Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **47** (2008) 77–82

Remark on Properties of Bases for Additive Logratio Transformations of Compositional Data

Karel HRON

Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science, Palacký University
Tomkova 40, 779 00 Olomouc, Czech Republic
e-mail: hronk@seznam.cz

(Received January 9, 2008)

Abstract

The statistical analysis of compositional data, multivariate data when all its components are strictly positive real numbers that carry only relative information and having a simplex as the sample space, is in the state-of-the-art devoted to represent compositions in orthonormal bases with respect to the geometry on the simplex and thus provide an isometric transformation of the data to an usual linear space, where standard statistical methods can be used (e.g. [2], [4], [5], [9]). However, in some applications from geosciences ([14]) or statistical aspects of multicriteria evaluation theory ([13]) it seems to be convenient to use another types of bases. This paper is devoted to describe its basic properties and illustrate the results on an example.

Key words: Aitchison geometry on the simplex; bases on the simplex; additive logratio transformations.

2000 Mathematics Subject Classification: 15A03, 62H99

1 Simplicial geometry

The concept of compositional data and its geometry on the simplex (called *Aitchison geometry*) is the starting point for building up statistical models for

such data. This short course follows earlier developements of compositional data ([1]) and cites the present results of the active research, as summarized in [8], [11] and [12].

Definition 1 A row vector, $\boldsymbol{x} = (x_1, \ldots, x_D)$, is called *D-parts composition* when all its components are strictly positive real numbers and they carry only relative information.

The assertion that D-parts composition (or only composition in short) carry only relative information means that all the relevant information is contained in the ratios among the parts, i.e. if c is a nonzero real number, $(x_1, \ldots, x_D)$ and $(cx_1, \ldots, cx_D)$ convey essentially the same information. A way to simplify the use of compositions is to represent them in closed form, i.e. as positive vectors with constant sum κ (usually 1 or 100 in case of percentages) of the parts. As a consequence, D-parts composition can be identified with the following vector:

Definition 2 For any composition $\boldsymbol{x}$, the *closure operation of* $\boldsymbol{x}$ *to the constant* κ is defined as

$$\mathcal{C}(\boldsymbol{x}) = \left(\frac{\kappa x_1}{\sum_{i=1}^{D} x_i}, \ldots, \frac{\kappa x_D}{\sum_{i=1}^{D} x_i} \right).$$

Proposition 1 *The sample space of compositional data is the simplex, defined as*

$$\mathcal{S}^D = \left\{ \boldsymbol{x} = (x_1, \ldots, x_D),\ x_i > 0,\ \sum_{i=1}^{D} x_i = \kappa \right\}.$$

The basics of Aitchison geometry on the simplex are mentioned below:

Definition 3 *Perturbation* of a composition $\boldsymbol{x} = \mathcal{C}(x_1, \ldots, x_D) \in \mathcal{S}^D$ by a composition $\boldsymbol{y} = \mathcal{C}(y_1, \ldots, y_D) \in \mathcal{S}^D$ is a composition

$$\boldsymbol{x} \oplus \boldsymbol{y} = \mathcal{C}(x_1 y_1, \ldots, x_D y_D).$$

Definition 4 *Power transformation* of a composition $\boldsymbol{x} \in \mathcal{S}^D$ by a constant $\alpha \in \mathbb{R}$ is a composition

$$\alpha \odot \boldsymbol{x} = \mathcal{C}(x_1^\alpha, \ldots, x_D^\alpha).$$

Proposition 2 *The simplex with the perturbation operation and the power transformation,* $(\mathcal{S}^D, \oplus, \odot)$*, is a vector space.*

The analogy between real vector space and the simplex leads to a definition of compositional (straight) line, based on operations of perturbation and power transformation, as the compositions $\boldsymbol{x}(t)$, $t \in \mathbb{R}$, satisfying

$$\boldsymbol{x}(t) = \boldsymbol{x}_0 \oplus (t \odot \boldsymbol{u}),$$

with starting point $\boldsymbol{x}_0$ and with direction given by the composition $\boldsymbol{u}$.

Let us remark, that the neutral element is the composition $\boldsymbol{n} = \mathcal{C}(1, \ldots, 1) = (\frac{1}{D}, \ldots, \frac{1}{D})$. The vector structure of $\mathcal{S}^D$ allows us to use the concepts of linear dependence and independence.

Definition 5 A set of m compositions in $\mathcal{S}^D$, $\boldsymbol{x}_1, \dots \boldsymbol{x}_m$, is said to be *linearly perturbation-dependent* if there exist scalars $\alpha_1, \dots, \alpha_m$ not all zero, such that

$$(\alpha_1 \odot \boldsymbol{x}_1) \oplus \cdots \oplus (\alpha_m \odot \boldsymbol{x}_m) = \boldsymbol{n}.$$

If no such scalars exist, the set is called *linearly perturbation-independent.*

In simplex $\mathcal{S}^D$, the maximum of perturbation-independent compositions is $D-1$. Thus, $\mathcal{S}^D$ is a vector space of dimension $D-1$.

Definition 6 If compositions $\boldsymbol{e}_1, \dots, \boldsymbol{e}_{D-1}$ are perturbation-independent, they constitute a *(simplicial) basis* of $\mathcal{S}^D$, i.e. each composition $\boldsymbol{x} \in \mathcal{S}^D$ can be expressed as

$$\boldsymbol{x} = (\alpha_1 \odot \boldsymbol{e}_1) \oplus \cdots \oplus (\alpha_{D-1} \odot \boldsymbol{e}_{D-1})$$

for some coefficients α_i, $i = 1, \dots, D-1$, that are termed *coordinates* with respect to the basis.

For deeper investigation of the bases on the simplex, we introduce further the concepts of inner product and norm in Aitchison geometry that enable us to use concepts of orthogonality and orthonormality of the bases.

Definition 7 *Inner product* of $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^D$,

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_a = \frac{1}{D} \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \ln \frac{x_i}{x_j} \ln \frac{y_i}{y_j} = \sum_{i=1}^{D} \ln \frac{x_i}{g(\boldsymbol{x})} \ln \frac{y_i}{g(\boldsymbol{y})},$$

and *norm* of $\boldsymbol{x} \in \mathcal{S}^D$,

$$\|\boldsymbol{x}\|_a = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_a},$$

where $g(\boldsymbol{x}) = (x_1 \dots x_D)^{\frac{1}{D}}$ denotes the geometric mean of the parts of the compositional vector in the argument.

It is easy to see, that using orthonormal bases on the simplex, all operations and metric concepts like perturbation, power transformation, inner product and norm are translated into coordinates as ordinary vector operations (sum of two vectors and multiplication of a vector by a scalar). See [6], [7] for details.

As consequence of the mentioned concepts we obtain the following definition:

Definition 8 The cosine of the *angle* $\angle(\boldsymbol{x}, \boldsymbol{y})_a$ *between two compositions* $\boldsymbol{x}$, $\boldsymbol{y} \in \mathcal{S}^D$, satisfying $\boldsymbol{x} \neq \boldsymbol{n}$, $\boldsymbol{x} \neq \boldsymbol{y}$, is expressible as

$$\cos \angle(\boldsymbol{x}, \boldsymbol{y})_a = \frac{\langle \boldsymbol{x}, \boldsymbol{y} \rangle_a}{\|\boldsymbol{x}\|_a \|\boldsymbol{y}\|_a}.$$

2 Bases for additive logratio transformations

Let us have a generating system of compositions in the simplex, $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_D$, where $\boldsymbol{w}_i = \mathcal{C}(1, 1 \ldots, \mathrm{e}, \ldots, 1)$ (the number e, base of natural logarithm, is placed in the i-th column, $i = 1, \ldots, D$). Then, taking any $D-1$ vectors, we obtain a basis, e.g. $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{D-1}$, and any vector $\boldsymbol{x} \in \mathcal{S}^D$ can be written as

$$\boldsymbol{x} = \ln \frac{x_1}{x_D} \odot (\mathrm{e}, 1, \ldots, 1, 1) \oplus \ln \frac{x_2}{x_D} \odot (1, \mathrm{e}, 1, \ldots, 1) \oplus \ln \frac{x_{D-1}}{x_D} \odot (1, 1 \ldots, 1, \mathrm{e}).$$

The mentioned basis has the following properties:

Theorem 1 *Let $\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{D-1}$ be the basis defined above. Then for $1 \leq i, j \leq D-1$, $i \neq j$,*

$$\langle \boldsymbol{w}_i, \boldsymbol{w}_j \rangle_a = -\frac{1}{D}, \quad \|\boldsymbol{w}_i\|_a^2 = \frac{D-1}{D}, \quad \cos \angle(\boldsymbol{w}_i, \boldsymbol{w}_j)_a = -\frac{1}{D-1}.$$

Proof We use the inner product in the form

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_a = \sum_{k=1}^{D} \ln \frac{x_k}{g(\boldsymbol{x})} \ln \frac{y_k}{g(\boldsymbol{y})}$$

for any $\boldsymbol{x}$, $\boldsymbol{y} \in \mathcal{S}^D$. Thus, in our case,

$$\langle \boldsymbol{w}_i, \boldsymbol{w}_j \rangle_a = \sum_{k=1, k \neq i,j}^{D} \ln \frac{1}{\sqrt[D]{\mathrm{e}}} \ln \frac{1}{\sqrt[D]{\mathrm{e}}} + 2 \ln \frac{\mathrm{e}}{\sqrt[D]{\mathrm{e}}} \ln \frac{1}{\sqrt[D]{\mathrm{e}}} = \frac{D-2}{D^2} - \frac{2(D-1)}{D^2} = -\frac{1}{D}.$$

Analogously

$$\|\boldsymbol{w}_i\|_a^2 = \sum_{k=1, k \neq i,j}^{D} \left(\ln \frac{1}{\sqrt[D]{\mathrm{e}}} \right)^2 + \left(\ln \frac{\mathrm{e}}{\sqrt[D]{\mathrm{e}}} \right)^2 = \frac{D-1}{D^2} + \frac{(D-1)^2}{D^2} = \frac{D-1}{D}.$$

The value for $\cos \angle(\boldsymbol{w}_i, \boldsymbol{w}_j)_a = -\frac{1}{D-1}$ is a simple consequence. □

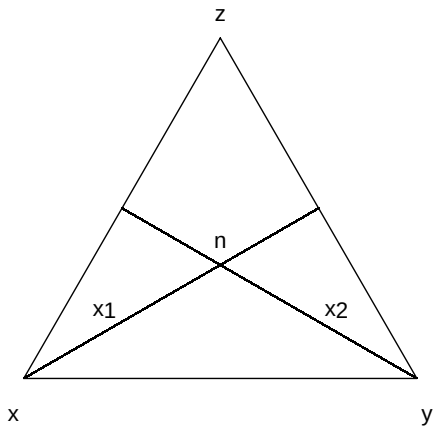
Example 1 In case of $D = 3$ we obtain $\boldsymbol{w}_1 = \mathcal{C}(e, 1, 1)$, $\boldsymbol{w}_2 = \mathcal{C}(1, e, 1)$, so thus $\|\boldsymbol{w}_1\|_a = \|\boldsymbol{w}_2\|_a = \frac{\sqrt{6}}{3}$ and $\angle(\boldsymbol{w}_1, \boldsymbol{w}_2)_a = 120°$. Compositional straight lines

$$\boldsymbol{x}_1(t) = \boldsymbol{n} \oplus (t \odot \boldsymbol{w}_1) = \mathcal{C}(\mathrm{e}^t, 1, 1),\ t \in \mathbb{R},$$

and

$$\boldsymbol{x}_2(s) = \boldsymbol{n} \oplus (s \odot \boldsymbol{w}_2) = \mathcal{C}(1, \mathrm{e}^s, 1),\ s \in \mathbb{R},$$

with neutral element $\boldsymbol{n} = \mathcal{C}(1, 1, 1)$ for starting points and directions given by $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$ are displayed on Figure 1. It is clear that their common composition is just the neutral element $\boldsymbol{n}$, obtained for $t = s = 0$.



Compositional lines $\boldsymbol{x}_1(t)$ and $\boldsymbol{x}_2(s)$ with neutral element $\boldsymbol{n}$ for starting points and directions given by $\boldsymbol{w}_1$ and $\boldsymbol{w}_2$.

The coefficients $\ln \frac{x_1}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D}$ of the above mentioned basis correspond to one member of the well known *additive logratio (alr) transformations* family, introduced by [1]. To obtain all the alr transformations, it is sufficient to choose by permutation another $D-1$ vectors from the generating system ([10]). We keep the basis chosen above, the considerations for the others are analogous. Thus, we denote by alr_D the transformation that gives the expression of a composition in additive logratio coordinates with the part x_D as ratioing part,

$$alr_D(\boldsymbol{x}) = \left(\ln \frac{x_1}{x_D}, \ln \frac{x_2}{x_D}, \ldots, \ln \frac{x_{D-1}}{x_D} \right) = \boldsymbol{y}.$$

The inverse of alr_D transformation, which gives the coordinates in the canonical basis of real space, is defined as

$$alr_D^{-1}(\boldsymbol{y}) = \mathcal{C}(\exp(y_1), \ldots, \exp(y_{D-1}), 1) = \boldsymbol{x}.$$

Let us emphasize that the alr_D (and also other transformations from the alr transformations family) is not isometric (its basis on the simplex is not orthonormal, see Theorem 1), i.e. metric concepts are not translated like as ordinary vector operations. On the other side, by many statistical methods this doesn't play a role and the remaining properties are sufficient (e.g. [1], [3], [9], for details). Moreover, the form of alr coordinates enables to use it by expert processes in multicriteria evaluation theory ([13]).

References

[1] Aitchison, J.: The statistical analysis of compositional data. *Chapman and Hall, London*, 1986.

[2] Aitchison, J., Greenacre, M.: *Biplots of compositional data.* Applied Statistics **51** (2002), 375–392.

[3] Billheimer, D.: *Compositional data in biomedical research.* In: Mateu-Figueras, G., Barceló-Vidal, C.: Compositional Data Analysis Workshop – CoDaWork'05, Proceedings, *Universitat de Girona*, 2005.

[4] Buccianti, A., Pawlowsky-Glahn, V.: *New perspectives on water chemistry and compositional data analysis.* Math. Geol. **37** (2005), 703–727.

[5] Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (eds): Compositional data analysis in the geosciences: From theory to practice. *Geological Society, London, Special Publications* **264**, 2006.

[6] Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueraz, G., Barceló-Vidal, C.: *Isometric logratio transformations for compositional data analysis.* Math. Geol. **35** (2003), 279–300.

[7] Egozcue, J. J., Pawlowsky-Glahn, V.: *Groups of parts and their balances in compositional data analysis.* Math. Geol. **37** (2005), 795–828.

[8] Egozcue, J. J., Pawlowsky-Glahn, V.: *Simplicial geometry for compositional data.* In: Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (eds.): Compositional data analysis in the geosciences: From theory to practice. *Geological Society, London, Special Publications* **264** (2006), 145–160.

[9] Filzmoser, P., Hron, K.: *Outlier detection for compositional data using robust methods.* Math. Geosci. **40** (2008), 233–248.

[10] Mateu-Figueras, G., Pawlowsky-Glahn, V., Barceló-Vidal, C.: *Distributions on the simplex.* In: Thió-Henestrosa, S., Martín-Fernández, J.A.: Compositional Data Analysis Workshop – CoDaWork'03, Proceedings, *Universitat de Girona*, 2003.

[11] Pawlowsky-Glahn, V., Egozcue, J. J.: *Geometric approach to statistical analysis on the simplex.* Stoch. Envir. Res. and Risk Ass. **15** (2001), 384–398.

[12] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, J.: Lecture notes on compositional data analysis. 2007, online.

[13] Talašová, J.: Fuzzy methods for multicriteria evaluation and decision making. *Publishing House of Palacký University, Olomouc*, 2006 (in Czech).

[14] Weltje, G. J.: *Ternary sandstone composition and provenance: an evaluation of the 'Dickinson model'.* In: Buccianti, A., Mateu-Figueras, G., Pawlowsky-Glahn, V. (eds.): Compositional data analysis in the geosciences: From theory to practice. *Geological Society, London, Special Publications* **264** (2006), 79–99.