Mathematical Geology, Vol. 34, No. 3, April 2002 (© 2002)

BLU Estimators and Compositional Data[1]

Vera Pawlowsky-Glahn[2] and Juan José Egozcue[3]

One of the principal objections to the logratio approach for the statistical analysis of compositional data has been the absence of unbiasedness and minimum variance properties of some estimators: they seem not to be BLU estimator. Using a geometric approach, we introduce the concept of metric variance and of a compositional unbiased estimator, and we show that the closed geometric mean is a c-BLU estimator (compositional best linear unbiased estimator with respect to the geometry of the simplex) of the center of the distribution of a random composition. Thus, it satisfies analogous properties to the arithmetic mean as a BLU estimator of the expected value in real space. The geometric approach used gives real meaning to the concepts of measure of central tendency *and* measure of dispersion *and opens up a new way of understanding the statistical analysis of compositional data.*

KEY WORDS: Aitchison distance, centered estimator, metric variance, perturbation, simplex, ternary diagram.

INTRODUCTION

The logratio approach to the statistical analysis of compositional data, proposed in Aitchison, 1982, has been the source of many discussions over the last two decades. The approach makes it possible to perform classical statistical analysis on transformed data and to backtransform the results, which is enormously advantageous because of the large number of methods available for multivariate normally distributed phenomena, and to demonstrate the robustness of those methods. But there has been a certain reluctance in using the new approach, which, besides the usual resistance to new theories, is due to the difficulty of interpretation of backtransformed results and to a lack of classical properties of backtransformed estimators and models, like unbiasedness and minimum variance.

[1] Received 3 July 2000; accepted 29 March 2001.

[2] Department of Informatics and Applied Mathematics, Universitat de Girona, Campus Montilivi – P1, E-17071 Girona, Spain; e-mail: vera.pawlowsky@udg.es

[3] Department of Applied Mathematics III, Universitat Politècnica de Catalunya, Campus Nord – C2, E-08034 Barcelona, Spain; e-mail: egozcue@ncsa.es

0882-8121/02/0400-0259/1 © 2002 International Association for Mathematical Geology

Here we propose an answer to these questions, based on the concepts of metric variance and centered estimator. The new measure of variability is an extension of the classical concepts of variance for random variables with sample space over the real line to a measure of variability for random vectors with sample space the simplex. To make this extension, we briefly recall the vector space structure of the simplex and the corresponding distance on the simplex; then we define the concept of metric variance of random compositions. On the basis of this concept, centered estimators in the simplex are introduced. Finally, it is shown that the closed geometric mean is a compositional unbiased linear minimum metric variance estimator of the center of the distribution of a random composition.

COMPOSITIONAL DATA AND THEIR SAMPLE SPACE

Recall that $\mathbf{x} = (x_1, x_2, \ldots, x_d)'$ is by definition a d-part composition if, and only if, all its components are strictly positive real numbers and their sum is a constant c. Zero components are excluded for reasons that will be discussed later. The constant c is 1 if measurements are made in parts per unit, or 100 if measurements are made in percent. The sample space of d-part compositional data with constant sum c is thus the simplex,

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

where the prime stands for transpose. Although mathematically less comfortable, we keep the constant c in the definition and in the notation, to avoid confusion arising from the fact that in geology it is more common to use $c = 100$ than $c = 1$. But, to simplify the mathematical developments, we include the constant in the closure operation as stated below.

Basic operations on the simplex (Aitchison, 1986) are the perturbation operation, defined for any two vectors $\mathbf{x}, \mathbf{y} \in \mathcal{S}_c^d$ as

$$\mathbf{x} \circ \mathbf{y} = \mathcal{C}(x_1 y_1, x_2 y_2, \ldots, x_d y_d)', \tag{1}$$

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

$$\alpha \diamond \mathbf{x} = \mathcal{C}\left(x_1^\alpha, x_2^\alpha, \ldots, x_d^\alpha\right)', \tag{2}$$

where $\mathcal{C}$ denotes the closure operation defined for a vector $\mathbf{z} = (z_1, z_2, \ldots, z_d)' \in \mathbb{R}^d_+$ as

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

VECTOR SPACE STRUCTURE OF THE SIMPLEX

As stated by Aitchison 2001, perturbation and power transformation induce a vector space structure in the simplex. In fact, for a set to have a vector space structure the following conditions (in mathematical terms) are required

1. The existence of an internal operation which satisfies the properties of a commutative group;
2. The existence of an external operation with respect to elements of a field, identified here for convenience with the real line, the usual sum, and product, $(\mathbb{R}, +, \cdot)$, which satisfies the following four conditions: (a) associativity with respect to the product operation in $\mathbb{R}$; (b) distributivity with respect to the internal operation; (c) distributivity with respect to the sum in $\mathbb{R}$; and (d) that the neutral element with respect to the product in $\mathbb{R}$ is also the neutral element with respect to the external operation.

In other words, for the simplex to be a vector space with respect to perturbation and power transformation, it is necessary and sufficient that for $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{S}^d_c$ and $\alpha, \beta \in \mathbb{R}$ the following conditions are satisfied:

1. Commutative group structure of $(\mathcal{S}^d_c, \circ)$ (Aitchison, 1992):
   (a) Commutative property: $\mathbf{x} \circ \mathbf{y} = \mathbf{y} \circ \mathbf{x}$;
   (b) Associative property: $(\mathbf{x} \circ \mathbf{y}) \circ \mathbf{z} = \mathbf{x} \circ (\mathbf{y} \circ \mathbf{z})$;
   (c) Existence of a neutral element with respect to perturbation:

$$\mathbf{e} = \mathcal{C}(1, 1, \ldots, 1)';$$

   this element plays the role of the origin of the vector space;
   (d) Existence of an inverse element of each $\mathbf{x} \in \mathcal{S}^d_c$:

$$\mathbf{x}^{-1} = \mathcal{C}\left(x_1^{-1}, x_2^{-1}, \ldots, x_d^{-1}\right)'$$

   satisfying the condition $\mathbf{x}^{-1} \circ \mathbf{x} = \mathbf{x} \circ \mathbf{x}^{-1} = \mathbf{e}$.

2. Properties of power transformation:
   (a) Associative property: $\alpha \diamond (\beta \diamond \mathbf{x}) = (\alpha \cdot \beta) \diamond \mathbf{x}$;
   (b) Distributive property 1: $\alpha \diamond (\mathbf{x} \circ \mathbf{y}) = (\alpha \diamond \mathbf{x}) \circ (\alpha \diamond \mathbf{y})$;
   (c) Distributive property 2: $(\alpha + \beta) \diamond \mathbf{x} = (\alpha \diamond \mathbf{x}) \circ (\beta \diamond \mathbf{x})$;
   (d) The neutral element with respect to multiplication in $\mathbb{R}$ acts as neutral element with respect to the external operation: $1 \diamond \mathbf{x} = \mathbf{x}$.

The proof of these properties is straightforward using the definition of perturbation and of power transformation given in Equations (1) and (2) combined with the classical properties of sum, product, and power in real space, and the fact that constants cancel whenever the closure operation is applied, and is therefore omitted.

Note that for a composition with zero components all the properties would hold, exception made of the existence of an inverse element with respect to the internal operation. Consequently, the simplex would not be a commutative group with respect to perturbation and we could not define a vector space structure on it, which is essential to our approach.

In the next section, whenever we use the term *simplex* and the notation $\mathcal{S}_c^d$, we shall understand the vector space $(\mathcal{S}_c^d, \circ, \diamond)$ over $(\mathbb{R}, +, \cdot)$.

METRIC VECTOR SPACE STRUCTURE OF THE SIMPLEX

Aitchison (1992) introduced several equivalent forms for a metric on the simplex, which were extensively discussed by Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (1998). Being equivalent, we have chosen to use for $\mathbf{x}, \mathbf{y} \in \mathcal{S}_c^d$ the following definition:

$$d_a(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{d} \sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2}, \tag{3}$$

which we shall hereafter call as *Aitchison distance*.

The Aitchison distance has, among others, the following properties, which have been reported in the previously mentioned publications without including the proofs. We include them, for the sake of completeness, in the appendix.

Proposition 1. *The Aitchison distance is perturbation invariant.*

$$d_a(\mathbf{x}, \mathbf{y}) = d_a(\mathbf{z} \circ \mathbf{x}, \mathbf{z} \circ \mathbf{y}), \quad \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{S}_c^d.$$

As a consequence,

$$d_a(\mathbf{x}, \mathbf{y}) = d_a(\mathbf{x} \circ \mathbf{y}^{-1}, \mathbf{e}),$$

which tells us that the distance between two compositions **x** and **y** is the same as the distance between the composition **x** perturbed by **y** and the baricenter **e** of the simplex.

Proposition 2. *The Aitchison distance is scale invariant.*

$$d_a(\alpha \diamond \mathbf{x}, \alpha \diamond \mathbf{y}) = |\alpha| \cdot d_a(\mathbf{x}, \mathbf{y}), \quad \mathbf{x}, \mathbf{y} \in \mathcal{S}_c^d, \alpha \in \mathbb{R}.$$

Proposition 3. *The Aitchison distance is invariant under permutation of the components of the composition.*

Recall that, given a vector space over $(\mathbb{R}, +, \cdot)$ (in this case the simplex $\mathcal{S}_c^d$) with a continuous external operation (the power transformation), if we can define on it a distance (the Aitchison distance) which is invariant with respect to the internal group operation (perturbation), then we have a metric vector space (Zamansky, 1967, p. 246). Thus, in what follows, we can consider the simplex as a metric vector space $(\mathcal{S}_c^d, \circ, \diamond)$ over $(\mathbb{R}, +, \cdot)$. To refer globally to the properties of $(\mathcal{S}_c^d, \circ, \diamond)$ we shall talk about the *Aitchison geometry on the simplex.*

Here we find again an argument against the inclusion of zero components in the definition of a composition. They not only conclude the definition of a group structure on the simplex with respect to perturbation; they are also not compatible with the Aitchison distance, as the logarithm of zero is undefined. But the Aitchison geometry helps us understanding better the role of compositions with zero components with respect to those without. Compositions with zero components play the role of points at infinity. The smaller one component (the closer to zero), the larger is the distance with respect to the origin. Martín-Fernández, Barceló-Vidal, and Pawlowsky-Glahn (2000) give a first approach to dealing with zeros in a manner coherent with the metric vector space structure discussed here.

With these elements in hand we can proceed to introduce the concepts of metric variance and center of the distribution of a random vector **X** with sample space $\mathcal{S}_c^d$.

CENTER AND METRIC VARIANCE

Let us introduce the rationale behind the definitions and properties of this section by recalling basic definitions and properties related to random variables in real space. Given a continuous random variable X, the expected value is $\mathrm{E}[X] = \int_{-\infty}^{+\infty} x f_X(x)\, dx$, where $f_X(x)$ stands for the density function of X, and the variance is $\mathrm{Var}[X] = \mathrm{E}[(X - \mathrm{E}[X])^2]$. The geometric interpretation of these concepts is well known, and is often given either as a motivation or as an illustration, but they are never directly defined in terms of distances. Nevertheless, the expected value can be defined as that value $\mathrm{E}[X] = \mu$, which minimizes the expected squared Euclidean distance $\mathrm{E}[d_e(X, \xi)^2]$, and the variance can be defined as the expected value of the squared Euclidean distance around μ, $\mathrm{Var}[X] = \mathrm{E}[d_e(X, \mu)^2]$. This geometric approach is more natural, as it clearly reveals the meaning of the expected value as a measure of central tendency and the variance as a measure of dispersion. Aitchison (2001) uses this philosophy to introduce the center of a random vector which support is the simplex, i.e., of a random composition, by substituting the Euclidean distance with the Aitchison distance defined in Equation (3). Here, we simply pursue this approach further.

Consider a random composition $\mathbf{X}$ with sample space $\mathcal{S}_c^d$ and density function $f(\mathbf{x})$. Thus, $f(\mathbf{x})$ is a nonnegative function inside $\mathcal{S}_c^d$ and zero outside, and its integral over $\mathcal{S}_c^d$ is one.

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: $\mathrm{Mvar}[\mathbf{X}, \xi] = \mathrm{E}[{d_a}^2(\mathbf{X}, \xi)]$, provided that the last expectation exists.

Assuming the metric variance of $\mathbf{X}$ exists, we can introduce the centre of its distribution as follows.

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 called *center of* $\mathbf{X}$ and is denoted by $\mathrm{cen}[\mathbf{X}]$ or by γ for short.

As mentioned before, the definition of center as the element $\xi \in \mathcal{S}_c^d$ which minimizes $\mathrm{E}[{d_a}^2(\mathbf{X}, \xi)]$ has been given in Aitchison (2001), although a similar approach can already be found in Aitchison [1997, Eq. (11)], where it appears as the element $\xi \in \mathcal{S}_c^d$ which minimizes the Kullback-Leibler directed divergence, an information-theoretic measure of the divergence of $\mathbf{X}$ from ξ. The result in both cases is the same.

Following our strategy to paraphrase standard statistical concepts, to call *metric variance* the metric variance around $\mathrm{cen}[\mathbf{X}]$ is natural. We state this as a definition for ease of reference.

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

Important properties of the center and metric variance of a random composition $\mathbf{X}$ follow. Proofs, although straightforward, are included in the appendix.

Proposition 4. *The center γ of $\mathbf{X}$ with respect to the Aitchison distance is the closed geometric mean (Aitchison, 2001, Eq. (18)).*

$$\gamma = \mathcal{C}(\exp\{\mathrm{E}[\ln X_1]\},\ \exp\{\mathrm{E}[\ln X_2]\}, \ldots,\ \exp\{\mathrm{E}[\ln X_d]\})'.$$

Note that this last proposition implies that, for $i, j = 1, \ldots, d, i \neq j$,

$$\mathrm{E}\left[\ln \frac{X_i}{X_j}\right] = \ln \frac{\gamma_i}{\gamma_j},$$

a property which is useful in some proofs in the appendix.

Proposition 5. *The center cen*$[\mathbf{X}]$ *of* $\mathbf{X}$ *is the* agl *transform of the vector of expected values of* alr$\mathbf{X}$, *where* alr *stands for the additive logratio transformation (Aitchison, 1986, p. 135) and* agl *for its inverse. Thus,*

$$\mathrm{cen}[\mathbf{X}] = \mathrm{agl}(\mathrm{E}[\mathrm{alr}\mathbf{X}]).$$

Proposition 6. *The metric variance is equal to the total variance defined in (Aitchison, 1997 Eq. (22)),* i.e.,

$$\mathrm{Mvar}[\mathbf{X}] = \frac{1}{d} \sum_{i<j} \mathrm{Var}\left[\ln \frac{X_i}{X_j}\right].$$

Given the absolute parallelism between expected value and variance in real space with the Euclidean distance, and the center and metric variance in the simplex with the Aitchison distance, it is not surprising that concepts equivalent to the linearity of the expectation operator and translation invariance of the variance are straightforward to prove on the simplex (see Appendix for details). In fact, the following basic properties hold:

Proposition 7. *For two random compositions* $\mathbf{X}, \mathbf{Y} \in \mathcal{S}_c^d$ *(Aitchison, 2001)*

$$\mathrm{cen}[\mathbf{X} \circ \mathbf{Y}] = \mathrm{cen}[\mathbf{X}] \circ \mathrm{cen}[\mathbf{Y}],$$

and, in general, for N random compositions $\mathbf{X}_n \in \mathcal{S}_c^d$, $n = 1, 2, \ldots, N$,

$$\mathrm{cen}[\mathbf{X}_1 \circ \mathbf{X}_2 \circ \cdots \circ \mathbf{X}_N] = \mathrm{cen}[\mathbf{X}_1] \circ \mathrm{cen}[\mathbf{X}_2] \circ \cdots \circ \mathrm{cen}[\mathbf{X}_N].$$

Proposition 8. *For a random composition* $\mathbf{X} \in \mathcal{S}_c^d$*, a perturbation* $b \in \mathcal{S}_c^d$ *and a scalar* $a \in \mathbb{R}$*,*

$$\mathrm{cen}[(a \diamond \mathbf{X}) \circ b] = (a \diamond \mathrm{cen}[\mathbf{X}]) \circ b.$$

Proposition 9. *For two independent random compositions* $\mathbf{X}, \mathbf{Y} \in \mathcal{S}_c^d$*,*

$$\mathrm{Mvar}[\mathbf{X} \circ \mathbf{Y}] = \mathrm{Mvar}[\mathbf{X}] + \mathrm{Mvar}[\mathbf{Y}].$$

Proposition 10. *For a random composition* $\mathbf{X} \in \mathcal{S}_c^d$*, a perturbation* $b \in \mathcal{S}_c^d$ *and a scalar* $a \in \mathbb{R}$*,*

$$\mathrm{Mvar}[(a \diamond \mathbf{X}) \circ b] = a^2 \mathrm{Mvar}[\mathbf{X}].$$

Thus, the metric variance, defined as the expected value of the distance to the center, allows us a geometric interpretation of basic statistical properties of random vectors in the simplex. It also satisfies the conditions of being invariant under perturbation, equivalent to invariance under translation in real space.

BLU-ESTIMATION

Assume a random composition $\mathbf{X}$ with sample space $\mathcal{S}_c^d$ and density function $f(\mathbf{x})$ and a random sample of size N, $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_N$. Recall that, in this context, a random sample is a set of independent random compositions, all of them with the same distribution as $\mathbf{X}$.

We know from previously cited developments by Aitchison that to estimate the center the closed geometric mean of the sample can be used. To better understand the properties of this estimator, let us first introduce general definitions concerning the concepts of unbiasedness, optimality, and linearity in the simplex. These definitions are specific to compositional parameters and are parallel to ordinary (Euclidean) ones. For this reason, our notation for each compositional-space concept consists of placing a *c*- before the name of the corresponding real-space concept, in order to relate both concepts and, at the same time, distinguish between them.

Consider a compositional estimator $\hat{\theta}$ of the unknown compositional parameter θ of the distribution of $\mathbf{X}$ with parameter space $\mathcal{S}_c^d$, *i.e.* $\hat{\theta} \in \mathcal{S}_c^d$ and $\theta \in \mathcal{S}_c^d$. By the definition of an estimator, $\hat{\theta}$ is a function of the random sample and thus a random vector itself.

Definition 4. $\hat{\theta}$ is a *c-centered or c-unbiased compositional estimator* of $\theta \in \mathcal{S}_c^d$ if and only if $\mathrm{cen}[\hat{\theta}] = \theta$, *or*, equivalently, if and only if $\mathrm{cen}[\hat{\theta} \circ \theta^{-1}] = \mathbf{e}$, the neutral element of the internal operation on the simplex.

This definition is closely related to centered or unbiased estimators of parameters which are not compositional. To recognize that fact, simply recall the classical definition, which states that

> $\hat{\psi}$ is a *centered or unbiased estimator* of $\psi \in \mathbb{R}^n$ if, and only if, $\mathrm{E}[\hat{\psi}] = \psi$, or, equivalently, if, and only if, $\mathrm{E}[\hat{\psi} - \psi] = 0$, the neutral element of the internal operation in $\mathbb{R}^n$.

Definition 5. Given a class Θ of c-unbiased compositional estimators of $\theta \in \mathcal{S}_c^d$, $\hat{\theta} \in \Theta$ is said to be *c-best* with respect to the Aitchison distance within the class Θ if, and only if, it is c-unbiased and $\mathrm{Mvar}[\hat{\theta}] < \mathrm{Mvar}[\hat{\theta}_i]$ for all $\hat{\theta}_i \in \Theta$; i.e., $\hat{\theta}$ is c-unbiased and has minimum metric variance within Θ.

Note that this definition is again the counterpart for compositional parameters of the concept of *best* or *most efficient estimator* given in standard textbooks.

Obviously, other standard characterizations of estimators, usual in the context of random variables which support is the real line, can be given, simply by substituting the Euclidean distance with the Aitchison distance, and the expected value by the center, but that goes beyond the purpose of this paper. Therefore, let us proceed to show that the closed geometric mean of the sample is a c-best c-unbiased estimator of the center $\mathrm{cen}[\mathbf{X}]$ of a random composition $\mathbf{X}$ within the class of linear estimators of $\mathrm{cen}[\mathbf{X}]$, where linear is understood in the following sense:

Definition 6. A compositional estimator $\hat{\theta}$ of θ is said to be *c-linear* if, and only if

$$\hat{\theta} = (\alpha_1 \diamond \mathbf{X}_1) \circ (\alpha_2 \diamond \mathbf{X}_2) \circ \cdots \circ (\alpha_N \diamond \mathbf{X}_N).$$

Note that this definition makes sense given the vector space structure of $(\mathcal{S}_c^d, \circ, \diamond)$. To simplify notation, let us introduce the symbol $\bigcirc_{n=1}^N$ for perturbation over a set of indices, analogous to $\Sigma_{n=1}^N$ for summation or $\Pi_{n=1}^N$ for product. Thus, a linear estimator in the simplex will be written $\hat{\theta} = \bigcirc_{n=1}^N(\alpha_n \diamond \mathbf{X}_n)$ and, taking $\alpha_n = 1/N$, for all $n = 1, 2, \ldots, N$, the closed geometric mean is obtained. Note that in this case, using the distributive property 1 of the power transformation, we can write $\bigcirc_{n=1}^N(1/N \diamond \mathbf{X}_n) = 1/N \diamond (\bigcirc_{n=1}^N \mathbf{X}_n)$. Now, the following propositions can be set forth, the proofs of which are included in the Appendix.

Proposition 11. *The closed geometric mean* $\hat{\gamma} = \frac{1}{N} \diamond (\bigcirc_{n=1}^N \mathbf{X}_n)$ *is a c-linear and c-unbiased estimator of* $\mathrm{cen}[\mathbf{X}]$.

Proposition 12. *The closed geometric mean* $\hat{\gamma} = \bigcirc_{n=1}^N (\frac{1}{N} \diamond \mathbf{X}_n)$ *is the c-best estimator of* $\gamma = \mathrm{cen}[\mathbf{X}]$ *within the class of c-linear c-unbiased estimators of* γ. *Moreover,* $\mathrm{Mvar}[\hat{\gamma}] = (1/N)\mathrm{Mvar}[\mathbf{X}]$.

Propositions 11 and 12 clearly establish that the closed geometric mean is the *c-best c-linear c-unbiased estimator*, or *c-BLU* estimator for short, of the center of $\mathbf{X}$ in the context of the simplex, whenever the Aitchison distance is considered.

Obviously, given a realization of a random sample, the closed geometric mean could be computed even in presence of zero components. Nevertheless, the result would be a center with as many zero components as different components with at least one zero, independently of the total number of observed zeros. Thus, one single observation with one single zero would be enough to get a center with a zero in the same position no matter the sample size, a result that clearly makes no sense if we look for a measure of central tendency.

CONCLUSIONS

The existence of an appropriate metric vector space structure in the simplex suggests a different approach to the statistical analysis of compositional data based on geometric reasoning. On the basis of this approach, which is completely parallel to the usual one in Euclidean space, it is straightforward to define reasonable properties for estimators of compositional parameters. In particular, it can be shown that the closed geometric mean is a linear estimator with respect to the Aitchison geometry on the simplex, as well as a *c*-unbiased and minimum metric variance estimator with respect to the Aitchison distance, or *c-BLU* estimator for short.

ACKNOWLEDGMENTS

The authors thank the reviewers for their encouraging comments. This research has been supported by the Dirección General de Enseñanza Superior (DGES) of the Spanish Ministry for Education and Culture through the project PB96-0501-C02-01.

REFERENCES

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

Aitchison, J., 1986, The statistical analysis of compositional data: Chapman and Hall, London, UK, 416 p.

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

Aitchison, J., 1997, The one-hour course in compositional data analysis or compositional data analysis is simple, *in* Pawlowsky-Glahn, V., ed., Proceedings of IAMG'97—The third annual conference of the International Association for Mathematical Geology: International Center for Numerical Methods in Engineering (CIMNE), Barcelona (E), p. 3–35.

Aitchison, J., in press, Simplicial inference, *in* Viana, M., and Richards, D., eds., Algebraic methods in statistics: American Mathematical Society, New York.

Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 1998, A critical approach to non-parametric classification of compositional data, *in* Rizzi, A., Vichi, M., and Bock, H.-H., eds., Advances in data science and classification: Springer, Berlin (D), p. 49–56.

Martín-Fernández, J. A., Barceló-Vidal, C., and Pawlowsky-Glahn, V., 2000, Zero replacement in compositional data sets, *in* Kiers, H., Rasson, J., Groenen, P., and Shader, M., eds., Studies in classification, data analysis, and knowledge organization: Springer, Berlin (D), p. 155–160.

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

APPENDIX

Proofs included in this appendix are essentially scholarly exercises. Their main interest lies in getting acquainted with standard operations in the simplex. Nevertheless, the fact that they force us to be aware, at each moment, of the space we are working in, helps in better understanding the essential properties and interpretation of concepts used.

Proof of proposition 1: Taking the square in Equation (3), it holds, for any $\mathbf{x}, \mathbf{y} \in \mathcal{S}_c^d$, that

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

Thus, for **x** and **y** perturbed by **z**, we have

$$\begin{aligned} d_a{}^2(\mathbf{z} \circ \mathbf{x}, \mathbf{z} \circ \mathbf{y}) &= \frac{1}{d} \sum_{i<j} \left(\ln \frac{z_i x_i}{z_j x_j} - \ln \frac{z_i y_i}{z_j y_j} \right)^2 \\ &= \frac{1}{d} \sum_{i<j} \left(\left(\ln \frac{x_i}{x_j} + \ln \frac{z_i}{z_j} \right) - \left(\ln \frac{y_i}{y_j} + \ln \frac{z_i}{z_j} \right) \right)^2 \\ &= \frac{1}{d} \sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{y_i}{y_j} \right)^2 \\ &= d_a{}^2(\mathbf{x}, \mathbf{y}). \end{aligned}$$

Proof of proposition 2: □

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

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

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

and taking the square root the desired result is obtained. □

Proof of proposition 3: The proof is straightforward taking into account that

$$d_a{}^2(\mathbf{x},\mathbf{y}) = \sum_{i<j}\frac{1}{d}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2 = \frac{1}{2d}\sum_{i,j=1}^{d}\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2. \qquad \text{(A1)}$$

Equality (A1) implies that all the possible pairs of indices appear in the summation, and therefore permutation of variables will not alter the result.

To prove equality (A1) just recall that for $i = j$ the corresponding terms are zero, and that

$$\left(\ln\frac{x_i}{x_j} - \ln\frac{y_i}{y_j}\right)^2 = \left(\ln\frac{x_j}{x_i} - \ln\frac{y_j}{y_i}\right)^2,$$

because an exchange of indices results in a change of sign, which is compensated by the square. □

Proof of proposition 4: By definition 2, γ is that element in $\mathcal{S}_c^d$ which minimizes the metric variance. For an arbitrary element $\xi \in \mathcal{S}_c^d$

$$\mathrm{Mvar}[\mathbf{X},\xi] = \mathrm{E}\left[d_a{}^2(\mathbf{X},\xi)\right] = \mathrm{E}\left[\frac{1}{d}\sum_{i<j}((\ln X_i - \ln X_j) - (\ln\xi_i - \ln\xi_j))^2\right],$$

and, using Equation (A1) to simplify operations, we obtain

$$\mathrm{Mvar}[\mathbf{X},\xi] = \mathrm{E}\left[\frac{1}{2d}\sum_{i,j=1}^{d}((\ln X_i - \ln X_j) - (\ln\xi_i - \ln\xi_j))^2\right]$$

$$= \frac{1}{2d}\sum_{i,j=1}^{d}\big(\mathrm{E}\left[(\ln X_i - \ln X_j)^2\right]$$

$$- 2\mathrm{E}[\ln X_i - \ln X_j](\ln\xi_i - \ln\xi_j) + (\ln\xi_i - \ln\xi_j)^2\big).$$

To minimize this expression, we take partial derivatives with respect to the components of ξ,

$$\frac{\partial}{\partial \xi_k} \mathrm{E}\left[{d_a}^2(\mathbf{X}, \xi)\right] = -\frac{2}{2d\xi_k} \sum_{j=1}^{d} (\mathrm{E}[\ln X_k - \ln X_j] - (\ln \xi_k - \ln \xi_j))$$

$$= -\frac{1}{d\xi_k} \sum_{j=1}^{d} ((\mathrm{E}[\ln X_k] - \ln \xi_k) - (\mathrm{E}[\ln X_j] - \ln \xi_j)).$$

Setting the partial derivatives equal to zero we obtain

$$d(\mathrm{E}[\ln X_k] - \ln \xi_k) = \sum_{j=1}^{d} (\mathrm{E}[\ln X_j] - \ln \xi_j) = \text{constant.}$$

Therefore, for all k,

$$\ln \xi_k = \mathrm{E}[\ln X_k] + C, \qquad \text{(A2)}$$

with C the resulting constant after dividing by d and changing the sign. Note that the constant C does not depend on k. Taking exponentials and applying the closure operation, given that constants cancel, the desired result is obtained, *i.e.*

$$\mathcal{C}(\exp\{\mathrm{E}[\ln X_1]\}, \exp\{\mathrm{E}[\ln X_2]\}, \ldots, \exp\{\mathrm{E}[\ln X_d]\})' = \gamma,$$

where γ stands for the solution obtained. Note that, applying Equation (A2) to γ, we obtain

$$\mathrm{E}\left[\ln \frac{X_i}{X_j}\right] = \mathrm{E}[\ln X_i] - \mathrm{E}[\ln X_j] = (\ln \gamma_i - C) - (\ln \gamma_j - C) = \ln \frac{\gamma_i}{\gamma_j}. \quad \square$$

Proof of proposition 5: The vector of expected values of alr**X** is

$$\begin{aligned} \mu &= (\mathrm{E}[\ln X_1 - \ln X_d], \mathrm{E}[\ln X_2 - \ln X_d], \ldots, \mathrm{E}[\ln X_{d-1} - \ln X_d])' \\ &= (\mathrm{E}[\ln X_1] - \mathrm{E}[\ln X_d], \mathrm{E}[\ln X_2] - \mathrm{E}[\ln X_d], \ldots, \mathrm{E}[\ln X_{d-1}] - \mathrm{E}[\ln X_d])' \end{aligned}$$

The agl backtransformation consists in taking exponentials, adding a last component equal to 1 and applying the closure operation, resulting in

$$\begin{aligned}\mathrm{agl}(\mu) &= \mathcal{C}(\exp\{\mathrm{E}[\ln X_1] - \mathrm{E}[\ln X_d]\}, \ldots, \exp\{\mathrm{E}[\ln X_{d-1}] - \mathrm{E}[\ln X_d]\}, 1)' \\ &= \mathcal{C}(\exp\{\mathrm{E}[\ln X_1]\}, \ldots, \exp\{\mathrm{E}[\ln X_{d-1}]\}, \exp\{\mathrm{E}[\ln X_d]\})'.\end{aligned}$$

which is precisely the expression of γ given in property 4. □

Proof of proposition 6: By definition, the metric variance is

$$\mathrm{Mvar}[\mathbf{X}] = \mathrm{E}\big[{d_a}^2(\mathbf{X}, \mathrm{cen}[\mathbf{X}])\big] = \int_{\mathcal{S}_c^d} \frac{1}{d} \sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{\gamma_i}{\gamma_j}\right)^2 f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x},$$

and the total variance defined in (Aitchison, 1997, Eq. (22)), can be expressed in terms of the joint density function as

$$\mathrm{totvar}[\mathbf{X}] = \frac{1}{d} \sum_{i<j} \mathrm{Var}\left[\ln \frac{X_i}{X_j}\right] = \frac{1}{d} \sum_{i<j} \int_{\mathcal{S}_c^d} \left(\ln \frac{x_i}{x_j} - \mu_{ij}\right)^2 f_{\mathbf{X}}(\mathbf{x})\, d\mathbf{x},$$

where $\mu_{ij} = \mathrm{E}[\ln X_i - \ln X_j] = \mathrm{E}[\ln X_i] - \mathrm{E}[\ln X_j] = \ln \gamma_i - \ln \gamma_j$ by property 4. □

Proof of proposition 7: To show that for two random compositions $\mathbf{X}, \mathbf{Y} \in \mathcal{S}_c^d$, $\mathrm{cen}[\mathbf{X} \circ \mathbf{Y}] = \mathrm{cen}[\mathbf{X}] \circ \mathrm{cen}[\mathbf{Y}]$, just note that by proposition 5 and standard properties of the alr and the agl transformations stated in Aitchison (1986),

$$\begin{aligned}\mathrm{cen}[\mathbf{X} \circ \mathbf{Y}] &= \mathrm{agl}(\mathrm{E}[\mathrm{alr}(\mathbf{X} \circ \mathbf{Y})]) \\ &= \mathrm{agl}(\mathrm{E}[\mathrm{alr}\mathbf{X} + \mathrm{alr}\mathbf{Y}]) \\ &= \mathrm{agl}(\mathrm{E}[\mathrm{alr}\mathbf{X}]) \circ \mathrm{agl}(\mathrm{E}[\mathrm{alr}\mathbf{X}]) \\ &= \mathrm{cen}[\mathbf{X}] \circ \mathrm{cen}[\mathbf{Y}].\end{aligned}$$

The second part is proved by induction. □

Proof of proposition 8: By proposition 7, $\mathrm{cen}[(a \diamond \mathbf{X}) \circ b] = \mathrm{cen}[a \diamond \mathbf{X}] \circ \mathrm{cen}[b]$. But, given that b is a constant, $\mathrm{cen}[b] = b$, as can be seen substituting in proposition 5 the random composition $\mathbf{X}$ by b. Finally, using again proposition 5 and the linearity of the expectation, we obtain $\mathrm{cen}[a \diamond \mathbf{X}] = a \diamond \mathrm{cen}[\mathbf{X}]$, and thus the desired property is obtained. □

Proof of proposition 9: Using proposition 6 and independency of random compositions **X**, **Y** the following identities hold:

$$\begin{aligned}\mathrm{Mvar}[\mathbf{X}\circ\mathbf{Y}] &= \frac{1}{d}\sum_{i<j}\mathrm{Var}\left[\ln\frac{X_iY_i}{X_jY_j}\right]\\ &= \frac{1}{d}\sum_{i<j}\left[\mathrm{Var}\left(\ln\frac{X_i}{X_j}\right)+\mathrm{Var}\left(\ln\frac{Y_i}{Y_j}\right)\right]\\ &= \mathrm{Mvar}[\mathbf{X}]+\mathrm{Mvar}[\mathbf{Y}]. \qquad\square\end{aligned}$$

Proof of proposition 10. The result is directly obtained from the definition of metric variance, properties 7, 1, and 2, and linearity of expectation. In fact,

$$\begin{aligned}\mathrm{Mvar}[(a\diamond\mathbf{X})\circ b] &= \mathrm{E}\left[{d_a}^2((a\diamond\mathbf{X})\circ b,\ \mathrm{cen}[(a\diamond\mathbf{X})\circ b])\right]\\ &= \mathrm{E}\left[{d_a}^2((a\diamond\mathbf{X})\circ b,\ (a\diamond\mathrm{cen}[\mathbf{X}])\circ b)\right]\\ &= \mathrm{E}\left[{d_a}^2((a\diamond\mathbf{X}),\ (a\diamond\mathrm{cen}[\mathbf{X}]))\right]\\ &= \mathrm{E}\left[a^2{d_a}^2(\mathbf{X},\ \mathrm{cen}[\mathbf{X}])\right]\\ &= a^2\mathrm{E}\left[{d_a}^2(\mathbf{X},\ \mathrm{cen}[\mathbf{X}])\right]\\ &= a^2\mathrm{Mvar}[\mathbf{X}]. \qquad\square\end{aligned}$$

Proof of proposition 11: $\hat{\gamma}=\frac{1}{N}\diamond(\bigcirc_{n=1}^{N}\mathbf{X}_n)$ is c-linear by definition, and using properties 7 and 8, as well as the fact that $(\mathcal{S}_c^d,\circ,\diamond)$ is a vector space, we obtain:

$$\mathrm{cen}[\hat{\gamma}]=\frac{1}{N}\diamond\left(\bigcirc_{n=1}^{N}\mathrm{cen}[\mathbf{X}_n]\right)=\frac{1}{N}\diamond(N\diamond\mathrm{cen}[\mathbf{X}])=\mathrm{cen}[\mathbf{X}]. \qquad\square$$

Proof of proposition 12: Consider an arbitrary c-linear c-unbiased estimator of the centre γ given by $\tilde{\gamma}=\bigcirc_{n=1}^{N}(\alpha_n\diamond\mathbf{X}_n)$. If $\tilde{\gamma}$ is c-unbiased, using property 7 and the distributive property 2 of the power transformation, we have that

$$\gamma=\mathrm{cen}[\hat{\gamma}]=\bigcirc_{n=1}^{N}(\alpha_n\diamond\mathrm{cen}[\mathbf{X}_n])=\bigcirc_{n=1}^{N}(\alpha_n\diamond\gamma)=\left(\sum_{n=1}^{N}\alpha_n\right)\diamond\gamma,$$

and thus $\sum_{n=1}^{N}\alpha_n=1$.

Consider now the metric variance of $\tilde{\gamma}$. After proposition 9 and 10, taking into account that we are assuming the sample to be random, and thus of independent random compositions, we have

$$\mathrm{Mvar}[\tilde{\gamma}] = \mathrm{Mvar}\Big[\bigcirc_{n=1}^{N} (\alpha_n \diamond \mathbf{X}_n)\Big] = \mathrm{Mvar}[\mathbf{X}] \sum_{n=1}^{N} \alpha_n^2,$$

which is minimum when, for all $n = 1, 2, \ldots, N$, $\alpha_n = 1/N$. Consequently, to have minimum metric variance we need $\tilde{\gamma} = \hat{\gamma}$ and $\mathrm{Mvar}[\hat{\gamma}] = (1/N)\mathrm{Mvar}[\mathbf{X}]$. □