1 &= \mathcal{C}\left[\exp\left(\tfrac{\sqrt{4}}{\sqrt{3\cdot 7}}, \tfrac{\sqrt{4}}{\sqrt{3\cdot 7}}, \tfrac{\sqrt{4}}{\sqrt{3\cdot 7}}, \tfrac{-\sqrt{3}}{\sqrt{4\cdot 7}}, \tfrac{-\sqrt{3}}{\sqrt{4\cdot 7}}, \tfrac{-\sqrt{3}}{\sqrt{4\cdot 7}}, \tfrac{-\sqrt{3}}{\sqrt{4\cdot 7}}\right)\right] \\
\mathbf{e}_2 &= \mathcal{C}\left[\exp\left(0, \quad 0, \quad 0, \quad \tfrac{\sqrt{2}}{\sqrt{2\cdot 4}}, \tfrac{\sqrt{2}}{\sqrt{2\cdot 4}}, \tfrac{-\sqrt{2}}{\sqrt{2\cdot 4}}, \tfrac{-\sqrt{2}}{\sqrt{2\cdot 4}}\right)\right] \\
\mathbf{e}_3 &= \mathcal{C}\left[\exp\left(\tfrac{\sqrt{2}}{\sqrt{1\cdot 3}}, \tfrac{-1}{\sqrt{2\cdot 3}}, \tfrac{-1}{\sqrt{2\cdot 3}}, \quad 0, \quad 0, \quad 0, \quad 0\right)\right] \\
\mathbf{e}_4 &= \mathcal{C}\left[\exp\left(0, \quad 0, \quad 0, \quad 0, \quad 0, \quad \tfrac{1}{\sqrt{1\cdot 2}}, \tfrac{-1}{\sqrt{1\cdot 2}}\right)\right] \\
\mathbf{e}_5 &= \mathcal{C}\left[\exp\left(0, \quad \tfrac{1}{\sqrt{1\cdot 2}}, \tfrac{-1}{\sqrt{1\cdot 2}}, \quad 0, \quad 0, \quad 0, \quad 0\right)\right] \\
\mathbf{e}_6 &= \mathcal{C}\left[\exp\left(0, \quad 0, \quad 0, \quad \tfrac{1}{\sqrt{1\cdot 2}}, \tfrac{-1}{\sqrt{1\cdot 2}}, \quad 0, \quad 0\right)\right]
\end{aligned} \qquad (3)$$

The values inside the square roots (previous equation) have not been simplified in order to facilitate the identification of the terms in Equation (2). For each sequential binary partition, (2) uniquely defines an associated orthonormal basis. However, the signs of the parts can be changed to obtain a basis which differs from the former in orientation. Also, different sequential binary partitions can be associated with orthogonal bases which only differ in the order of the elements or in some permutation of parts. In the example of Table 1, the three last binary partitions can be permuted and the associated orthonormal basis remains the same, except for the fact that elements $\mathbf{e}_4$, $\mathbf{e}_5$, $\mathbf{e}_6$ are also permuted accordingly.

Projections of an arbitrary composition $\mathbf{x} \in \mathcal{S}^n$ on unitary compositional vectors like those in Equation (2) are obtained from the inner products, $x_\ell^* = \langle \mathbf{x}, \mathbf{e}_\ell \rangle_a$ (Egozcue and others, 2003). They are the coordinates of $\mathbf{x}$ with respect to the basis elements $\mathbf{e}_\ell$, $\ell = 1, 2, \ldots, n-1$, and they are the log-ratios

$$\begin{aligned} x_\ell^* &= \sqrt{\frac{rs}{r+s}}\ \ln\left[\frac{g(x_{k+1},\ldots,x_{k+r})}{g(x_{k+r+1},\ldots,x_{k+r+s})}\right] \\ &= \ln\left[\frac{(x_{k+1}\cdots x_{k+r})^{\sqrt{s/(r(r+s))}}}{(x_{k+r+1}\cdots x_{k+r+s})^{\sqrt{r/(s(r+s))}}}\right], \end{aligned} \tag{4}$$

where $g(\cdot)$ denotes geometric mean of parts in the argument.

The form of log-ratios in (4) intuitively shows why x_ℓ^* has been called *balance* between the groups of parts $x_{k+1}, x_{k+2}, \ldots, x_{k+r}$ and $x_{k+r+1}, x_{k+r+2}, \ldots, x_{k+r+s}$, and why $\mathbf{e}_\ell$ has been called the *balancing element* for the two sets of parts (Egozcue and others, 2003).

Equation (5) shows the coordinates of $\mathbf{x}$ corresponding to the basis (3). Note that powers in the log-ratios are equal whenever the number of parts in each balanced group (numerator and denominator) are equal. Conversely, a different number of parts implies re-scaling powers of the ratio.

$$\begin{aligned} x_1^* &= \ln\left[\frac{(x_1x_2x_3)^{\sqrt{4/21}}}{(x_4x_5x_6x_7)^{\sqrt{3/28}}}\right], \quad x_2^* = \ln\left[\frac{(x_4x_5)^{1/2}}{(x_6x_7)^{1/2}}\right], \\ x_3^* &= \ln\left[\frac{x_1^{\sqrt{2/3}}}{(x_2x_3)^{\sqrt{1/6}}}\right], \quad x_4^* = \ln\left[\frac{x_6^{\sqrt{1/2}}}{x_7^{\sqrt{1/2}}}\right], \\ x_5^* &= \ln\left[\frac{x_2^{\sqrt{1/2}}}{x_3^{\sqrt{1/2}}}\right], \quad x_6^* = \ln\left[\frac{x_4^{\sqrt{1/2}}}{x_5^{\sqrt{1/2}}}\right]. \end{aligned} \tag{5}$$

Intra-Group Analysis: Subcompositions

An orthonormal basis of a sequential binary partition can be used to define an *orthonormal sub-basis associated with a group of parts*. Let $\mathbf{e}_i$, $i = 1, 2, \ldots, n-1$, be the orthonormal basis of a sequential binary partition and let $x_{k+1}, x_{k+2}, \ldots, x_{k+r}$ be a group of parts obtained in the ℓ-th sequential order of the binary partition. We focus on the r-part subcomposition, or R-subcomposition, defined by the subscripts in $R = \{k+1, k+2, \ldots, k+r\}$ (see Equation (A.1) in the appendix for details),

$$\text{sub}(\mathbf{x}; R) = \mathcal{C}[x_{k+1}, x_{k+2}, \ldots, x_{k+r}].$$

The basis element $\mathbf{e}_j$ is in the sub-basis of the R-group if $\text{sub}(\mathbf{e}_j; R) \neq \mathbf{n}_r$, where $\mathbf{n}_r = [1/r, 1/r, \ldots, 1/r]$ is the neutral element in $\mathcal{S}^r$. Basis elements such that $\text{sub}(\mathbf{e}_j; R) = \mathbf{n}_r$ are not associated with the R-group because they do not inform about its internal structure.

The basis elements obtained for a sequential order less than or equal to ℓ are not associated with the R-group. Only $r-1$ of the remaining elements $\mathbf{e}_{\ell+1}, \ldots, \mathbf{e}_{n-1}$ are associated with the R-group: those which parts with subscripts $k+1, k+1, \ldots, k+r$ are not equal.

Continuing with the example, let us inquire about the sub-basis associated with the group defined by $R = \{4, 5, 6, 7\}$ in Table 2. This group of parts is obtained at the first-order partition and, then, $\mathbf{e}_1$ in (3) is not associated with the R-group. Among the higher order basis elements only three are associated with the R-group: $\mathbf{e}_2$, $\mathbf{e}_4$, and $\mathbf{e}_6$. The other two, $\mathbf{e}_3$ and $\mathbf{e}_5$, are associated with group $\{1, 2, 3\}$.

The sub-basis of the R-group generates a subspace of $r-1$ dimensions $\mathcal{S}^n(R) \subset \mathcal{S}^n$. The main property of such a subspace is that orthogonal projections of compositions from $\mathcal{S}^n$ into it do not affect the R-subcomposition. In other words, ratios of parts in the R-subcomposition can be studied directly in $\mathcal{S}^n(R)$, after a projection that reduces the dimension from $n-1$ to $r-1$. This is coherent with an analogous assertion by Aitchison (1986) when introducing subcompositional analysis, although at that moment the algebraic–geometric structure of the simplex was not yet completely defined.

There are two possible ways to study the R-subcomposition of a data set: either to extract the R-subcomposition from raw data and then to carry out the analysis or, alternatively, to first obtain the coordinates (4) with respect to the basis (2) and then to extract those coordinates which correspond to the R-sub-basis. In this latter alternative, let the coordinates with respect to the R-sub-basis be $x_i^*, i \in R^*$, where R^* is a set of $r-1$ subscripts for the coordinates. Note these subscripts can be ordered arbitrarily. The projection of data into $\mathcal{S}^n(R)$ is given by $\oplus_{i \in R^*}(x_i^* \odot \mathbf{e}_i)$; it is still a compositional vector of n parts although it is in

the $r-1$-dimensional subspace $\mathcal{S}^n(R)$. In order to obtain an effective reduction of the dimensionality we have to represent it in $\mathcal{S}^r$. To do this, we look for a representation basis $\mathbf{h}_i \in \mathcal{S}^r$ such that, for any $\mathbf{x} \in \mathcal{S}^n$,

$$\mathrm{sub}(\mathbf{x}; R) = \bigoplus_{i \in R^*} (x_i^* \odot \mathbf{h}_i).$$

A sensible representation basis is $\mathbf{h}_i = \mathrm{sub}(\mathbf{e}_i; R)$, where $\mathbf{e}_i$ are the sub-basis elements associated with the R-group. Its advantage is that there is no need to re-calculate coordinates after taking subcomposition, as they are just those corresponding to the sub-basis elements associated with the R-group (see Equation (A.2) in the appendix).

From coordinates with respect to the sub-basis associated with the R-group, we easily reconstruct associated subcompositions. For instance, in the example, the subcomposition with $R = \{1, 2, 3\}$ is associated with elements $\mathbf{e}_3$ and $\mathbf{e}_5$ in the basis (3), i.e. $R^* = \{3, 5\}$. We have

$$\mathcal{C}[x_1, x_2, x_3] = \mathrm{sub}((x_3^* \odot \mathbf{e}_3) \oplus (x_5^* \odot \mathbf{e}_5); \{1, 2, 3\}) = (x_3^* \odot \mathbf{h}_3) \oplus (x_5^* \odot \mathbf{h}_5), \tag{6}$$

where the representation basis in $\mathcal{S}^3$ is

$$\mathbf{h}_3 = \mathrm{sub}(\mathbf{e}_3; \{1, 2, 3\}), \qquad \mathbf{h}_5 = \mathrm{sub}(\mathbf{e}_5; \{1, 2, 3\}).$$

Representation basis allows us to reconstruct the parts of the $\{1, 2, 3\}$-subcomposition as shown in (6). This apparently trivial fact is important when plotting data in a ternary diagram; for instance, the corners of the diagram should be labeled with the reconstructed part, in this case x_1, x_2, x_3, where the closure constant is obviated.

Inter-Group Analysis: Balances

In order to represent relationships between groups of parts, we use the appropriate coordinates, the balances. This generally causes a reduction of dimension—there are less groups than parts—and we accordingly need a representation in a lower-dimension simplex. Although apparently similar to a subcompositional analysis, normalization constants appear because of the possibly different number of parts in each group.

Assume we are interested in balances between $\ell + 1$ groups which constitute a partition of the whole set of n parts. This partition can be obtained as a binary partition of sequential order ℓ. Let be $\mathbf{e}_i$, $i = 1, \ldots, \ell$, the basis elements (2) associated with the binary partition up to the required sequential order. The

corresponding coordinates, x_i^*, $i = 1, \ldots, \ell$, are the balances between groups of parts and contain all the information about the relationships between such groups of parts. The explicit expression of the orthogonal projection of a composition on the balancing element, $x_i^* \odot \mathbf{e}_i$, for any $i \leq \ell$, allows us a better insight into what a balance is representing. We assume that in the partition of order i we separate two contiguous groups of parts, made of r and s parts respectively, as in (2). Then, from expressions (2) and (4),

$$x_i^* \odot \mathbf{e}_i = \mathcal{C}\left[\underbrace{\left(\prod_{i=k+1}^{k+r+s} x_i\right)^{\frac{1}{r+s}}}_{k \text{ repeated elements}}, \underbrace{\left(\prod_{i=k+1}^{k+r} x_i\right)^{\frac{1}{r}}}_{r \text{ repeated elements}}, \underbrace{\left(\prod_{i=k+r+1}^{k+r+s} x_i\right)^{\frac{1}{s}}}_{s \text{ repeated elements}}, \underbrace{\left(\prod_{i=k+1}^{k+r+s} x_i\right)^{\frac{1}{r+s}}}_{j \text{ repeated elements}}\right] \tag{7}$$

with $k + r + s + j = n$, where we realize that each original part in a group is substituted by the geometric mean of the parts included in that group. Outside the groups, each element is substituted by the geometric mean of all parts included in both groups. In order to interpret (7) we need to intuitively understand why each component of the original composition has been replaced by a geometric mean in this projection on the balancing element. Compositional vector (7) should not carry any intra-group information—we are dealing with three groups of parts—and, therefore, parts within the same group should be equal. A simple but relevant exercise shows that the closest composition of the form $[a, a, \ldots, a, x_r, x_{r+1}, \ldots, x_n]$ to $[x_1, x_2, \ldots, x_n]$, in the sense of Aitchison distance d_a, is that one with $a = g(x_1, \ldots, x_r)$, the geometric mean of the replaced components. This is an intuitive argument which would lead to Equation (7), favoring the geometric mean against other alternatives like the arithmetic mean.

Balancing elements $\mathbf{e}_i$, $i = 1, 2, \ldots, \ell$, constitute an orthogonal basis of a subspace and projection of $\mathbf{x}$ onto it is a composition of n parts in which only inter-group relationships are taken into account and information about intra-group ratios has been removed. The expression of such a projection is

$$\bigoplus_{i=1}^{\ell} (\mathbf{x}_i^* \odot \mathbf{e}_i) = \mathcal{C}\left[\underbrace{\left(\prod_{j=1}^{r_1} x_j\right)^{\frac{1}{r_1}}}_{r_1 \text{ repeated elements}}, \underbrace{\left(\prod_{j=1}^{r_2} x_{r_1+j}\right)^{\frac{1}{r_2}}}_{r_2 \text{ repeated elements}}, \ldots, \underbrace{\left(\prod_{j=1}^{r_\ell} x_{n+1-j}\right)^{\frac{1}{r_{\ell+1}}}}_{r_{\ell+1} \text{ repeated elements}}\right], \tag{8}$$

where $r_1, r_2, \ldots, r_{\ell+1}$ are, respectively, the number of parts of each one of the $\ell + 1$ groups obtained in the ℓ-order binary partition.

Equation (8) can be proved by induction. For $\ell = 1$, it reduces to (7) with $k = j = 0$. Assume (8) holds for order ℓ and that the t-th group is divided into two groups of, respectively, s_1 and s_2 parts ($s_1 + s_2 = r_t$) in the $(\ell + 1)$-order binary partition. We should perturb (8) with (7) taking $k + j = n - r_t, r = s_1$, and $s = s_2$. After removing irrelevant equal factors, we obtain the desired expression equivalent to (8) updated to order $\ell + 1$.

As an example, the sequential binary partition in Table 2 defines, at order $\ell = 2$, three groups of parts, namely $\{1, 2, 3\}$, $\{4, 5\}$, and $\{6, 7\}$. Information of balances between these three groups is conveyed by coordinates x_1^* and x_2^* in (5) with respect to the associated orthonormal basis (3). Projection of $\mathbf{x}$ onto the balancing elements $\mathbf{e}_1$ and $\mathbf{e}_2$ are

$$x_1^* \odot \mathbf{e}_1 = \mathcal{C}\left[\underbrace{(x_1x_2x_3)^{1/3}}_{\text{three repeated elements}}, \underbrace{(x_4x_5x_6x_7)^{1/4}}_{\text{four repeated elements}}\right],$$

$$x_2^* \odot \mathbf{e}_2 = \mathcal{C}\left[\underbrace{(x_4x_5x_6x_7)^{1/4}}_{\text{three repeated elements}}, \underbrace{(x_4x_5)^{1/2}}_{\text{three repeated elements}}, \underbrace{(x_6x_7)^{1/2}}_{\text{two repeated elements}}\right].$$

The perturbation of these two projections is the orthogonal projection of $\mathbf{x}$ onto the subspace generated by the two balancing elements $\mathbf{e}_1, \mathbf{e}_2$. This projection is

$$\bigoplus_{i=1}^{2}(x_i^* \odot \mathbf{e}_i) = \mathcal{C}\left[\underbrace{(x_1x_2x_3)^{1/3}}_{\text{three repeated elements}}, \underbrace{(x_4x_5)^{1/2}}_{\text{two repeated elements}}, \underbrace{(x_6x_7)^{1/2}}_{\text{two repeated elements}}\right]. \quad (9)$$

Balance coordinates x_i^*, $i = 1, \ldots, \ell$, can be represented in a lower-dimension simplex with $\ell + 1$ parts and dimension ℓ. In order to do so, we need to choose an orthonormal basis and to assign to each balance coordinate x_i^*, $i = 1, \ldots, \ell$, a vector of a representation basis $\mathbf{h}_i$ as was done for subcompositions. Then, the balances between groups can be represented in $\mathcal{S}^{\ell+1}$ as $\bigoplus_{i=1}^{\ell}(x_i^* \odot \mathbf{h}_i)$. Although the representation basis is arbitrary, we propose selecting it in the following way: At the sequential partition of order ℓ we have got the desired groups of parts. Each of these groups can be treated as a single part and, then, the sequential binary partition up to the order ℓ is readily identified with a sequential binary partition of $\ell + 1$ parts; the associated orthonormal basis of $\mathcal{S}^{\ell+1}$ can thus be taken to be the representation basis $\mathbf{h}_i$, $i = 1, \ldots, \ell$.

Coming back to the sequential binary partition in Table 2 at order 2, we have the three groups of parts $\{1, 2, 3\}$, $\{4, 5\}$, and $\{6, 7\}$. Three balances between these three groups can be defined, but two of them are enough to describe inter-group

ratios. The selected sequence of partitions in Table 2, up to order 2, determines the respective balance coordinates x_1^* and x_2^* in (5). Note that these balances can be represented in a simplex of three parts. As proposed, we choose the representation basis in $\mathcal{S}^3$ defined by the sequential binary partition $[g_1||g_2|g_3]$, where grouped parts have been denoted by g_i. Consequently, the representation vectors are

$$\mathbf{h}_1 = \mathcal{C}\left[\exp\left(\frac{2}{\sqrt{6}}, \frac{-1}{\sqrt{6}}, \frac{-1}{\sqrt{6}}\right)\right], \quad \mathbf{h}_2 = \mathcal{C}\left[\exp\left(0, \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right)\right], \tag{10}$$

where the first vector represents the balance between the first group and the second and third ones as a whole, and the second vector represents the balance between the second and third groups.

The grouped parts in $\mathcal{S}^3$ are then

$$x_1^* \odot \mathbf{h}_1 = \left(\frac{\sqrt{3\cdot 4}}{\sqrt{3+4}} \cdot \frac{3}{\sqrt{6}}\right) \odot \mathcal{C}\left[(x_1x_2x_3)^{\frac{1}{3}}, (x_4x_5x_6x_7)^{\frac{1}{4}}, (x_4x_5x_6x_7)^{\frac{1}{4}}\right],$$

$$x_2^* \odot \mathbf{h}_2 = \left(\frac{\sqrt{2\cdot 2}}{\sqrt{2+2}} \cdot \frac{2}{\sqrt{2}}\right) \odot \mathcal{C}\left[(x_4x_5x_6x_7)^{\frac{1}{4}}, (x_4x_5)^{\frac{1}{2}}, (x_6x_7)^{\frac{1}{2}}\right],$$

$$\begin{aligned}[g_1, g_2, g_3] &= x_1^* \odot \mathbf{h}_1 \oplus x_2^* \odot \mathbf{h}_2 \\ &= \mathcal{C}\left[(x_1x_2x_3)^{\frac{2}{\sqrt{14}}}, (x_4x_5)^{\frac{3}{\sqrt{56}}+\frac{1}{\sqrt{8}}}(x_6x_7)^{\frac{3}{\sqrt{56}}-\frac{1}{\sqrt{8}}}, (x_4x_5)^{\frac{3}{\sqrt{56}}-\frac{1}{\sqrt{8}}}(x_6x_7)^{\frac{3}{\sqrt{56}}+\frac{1}{\sqrt{8}}}\right],\end{aligned}$$

where some algebra is required to compute all constants. The main tool to obtain such an expression is the fact that closure allows us to multiply all components in the vector by a common factor. We realize that the expression of reconstructed parts in $\mathcal{S}^3$ is not simple, but all parts in a group appear in this expression with the same power. However, we prefer to look at the projections on the balancing elements (7) or on the whole subspace of balance (8) because they reduce to compositions in which all parts of a group are equal and depend only on the geometric mean of the parts in the group, as illustrated in (9). These expressions do not provide a simple way of labeling the vertices of a ternary diagram when representing these balances in $\mathcal{S}^3$. We should recognize that the lack of simplicity of reconstructed parts in reduced dimension is mainly due to the possibly different number of parts in each group, which imply normalization powering in order to preserve distances.

An alternative example shows that an equal number of parts in each group results in simple expressions of reconstructed parts. Consider $\mathbf{x} \in \mathcal{S}^6$ and the three

groups obtained at the second-order binary partition $[x_1, x_2||x_3, x_4|x_5, x_6]$. The first and second balances are

$$x_1^* = \ln\left[\frac{(x_1x_2)^{\sqrt{4/(2\cdot 6)}}}{(x_3x_4x_5x_6)^{\sqrt{2/(4\cdot 6)}}}\right], \quad x_2^* = \ln\left[\frac{(x_3x_4)^{1/2}}{(x_5x_6)^{1/2}}\right],$$

where again we have avoided simplification of fractions. If we use $\mathbf{h}_1, \mathbf{h}_2$ (10) as a representation basis, we obtain a simpler expression of reconstructed parts in $\mathcal{S}^3$

$$(x_1^* \odot \mathbf{h}_1) \oplus (x_2^* \odot \mathbf{h}_2) = \mathcal{C}\big[(x_1x_2)^{1/\sqrt{2}}, (x_3x_4)^{1/\sqrt{2}}, (x_5x_6)^{1/\sqrt{2}}\big].$$

After these two examples, it appears that the simplest labeling of parts in reduced dimension, for instance in a ternary diagram, is just the product of grouped parts—or some generic function of the product—so ignoring scaling powers. As shown in the next section, scaling powers are needed to preserve distances. Balances fully represent inter-group relationships, and Aitchison distances between sample compositions in this representation are dominated by distances in the original sample space, as desired.

Finally, a conceptual example can improve our understanding of what a balance is and how balances do appear in a process of exponential decay or growth of mass. Imagine that n different radiogenic isotopes disintegrate with no interaction, i.e. the products of disintegration are not accounted for and they do not correspond to any originally considered isotope. Assume that in a time t the remaining mass of the ith isotope is $z_i = \exp[\alpha_i + \lambda_i t]$ for $i = 1, 2, \ldots, n$. In the mass decay case the constants λ_i are negative but this is irrelevant for our purposes. It is known that the corresponding composition of masses, $\mathcal{C}[z_1, \ldots, z_n]$, follows a line in the simplex (Egozcue and others, 2003). Assume now that the involved isotopes are classified into two groups of affine isotopes in a first-order partition, e.g. because the decay constants λ_i are similar. Let $R_0 = \{1, 2, \ldots, r_0\}$ and $Q_0 = \{r_0 + 1, r_0 + 2, \ldots, n\}$ be the subscripts of these two groups. Furthermore, assume that the group Q_0 is also partitioned into two groups of isotopes, the second-order partition, which subscripts are $R_1 = \{r_0 + 1, \ldots, r_0 + r_1\}$ and $R_2 = \{r_0 + r_1 + 1, \ldots, n\}$, which have respectively r_1 and r_2 elements. Suppose we are mainly interested in the compositional relations of these three groups obtained in the second-order partition and intra-group compositions of mass are not aimed at in such a study.

Based on the affinity of isotopes within a group, we accept to approach the mass of these isotopes within the group R_j by $z_i \simeq \exp[\beta_j + \nu_j t]$, where $\beta_j = \sum_{i\in R_j} \alpha_i/r_j$ and $\nu_j = \sum_{i\in R_j} \lambda_i/r_j$, i.e. the masses of all isotopes within a group behave equal to the average constant of decay. This assumption removes all intra-group compositional information. After this simplification and in order

to study inter-group behavior of masses, we express the exponential decay of masses as

$$
\begin{aligned}
z_i &\simeq \exp\left[\beta_0 + \nu_0 t\right] \qquad \exp\left[0\, t\right], && \text{if } i \in R_0,\\
z_i &\simeq \exp\left[\tfrac{r_1(\beta_1+\nu_1 t)+r_2(\beta_2+\nu_2 t)}{r_1+r_2}\right] \exp\left[\beta_1 + \nu_1 t - \tfrac{r_1(\beta_1+\nu_1 t)+r_2(\beta_2+\nu_2 t)}{r_1+r_2}\right], && \text{if } i \in R_1,\\
z_i &\simeq \exp\left[\tfrac{r_1(\beta_1+\nu_1 t)+r_2(\beta_2+\nu_2 t)}{r_1+r_2}\right] \exp\left[\beta_2 + \nu_2 t - \tfrac{r_1(\beta_1+\nu_1 t)+r_2(\beta_2+\nu_2 t)}{r_1+r_2}\right], && \text{if } i \in R_2.
\end{aligned}
$$

This expression decomposes the exponential mass decay of each isotope into the product of two exponential models. From the compositional point of view, the process is expressed as the perturbation of two linear processes. The first column of exponentials accounts for mass decay in the group R_0, with initial mass $\exp[\beta_0]$ and rate ν_0, and the group $Q_0 = R_1 \cup R_2$, with initial average mass $\exp[(r_1\beta_1 + r_2\beta_2)/(r_1 + r_2)]$ and average rate $(r_1\nu_1 + r_2\nu_2)/(r_1 + r_2)$. For a fixed t—and after closure—we readily assign the values of this first model to (7) when applied to the partition into the two groups R_0 and Q_0. In fact, the terms of this first column are the geometric means of the parts within their respective groups R_0 and Q_0. The second column of exponentials represents a compositional process, that again corresponds to (7), now applied to the partition of Q_0 into the groups R_1 and R_2, while R_0 remains unchanged. In fact, (7) can be re-written as

$$
\mathcal{C}\left[\underbrace{1}_{r_0 \text{ terms}}, \underbrace{\frac{g(x_i;\, i \in R_1)}{g(x_k;\, k \in Q_0)}}_{r_1 \text{ terms}}, \underbrace{\frac{g(x_j;\, j \in R_2)}{g(x_k;\, k \in Q_0)}}_{r_2 \text{ terms}}\right],
$$

where $g(\cdot)$ is the geometric mean of the arguments. The quotients of geometric means are now identified to the minus signs in the exponentials of this second column. The conclusion is that the simple decomposition of the exponential decay of mass, just corresponds to a decomposition into two decay of mass (orthogonal) processes from the compositional point of view. Generalization to higher-order sequential partitions is straightforward.

Subcompositional and Balance Dominance for Distances

An important property of the analysis of groups of parts is that the desired properties of distances in the simplex are preserved. Egozcue and others (2003) have shown that the Aitchison distance in $\mathcal{S}^n$ between two compositions $\mathbf{x}$ and $\mathbf{y}$ can be expressed as an ordinary Euclidean distance in terms of coordinates with

respect to an orthonormal basis, like those given in (4), resulting in

$$d_a^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n-1} (x_i^* - y_i^*)^2. \tag{11}$$

Let us assume that we are interested in the $\ell + 1$ groups defined in the ℓ-sequential order partition. The basis associated with the sequential binary partition can be separated into groups of elements: (a) Those containing balancing elements between groups of the partition; for instance, $\mathbf{e}_i$, $i = 1, \ldots, \ell$. (b) Those sub-bases associated with each group in the partition; let $\mathbf{e}_j$, $j \in R_k^*$ be the sub-basis associated with the R_k-group of r_k parts. The number of subscripts in R_k^* is then $r_k - 1$, being $\sum_{j=1}^{\ell+1}(r_k - 1) = n - \ell - 1$ and $\bigcup_{k=1}^{\ell+1} R_k = \{\ell + 1, \ell + 2, \ldots, n\}$.

Now the squared distance (11) is readily decomposed into squared terms associated with inter-group balances, and terms associated with intra-group coordinates,

$$d_a^2(\mathbf{x}, \mathbf{y}) = \underbrace{\sum_{i=1}^{\ell} (x_i^* - y_i^*)^2}_{\text{inter-group distance}} + \sum_{k=1}^{\ell+1} \underbrace{\sum_{j \in R_k^*} (x_j^* - y_j^*)^2}_{\text{intra-}k\text{-group distance}} . \tag{12}$$

The sum of terms for $i = 1, \ldots, \ell$ is the inter-group contribution of balances to the square distance. The sum of terms for $j \in R_k^*$ is the intra-R_k-group contribution; it is equal to the Aitchison distance between $\mathbf{x}$ and $\mathbf{y}$ measured in the R_k-subcomposition.

A first conclusion is that subcompositional contribution to the square distance (12) is dominated by the distance in $\mathcal{S}^n$ as stated by Aitchison (1992), i.e.

$$d_a^2(\mathbf{x}, \mathbf{y}) \geq \sum_{j \in R_k^*} (x_j^* - y_j^*)^2.$$

A second conclusion is that balance or inter-group contribution to the square distance (12) is also dominated by it, i.e.

$$d_a^2(\mathbf{x}, \mathbf{y}) \geq \sum_{i=1}^{\ell} (x_i^* - y_i^*)^2.$$

These two properties give the necessary consistency to statistical analysis of compositional data when dealing with grouped parts, both as balances between them and as subcompositions. Moreover, all terms in (12) are square distances measured along orthogonal directions—the directions of the elements of the orthonormal

basis—and, therefore, decomposition (12) should be interpreted as the Pythagoras theorem: square distances are obtained by adding square distances of orthogonal contributions.

SUBCOMPOSITIONS AND BALANCES AS ORTHOGONAL PROJECTIONS

To better understand balance coordinates and subspace associated with a group—as well as their properties—a more formal description than the intuitive presentation in the previous section is needed. Here we intend to give such a description. The main concepts are *subspace associated with a group of parts* and *subspace associated with a balance of two groups of parts*. The first one gives rise to projections used in intra-group or subcompositional analysis; the second one allows us to face inter-group analysis as an orthogonal projection.

The concept of subspace associated with an R-group has been used in the previous section. Now, we redefine it formally. Let be R a non-empty set of indexes from the set $\{1, 2, \ldots, n\}$, and r the number of indexes in R, i.e. $1 \leq r = \text{Card}(R) \leq n - 1$.

Definition 1. A composition $\mathbf{x} \in \mathcal{S}^n$ is associated with the R-group if

$$\text{sub}(\mathbf{x}; R) \neq \mathbf{n}_r, \quad \text{sub}(\mathbf{x}; \bar{R}) = \mathbf{n}_{n-r}, \tag{13}$$

where $R \cup \bar{R} = \{1, 2, \ldots, n\}$, $R \cap \bar{R} = \emptyset$, and $\mathbf{n}_r$, $\mathbf{n}_{n-r}$ are the neutral elements in $\mathcal{S}^r$ and $\mathcal{S}^{n-r}$ respectively. The set of compositions associated with the R-group, complemented with the neutral element of $\mathcal{S}^n$, $\mathbf{n}_n$, is denoted by $\mathcal{S}^n(R)$.

As an example, consider a composition $\mathcal{C}[1, 1, 1, 2]$ in $\mathcal{S}^4$. It is associated with groups $\{1, 4\}$ and $\{1, 2, 4\}$, while it is not associated with group $\{1, 3\}$.

Proposition 1. *$\mathcal{S}^n(R)$ is an $(r - 1)$-dimensional subspace of $\mathcal{S}^n$.*

Proof: If $r = 1$ we have $\mathbf{n}_r = [1]$ and, for all compositions, $\mathbf{x} \in \mathcal{S}^n$, $\text{sub}(\mathbf{x}; R) = \mathbf{n}_r$ against (13). Then, the only element in $\mathcal{S}^n(R)$ is $\mathbf{n}_n$; therefore, it is a degenerate subspace of null dimension. For $2 \leq r \leq n - 1$, by Definition 1, $\mathcal{S}^n(R)$ is closed under perturbation and powering and, therefore, $\mathcal{S}^n(R)$ is a subspace of $\mathcal{S}^n$. In order to determine the dimension of $\mathcal{S}^n(R)$, let $\mathbf{e}_i$, $i = 1, 2, \ldots, r - 1$, be an orthonormal basis of $\mathcal{S}^r$ and assume that the R-parts have been placed at the r-first positions. We complete these vectors to n parts by adding $n - r$ equal constants a_i to obtain $\mathbf{z}_i = [e_{i1}, e_{i2}, \ldots, e_{ir}, a_i, \ldots, a_i]$ such that $\|\mathbf{z}_i\|_a = 1$. The constant a_i is fully determined by this condition for each $i = 1, 2, \ldots, r - 1$. Compositions $\mathbf{z}_i$ are orthonormal and they are in $\mathcal{S}^n(R)$. Therefore, the dimension of $\mathcal{S}^n(R)$ is

greater than or equal to $r-1$. If that dimension were $s_1 \geq r$, the same argument assures that dimension of $\mathcal{S}^n(\bar{R})$ is $s_2 \geq n-r$, but $s_1 + s_2 \geq n > n-1$. Since $\mathcal{S}^n(\bar{R})$ and $\mathcal{S}^n(R)$ are orthogonal, the sum of their dimensions should be less than or equal to $n-1$. Therefore, dimension of $\mathcal{S}^n(R)$ is $r-1$. □

Proposition 1 points out that compositions associated with an R-group have equal parts for subscripts in $\bar{R}$, provided that parts with subscripts in R are not equal.

Definition 2. $\mathcal{S}^n(R)$ is called subspace associated with the R-group.

Proposition 2. *Let R define a group of r parts, $2 \leq r \leq n-1$, obtained in a sequential binary partition at the order ℓ, $\ell < n-1$, and let $\mathbf{e}_i$, $i = 1, 2, \ldots, n-1$ be the associated orthonormal basis. Let $x_1^*, x_2^*, \ldots, x_{n-1}^*$ be the coordinates of $\mathbf{x} \in \mathcal{S}^n$ with respect to this basis. It holds that*

(a) *there are $r-1$ elements in the basis that constitute a sub-basis for $\mathcal{S}^n(R)$, let them be $\mathbf{e}_j$, $j \in R^*$;*
(b) *the orthogonal projection of $\mathbf{x}$ on $\mathcal{S}^n(R)$ has coordinates x_j^* for $j \in R^*$ and null otherwise;*
(c) *$\mathbf{h}_j = \text{sub}(\mathbf{e}_j; R)$, $j \in R^*$, constitute an orthonormal basis of $\mathcal{S}^r$;*
(d) *the coordinates of $\text{sub}(\mathbf{x}; R)$ with respect to the basis $\mathbf{h}_j$, $j \in R^*$, are x_j^*.*

Proof: Basis elements associated with the sequential binary partition up to order ℓ are not associated with the R-group. To attain a partition of the R-group in single part sub-groups, $r-1$ binary partitions inside the R-group are required and those binary partitions generate the sub-basis mentioned in (a). Statement (b) is a direct consequence of (a). Taking R-subcomposition on vectors associated with the R-group do not modify their inner products as a consequence of Definition 1; this implies (c). Statement (d) is obtained from (c) and basic properties of subcompositions (A2). □

Example 1. Assume $n = 5$, $R = \{1, 3, 4\}$, and thus $r = 3$. The subspace $\mathcal{S}^5(R)$, which dimension is 2, is generated by any two independent vectors associated with the R-group, e.g. $\mathcal{C}[1, 1, 2, 2, 1]$ and $\mathcal{C}[2, 1, 1, 2, 1]$. An orthonormal basis of $\mathcal{S}^5(R)$ is readily obtained as described in the previous section; an example is

$$\mathbf{e}_1 = \mathcal{C}\left[\exp\left(\frac{1}{\sqrt{6}},\ 0,\ \frac{1}{\sqrt{6}},\ -\frac{\sqrt{2}}{\sqrt{3}},\ 0\right)\right],$$

$$\mathbf{e}_2 = \mathcal{C}\left[\exp\left(\frac{1}{\sqrt{2}},\ 0,\ -\frac{1}{\sqrt{2}},\ 0, 0\right)\right],$$

in agreement with (2) up to the order of parts; in this case we take $R^* = \{1, 2\}$. The representation basis in the R-subcomposition is

$$\mathbf{h}_1 = \mathcal{C}\left[\exp\left(\frac{1}{\sqrt{6}}, \frac{1}{\sqrt{6}}, -\frac{\sqrt{2}}{\sqrt{3}}\right)\right], \quad \mathbf{h}_2 = \mathcal{C}\left[\exp\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0\right)\right].$$

A consequence of Proposition 2 is that the orthogonal projection on the associated subspace to the R-group does not affect the R-subcomposition. This means that such a projection filters out all information that is not related to the R-subcomposition and, in practice, we can identify the projection with the operation of taking R-subcomposition. This is reformulated in the next proposition, where the orthonormal basis is not necessarily that associated with a sequential binary partition.

Proposition 3. *Let $\{\mathbf{e}_j,\ j \in R^*\}$ be a set of $r-1$ orthonormal elements in $\mathcal{S}^n$. This set is an orthonormal basis of $\mathcal{S}^n(R)$ if and only if, for all $\mathbf{x} \in \mathcal{S}^n$, it satisfies*

$$\mathrm{sub}(\mathbf{x}; R) = \mathrm{sub}\Big(\bigoplus_{j \in R^*} (x_j^* \odot \mathbf{e}_j); R\Big), \tag{14}$$

where $x_j^* = \langle \mathbf{x}, \mathbf{e}_j \rangle_a$.

Proof: Assume $\mathbf{e}_j$, $j \in R^*$, constitute an orthonormal basis of $\mathcal{S}^n(R)$ and complete it to an orthonormal basis in $\mathcal{S}^n$ with $n-r$ elements $\mathbf{e}_k, k \in \bar{R}^*$. These vectors satisfy $\mathrm{sub}(\mathbf{e}_k; R) = \mathbf{n}_n$, because the dimension of $\mathcal{S}^n(R)$ is $r-1$. Property (14) is obtained by taking the R-subcomposition in the expression of $\mathbf{x}$ and using (A.2) of the appendix.

Assume now that (14) holds. Since $\mathbf{e}_j$, $j \in R^*$, are $r-1$ orthonormal vectors, they constitute an orthonormal basis of $\mathcal{S}^n(R)$ only if they are in $\mathcal{S}^n(R)$. To show this, we complete the orthonormal basis of $\mathcal{S}^n$ with $n-r$ elements $\mathbf{e}_k$, $k \in \bar{R}^*$. Each composition $\mathbf{x} \in \mathcal{S}^n$ is expressed as $\mathbf{x} = \bigoplus_i (x_i^* \odot \mathbf{e}_i)$. Particularly, this is true for $\mathbf{e}_k$, $k \in \bar{R}^*$, and (14) is reduced to $\mathrm{sub}(\mathbf{e}_k; R) = \mathbf{n}_r$. This implies that $\mathbf{e}_k$, $k \in \bar{R}^*$, are not in $\mathcal{S}^n(R)$, and they constitute an orthonormal basis of the orthogonal complement of $\mathcal{S}^n(R)$. Since we assume $\mathbf{e}_j$, $j \in R^*$, are orthogonal to $\mathbf{e}_k$, $k \in \bar{R}^*$, they are in $\mathcal{S}^n(R)$. □

Example 2. Assume $n = 3$ and $R = \{1, 2\}$. An orthonormal basis of $\mathcal{S}^3$ is

$$\mathbf{e}_1 = \mathcal{C}\left[\exp\left(\sqrt{\frac{1}{6}}, \sqrt{\frac{1}{6}}, -\sqrt{\frac{2}{3}}\right)\right], \quad \mathbf{e}_2 = \mathcal{C}\left[\exp\left(\sqrt{\frac{1}{2}}, -\sqrt{\frac{1}{2}}, 0\right)\right], \tag{15}$$

which is associated with the sequential binary partition $[x_1|x_2||x_3]$. Element $\mathbf{e}_2$ is associated with the R-subcomposition and it is an unitary basis of the one-dimensional space $\mathcal{S}^3(R)$. Consider a composition $\mathbf{x} = \mathcal{C}[5, 2, 3]$ and its R-subcomposition $\mathcal{C}[5, 2]$. Note that the projection on $\mathcal{S}^3(R)$ is $x_2^* = \langle \mathbf{x}, \mathbf{e}_2 \rangle_a \odot \mathbf{e}_2 = \mathcal{C}[5, 2, \sqrt{10}]$ which, after taking subcomposition returns again $\mathcal{C}[5, 2]$. The coordinates of $\mathbf{x} = \mathcal{C}[5, 2, 3]$ in the basis (15) are $x_1^* = 6^{-1/2} \ln(10/9)$ and $x_2^* = 2^{-1/2} \ln(5/2)$. If we only retain x_2^*, which is associated with the $\{1, 2\}$-subcomposition, we readily reconstruct a composition in $\mathcal{S}^3$ by $\langle \mathbf{x}, \mathbf{e}_2 \rangle_a \odot \mathbf{e}_2$, whose $\{1, 2\}$-subcomposition is $\mathcal{C}[5, 2]$ as was obtained previously.

An immediate consequence of this is that the ratios and log-ratios of parts in an R-subcomposition are, as expected, equal to the ratios and log-ratios of R-parts in the orthogonal projection onto $\mathcal{S}^n(R)$.

The elements of bases associated with sequential binary partitions can also be interpreted as balancing elements between groups. The element of the basis obtained at a given sequential order of binary partition is the balancing element between the two new groups obtained at this step. This allows the analysis of grouped components just using the corresponding balance coordinates. Next definitions and properties formalize these ideas.

We now deal with two non-overlapping, non-empty, groups of parts of compositions in $\mathcal{S}^n$. The groups are represented by the corresponding sets of indexes R_1 and R_2, $R_1 \cap R_2 = \emptyset$, $\text{Card}(R_i) = r_i$, $i = 1, 2$, and $r_1 + r_2 \leq n$. Let Q denote the complement of $R_1 \cup R_2$ in $\{1, 2, \ldots, n\}$, $\text{Card}(Q) = q$, $r_1 + r_2 + q = n$.

Definition 3. A unitary composition $\mathbf{e}$ in $\mathcal{S}^n$, $\|\mathbf{e}\|_a = 1$, is said to be an (R_1, R_2)-balancing element if: (a) $\text{sub}(\mathbf{e}; R_i) = \mathbf{n}_{r_i}$ for $i = 1, 2$; (b) $\text{sub}(\mathbf{e}; Q) = \mathbf{n}_q$; and (c) $\text{sub}(\mathbf{e}; R_1 \cup R_2) \neq \mathbf{n}_{r_1+r_2}$.

This means that (R_1, R_2)-balancing elements have equal components in the parts corresponding to the indexes in R_1; this also holds for indexes in R_2 and in Q, if non-empty; but for indexes in R_1 and R_2 components are not equal. This is equivalent to say that the (R_1, R_2)-balancing element is in $\mathcal{S}^n(R_1 \cup R_2)$, but that it is neither in $\mathcal{S}^n(R_1)$ nor in $\mathcal{S}^n(R_2)$. The (R_1, R_2)-balancing element can be considered as an orthonormal basis of a one-dimensional subspace. This concept was previously introduced by Egozcue and others (2003).

Definition 4. Let $\mathbf{e}$ be an (R_1, R_2)-balancing element. For $\mathbf{x} \in \mathcal{S}^n$, the inner product $\langle \mathbf{x}, \mathbf{e} \rangle_a = x^*_{(R_1,R_2)}$ is called the (R_1, R_2)-balance of $\mathbf{x}$.

Proposition 4. *The (R_1, R_2)-balancing element is unique up to a (-1) powering, which means change in orientation.*

Proof: From Definition 3, the property of being a balancing element is not lost by (-1) powering. First assume $Q = \emptyset$ and, therefore, $r_1 + r_2 = n$. Orthogonal

subspaces $\mathcal{S}^n(R_j)$ have dimensions $r_j - 1$, $j = 1, 2$. Consequently, their orthogonal complement in $\mathcal{S}^n$ is one dimensional. Since the (R_1, R_2)-balancing element is not associated with the R_j-group, $j = 1, 2$, from Definitions 1 and 3 it follows that it is a basis for that one-dimensional subspace; as the balancing element is unitary, it is unique up to (-1) powering.

Alternatively, assume $q \geq 1$. Then, the orthogonal complement of $\mathcal{S}^n(R_j)$, $j = 1, 2$, and of $\mathcal{S}^n(Q)$, has dimension $2 = (n-1) - (r_1 - 1) - (r_2 - 1) - (q - 1)$. Both the (R_1, R_2)-balancing element and the $(R_1 \cap R_2, Q)$-balancing element are in this two-dimensional subspace. These two balancing elements are orthogonal, because they are associated with a sequential binary partition. Therefore, the (R_1, R_2)-balancing element is the only element of a basis for the one-dimensional subspace orthogonal to $\mathcal{S}^n(R_j)$, $j = 1, 2$, to $\mathcal{S}^n(Q)$ and to the $(R_1 \cap R_2, Q)$-balancing element. □

Proposition 5. *The* (R_1, R_2)*-balance of* $\mathbf{x}$ *is*

$$x^*_{(R_1,R_2)} = \sqrt{\frac{r_1 r_2}{r_1 + r_2}}\ \ln\left[\frac{g(x_j; j \in R_1)}{g(x_k; k \in R_2)}\right] = \ln\left[\frac{\left(\prod_{j\in R_1} x_j\right)^{\sqrt{r_2/(r_1(r_1+r_2))}}}{\left(\prod_{k\in R_2} x_k\right)^{\sqrt{r_1/(r_2(r_1+r_2))}}}\right], \tag{16}$$

where $g(\cdot)$ *denotes geometric mean of parts in the argument.*

Proof: Just agreeing with (4). □

Example 3. (Example 2 continued) The unitary composition $\mathbf{e}_1$ in (15) is a balancing element for $R_1 = \{1, 2\}$ and $R_2 = \{3\}$. In this case $Q = \emptyset$. Moreover, using (16), the $(\{1, 2\}, \{3\})$-balance is $x^*_1 = 6^{-1/2} \ln(10/9)$ as obtained previously. Note that other coordinates do not play any role in the balance. See other examples in (2).

A consequence of Definition 3 is that orthonormal bases associated with a sequential binary partition are made of balancing elements. Each element corresponds to groups obtained in each sequential order binary partition. The construction of bases associated with sequential binary partitions allows us to state the following proposition.

Proposition 6. *Let* $\mathbf{x}$ *be a composition in* $\mathcal{S}^n$ *and define two non-overlapping, non-empty, groups of parts* R_1 *and* R_2*. Let* $\{\mathbf{e}_i, i = 1, \ldots, n-1\}$ *be an orthonormal basis associated with a sequential binary partition in which, for a given sequential order, the* $R_1 \cup R_2$*-group is obtained, and the* R_1*-group and* R_2*-group are obtained subsequently. Then, this basis contains the* (R_1, R_2)*-balancing element. Moreover,* $\text{sub}(\mathbf{x}; R_1 \cup R_2)$ *has the following coordinates with respect to the basis associated with the* $R_1 \cup R_2$*-group:* x^*_j, $j \in R^*_1$*, are the* $r_1 - 1$ *coordinates*

of sub($\mathbf{x}$; R_1)*;* $x_k^* \in R_2^*$*, are the* $r_2 - 1$ *coordinates of* sub($\mathbf{x}$; R_2)*; and* $x^*_{(R_1,R_2)}$ *that is the* (R_1, R_2)*-balance*.

Proposition 6 means that the (R_1, R_2)-balance is the only information we need to recover the $(R_1 \cup R_2)$-subcomposition from the R_1 and R_2-subcompositions; it conveys the information about the weight of each group of parts inside their union, $(R_1 \cup R_2)$.

Example 4. Assume $n = 5$ and $R_1 = \{1, 2\}$ and $R_2 = \{3, 4\}$; then $Q = \{5\}$ and $\mathcal{S}^5(R_1 \cup R_2)$ is a three-dimensional subspace. Bases for $\mathcal{S}^5(R_i)$, $i = 1, 2$, are one dimensional and the two basis elements are

$$\mathbf{e}_1 = \mathcal{C}\left[\exp\left(\sqrt{\frac{1}{2}}, -\sqrt{\frac{1}{2}}, \ 0, \ 0, \ 0\right)\right], \quad \mathbf{e}_2 = \mathcal{C}\left[\exp\left(0, \ 0, \sqrt{\frac{1}{2}}, -\sqrt{\frac{1}{2}}, \ 0\right)\right],$$

where $R_1^* = \{1\}$, $R_2^* = \{2\}$, and the remaining dimension corresponds to the (R_1, R_2)-balancing element

$$\mathbf{e}_{(R_1,R_2)} = \mathcal{C}\left[\exp\left(\frac{1}{2}, \frac{1}{2}, -\frac{1}{2}, -\frac{1}{2}, \ 0\right)\right].$$

The corresponding balance is

$$x^*_{(R_1,R_2)} = \ln\left[\frac{(x_1x_2)^{1/2}}{(x_3x_4)^{1/2}}\right].$$

The sub-basis associated with the $R_1 \cup R_2$-group, $\mathbf{e}_{(R_1,R_2)}$, $\mathbf{e}_1$, $\mathbf{e}_2$, is obtained at the sequential orders 2, 3, 4 of the sequential binary partition $[x_1||x_2|||x_3|x_4||||x_5]$.

GROUPED PARTS ANALYSIS: METHOD AND EXAMPLES

The aim of the two previous sections has been to introduce some algebraic and geometric ideas on orthonormal basis and their respective coordinates when analyzing subcompositions and balances of grouped parts. However, one is interested on how to use these tools in a particular problem. This section is intended to describe an easy—albeit general—way of carrying out these analysis in practice and how to interpret them. A simulated example will be used to guide the reader through this methodological discussion.

Let us state a fictitious and simplified problem in compositional data analysis. Assume that, in an oil field, we have sampled the ternary composition of rocks in two different layers from some well-logs. We assume the volume of the rock is composed of three parts: water, part 1; oil, part 2; and solid, part 3. For layer

X, the sample compositions are $\mathbf{x}_i = [x_{i1}, x_{i2}, x_{i3}]$, with sample size $m = 20$; similarly, samples from layer Y are $\mathbf{y}_i = [y_{i1}, y_{i2}, y_{i3}]$, again with sample size $m = 20$. Previous experiences point out that these compositions are distributed as additive logistic normal (ALN; see Aitchison, 2003a; Mateu-Figueras, 2003). Moreover, variabilities of both samples are different and approximately known. We also fully assume that Aitchison distances are appropriate to model distances between compositions. This means, for instance, that compositions with an intrinsic null component are very far (infinite distance) from current compositions; or that [0.20, 0.01, 0.79] (water, oil, solid) is quite different from [0.20, 0.02, 0.78] (double oil content), whereas [0.15, 0.15, 0.70] and [0.15, 0.16, 0.69] are quite similar, although the per-one (absolute) difference in oil content is exactly the same in both cases.

Our goal is to decide whether the centers of compositions of three parts in both layers are equal or not. However, in order to study the possibilities of oil extraction we are also interested in solid–liquid ratios. Therefore, we propose to test for equal means the balance between the groups of parts $\{1, 2\} = \{\text{water, oil}\} = \{\text{grouped liquids}\}$ and $\{3\} = \{\text{solid}\}$ obtained in the first-order sequential binary partitions $[x_1|x_2||x_3]$, respectively $[y_1|y_2||y_3]$. This test is an univariate one. The standard approach would be the amalgamation of water and oil (liquid = water + oil), which also treats the hypothesis testing of equal centers in both layers as an univariate problem (a single coordinate in $\mathcal{S}^2$).

The three proposed tests would thus be as follows:

(A) The bivariate one in $\mathcal{S}^3$, which states that, on (geometric) average, both layers X and Y are equal

$$\mathrm{H}_0 : \mathrm{Cen}[\mathbf{x}] = \mathrm{Cen}[\mathbf{y}], \qquad \mathrm{H}_1 : \mathrm{Cen}[\mathbf{x}] \neq \mathrm{Cen}[\mathbf{y}],$$

where H_0 is the null hypothesis and H_1 the alternative.

(B) The balance test, which states that, on (geometric) average, the ratio of grouped liquids and solid in both layers X and Y is equal

$$\mathrm{H}_0 : \mathrm{E}[x_1^*] = \mathrm{E}[y_1^*], \qquad \mathrm{H}_1 : \mathrm{E}[x_1^*] \neq \mathrm{E}[y_1^*],$$

where $x_1^* = (1/\sqrt{6})\ \ln((x_1 x_2)/x_3^2)$ in layer X. Analogously, for layer Y.

(C) The test on the log-ratio obtained after amalgamation, which states that, on (geometric) average, the ratio of (liquid = water + oil) and solid in both layers X and Y is equal

$$\mathrm{H}_0 : \mathrm{E}[x_{\mathrm{am}}^*] = \mathrm{E}[y_{\mathrm{am}}^*], \quad \mathrm{H}_1 : \mathrm{E}[x_{\mathrm{am}}^*] \neq \mathrm{E}[y_{\mathrm{am}}^*],$$

where $x^*_{\text{am}} = (1/\sqrt{2})\ \ln((x_1 + x_2)/x_3)$ is the only orthonormal coordinate in $\mathcal{S}^2$ after amalgamation in layer X and, analogously, for layer Y.

These tests can be carried out using several statistics. For reasons of homogeneity, we used the generalized likelihood-ratio test for the three cases, assuming normality with known, but different, variances–covariances. The sample data and main parameters are shown in Table 3. In the bivariate test, both couples (x^*_1, x^*_2) and (y^*_1, y^*_2) are assumed independent and their respective standard deviations are (0.4, 0.4) and (0.05, 0.05). Means of these coordinates and the corresponding centers are shown in Table 3 under the heading *theoretical*. Although the standard way of defining ALN parameters is based on the variances–covariances of alr log-ratios (Aitchison, 2003a), we use the parametrization in terms of orthonormal coordinates as introduced by Mateu-Figueras (2003).

For the univariate test on balances, the marginals for x^*_2 and y^*_2 are used. For the univariate test on the amalgamated compositions, the coordinates x^*_{am} and y^*_{am} were assumed to be normally distributed, being the variances 0.12 for layer X and 0.0027 for layer Y. We remark that the hypothesis of normality of log-ratios including amalgamated parts is incompatible with the assumed ALN distribution of the three-part composition. Although this is an additional inconvenience in using amalgams, it is not a point in this discussion. The theoretical means of x^*_{am} and y^*_{am} are -1.24 and -1.23, respectively.

From the theoretical parameters used in the simulation, we easily conclude that the two samples should be centered at quite different points, although part 3 (solid) is approximately the same in the two layers. This is confirmed by the results of the three tests shown in Table 3.

The bivariate test takes into account the log-ratios of the three parts and rejects null hypothesis (A, earlier) because the centers are very different (with respect to dispersion). The univariate test using balance rejects null hypothesis (B) accordingly. The test statistic is using the information of log-ratios relating the group {(water,oil)} with the group {(solid)}—i.e. the inter-group information—but not the intra-group log-ratio corresponding to the subcomposition {water,oil}. This means, that the test statistic is a function of the log-ratios water–solid and oil–solid, but not of the log-ratio water–oil. As the ratios of inter-group components, water–solid and oil–solid, are quite different, it also rejects the null hypothesis. Contrarily, the univariate test on the amalgamated coordinate does not reject null hypothesis (C). In fact, the statistic used ignores both the intra-group {water,oil} or subcompositional information and also most of the inter-group ratios. The remaining information, the amalgamated coordinates, x^*_{am} and y^*_{am}, is quite similar in both layers and the test passes the null hypothesis.

This example reveals that the amalgamated test may strongly disagree with the bivariate (A) and balance (B) tests. Although the centers of the two layers are quite different, this feature is lost in the amalgamation. The question arises,

Table 3. Simulated Samples and Parameters for Testing a Lattice of Hypothesis

20 samples						Center		Mean of coordinates	
Layer A			LayerB			Theoretical			
x_1	x_2	x_3	y_1	y_2	y_3			Layer A	
0.0193	0.0754	0.7509	0.0031	0.1425	0.8544	0.0736	0.0736	0.8528	0.0000	−2.0000
0.1003	0.0554	0.8443	0.0033	0.1561	0.8407			Layer B		
0.0637	0.1466	0.7897	0.0030	0.1462	0.8508	0.0032	0.1457	0.8511	−2.7000	−3.0000
0.1009	0.1010	0.7982	0.0032	0.1419	0.8548			Sample		
0.1922	0.1364	0.6714	0.0031	0.1649	0.8320			Layer A		
0.1104	0.0343	0.8553	0.0036	0.1541	0.8423	0.0762	0.0742	0.8496	0.0196	−1.9798
0.0753	0.0599	0.8648	0.0033	0.1495	0.8472			Layer B		
0.0515	0.0612	0.8873	0.0029	0.1395	0.8576	0.0033	0.1465	0.8503	−2.6886	−2.9884
0.0559	0.0701	0.8740	0.0032	0.1485	0.8483					
0.1023	0.0820	0.8157	0.0032	0.1415	0.8553					
0.1720	0.0954	0.7326	0.0034	0.1483	0.8483					
0.0509	0.0686	0.8805	0.0032	0.1320	0.8648					
0.0483	0.0484	0.9033	0.0035	0.1525	0.8440					
0.0550	0.0477	0.8973	0.0033	0.1488	0.8479			Bivariate test		
0.0471	0.0834	0.8695	0.0031	0.1434	0.8535	$\chi^2_2 = 1028$		p-value $= 0.000$		
0.0896	0.1006	0.8098	0.0034	0.1434	0.8532			Univariate test on balance		
0.0632	0.1280	0.8087	0.0034	0.1514	0.8452	$\chi^2_1 = 125.2$		p-value $= 0.000$		
0.0801	0.0967	0.8232	0.0032	0.1320	0.8648			univariate test of log-ratio on amalgam		
0.0582	0.0257	0.9162	0.0036	0.1507	0.8457	$\chi^2_1 = 0.002$		p-value $= 0.962$		

which of these tests are appropriate for the analysis. The answer is that they are best suited to answering different questions and they should not be compared. The amalgamated problem ignores the existence of two different parts in the liquid group. Only liquid–solid ratio is considered, for instance, to evaluate drilling costs roughly based on this proportion. But the amalgamated data cannot explain anything about compositional problems involving oil and water separately. Ratios involving these parts are lost after amalgamation.

Alternatively, tests (A) and (B) may be steps in a lattice of hypothesis tests trying to decide a parsimonious model for the centers of the two layers (Aitchison, 2003a, p. 149). The interest is centered in the whole composition of three parts (water, oil, solid) and, possibly, in their subcompositions. An example of such a lattice may start testing (A). Acceptance of such a null hypothesis represents the simplest model, centers of the two layers are equal. If (A) is rejected, we proceed with (B). Acceptance of (B) leads to conclude that the ratio of the center of water–oil subcomposition over solid is not responsible of differences between layers X and Y. The lattice may conclude by testing the equal center of the subcomposition {water,oil}. Although this test is not presented here, parameter values directly suggest rejection of the null hypothesis.

When following this lattice of hypothesis testing, the tests involved should be compatible and based on the same composition, i.e. the log-ratios involved cannot change from test to test. This excludes test (C) from such a lattice and test (B) appears as a natural alternative.

Let us turn back to a more general methodological framework by stating a general but common problem in compositional data analysis. Each individual in a sample is characterized by a row vector of n parts $\mathbf{x} = [x_1, x_2, \ldots, x_n]$. We would like to estimate basic statistics: the center, some measure of dispersion or fit a probability distribution. Additionally, some standard statistics may be required, e.g. testing equal centers of two samples, cluster analysis, etc. Suppose we also wish to compare the behavior of two, three, or more groups of parts, and we wish to undertake compositional analysis within one or more of the groups (subcompositions). Finally, summary representations of grouped-part data and subcompositional data are required to complete the exploratory analysis and to show results.

The implicit or explicit goal when grouping parts of a composition is to reduce dimension to facilitate interpretation. There are three ways of doing this: analyze a subcomposition, analyze inter-group balances or, alternatively, amalgamate some parts. This latter alternative, being standard, implies changing the initial problem and original parts, after amalgamation, do not play any role in the new problem. This is not the case in subcompositional and balance analysis, where the initial problem remains unaltered when we deal with it in a reduced dimension. The reason is simple: within the Aitchison geometry, amalgamation is a non-linear projection, while subcompositional and balance analysis plays the same role as usual orthogonal, and thus linear, projections in multivariate analysis.

Standard techniques in compositional data analysis proceed in two or more separate steps, although their order may be changed. Analysis unrelated to groups is then carried out as one of the steps. The other steps deal with questions related to groups. Some of them are related to subcompositions and they are treated by extracting subcompositions from raw data and then applying the appropriate statistical techniques. Other questions are related to groups and they are faced by amalgamating parts in each group and, then, the compositional data analysis on the amalgamated sample is carried out (Aitchison, 2003, pp. 38–40). As mentioned in the "Introduction" section and previous examples, an analysis based on amalgamated groups can lead to conclusions which disagree with those obtained with non-amalgamated data.

The present alternative starts seeking for an intuitive and natural partition of compositional vectors into groups of parts. These groups should be the target of the analysis, both from the subcompositional and balance points of view. An adequate reordering of parts, keeping affine parts contiguous, may facilitate this task. It is also advisable to decide a sequential order of importance, when separating two groups of parts, to arrive to the desired partition. Then, the process of subdividing the partition is continued until a partition in individual parts is obtained. This latter binary partitions can be done arbitrarily if no additional intuitive criterion is given. These decisions are equivalent to the design of a sequential binary partition of the ordered parts. Obviously, these decisions depend on the particular stated problem and the preferences of the analyst.

From the sequential binary partition, we obtain the associated orthonormal basis and all sample data can be represented by their coordinates in such a basis. Note that the particular expression of the used orthonormal basis is not necessary; coordinates for each individual can be directly obtained using log-ratios (4). Standard multivariate analysis dealing with original parts can be carried out using this coordinate representation of the data. Aitchison distances between individuals are now obtained as ordinary Euclidean distances between vectors of coordinates. The center of a sample, expressed in coordinates, is readily obtained by averaging these sample coordinates. A cluster analysis can also be carried out on the coordinates with standard methods when analyzing real multivariate data. Testing equal centers of two sub-populations, or testing goodness of fit to an ALN distribution are reduced to standard tests on equal means, respectively on multivariate normality.

Analysis dealing with subcompositions corresponding to groups defined in the sequential binary partition can be performed on the coordinates associated with those groups; neither the closure of data in each step, nor re-computation of coordinates, is required. This is due to the fact that the basis associated with the sequential binary partition contains the sub-basis associated with the group, and they constitute the coordinates of the subcomposition.

The representation of data in coordinates calculated with respect to the basis associated with the sequential binary partition is also readily used to analyze

balances between groups. We only need to identify those elements in the basis which are balancing elements for our groups and, then, to select the corresponding coordinates. In this way, the analysis of disaggregated parts and grouped parts are immediately compatible because they are based on a single coordinate system, which maintains the metric properties of compositional data.

CONCLUSIONS

The introduction of orthonormal bases in the simplex and the corresponding coordinates by Egozcue and others (2003) allows one to select suitable orthonormal bases in order to facilitate interpretation of results. We introduce the sequential binary partitioning of parts of a composition as a tool to design a particular basis in the simplex. Such bases make the corresponding coordinates directly interpretable as balances between two groups of parts appearing in some order of the sequential binary partition.

An important point in these techniques is that they allow for a simultaneous and compatible analysis of intra-group of parts (subcompositional analysis) and of inter-group of parts (balance analysis). Both points of view are reduced to orthogonal projections into subspaces of the simplex, thus guaranteeing consistency of distances and statistical analysis when working in a reduced dimension.

We conclude that amalgamation of parts changes the original problem and cannot be considered as a compatible reduction of dimension. However, the newly stated problem may make full sense by itself, i.e. anamalgamated part should be clearly interpretable and the ratio of it with respect to other parts should be relevant in the new problem.

APPENDIX: SUMMARY OF AITCHISON GEOMETRY IN THE SIMPLEX

The fundamental ideas leading to what we call now *Aitchison geometry in the simplex* were already set by Aitchison (1982, 1986). Recent results were introduced in several papers (Aitchison and others, 2002; Billheimer, Guttorp, and Fagan, 2001; Pawlowsky-Glahn and Egozcue, 2001, 2002; Egozcue and others, 2003).

Simplex of n-parts. Compositional vectors, or simply compositions, of n parts are real vectors whose positive components add up to a closure constant $\kappa > 0$. This set is called simplex of n parts and is formally written

$$\mathcal{S}^n = \left\{ [x_1, x_2, \ldots, x_n] \;\middle|\; x_i > 0, i = 1, \ldots, n\,;\; \sum_{i=1}^{n} x_i = \kappa \right\},$$

where square brackets are used to denote row vectors. A first assumption in compositional data analysis is that the value of κ is irrelevant and only the ratios between

components convey compositional information. The closure operation transforms a positive vector of n positive components into a vector in $\mathcal{S}^n$, maintaining all ratios between components. It is defined as

$$\mathcal{C}(\mathbf{z}) = \left[\frac{\kappa z_1}{\sum_{i=1}^n z_i}, \frac{\kappa z_2}{\sum_{i=1}^n z_i}, \ldots, \frac{\kappa z_n}{\sum_{i=1}^n z_i} \right],$$

where $\mathbf{z} = [z_1, \ldots, z_n]$ is any positive n-vector. Note that components cannot be null once accepted that compositional information is based on ratios of components.

Perturbation, internal operation in $\mathcal{S}^n$. Perturbation in the simplex plays a role analogous to the sum in real spaces. If $\mathbf{x}$ and $\mathbf{y}$ are in $\mathcal{S}^n$, the perturbation is

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}[x_1 y_1, x_2 y_2, \ldots, x_n y_n].$$

Perturbation is a commutative group operation, i.e. it is associative and commutative; the neutral element is $\mathbf{n}_n = \mathcal{C}[1, 1, \ldots, 1]$ and the opposite element of $[x_1, \ldots, x_n]$ is $\ominus\mathbf{x} = \mathcal{C}[1/x_1, \ldots, 1/x_n]$. Perturbation has been denoted by other symbols like $(\circ)$ in other publications, e.g. Aitchison (1986).

Powering, external operation. It operates real numbers, $\alpha \in \mathbb{R}$, with compositional vectors, $\mathbf{x} \in \mathcal{S}^n$. It is analogous to constant multiplication in vector spaces. It is defined as

$$\alpha \odot \mathbf{x} = \mathcal{C}\left[x_1^\alpha, x_2^\alpha, \ldots, x_n^\alpha\right].$$

Powering, also called power transformation, has been denoted by other symbols $(\diamond, \otimes)$ in other publications, e.g. Aitchison and others (2002), Pawlowsky-Glahn and Egozcue, (2001, 2002), and Egozcue and others (2003).

Perturbation and powering give to $\mathcal{S}^n$ a linear vector space structure, which dimension is $n - 1$. Powering is distributive with respect to perturbation, and the unitary element in $\mathbb{R}$ is 1. Moreover, the opposite element of $\mathbf{x}$ can be expressed as $\ominus\mathbf{x} = (-1) \odot \mathbf{x}$.

R-subcomposition. Let R be a subset of the indexes $\{1, 2, \ldots, n\}$ with $r = \text{Card}(R)$ and let $\mathbf{x} \in \mathcal{S}^n$ be a composition. We define R-subcomposition, $\mathbf{u}$, the closed set of r parts of $\mathbf{x}$, $1 < r < n$, which subscripts are in R:

$$\mathbf{u} = \text{sub}(\mathbf{x}; R) = \mathcal{C}[x_i, x_j, \ldots, x_k], \quad i, j, \ldots, k \in R, \quad \mathbf{u} \in \mathcal{S}^r. \tag{A.1}$$

The R-subcomposition operation is linear with respect to perturbation and powering, i.e. for any real constants α, β, and for $\mathbf{x}_i \in \mathcal{S}^n$, $i = 1, 2$,

$$\text{sub}((\alpha \odot \mathbf{x}_1) \oplus (\beta \odot \mathbf{x}_2); R) = (\alpha \odot \text{sub}(\mathbf{x}_1; R)) \oplus (\beta \odot \text{sub}(\mathbf{x}_2; R)). \tag{A.2}$$

Straight lines in $\mathcal{S}^n$. Let be $\mathbf{x}_0$ and $\mathbf{z}$ compositions in $\mathcal{S}^n$. For real values α, compositions $\mathbf{x} = \mathbf{x}_0 \oplus (\alpha \odot \mathbf{z})$ define a straight line, which direction is given by $\mathbf{z}$, and contains the point $\mathbf{x}_0$.

Inner product in $\mathcal{S}^n$. Let $\mathbf{x} = [x_1, \ldots, x_n]$ and $\mathbf{y} = [y_1, \ldots, y_n]$ be compositions in $\mathcal{S}^n$. Their inner product is

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \sum_{i=1}^{n} \ln \frac{x_i}{g(\mathbf{x})} \ln \frac{y_i}{g(\mathbf{y})} = \frac{1}{n} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j}, \tag{A.3}$$

where $g(\cdot)$ denotes geometric mean of components. The subscript a refers to *Aitchison geometry*, in order to distinguish it from the standard inner product in $\mathbb{R}^n$. It satisfies the standard properties: distributive with respect to perturbation, linear-powering in both arguments and commutative. The inner product gives an Euclidean structure to the simplex or, equivalently, a finite-dimensional Hilbert-space structure.

Norm and distance. The inner product is used to define a norm and a distance in the simplex. For $\mathbf{x}$ and $\mathbf{y}$ compositions in $\mathcal{S}^n$ we obtain

$$\|\mathbf{x}\|_a^2 = \langle \mathbf{x}, \mathbf{x} \rangle_a, \quad d_a(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} \ominus \mathbf{y}\|_a.$$

Norm and distance configure $\mathcal{S}^n$ as a finite-dimensional normed (Banach) and metric space. The definition of Aitchison distance (Aitchison, 1986) was previous to inner product and norm. It was defined taking into account remarkable properties that are naturally associated with compositional data. Among those properties, invariance under perturbation $d_a(\mathbf{z} \oplus \mathbf{x}, \mathbf{z} \oplus \mathbf{y}) = d_a(\mathbf{x}, \mathbf{y})$; invariance under permutation of parts; and subcompositional dominance, $d_a(\mathbf{x}, \mathbf{y}) \geq d_a(\text{sub}(\mathbf{x}; R), \text{sub}(\mathbf{y}; R))$, play a major role.

Linear bases and coordinate transformations. Elements of $\mathcal{S}^n$ are readily expressed as perturbation-linear combinations of compositional vectors in a basis or a set of generators of the space. Coefficients (or coordinates) of compositions in these combinations represent elements of the simplex as real vectors. Real vectors of these coefficients are called transformations because they map $\mathcal{S}^n$ onto $\mathbb{R}^{n-1}$ or a subspace of $\mathbb{R}^n$. The use of orthonormal bases is natural, but, historically, oblique bases and oblique generator systems have been implicitly used. We mention the alr basis and the clr system of generators and their respective transformation into coefficients. Orthonormal bases and ilr transformation are also given.

The vectors,

$$\mathbf{e}_i^{\text{alr}} = \sqrt{\frac{n-1}{n}} \odot \mathcal{C}\left[\exp\left(\frac{-1}{n-1}, \ldots, \frac{-1}{n-1}, \underbrace{1}_{i\text{-th element}}, \frac{-1}{n-1}, \ldots, \frac{-1}{n-1}\right)\right],$$

for $i = 1, 2, \ldots, n-1$, constitute a unitary basis of $\mathcal{S}^n$ because these vectors are perturbation-linear independent. However, they are not orthogonal. In order to express $\mathbf{x}$ as a perturbation-linear combination of $\mathbf{e}_i^{\text{alr}}$, the coefficients are obtained as inner products with the dual basis elements. This dual basis (unitary) is

$$\mathbf{e}_i^{\text{dalr}} = \frac{1}{\sqrt{2}} \odot \mathcal{C}\left[\exp\left(0, \ldots, 0, \underbrace{1}_{i\text{-th element}}, 0, \ldots, 0, -1\right)\right], \quad i = 1, 2, \ldots, n-1.$$

This basis is not orthogonal because, for $i \neq j$, $\langle \mathbf{e}_i^{\text{dalr}}, \mathbf{e}_j^{\text{dalr}} \rangle_a = 1/2$; this means that each couple of vectors in both bases—alr and dalr—form angles of $\pi/3$ irrespective of the number of parts. The expression of a composition is then

$$\mathbf{x} = \bigoplus_{i=1}^{n-1} \left(c_i^{\text{alr}} \odot \mathbf{e}_i^{\text{alr}}\right), \quad c_i^{\text{alr}} = \left\langle \mathbf{x}, \mathbf{e}_i^{\text{dalr}} \right\rangle_a = \ln \frac{x_i}{x_n}.$$

The additive-log-ratio transformation (alr) of a composition $\mathbf{x} \in \mathcal{S}^n$ leads to the real $(n-1)$-vector of coordinates

$$\text{alr}(\mathbf{x}) = \left[c_1^{\text{alr}}, c_2^{\text{alr}}, \ldots, c_{n-1}^{\text{alr}}\right] = \left[\ln \frac{x_1}{x_n}, \ldots, \ln \frac{x_{n-1}}{x_n}\right], \qquad \text{alr}(\mathbf{x}) \in \mathbb{R}^{n-1}.$$

Centered log-ratio transformation (clr) is obtained when $\mathbf{x}$ is expressed as a perturbation-linear combination of a unitary generator system

$$\mathbf{e}_i^{\text{clr}} = \sqrt{\frac{n}{n-1}} \odot \mathcal{C}\left[\exp\left(0, \ldots, 0, \underbrace{1}_{i\text{-th element}}, 0, \ldots, 0\right)\right], \quad i = 1, 2, \ldots, n,$$

and then $\mathbf{x} = \bigoplus_{i=1}^{n} (\alpha_i \odot \mathbf{e}_i^{\text{clr}})$. Although the coefficients α_i of a generator system are not unique, they are defined as

$$\text{clr}(\mathbf{x}) = [\alpha_1, \alpha_2, \ldots, \alpha_n] = \left[\ln \frac{x_1}{g(\mathbf{x})}, \ldots, \ln \frac{x_n}{g(\mathbf{x})}\right], \quad \text{clr}(\mathbf{x}) \in \mathbb{R}^n,$$

where $g(\cdot)$ denotes the geometric mean of the arguments. This choice of the coefficients α_i has some remarkable properties, e.g. it gives a straightforward way of computing Aitchison norms and distances.

$$\|\text{clr}(\mathbf{x})\| = \|\mathbf{x}\|_a; \qquad d(\text{clr}(\mathbf{x}_1), \text{clr}(\mathbf{x}_2)) = d_a(\mathbf{x}_1, \mathbf{x}_2),$$

where $\|\cdot\|$ and $d(\cdot, \cdot)$ are the ordinary norm and distance in $\mathbb{R}^n$, respectively.

Orthonormal bases in $\mathcal{S}^n$ allow the expression of compositions in Cartesian coordinates. Many orthonormal bases can be obtained, see for instance "Orthonormal basis of a sequential binary partition" section. A standard one is

$$\mathbf{e}_i = \mathcal{C}\left[\exp\left(\underbrace{\sqrt{\frac{1}{i(i+1)}},\ldots,\sqrt{\frac{1}{i(i+1)}}}_{i \text{ elements}},-\sqrt{\frac{i}{(i+1)}},0,\ldots,0\right)\right], \quad \text{(A.4)}$$

for $i = 1, 2, \ldots, n-1$. Then, $\mathbf{x}$ is reconstructed using the corresponding coordinates

$$\mathbf{x} = \bigoplus_{i=1}^{n-1} (c_i \odot \mathbf{e}_i), \qquad c_i = \langle \mathbf{x}, \mathbf{e}_i \rangle_a = \sqrt{\frac{i}{i+1}}\ \ln\left[\frac{g(x_1,\ldots,x_i)}{x_{i+1}}\right].$$

Transformation of $\mathbf{x} \in \mathcal{S}^n$ into its coordinates with respect to the orthogonal basis has been called *isometric log-ratio transformation,* (ilr), and it is

$$\mathbf{x}^* = \text{ilr}(\mathbf{x}) = [c_1, c_2, \ldots, c_{n-1}], \quad \text{ilr}(\mathbf{x}) \in \mathbb{R}^{n-1}.$$

Coordinates of a composition with respect to (A.4) or any other given orthonormal basis are denoted using (*) in this development, e.g., $c_i = x_i^*, i = 1, 2, \ldots, n-1$.

The main property of the representation of compositions by their coordinates with respect to an orthonormal basis is that the whole Aitchison geometry of compositions in the simplex is reduced to the ordinary Euclidean geometry in $\mathbb{R}^{n-1}$ for their coordinates. This allows direct statistical analysis in the simplex (Pawlowsky-Glahn, 2003). If $\mathbf{x}, \mathbf{y} \in \mathcal{S}^n$ and $\mathbf{x}^*, \mathbf{y}^*$ are their respective coordinate vectors with respect to an orthonormal basis, the isometry implies:

$$\text{ilr}(\mathbf{x} \oplus \mathbf{y}) = \mathbf{x}^* + \mathbf{y}^*; \qquad \text{ilr}(\alpha \odot \mathbf{x}) = \alpha \cdot \mathbf{x}^*,$$
$$\|\mathbf{x}^*\| = \|\mathbf{x}\|_a; \qquad d(\mathbf{x}^*, \mathbf{y}^*) = d_a(\mathbf{x}, \mathbf{y}); \qquad \langle \mathbf{x}^*, \mathbf{y}^* \rangle = \langle \mathbf{x}, \mathbf{y} \rangle_a,$$

where norm, distance and inner product without subscript refer to ordinary Euclidean ones in $\mathbb{R}^{n-1}$.

ACKNOWLEDGMENTS

The authors would like to thank the referees, D. Billheimer, H. Schaeben, and R. Howarth for their constructive comments, which helped to improve the paper. This research has received financial support from the *Dirección General*

de Investigación of the Spanish Ministry for Science and Technology through the project BFM2003-05640/MATE and from the *Departament d'Universitats, Recerca i Societat de la Informació* of the *Generalitat de Catalunya* through the project 2003XT 00079.

REFERENCES

Aitchison, J., 1982, The statistical analysis of compositional data (with discussion): J. R. Stat. Soc. B (Stat. Methodol.), v. 44, no. 2, p. 139–177.

Aitchison, J., 1992, On criteria for measures of compositional difference: Math. Geol., v. 24, no. 4, pp. 365–379.

Aitchison, J., 2003a, The statistical analysis of compositional data (reprint): Blackburn Press, Caldwell, NJ, 416 p.

Aitchison, J., 2003b, Compositional data analysis: Where are we and where should we be heading? See Thió-Henestrosa and Martín-Fernández (2003).

Aitchison, J., Barceló-Vidal, C., Egozcue, J. J., and Pawlowsky-Glahn, V., 2002, A concise guide for the algebraic–geometric structure of the simplex, the sample space for compositional data analysis, *in* Bayer U., Burger H., and Skala W., eds., Proceedings of IAMG'02—The Eighth Annual Conference of the International Association for Mathematical Geology, Terra Nostro, no. 3, p. 387–392.

Barceló-Vidal, C., 2000, Fundamentación matemática del análisis de datos composicionales: Technical Report IMA 00-02-RR, Departament d'Informática i Matemática Aplicada, Universitat de Girona, Spain, 77 p.

Barceló-Vidal, C., Martín-Fernández, J. A., and Pawlowsky-Glahn, V., 2001, Mathematical foundations of compositional data analysis, *in* Ross G., ed., Proceedings of IAMG'01—The Sixth Annual Conference of the International Association for Mathematical Geology, CO-ROM, 20 p.

Billheimer, D., Guttorp, P., and Fagan, W., 2001, Statistical interpretation of species composition: J. Am. Stat. Assoc., v. 96, p. 1205–1214.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G., and Barceló-Vidal, C., 2003, Isometric logratio transformations for compositional data analysis: Math. Geol., v. 35, no. 3, p. 279–300.

Mateu-Figueras, G., 2003, Models de distribució sobre el símplex. Ph.D. thesis, Universitat Politècnica de Catalunya, Barcelona, Spain.

Pawlowsky-Glahn, V., 2003, Statistical modelling on coordinates. See Thió-Henestrosa and Martín-Fernández (2003).

Pawlowsky-Glahn, V. and Egozcue, J. J., 2001, Geometric approach to statistical analysis on the simplex: Stochastic Environ. Res. Risk Assess. (SERRA), v. 15, no. 5, p. 384–398.

Pawlowsky-Glahn, V. and Egozcue, J. J., 2002, BLU estimators and compositional data: Math. Geol., v. 34, no. 3, p. 259–274.

Thió-Henestrosa, S. and Martín-Fernández, J. A., eds., 2003, Compositional Data Analysis Workshop—CoDaWork'03, Proceedings, Universitat de Girona, CD-ROM, ISBN 84-8458-111-X, http://ima.udg.es/Activitats/CoDaWork03/.