About covariance and correlation in the simplex

V. Pawlowsky-Glahn*and J. J. Egozcue†

Abstract

Historically, the problem with the statistical analysis of compositional data has been described as a problem of *spurious correlation* induced by the constant sum constraint. It has been explained as a consequence of the singularity of the covariance matrix, which induces a bias towards negative correlation. This facts are undeniable. But, to our understanding, the real problem is the use of an unsuitable geometry on the simplex, based on the usual sum, product, Euclidean distance, Euclidean norm and Euclidean inner product defined in real space. We suggest to substitute the Euclidean geometry associated to real space by the Aitchison geometry defined on the simplex. The latter is based on the concepts of perturbation, power transformation, Aitchison distance, Aitchison norm and Aitchison inner product. It allows us to define appropriate measures of single and joint variability of random compositions, leading naturally to the concept of correlation as a measure of the strength of the *linear* relationship (in the Aitchison sense) between two random compositions.

Key words: Aitchison geometry, compositional data, finite dimensional Hilbert space, inner product, ternary diagram.

1 Introduction

The preface of the monograph on the statistical analysis of compositional data by Aitchison (1986) begins with the following words:

> As long ago as 1987 Karl Pearson, in a now classic paper on spurious correlation, first pointed out dangers that may befall the analyst who attempts to interpret correlations between ratios whose numerators and denominators contain common parts.

The statistical methodology presented in the monograph, based on the essential idea that in compositional problems size of specimens is irrelevant, and thus ratios should be used, has been a milestone in the development of an appropriate methodology, suitable to this type of data. Nevertheless, the fact that the actual modeling is performed

*Dept. d'Informàtica i Mathemàtica Aplicada, Universitat de Girona – Girona (E); Phone: ++34-972.418.421; Fax: ++34-972.418.399; e-mail: vera.pawlowsky@udg.es

†Dept. de Mathemàtica Aplicada III, ETSECCPB, UPC – Barcelona (E)

in a transformed space, *e.g.* using the alr or the clr transformation, has provoked resistance due to the loss of classical properties like *unbiasedness* or *minimum variance* of estimators, or *correlation* of raw components.

In (Pawlowsky-Glahn and Egozcue, 2001a, 2001b) we have used an alternative—geometric—approach, which leads to concepts like *metric center* and *metric variance* defined on the simplex. They are analogous to the usual ones, and they are easy to interpret. The same methodology can be used to define covariance and correlation in the simplex from a geometric perspective, and this is the purpose of this contribution. But before we proceed, we recall briefly the *Aitchison geometry* on the simplex, *i.e.* the finite dimensional real Hilbert space structure of the simplex $\mathcal{S}_c^d$.

2 The Aitchison geometry on the simplex

2.1 Metric vector space structure

By definition, the sample space of a d-part composition with constant sum c, where c is 1 if measurements are made in parts per unit, or 100 if measurements are made in percent, is the simplex

$$\mathcal{S}_c^d = \{\mathbf{x} = (x_1, x_2, ..., x_d)' | x_i > 0, i = 1, 2, ..., d; \sum_{i=1}^{d} x_i = c\},$$

where the prime stands fro transpose. Let $\mathcal{C}$ denote the closure operation, defined for a vector with strictly positive components, $\mathbf{z} = (z_1, z_2, ..., z_d)'$, as

$$\mathcal{C}(\mathbf{z}) = \mathcal{C}\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_d \end{pmatrix} = \begin{pmatrix} \frac{c \cdot z_1}{z_1 + z_2 + \cdots + z_d} \\ \frac{c \cdot z_2}{z_1 + z_2 + \cdots + z_d} \\ \vdots \\ \frac{c \cdot z_d}{z_1 + z_2 + \cdots + z_d} \end{pmatrix}.$$

As stated in Aitchison (2001), the simplex is a metric vector space with the perturbation operation, defined for any two vectors $\mathbf{x}, \mathbf{y} \in \mathcal{S}_c^d$ as

$$\mathbf{x} \circ \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, ..., x_d y_d)',$$

the power transformation, defined for a vector $\mathbf{x} \in \mathcal{S}_c^d$ and a scalar $\alpha \in \Re$ as

$$\alpha \diamond \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, ..., x_d^\alpha)',$$

and the Aitchison distance, defined for any two vectors $\mathbf{x}, \mathbf{y} \in \mathcal{S}_c^d$ as

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{d} \sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}. \tag{1}$$

For a proof of this assertion, see (Pawlowsky-Glahn and Egozcue, 2001a).

2.2 Finite dimensional Hilbert space structure

To define a Hilbert space structure on a metric vector space, we need an inner product which induces the distance given. As stated in (Aitchison, 2001), the norm associated to the distance given in equation (1), which we shall call the *Aitchison norm* and denote by $\|.\|_a$, is

$$\|\mathbf{x}\|_a = d_a(\mathbf{x}, \mathbf{e}), \quad \text{or} \quad \|\mathbf{x} \circ \mathbf{y}^{-1}\|_a = d_a(\mathbf{x}, \mathbf{y}), \tag{2}$$

and the *Aitchison inner product*, denoted by $\langle ., . \rangle_a$, is

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2} \left(\|\mathbf{x}\|_a^2 + \|\mathbf{y}\|_a^2 - \|\mathbf{x} \circ \mathbf{y}^{-1}\|_a^2 \right), \tag{3}$$

where $\mathbf{y}^{-1}$ denotes the inverse of composition $\mathbf{y}$ with respect to the perturbation operation. For completeness, the following property is proven in the appendix:

Proposition 1 *The inner product of two compositions defined in equation (3) can be expressed as*

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{d} \sum_{i<j} \left(\ln \frac{x_i}{x_j} \right) \left(\ln \frac{y_i}{y_j} \right),$$

and satisfies the conditions necessary to be a positive, non degenerate hermitic form on the simplex.

The principal consequence of proposition 1 is that the norm associated to $\langle \mathbf{x}, \mathbf{y} \rangle_a$, which can be written $\|\mathbf{x}\|_a = \langle \mathbf{x}, \mathbf{x} \rangle_a^{1/2}$, defines a topology in $\mathcal{S}_c^d$, and that with respect to this topology $\mathcal{S}_c^d$ is complete.

The proof of completeness is based on the following properties:

1. Any vector $\mathbf{x}$ in $\mathcal{S}_c^d$ can be obtained as a perturbation of d vectors $\mathbf{u}_i$, where the $\mathbf{u}_i$ are obtained as the closure of a vector with all elements equal to one except the i-th one, which is equal to the number e. The coefficients have to be $\ln x_i$:

   $$\mathbf{x} = \bigcirc_{i=1}^{d} (\ln x_i \diamond \mathbf{u}_i).$$

   Thus, the dimension of $\mathcal{S}_c^d$ is at the most d and is thus finite.

2. Every normed space of finite dimension is complete (Berberian, 1961).

Stated properties can be summed up in the statement that $\mathcal{S}_c^d$ is a finite dimensional Hilbert space (Zamansky, 1967). Although in mathematical textbooks finite dimensional Hilbert spaces are refered to as Euclidean spaces, we prefer to avoid the latter terminology to prevent confusion between the usual geometry in real space and the Aitchison geometry in the simplex.

Note that the dimension of $\mathcal{S}_c^d$ is not d, but actually $(d-1)$. To prove that this is so, note that the alr transformation, as defined in (Aitchison, 1986), is a linear application from the vector space $(\mathcal{S}_c^d, \circ, \diamond)$ onto the vector space $(\Re^{d-1}, +, \cdot)$ whose kernel reduces to the neutral element $\mathbf{e}$ of perturbation. Thus, the alr transformation is an isomorphism and, both vector spaces being finite, they have the same dimension, which is necessarily $(d-1)$, the dimension of $\Re^{d-1}$.

One could think that the isomorphism between both vector spaces would be helpful in finding a basis for the simplex. In fact, as an isomorfism transforms forth and back

independent vectors, it seems reasonable to take the canonical basis in $\Re^{d-1}$, consisting of $d-1$ vectors of dimension $d-1$ with all elements equal to zero and the i-th equal to one, and apply the agl transformation, inverse of the alr transformation, to obtain the set of vectors $\mathbf{u}_i$, $i = 1, ..., d-1$, defined above, except for the last one. The resulting set of compositional vectors is certainly a basis, but not an orthogonal one, *i.e.*, $\langle \mathbf{u}_i, \mathbf{u}_j \rangle_a = -1/d \neq 0$. As a consequence thereof, the following statement holds:

Proposition 2 *The* alr *transformation from* $\mathcal{S}_c^d$ *onto* $\Re^{d-1}$ *is not an isometry.*

This fact is certainly the cause of many misunderstandings concerning the use of standard statistical methods with alr-transformed data. At the same time, it justifies the introduction of a metric approach which complements the logratio approach presented in (Aitchison, 1986). We have to be aware that an isomorphism guarantees probabilistic assessments to be preserved, whereas an isometry is required for distance based methods, like *e.g.* those based on the mean square error.

Let us recall, before proceeding to statistical concepts, a few properties derived from the previous statements.

1. An inner product satisfies the Cauchy-Schwarz inequality: $|\langle \mathbf{x}, \mathbf{y} \rangle_a|^2 \leq \langle \mathbf{x}, \mathbf{x} \rangle_a \langle \mathbf{y}, \mathbf{y} \rangle_a$.
2. For the norm associated to the inner product it holds
   (a) $\|\mathbf{x}\|_a = 0 \iff \mathbf{x} = \mathbf{e}$;
   (b) $\|\mathbf{x} \circ \mathbf{y}\|_a \leq \|\mathbf{x}\|_a + \|\mathbf{y}\|_a$, for all $\mathbf{x}, \mathbf{y} \in \mathcal{S}_c^d$;
   (c) $\|\alpha \diamond \mathbf{x}\|_a = |\alpha| \cdot \|\mathbf{x}\|_a$, for all $\alpha \in \Re$, $\mathbf{x} \in \mathcal{S}_c^d$.

3 Metric covariance and metric correlation

As mentioned previously, Pawlowsky-Glahn and Egozcue (2001a, 2001b), developing an original approach by Aitchison (2001), introduced the concepts of center and metric variance as the natural counterpart for random compositions to the concepts of expected value and variance for random variables in real space. The approach, based on the simple idea to substitute Euclidean distance by Aitchison distance, led to the following definitions for a random composition $\mathbf{X}$ with sample space $\mathcal{S}_c^d$:

Definition 1 *The dispersion or metric variance around* $\xi \in \mathcal{S}_c^d$ *is the expected value of the squared distance between* $\mathbf{X}$ *and* ξ*:* $\mathrm{Mvar}[\mathbf{X}, \xi] = \mathrm{E}\left[{d_a}^2(\mathbf{X}, \xi)\right]$.

Definition 2 *The center of the distribution of* $\mathbf{X}$ *is that element* $\xi \in \mathcal{S}_c^d$ *which minimizes* $\mathrm{Mvar}[\mathbf{X}, \xi]$. *It is denoted by* $\mathrm{cen}(\mathbf{X})$ *or by* γ *for short.*

Definition 3 *The metric variance around the center* $\mathrm{cen}(\mathbf{X}) = \gamma$ *of the distribution of* $\mathbf{X}$ *is given by* $\mathrm{Mvar}[\mathbf{X}, \gamma] = \mathrm{E}\left[{d_a}^2(\mathbf{X}, \gamma)\right]$. *It is called* metric variance *and denoted* $\mathrm{Mvar}[\mathbf{X}]$ *for short.*

An extensive development of the properties derived from this definitions can be found in (Pawlowsky-Glahn and Egozcue, 2001a, 2001b). In the latter contribution the theoretical foundation of consistency of the use of the expected value operator can be found.

Now, given the inner product associated to the Aitchison distance defined on the simplex in equation (3), the same rationale leads to introduce the natural counterparts of covariance and correlation for random compositions.

Definition 4 *The* metric covariance *of two random compositions* $\mathbf{X}$ *and* $\mathbf{Y}$ *in* $\mathcal{S}_c^d$*, centered respectively at* $\mathrm{cen}(\mathbf{X})$ *and* $\mathrm{cen}(\mathbf{Y})$*, is defined as*

$$\mathrm{Mcov}[\mathbf{X} \circ \mathrm{cen}(\mathbf{X})^{-1}, \mathbf{Y} \circ \mathrm{cen}(\mathbf{Y})^{-1}] = \mathrm{E}\left[\langle \mathbf{X} \circ \mathrm{cen}(\mathbf{X})^{-1}, \mathbf{Y} \circ \mathrm{cen}(\mathbf{Y})^{-1} \rangle_a\right].$$

It is denoted by $\mathrm{Mcov}[\mathbf{X}, \mathbf{Y}]$ *for short.*

Definition 5 *The* metric correlation *of two random compositions* $\mathbf{X}$ *and* $\mathbf{Y}$ *in* $\mathcal{S}_c^d$ *is defined as*

$$\rho_m[\mathbf{X}, \mathbf{Y}] = \frac{\mathrm{Mcov}[\mathbf{X}, \mathbf{Y}]}{(\mathrm{Mvar}[\mathbf{X}]\mathrm{Mvar}[\mathbf{Y}])^{1/2}}$$

Whenever it is clear from the context, the explicit reference to the random compositions $\mathbf{X}$ *and* $\mathbf{Y}$ *is omitted and the metric correlation is denoted simply by* ρ_m*.*

From the definitions of distance, norm and inner product in equations (1), (2) and (3), it is straightforward to show that the definition of metric covariance reduces to the definition of metric variance whenever $\mathbf{Y}$ is substituted by $\mathbf{X}$. Useful properties of the metric covariance and correlation follow. The proofs are omitted as they are direct applications of properties of the inner product stated before and of the linearity of the expected value operator.

Proposition 3

$$\mathrm{Mcov}[\mathbf{X}, \mathbf{Y}] = \frac{1}{d} \sum_{i<j} \mathrm{E}\left[\left(\ln \frac{X_i}{X_j} - \ln \frac{\gamma_{xi}}{\gamma_{xj}}\right)\left(\ln \frac{Y_i}{Y_j} - \ln \frac{\gamma_{yi}}{\gamma_{yj}}\right)\right]$$

where $\gamma_x = \mathrm{cen}(\mathbf{X})$ *and* $\gamma_y = \mathrm{cen}(\mathbf{Y})$*.*

Proposition 4

$$\mathrm{Mcov}[\mathbf{X}, \mathbf{Y}] = \frac{1}{2}\left(\mathrm{Mvar}[\mathbf{X}] + \mathrm{Mvar}[\mathbf{Y}] - \mathrm{Mvar}[\mathbf{X} \circ \mathbf{Y}^{-1}]\right).$$

From property 4 the following 'standard' relationships between the metric variance of the perturbation of two random compositions, their metric variances and their covariance, are obtained. They are the counterpart of properties which relate the variance of the sum and difference of random variables with their variance and covariance. Thus it holds

$$\begin{aligned} \mathrm{Mvar}[\mathbf{X} \circ \mathbf{Y}^{-1}] &= \mathrm{Mvar}[\mathbf{X}] + \mathrm{Mvar}[\mathbf{Y}] - 2\mathrm{Mcov}[\mathbf{X}, \mathbf{Y}] \\ \mathrm{Mvar}[\mathbf{X} \circ \mathbf{Y}] &= \mathrm{Mvar}[\mathbf{X}] + \mathrm{Mvar}[\mathbf{Y}] - 2\mathrm{Mcov}[\mathbf{X}, \mathbf{Y}^{-1}]. \end{aligned}$$

4 Conclusions

Metric covariance and metric correlation are concepts that relate random compositions, thus random vectors, and not random variables in an univariate sense. It is clear that the definitions and properties given hold for subcompositions, but they will always be related to random compositions with at least two components and cannot be reduced to a kind of relationship between individual components, although for that case we have already the concept of *perfect proportionality* defined in (Aitchison, 2001), which is given whenever the usual Euclidean variance of the logratio of two components is constant.

Acknowledgments

This research has been supported by the Dirección General de Enseñanza Superior (DGES) of the Spanish Ministry for Education and Culture through the projects PB96-0501-C02-01 and BFM2000-0540.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). 416 p.

Aitchison, J. (2001). Simplicial inference. In M. Viana and D. Richards (Eds.), *Algebraic Methods in Statistics*, Contemporary Mathematics Series, pp. (in press). American Mathematical Society, New York, NY (USA).

Berberian, S. K. (1961). *Introduction to Hilbert Space*. Oxford University Press, New York, NY (USA). 206 p.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001a). About BLU estimators and compositional data.

Pawlowsky-Glahn, V. and J. J. Egozcue (2001b). Geometric approach to statistical analysis on the simplex. *SERRA*, (accepted for publication).

Zamansky, M. (1967). *Introducción al álgebra y análisis moderno*. Montaner y Simón, Barcelona (E). 437 p.

Appendix

Proof of proposition 1. Consider equation (3),

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2} \left(\|\mathbf{x}\|_a^2 + \|\mathbf{y}\|_a^2 - \|\mathbf{x} \circ \mathbf{y}^{-1}\|_a^2 \right),$$

and rewrite the norms in terms the Aitchison distance

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2} \left({d_a}^2(\mathbf{x}, \mathbf{e}) + {d_a}^2(\mathbf{y}, \mathbf{e}) - {d_a}^2(\mathbf{x}, \mathbf{y}) \right)$$

$$= \frac{1}{2}\left(\frac{1}{d}\sum_{i<j}\left(\ln\frac{x_i}{x_j}\right)^2 + \frac{1}{d}\sum_{i<j}\left(\ln\frac{y_i}{y_j}\right)^2 - \frac{1}{d}\sum_{i<j}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2\right).$$

Developing the square and simplifying the corresponding sums of squares, the desired result is obtained. Thus, we can write

$$\langle \mathbf{x}, \mathbf{y}\rangle_a = \frac{1}{d}\sum_{i<j}\left(\ln\frac{x_i}{x_j}\right)\left(\ln\frac{y_i}{y_j}\right),$$

an expression that is more suitable to proof that it is actually an inner product. According to Zamansky (1967), to have an inner product we need an hermitic form, which is an application $\varphi(.,)$, defined from $\mathcal{S}_c^d \times \mathcal{S}_c^d$ onto $\Re$, which satisfies the following conditions (we write only those required for the real case, ommitting the corresponding properties for a complex field):

1. $\varphi(\mathbf{x}\circ\mathbf{x}', \mathbf{y}) = \varphi(\mathbf{x}, \mathbf{y}) + \varphi(\mathbf{x}', \mathbf{y})$, $\varphi(\mathbf{x}, \mathbf{y}\circ\mathbf{y}') = \varphi(\mathbf{x}, \mathbf{y}) + \varphi(\mathbf{x}, \mathbf{y}')$;
2. $\varphi(\alpha\diamond\mathbf{x}, \mathbf{y}) = \alpha\cdot\varphi(\mathbf{x}, \mathbf{y})$, $\varphi(\mathbf{x}, \alpha\diamond\mathbf{y}) = \alpha\cdot\varphi(\mathbf{x}, \mathbf{y})$;
3. $\varphi(\mathbf{x}, \mathbf{y}) = \varphi(\mathbf{y}, \mathbf{x})$.

To show that $\langle \mathbf{x}, \mathbf{y}\rangle_a$ is an hermitic form, note that

$$\langle \mathbf{x}, \mathbf{y}\rangle_a = \frac{1}{d}\sum_{i<j}\left(\ln\frac{x_i}{x_j}\right)\left(\ln\frac{y_i}{y_j}\right) = \frac{1}{d}\sum_{i<j}\left(\left(\ln\frac{y_i}{y_j}\ln\frac{x_i}{x_j}\right)\right) = \langle \mathbf{y}, \mathbf{x}\rangle_a$$

due to commutativity of the product in real space. Thus, condition 3 is satisfied, implying that the form is simmetric, and to proof conditions 1 and 2 we only need to proof the first part. Consider now

$$\begin{aligned}
\langle \mathbf{x}\circ\mathbf{x}', \mathbf{y}\rangle_a &= \frac{1}{d}\sum_{i<j}\left(\ln\frac{x_i x_i'}{x_j x_j'}\right)\left(\ln\frac{y_i}{y_j}\right) \\
&= \frac{1}{d}\sum_{i<j}\left(\ln\frac{x_i}{x_j} + \ln\frac{x_i'}{x_j'}\right)\left(\ln\frac{y_i}{y_j}\right) \\
&= \frac{1}{d}\sum_{i<j}\left(\ln\frac{x_i}{x_j}\right)\left(\ln\frac{y_i}{y_j}\right) + \frac{1}{d}\sum_{i<j}\left(\ln\frac{x_i'}{x_j'}\right)\left(\ln\frac{y_i}{y_j}\right),
\end{aligned}$$

which proofs condition 1, whereas

$$\begin{aligned}
\langle \alpha\diamond\mathbf{x}, \mathbf{y}\rangle_a &= \frac{1}{d}\sum_{i<j}\left(\ln\frac{x_i^\alpha}{x_j^\alpha}\right)\left(\ln\frac{y_i}{y_j}\right) \\
&= \frac{1}{d}\sum_{i<j}\left(\alpha\ln\frac{x_i}{x_j}\right)\left(\ln\frac{y_i}{y_j}\right) \\
&= \alpha\frac{1}{d}\sum_{i<j}\left(\ln\frac{x_i}{x_j}\right)\left(\ln\frac{y_i}{y_j}\right)
\end{aligned}$$

proofs condition 2. Clearly, $\langle \mathbf{x}, \mathbf{y}\rangle_a$ is a positive form, as the condition for this to be so is that $\langle \mathbf{x}, \mathbf{x}\rangle_a \geq 0$ for all $\mathbf{x}$, which is trivially satisfied. Furthermore, to be non

degenerate, the necessary and sufficient condition is that $\langle \mathbf{x}, \mathbf{x} \rangle_a = 0 \iff \mathbf{x} = \mathbf{e}$, where $\mathbf{e}$ is the neutral element of the inner operation on the simplex. Again, the verification of this condition is straightforward, as

$$\langle \mathbf{x}, \mathbf{x} \rangle_a = \frac{1}{d} \sum_{i<j} \left(\ln \frac{x_i}{x_j} \right) \left(\ln \frac{x_i}{x_j} \right) = 0$$

if, and only if, for all i, j, $\ln x_i - \ln x_j = 0$. But this is only satisfied if, for all i, $x_i = 1/d$, which is equivalent to say that $\mathbf{x} = \mathbf{e}$. □