Acta Univ. Palacki. Olomuc., Fac. rer. nat.,
Mathematica **48** (2009) 53–60

Analytical Representation of Ellipses in the Aitchison Geometry and Its Application

Karel HRON

Department of Mathematical Analysis and Applications of Mathematics
Faculty of Science, Palacký University
tř. 17. listopadu 12, 771 46 Olomouc, Czech Republic
e-mail: hronk@seznam.cz

(Received January 30, 2009)

Abstract

Compositional data, multivariate observations that hold only relative information, need a special treatment while performing statistical analysis, with respect to the simplex as their sample space ([1], [2], [3], [8], [9], [10], [11], [18]). For the logratio approach to the statistical analysis of compositional data the so called Aitchison geometry was introduced and confirmed to be the meaningful one. It was shown in [7], [17] that it is quite easy to express simple geometric objects like compositional lines, this is however not the case for ellipses, although they play a fundamental role within most statistical methods, for example in outlier detection ([8]). The aim of the paper is to introduce a way, based on coordinate representations of compositions, how to obtain an analytical representation of ellipses in the Aitchison geometry.

Key words: Aitchison geometry on the simplex, coordinates, ellipse.

2000 Mathematics Subject Classification: 14P99, 15A03, 15A63, 62H99, 62J05

1 Compositional data

At first, we briefly summarize all the basic properties of compositional data as well as the geometry on the simplex, called in the following Aitchison geometry. More detailed insight is available e.g. in [7]:

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 compositions (or only compositions 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 ([7]). As a consequence, D-parts compositions 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*

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

The basics of the 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

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

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

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

The inner product of $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{S}^D$ can be expressed as

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

Proposition 2 *The simplex with the perturbation operation and the power transformation,* $(\mathcal{S}^D, \oplus, \odot)$*, is a linear vector space. Moreover, the Aitchison inner product induces a* $(D-1)$*-dimensional Hilbert space.*

Definition 4 If compositions $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_{D-1}$ are independent (in terms of the Aitchison geometry), 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, $i = 1, \ldots, D-1$, that are termed *coordinates* with respect to the basis.

Obviously, 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] and [17] for details. For a composition $\boldsymbol{x}$, we denote $h(\boldsymbol{x})$ its representation in coordinates. Thus, for $\alpha, \beta \in \mathbb{R}$ it holds that

$$h(\alpha \odot \boldsymbol{x} \oplus \beta \odot \boldsymbol{y}) = \alpha \cdot h(\boldsymbol{x}) + \beta \cdot h(\boldsymbol{y});$$

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle_A = \langle h(\boldsymbol{x}), h(\boldsymbol{y}) \rangle_E, \quad \|\boldsymbol{x}\|_A = \sqrt{\langle \boldsymbol{x}, \boldsymbol{x} \rangle_A} = \|h(\boldsymbol{x})\|_E. \tag{1}$$

Example 1 Let us denote the coordinate representation of $\boldsymbol{x}$ as $\boldsymbol{z} = (z_1, \ldots, z_{D-1})$. Coefficients for a chosen simplicial basis ([5]) can be expressed as

$$z_i = \sqrt{\frac{i}{i+1}} \ln \frac{\sqrt[i]{\prod_{j=1}^{i} x_j}}{x_{i+1}} \quad \text{for } i = 1, \ldots, D-1.$$

The inverse transformation, i.e. $h^{-1}(\boldsymbol{z}) = \boldsymbol{x} = \mathcal{C}(x_1, \ldots, x_D)$, is then obtained using

$$x_i = \exp\left(\sum_{j=i}^{D} \frac{z_j}{\sqrt{j(j+1)}} - \sqrt{\frac{i-1}{i}} z_{i-1} \right) \quad \text{with } z_0 = z_D = 0 \text{ for } i = 1, \ldots, D.$$

2 Simplicial ellipses

A $(D-1)$-dimensional real vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_{D-1})$ and a positive definite real matrix $\boldsymbol{\Sigma} = (s_{ij})$ determine an ellipse $\mathcal{E}_{D-1}(\boldsymbol{z})$ with centre $\boldsymbol{\mu}$,

$$\mathcal{E}_{D-1}(\boldsymbol{z})\colon (\boldsymbol{z} - \boldsymbol{\mu})\, \boldsymbol{\Sigma}\, (\boldsymbol{z} - \boldsymbol{\mu})^T = c^2, \quad c > 0. \tag{2}$$

The ellipse $\mathcal{E}_{D-1}(\boldsymbol{z})$ can be equivalently expressed in analytical form

$$\sum_{i=1}^{D-1} \sum_{j=1}^{D-1} s_{ij} z_i z_j - 2 \sum_{i=1}^{D-1} \sum_{j=1}^{D-1} s_{ij} \mu_i z_j + k = 0$$

with $k = \boldsymbol{\mu} \boldsymbol{\Sigma} \boldsymbol{\mu}^T - c^2$. Using (2) and spectral decomposition of the matrix $\boldsymbol{\Sigma}$,

$$\boldsymbol{\Sigma} = \sum_{i=1}^{D-1} \lambda_i \boldsymbol{f}_i^T \boldsymbol{f}_i,$$

where λ_i and $\boldsymbol{f}_i$ denote eigenvalues (in descending order) and orthonormal eigenvectors of $\boldsymbol{\Sigma}$, respectively, the ellipse $\mathcal{E}_{D-1}(\boldsymbol{z})$ can also be expressed in terms of the Euclidean inner product as

$$\sum_{i=1}^{D-1} \lambda_i (\langle \boldsymbol{f}_i, \boldsymbol{z} \rangle_E)^2 - 2 \sum_{i=1}^{D-1} \lambda_i \langle \boldsymbol{f}_i, \boldsymbol{\mu} \rangle_E \langle \boldsymbol{f}_i, \boldsymbol{z} \rangle_E + k = 0. \tag{3}$$

It is easy to see that the spectral decomposition of $\mathbf{\Sigma}$ represents only one chosen, nevertheless the most convenient, decomposition of $\mathbf{\Sigma}$ in order to obtain (3). Namely, the vectors $\boldsymbol{f}_i$ determine the ellipse axes' directions, and their lengths are functions of the eigenvalues λ_i.

Let $h(\boldsymbol{x}) = \boldsymbol{z}$, $h(\boldsymbol{\gamma}) = \boldsymbol{\mu}$ and $h(\boldsymbol{e}_i) = \boldsymbol{f}_i$, $i = 1, \ldots, D-1$. Considering (1), the simplicial counterpart to $\mathcal{E}_{D-1}(\boldsymbol{z})$, denoted in the following as $\mathcal{E}_D^{\mathcal{S}}(\boldsymbol{x})$, is given by

$$\sum_{i=1}^{D-1} \lambda_i (\langle \boldsymbol{e}_i, \boldsymbol{x} \rangle_A)^2 - 2 \sum_{i=1}^{D-1} \lambda_i \langle \boldsymbol{e}_i, \boldsymbol{\gamma} \rangle_A \langle \boldsymbol{e}_i, \boldsymbol{x} \rangle_A + k = 0. \tag{4}$$

The following theorem is thus a simple consequence of the above mentioned considerations and definition of the Aitchison inner product:

Theorem 1 *The analytical form of the simplicial ellipse $\mathcal{E}_D^{\mathcal{S}}(\boldsymbol{x})$ is uniquely determined as*

$$\sum_{i=1}^{D-1} \sum_{j=i+1}^{D} \sum_{k=1}^{D-1} \sum_{l=k+1}^{D} a_{ijkl} \ln \frac{x_i}{x_j} \ln \frac{x_k}{x_l} + \sum_{i=1}^{D-1} \sum_{j=i+1}^{D} b_{ij} \ln \frac{x_i}{x_j} + k = 0,$$

where

$$a_{ijkl} = \frac{1}{D^2} \sum_{m=1}^{D-1} \lambda_m \ln \frac{e_{mi}}{e_{mj}} \ln \frac{e_{mk}}{e_{ml}},\ b_{ij} = -\frac{2}{D} \sum_{m=1}^{D-1} \lambda_m \langle \boldsymbol{e}_i, \boldsymbol{\gamma} \rangle_A \ln \frac{e_{mi}}{e_{mj}}$$

and

$$k = \sum_{i=1}^{D-1} \lambda_i (\langle \boldsymbol{e}_m, \boldsymbol{\gamma} \rangle_A)^2 - c^2.$$

The compositions $\boldsymbol{\gamma}$ and $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{iD})$ represent centre of $\mathcal{E}_D^{\mathcal{S}}(\boldsymbol{x})$ and the ellipse axes' directions, respectively. Theorem 1 provides a procedure how to construct an analytical representation of an ellipse on the simplex, obtained as a result of statistical computations in coordinates.

Example 2 A simplicial ellipse in coordinates (see Example 1) is given by

$$\boldsymbol{\mu} = (1, 1), \quad \mathbf{\Sigma} = \begin{pmatrix} 2.5, & 1.5 \\ 1.5, & 2.5 \end{pmatrix}, \quad c = 1,$$

i.e. with centre $\boldsymbol{\mu} = (1, 1)$, eigenvalues $\lambda_1 = 4$, $\lambda_2 = 1$ and axis directions

$$\boldsymbol{f}_1 = \frac{1}{\sqrt{2}}(1, 1) \quad \text{and} \quad \boldsymbol{f}_2 = \frac{1}{\sqrt{2}}(1, -1).$$

The analytical form of the ellipse $\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x})$ in coordinates, i.e. $\mathcal{E}_2(\boldsymbol{z})$, is thus

$$2.5z_1^2 + 2.5z_2^2 + 3z_1z_2 - 8z_1 - 8z_2 + 7 = 0.$$

Using (4) and Theorem 1 and after an adjustment we obtain the analytical form of $\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x})$ as

$$0.56\ln^2\frac{x_1}{x_2}+0.84\ln^2\frac{x_1}{x_3}+0.27\ln^2\frac{x_2}{x_3}+1.13\ln\frac{x_1}{x_2}\ln\frac{x_1}{x_3}+0.02\ln\frac{x_1}{x_2}\ln\frac{x_2}{x_3}$$
$$+0.56\ln\frac{x_1}{x_3}\ln\frac{x_2}{x_3}-3.77\ln\frac{x_1}{x_2}-5.15\ln\frac{x_1}{x_3}-1.38\ln\frac{x_2}{x_3}+7=0.$$

Here, the centre $\boldsymbol{\gamma}=(0.72,0.18,0.10)$ and axis directions are $\boldsymbol{e}_1=(0.61,0.23,0.16)$ and $\boldsymbol{e}_2=(0.36,0.13,0.51)$, respectively. Fig. 1 shows the simplicial ellipse displayed in a ternary diagram as well as in coordinates.

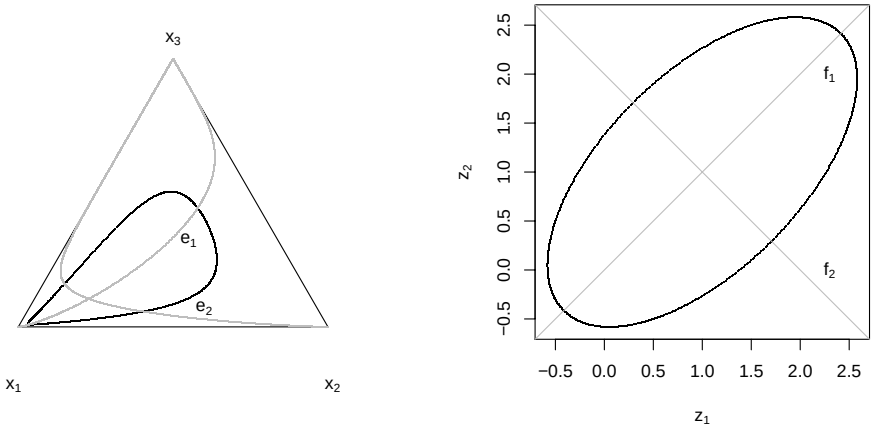

Fig. 1. The simplicial ellipse displayed in a ternary diagram (left) and in coordinates (right) together with the directions of the ellipse axes.

Let us remark that the existence of an analytical expression for ellipses on the simplex opens also a possibility for further generalizations in many directions, e.g. [13], [14].

3 Application in a statistical method

Ellipses frequently occur as a result of many statistical methods. In the case of compositional data one has to be careful to check whether the given problem is solvable in coordinates and how the results can be interpreted back on the simplex. One such problem is to find a regression line (in the compositional sense) that represents the main trend in the data, e.g. using the first principal component or equivalently the total least squares problem, computed in coordinates ([4], [15], [19]). In its simplest form it attempts to fit a line that explains the set of n two-dimensional data points (e.g. three-part compositions in coordinates) in such a way that the sum of squared distances from data points to the estimated line is minimal. In [12] it was shown that in this case, the problem is also solvable iteratively using the theory of linear regression models

with type-II constraints (in the constraints not only parameters of the model but also additional parameters occur, [16]), see [12] for details. Moreover, this approach enables to perform deeper statistical analysis like confidence regions and hypotheses testing. Considering the first mentioned possibility, under the assumption of normality we can construct confidence ellipses for locations of the unknown errorless results of the measurement, i.e. for the locations of each of n points $\boldsymbol{z}_i = h(\boldsymbol{x}_i) = (z_{i1}, z_{i2})$, $i = 1, \ldots, n$. The numerical results in the text below correspond to the statistical analysis of the well known Aphyric Skye Lavas data set that comes from [1, p. 360] and represents percentages of three variables ($Na_2O + K_2O$, Fe_2O_3 and MgO) related to the chemistry of 23 lava samples.

The confidence ellipses for the single errorless results of the measurement (true concentrations of the chemical compounds) in coordinates are constructed in such a way that their centers $\boldsymbol{\mu}_i$ (and $\boldsymbol{\gamma}_i$ in the Aitchison geometry) lie on the regression line $z_2 = \beta_1 + \beta_2 z_1$, where the parameters β_1, β_2 are estimated using the iterative algorithm described in [12]. Thus, we can assert that the unknown errorless results lie in the ellipses with the prescribed probability $1 - \alpha$. The directions $\boldsymbol{f}_1$ of the main half-axes of such ellipses follow the direction given by the estimated line, $\boldsymbol{f}_1 = (0.8903, -0.4554)$, thus $\boldsymbol{f}_2 = (0.4554, 0.8903)$ for the adjacent half-axes.

Although it might not to be visible from the ternary diagram, the unitary directions of the ellipses' main and adjacent half-axes are also the same and for all of them we have $\boldsymbol{e}_1 = (0.4515, 0.1282, 0.4203)$, $\boldsymbol{e}_2 = (0.5654, 0.2969, 0.1377)$; note that, of course, $\langle \boldsymbol{e}_1, \boldsymbol{e}_2 \rangle_A = 0$. Concretely, for a 95%-confidence ellipse, belonging to $\boldsymbol{x}_1$, we obtain the center of this ellipse in coordinates and on the simplex as

$$\boldsymbol{\mu}_1 = (0.0122, 1.7471), \quad \boldsymbol{\gamma}_1 = (0.4763, 0.4682, 0.0556),$$

respectively. Here c^2 equals $2\mathrm{F}_{2,21}(0.95)$, 95%-quantile of the F distribution with 2 and 21 degrees of freedom, see again [12] for details. The analytical representation of the ellipse in coordinates equals to

$$570.53 z_1^2 + 233.99 z_2^2 - 466.44 z_1 z_2 + 801.04 z_1 - 811.93 z_2 + 697.45 = 0$$

(the matrix $\boldsymbol{\Sigma}$ was obtained as inverse of the covariance matrix of the centre $\boldsymbol{\mu}_1$, [12]) and back-transformed to the simplex we obtain for $\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x})$

$$126.79 \ln^2 \frac{x_1}{x_2} + 25.81 \ln^2 \frac{x_1}{x_3} + 115.58 \ln^2 \frac{x_2}{x_3} + 37.02 \ln \frac{x_1}{x_2} \ln \frac{x_1}{x_3} - 216.55 \ln \frac{x_1}{x_2} \ln \frac{x_2}{x_3}$$
$$+ 14.61 \ln \frac{x_1}{x_3} \ln \frac{x_2}{x_3} + 377.61 \ln \frac{x_1}{x_2} - 142.66 \ln \frac{x_1}{x_3} - 520.28 \ln \frac{x_2}{x_3} + 697.45 = 0.$$

Here, the composition $\boldsymbol{x}_1 = (0.52, 0.42, 0.06)$ is not contained in the corresponding confidence ellipse, because $\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x}_1) = 9.85 > 0$. The corresponding results of all compositions $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are collected in Table 1. Note that there are many positive values, meaning that the data point is outside the ellipse.

This indicates a poor fit of the model to the data. As a consequence, a more complex model could be selected.

obs. number i	1	2	3	4	5	6
$\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x}_i)$	9.85	−6.40	−4.20	−6.05	3.70	−3.69
obs. number i	7	8	9	10	11	12
$\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x}_i)$	1.38	13.92	7.42	6.81	21.94	31.44
obs. number i	13	14	15	16	17	18
$\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x}_i)$	13.26	−0.33	−5.71	−1.95	67.19	−3.80
obs. number i	19	20	21	22	23	
$\mathcal{E}_3^{\mathcal{S}}(\boldsymbol{x}_i)$	−6.85	−3.61	−6.44	22.36	−5.20	

Tab. 1. Overview of results for the Aphyric Skye Lavas data. The values correspond to the observed compositions $\boldsymbol{x}_i$, $i = 1, \ldots, 23$, substituted in the corresponding confidence ellipses. A value less than zero indicates that the data point is contained inside the ellipse and for values greater than zero outside. Exact zero values would mean that the composition lies on the boundary.

Detailed interpretation of the logratios' coefficients in the analytical representation of ellipses on the simplex is the topic of the author's research and will be presented in the future.

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] 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.

[4] Daunis-i-Estadella, J., Barceló-Vidal, C., Buccianti, A.: *Exploratory compositional data analysis.* 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), 161–174.

[5] 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.

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

[7] 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.

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

[9] Filzmoser, P., Hron, K.: *Correlation analysis for compositional data.* Math. Geosci. (to appear).

[10] Filzmoser, P., Hron, K., Reimann, C.: *Principal component analysis for compositional data with outliers.* Environmetrics (to appear).

[11] Filzmoser, P., Hron, K., Reimann, C., Garrett, R.: *Robust factor analysis for compositional data.* Computers & Geosciences (to appear).

[12] Fišerová, E., Hron, K.: *Total least squares solution for compositional data using linear models.* Journal of Applied Statistics (to appear).

[13] Jukl, M.: *Linear forms on free modules over certain local ring.* Acta Univ. Palacki. Olomuc., Fac. rer. nat., Math. **110** (1993), 49–62.

[14] Jukl, M.: *Inertial law of quadratic forms on modules over plural algebra.* Mathematica Bohemica **3** (1995), 255–263.

[15] Kendall, M. G., Stuart, A.: The advanced theory of statistics, vol 2. *Charles Griffin, London,* 1967.

[16] Kubáček, L., Kubáčková, L.: *One of the calibration problems.* Acta Univ. Palacki. Olomuc., Fac. rer. nat., Math. **36** (1997), 117–130.

[17] Pawlowsky-Glahn, V., Egozcue, J. J., Tolosana-Delgado, J.: *Lecture notes on compositional data analysis.* http://hdl.handle.net/10256/297, 2007.

[18] Pearson, K.: *Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs.* Proceedings of the Royal Society of London **60** (1897), 489–502.

[19] Schuermans, M., Markovsky, I., Wentzell, P. D., Van Huffel, S.: *On the equivalence between total least squares and maximum likelihood PCA.* Analytica Chimica Acta **544** (2005), 254–267.