SORT 42 (2) July-December 2018, 101-124

DOI: 10.2436/20.8080.02.71

Evidence functions: a compositional approach to information

J.J. Egozcue[1] and V. Pawlowsky-Glahn[2]

Abstract

The discrete case of Bayes' formula is considered the paradigm of information acquisition. Prior and posterior probability functions, as well as likelihood functions, called evidence functions, are compositions following the Aitchison geometry of the simplex, and have thus vector character. Bayes' formula becomes a vector addition. The Aitchison norm of an evidence function is introduced as a scalar measurement of information. A fictitious fire scenario serves as illustration. Two different inspections of affected houses are considered. Two questions are addressed: (a) which is the information provided by the outcomes of inspections, and (b) which is the most informative inspection.

MSC: 60A10, 60E10, 62E10

Keywords: Evidence function, Bayes' formula, Aitchison geometry, compositions, orthonormal basis, simplex, scalar information

1. Introduction

Each summer fires in forests and suburban areas affect houses, industries, and the whole environment. When this occurs, authorities need to get a quick diagnostic of damages, both for mitigation of effects, evaluation of economic costs and, especially, for evacuation of population from houses and planning of further actions. Airborne photography and visual inspection of houses are emergency means to classify houses into categories, usually corresponding to (a) buildings that can be reoccupied by the previously evacuated people, (b) buildings that require some repairs, (c) buildings that are largely damaged or (d) buildings that are collapsed. The impact of such diagnostics is critical, as the damnified population can or cannot recover their homes, do or do not receive economic compensations, depending on the result of the inspection. Typical questions are: How uncertain/informative are the results of an inspection? Which type of inspection is more reliable? What is the amount of information after inspections? These questions are

[1] Dept. Civil and Environmental Engineering, Universitat Politècnica de Catalunya, Barcelona, Spain. E-mail: juan.jose.egozcue@upc.edu

[2] Dept. of Computer Sciences, Applied Mathematics, and Statistics, Universitat de Girona, Spain. E-mail: vera.pawlowsky@udg.edu

Received: August 2018

related to the quantification of information provided by an experiment (inspections) and, therefore, should be answered by the statistical theory of information.

The above scenario of fires is not the only one where the questions on information provided by experiments are relevant. A very similar situation corresponds to many hazardous situations like earthquakes, floods, hurricanes, terrorist attacks... Also, clinic diagnostic of deseases, military actions or, in general, operational decisions under uncertainty correspond to the same type of scenario, which can be modelled as a collection of uncertain states or events, frequently assumed non-overlapping, to which some prior probabilities describing uncertainty on the true event are assigned; then, one or more experiments (diagnostic tests, inspections) are carried out, trying to reduce uncertainty; finally, after the results of the experiments, the updating of the probabilities (posterior probabilities) may allow to use the information available in decision making schemes. This scheme has been well known for decades, and still maintains its validity (e.g. Benjamin and Cornell, 1960).

The previous questions have been addressed from different points of view in information theory, specially following the line proposed by Lindley (1956). However, information theory was born from the study of coding and communication (Shannon, 1948, Shannon and Weaver, 1949, McMillan, 1953) and built on an early contribution by Hartley (1928), where logarithms of probabilities were identified as a measure of information. The initial development of the theory in the framework of communications and its particular syntaxis may be the reason why the statistical theory of information was developed some years later (e.g. Kullback and Leibler, 1951a, Kullback, 1997, Lindley, 1956, Khinchin, 1957, Ash, 1990). In medicine, diagnostic tests were studied, for instance, by Aitchison and Kay (1975) (see also Aitchison, Kay and Lauder, 2005).

The statistical theory of information is directly related to the concept of entropy. This is viewed as an average of measures of uncertainty (Shannon, 1948, McMillan, 1953) which is common to all branches of information theory. More rarely, information acquisition is linked to the Bayes' formula (Lindley, 1956) and its extensions, for example Dempster's rule in the theory of beliefs (Yager and Liu, 2008).

The aim of the present contribution is rethinking the bases of information theory from the point of view of compositional data analysis (Aitchison, 1986, Pawlowsky-Glahn and Buccianti, 2011, Pawlowsky-Glahn, Egozcue and Tolosana-Delgado, 2015). For completeness, Appendix A is a summary of Aitchison geometry for compositions, introducing notation and basic tools used. The main proposal is that information is a vector magnitude identified as a composition. These compositions are here called evidence functions, e-functions for short, and include traditional (discrete) probability functions and also likelihood functions. The Aitchison norm of e-functions as compositions (see Appendix A) is used as a scalar measure of information called e-information. This is in contrast to Shannon information and its related magnitudes, which were developed as scalar measures of information. Other points which are relevant to this proposal are:

- The Bayes' formula (discrete case) is the paradigm of information acquisition;

- The Bayes' formula is a vector additive Abelian group operation in the simplex endowed with the Aitchison geometry;
- Discrete probability functions (prior, posterior) and discrete likelihood functions are compositions and, consequently, they share the same properties.

Section 2 reviews concepts of compositional geometry and identifies evidence functions involved in Bayesian updating as compositions (see also Appendix A). Section 3 introduces a scalar measure of information, namely the Aitchison norm of an evidence function. Its properties characterize it as a proper measure of information. Section 4 discusses the acquisition of information through a fictitious fire scenario and inspections of affected houses.

2. Bayes theorem, evidence functions and compositions

Consider the fire scenario in which a number of isolated, but close, houses have been affected. It is assumed that these houses can be in $D = 4$ states, denoted A_i, $i = 1,2,\dots,D$, which can be identified with *service* or *no damage* (Nod), *moderate damage* (Mod), *severe damage* (Sev) and *ruin or collapse* (Col). These states are assumed non-overlapping. Based on previous urban studies, there is a perception that, after the fire, most houses will remain in service (80%) or with little damage (15%), meanwhile some of them will be largely damaged (4%) or in ruin (1%). In the Bayesian terminology, the vector of probabilities $\mathbf{p} = (p_1,\dots,p_D) = (0.80, 0.15, 0.04, 0.01)$, is known as prior or initial probabilities (this prior is reported in Table 1 as $\mathbf{p}^{(1)}$). The vector $\mathbf{p}$ is a composition. In fact, expressed as proportions or as percentages, the information is exactly the same; in particular, ratios between components remain the same. Moreover, the set of odds obtained by the ratios between components contains all the relative information and could be used to retrieve the numerical value of $\mathbf{p}$. These simple features characterize $\mathbf{p}$ as a D-part composition. The fact that the relative information contained in $\mathbf{p}$ remains unaltered when it is multiplied by a positive constant corresponds to the *scale invariance principle* of compositional data, and to its consequence, namely that the relative information is provided by the ratios of components (Aitchison, 1986, 1994). More recently, compositional equivalence has been defined as the condition that vectors of positive components which are proportional are compositionally equivalent (Barceló-Vidal and Martín-Fernández, 2016, Pawlowsky-Glahn et al., 2015, Barceló-Vidal, Martín-Fernández and Pawlowsky-Glahn, 2001). The generated equivalence classes can always be represented in a unitary D-part simplex, denoted $\mathbb{S}^D$, so that the sum of the parts is one, as in the usual normalization of probability. For simplicity, the projection of a non-normalized composition onto $\mathbb{S}^D$ is denoted by the closure operator $\mathscr{C}$.

Frequently, only some parts of the vector $\mathbf{p}$ are considered. For instance, only reparable buildings, i.e. only the three first parts, are taken into account. This restriction is a *subcomposition*. A subcomposition like $\mathscr{C}(p_1, p_2, p_3)$ corresponds to a conditional probability vector $(p_1/p_c, p_2/p_c, p_3/p_c)$ with $p_c = p_1 + p_2 + p_3$. This suggests that the identification of vectors of probabilities with compositions is natural.

Returning to the fire scenario, assume that a visual inspection of affected houses has been devised. The inspectors, after a quick visit of a building, decide to assign a color code according to their perception: green for service or no damage, orange for moderate damage, red for severe damage, and black for ruin or collapse. Obviously, this kind of assessment is quite uncertain, and the color codes do not correspond exactly to the real state of the building. Let R be the result of an inspection (e.g. orange: moderate damage in a visual inspection). For each possible result R, the conditional probabilities $q_i = \Pr(R|A_i)$, $i = 1, 2, \ldots, D$, characterize the experiment. In fact, the likelihood function associated with R, $\mathbf{q} = (q_1, q_2, \ldots, q_D)$, allows to apply Bayes' formula to obtain final or posterior probabilities $\mathbf{f} = (f_1, f_2, \ldots, f_D)$ as

$$\mathbf{f} = C \cdot (p_1 q_1, p_2 q_2, \ldots, p_D q_D) \,, \quad C = \frac{1}{\Pr(R)} = \left(\sum_{k=1}^{D} p_k q_k \right)^{-1} \,, \tag{1}$$

with $\mathbf{p}$ the vector of prior probabilities. This expression of the final probabilities, after the observation of R, matches exactly the definition of perturbation in the simplex, as pointed out by Aitchison (1986). Perturbation is an Abelian group operation in the simplex, and it is the addition in the Aitchison geometry for compositions (Pawlowsky-Glahn and Egozcue, 2001, Pawlowsky-Glahn et al., 2015), that is, Bayesian updating is a shift of the prior probabilities to the final probabilities by the likelihood. The simplex $\mathbb{S}^D$, endowed with perturbation ($\oplus$, group operation), powering ($\odot$, external multiplication) and Aitchison inner product, is a $(D-1)$-dimensional Euclidean space (Billheimer, Guttorp and Fagan, 2001, Pawlowsky-Glahn and Egozcue, 2001) (see Appendix A for detailed definitions). Therefore, denoting perturbation by $\oplus$, the Bayes formula is simply

$$\mathbf{f} = \mathbf{p} \oplus \mathbf{q} \,, \tag{2}$$

where no reference to the normalizing constant is necessary due to the compositional equivalence. Commonly, it is assumed that the difference between vectors of probabilities (initial or prior, final or posterior) and the likelihood function is that the latter is not normalized. The three symbols $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{f}$ are considered as compositions: in fact, the normalization of probabilities is irrelevant and the *likelihood principle* (Birnbaum, 1962), preconizes equal inferences for proportional likelihood functions, thus the likelihood itself is a composition.

The standard information theory (e.g. Gray, 2011), assigns a measure of uncertainty to a vector of probabilities called (Shannon) *entropy*,

$$\mathscr{H}_S(\mathbf{p}) = -\sum_{i=1}^{D} p_i \log p_i \,. \tag{3}$$

The terms $\log(1/p_i)$, $i = 1,2\ldots,D$, were proposed by Hartley (1928) as information provided by the observation of the event A_i. Defining a random variable which takes the values $\log(1/p_i)$ with probability p_i, Equation 3 is the mean of such random variable. Then, within the framework of the standard information theory, differences of entropies, for instance, after and before observing the result of an experiment, gives a measure of information. There are several ways of measuring these differences of uncertainties or entropies. The most popular is the Kullback-Leibler divergence (Kullback, 1997) which considers the differences $\log(1/f_i) - \log(1/p_i)$ and takes the mean using the posterior probabilities f_i

$$\mathscr{I}_{KL}(\mathbf{f}:\mathbf{p}) = \sum_{i=1}^{D} f_i \log \frac{f_i}{p_i} \ ,$$

using the notation $\mathbf{p}$ (prior) and $\mathbf{f}$ (final) in the Bayes' formula (1). Following Lindley (1956), the information, assigned to a vector of probabilities like $\mathbf{p}$, is

$$\mathscr{I}_S(\mathbf{p}) = \sum_{i=1}^{D} p_i \ \log p_i = -\mathscr{H}_S(\mathbf{p}) \ . \tag{4}$$

These measures of information, and many other entropy divergences (e.g. Martín-Fernández, 2001, and references therein) are not invariant under scaling of $\mathbf{p}$ and $\mathbf{f}$ and, therefore, the computation of $\mathscr{I}_S$ or $\mathscr{I}_{KL}$ requires that $\mathbf{p}$ and $\mathbf{f}$ are normalized, i.e. their components sum to 1. This is a major inconvenience for likelihood functions which, in general, are not normalized. A symmetrized and compositional version of the Kullback-Leibler divergence is given by Martín-Fernández (2001).

From the compositional point of view, the three compositions, $\mathbf{p}$, $\mathbf{q}$ and $\mathbf{f}$, live in the same space, $\mathbb{S}^D$, equipped with the Aitchison geometry (see discussion in the continuous case by Egozcue et al., 2013). Furthermore, the three compositions model the uncertainty on the actual event A_i or, from the opposite point of view, the evidence in favour of these events. This motivates calling the three compositions *evidence functions* or e-functions for short.

With this terminology, evidence functions are vectors and Bayes updating is just vector addition (perturbation) in the space of e-functions. Figure 1, illustrates these facts. In the left panel the three evidence functions (prior, likelihood and posterior) are represented as probabilities. The likelihood corresponds to the visual observation of moderate damage (vMod), the prior corresponds to the subjective impression of almost complete destruction of houses in the neighbourhood ($\Pr(A_4) = 0.7$, $\Pr(A_1) = \Pr(A_2) = \Pr(A_3) = 0.1$), which was selected for clarity of the picture. The right panel shows the three evidence functions as vectors, in which the posterior is the vector sum of the prior and the likelihood. The simplicity of the vectorial representation contrasts with the difficulties in comparing the proportions in the left panel.

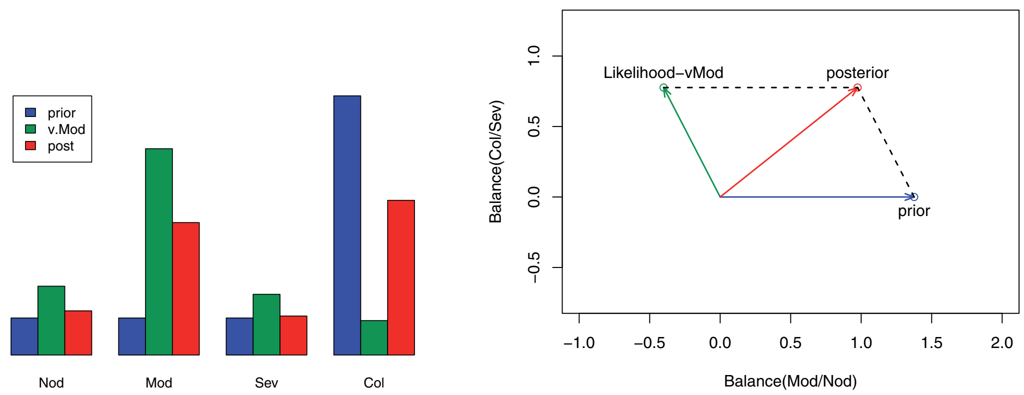

Figure 1: *Left panel: evidence functions, prior (blue), likelihood, corresponding to R =vMod (green), posterior (red) for the actual states Nod, Mod, Sev, Col. Right panel: the Bayes' formula in the two first coordinates; it appears as a vector addition. See definition of coordinates in Section 4.*

The consequences of the vectorial character of evidence functions are multiple. Bayes' formula, (1) and (2), has the equivalent expression in ilr coordinates or in clr coefficients (see Appendix A), that is

$$\text{ilr}(\mathbf{f}) = \text{ilr}(\mathbf{q}) + \text{ilr}(\mathbf{p}) \quad , \quad \text{clr}(\mathbf{f}) = \text{clr}(\mathbf{q}) + \text{clr}(\mathbf{p}) \ ,$$

where the additive character of the Bayes updating is explicit. The size of a vector is described by its norm (or a monotone function of it), regardless of its direction, a fact which motivates the definition of a scalar measure of information (Section 3). Vectors in a Euclidean space can be parallel, orthogonal, unitary; they can be projected one onto other, approximated by linear combinations of other vectors; distances between them are available, they can be expressed in coordinates. Remarkably, all these concepts and operations can be applied to or performed on evidence functions and, consequently, to information: *information* represented by *evidence functions* is a vectorial magnitude.

The parallelogram property of vectors in Euclidean spaces can be rephrased in terms of Bayesian updating. Consider the result of an experiment which provides a likelihood function $\mathbf{q}$. Imagine that two different priors, $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$, are proposed, for instance, in the fire scenario the prior initially mentioned, $\mathbf{p}^{(1)} = (p_1^{(1)}, \ldots, p_D^{(1)}) = (0.80, 0.15, 0.04, 0.01)$, and that used in Figure 1, denoted $\mathbf{p}^{(2)} = (p_1^{(2)}, \ldots, p_D^{(2)}) = (0.1, 0.1, 0.1, 0.7)$ (Table 1). The Aitchison distance between $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$, $\text{d}_a(\mathbf{p}^{(1)}, \mathbf{p}^{(2)})$, can be easily computed using any of the expressions in Equation (12) of Appendix A. In the example, this Aitchison distance is approximately 4.65 and the norms are $\|\mathbf{p}^{(1)}\|_a = 3.24$ and $\|\mathbf{p}^{(2)}\|_a = 1.69$, that is, $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ are neither orthogonal nor parallel (see Table 1). In fact, the two priors were designed to represent very different situations: $\mathbf{p}^{(1)}$ assumes that the zone, being largely affected by fire, has not been completely destroyed; for $\mathbf{p}^{(2)}$ houses which are completely destroyed are a large majority. These two priors $\mathbf{p}^{(1)}$, $\mathbf{p}^{(2)}$ can be updated with the same likelihood $\mathbf{q}$, thus obtaining two different final probabilities $\mathbf{f}^{(1)} = \mathbf{p}^{(1)} \oplus \mathbf{q}$, $\mathbf{f}^{(2)} = \mathbf{p}^{(2)} \oplus \mathbf{q}$. Elementary properties of Aitchison geometry, as a Euclidean geometry, state that the perturbation difference between $\mathbf{f}^{(1)}$ and $\mathbf{f}^{(2)}$ is that of the priors, that is

$$\mathbf{p}^{(1)} \ominus \mathbf{p}^{(2)} = (\mathbf{p}^{(1)} \oplus \mathbf{q}) \ominus (\mathbf{p}^{(2)} \oplus \mathbf{q}) = \mathbf{f}^{(1)} \ominus \mathbf{f}^{(2)} .$$

Hence, due to the parallelogram property of vectors in Euclidean spaces, the relation between the Aitchison distances is

$$\mathrm{d}_a(\mathbf{p}^{(1)}, \mathbf{p}^{(2)}) = \mathrm{d}_a(\mathbf{f}^{(1)}, \mathbf{f}^{(2)}) = \mathrm{d}_a(\mathbf{p}^{(1)} \oplus \mathbf{q}, \mathbf{p}^{(2)} \oplus \mathbf{q}) .$$

This means that the difference of prior e-functions is equal to the difference of posterior e-functions, provided that the likelihood was the same in the application of Bayes' formula. This result is different when using Shannon information or Kullback-Leibler divergence, for which the information provided by an experiment depends on the prior, a property which is well accepted in Bayesian statistics. These facts, are illustrated in Section 4.

3. Scalar information in an evidence function

Which of two results of an experiment is the most informative? This natural question cannot be answered if information is a vector, as real vectors cannot be ordered. A scalar measure of information associated with e-functions is needed, despite their vectorial character. The norm of an e-function, as a composition represented in $\mathbb{S}^D$, is the natural candidate for a scalar measure of information. Consequently, the scalar information contained in an e-function, $\mathbf{f} = (f_1, f_2, \ldots, f_D) \in \mathbb{S}^D$, is defined as

$$\mathscr{I}_e(\mathbf{f}) = \|\mathbf{f}\|_a , \tag{5}$$

where $\|\cdot\|_a$ denotes the Aitchison norm of a composition (Appendix A, Eq. 13). Therefore, the scalar information $\mathscr{I}_e$ has all standard properties of a vector norm. Some properties, which have a meaningful interpretation in the framework of information, are detailed below (Egozcue and Pawlowsky-Glahn, 2011). It is worth comparing the following properties of $\mathscr{I}_e$ with those which are satisfied by the Shannon entropy, $\mathscr{H}_S$, for instance, those proposed by Shannon (1948), Khinchin (1957) or Ash (1990). Entropy is conceived as a measure of uncertainty, and information is then defined from differences between initial and final entropy (Kullback and Leibler, 1951a, Khinchin, 1957, Ash, 1990), or even as negative entropy (Eq. 4) (Lindley, 1956).

Null e-Information. A flat e-function $\mathbf{n} = (1/D, 1/D, \ldots, 1/D)$ does not provide any information, as $\mathscr{I}_e(\mathbf{n}) = 0$ corresponds to the neutral element in $\mathbb{S}^D$. This property is shared by all definitions of measures of information, alternatively entropy. Note that, reciprocally, for any e-function, $\mathscr{I}_e(\mathbf{f}) = 0$ implies $\mathbf{f} = \mathbf{n}$. This is due to the fact that $\|\mathbf{f}\|_a$ is the Aitchison distance (not a divergence) from $\mathbf{f}$ to the neutral element $\mathbf{n}$.

Continuity. Also common to all definitions of information is the continuity of information/entropy with respect to each component of the e-function. The ilr coordinates (Eq. 10 in Appendix A) are continuous functions of the components of the e-functions. Also the Aitchison norm (Eq. 13 in Appendix A) is a continuous function of the ilr coordinates. Then, $\mathscr{I}_e(\mathbf{f})$ is a continuous function of the $\mathbf{f}$ components. The only critical points are those in which one or more parts $f_i = 0$, as null components place the value of $\mathscr{I}_e(\mathbf{f})$ at infinity. Knowledge of the impossibility of event A_i represents the strongest information. It forces the change of sample space just by removing event A_i. This is opposite to the case of Shannon information, where $-\log f_i$ is minus infinity before averaging, while $\mathscr{I}_S$ (Eq. 4) remains unaltered after averaging with null probability.

Monotonicity. A set of properties was used to introduce (Shannon) *entropy*, $\mathscr{H}_S$, in an axiomatic way. Simultaneously, entropy was taken as opposite to information (for the Shannon case $\mathscr{I}_S = -\mathscr{H}_S$). Following Ash (1990), the monotonicity property for Shannon entropy $\mathscr{H}_S$ is that, if d and D, $d < D$, are the number of parts of two neutral compositions then

$$\mathscr{H}_S(1/d, 1/d, \ldots, 1/d) < \mathscr{H}_S(1/D, 1/D, \ldots, 1/D) ,$$

which, loosely speaking, means that uncertainty or entropy increases with the number of components, here written for neutral compositions. This statement is not really useful for a measure of information which attains a null value at neutral elements, like $\mathscr{I}_e(1/d, 1/d, \ldots, 1/d) = 0$. In the case of $\mathscr{I}_e$ this kind of monotonicity is captured by the subcompositional dominance property of the Aitchison distance (e.g. Aitchison, 1983, Egozcue and Pawlowsky-Glahn, 2018), which is formulated as follows. Let $\mathbf{x}$ and $\mathbf{y}$ be compositions in $\mathbb{S}^D$ and their corresponding d-part ($d < D$) subcompositions $\mathbf{x}_d$ and $\mathbf{y}_d$. Then, $\mathrm{d}_a(\mathbf{x}, \mathbf{y}) \geq \mathrm{d}_a(\mathbf{x}_d, \mathbf{y}_d)$. When $\mathbf{y} = (1/D, \ldots, 1/D)$ (the neutral element), distances become norms and

$$\mathscr{I}_e(\mathbf{x}) \geq \mathscr{I}_e(\mathbf{x}_d) \quad , \quad D > d ,$$

which means that the information contained in a d-part subcomposition of an e-function is always less than or equal to the information contained in the (D-part) original e-function.

Null information extension. In Shannon entropy/information theory, extending the e-function with zeroes does not decrease entropy or increase information (e.g. Khinchin, 1957). This is a direct consequence of the fact that, for $p_{D+1} = 0$ the term $p_{D+1} \ln p_{D+1}$ is assumed null, and the previous information in Equation (4) remains unaltered after adding the term. This situation is completely different for $\mathscr{I}_e$. It can be proven that

$$\mathscr{I}_e(\mathbf{x}) = \mathscr{I}_e(\mathbf{x}, x_{D+1}) \quad \text{if and only if} \quad x_{D+1} = \mathrm{g_m}(\mathbf{x}) \ ,$$

that is, adding a part, x_{D+1}, equal to the geometric mean of the previous e-function does not alter $\mathscr{I}_e$. In fact, the extended composition can be represented using a system of ilr coordinates valid for $\mathbf{x}$, plus a new coordinate

$$b_D = \sqrt{\frac{D}{D+1}} \log \frac{x_{D+1}}{\mathrm{g_m}(\mathbf{x})} = 0 \ ,$$

which corresponds to completing a previous Sequential Binary Partition (SBP) (see Appendix A) for $\mathbf{x}$ with a sign code row $(-1, -1, \ldots, -1, +1)$. When computing the square Aitchison norm of $(\mathbf{x}, x_{D+1})$ a null term (Eq. 13 in Appendix A) is added.

The idea that extending a likelihood function, or other e-function, with zeros does not change the information provided by the experiment is counterintuitive: the result of the experiment informs the analyst that one or more categories are impossible, which would imply a great amount of information (infinite if using $\mathscr{I}_e$ as an information measure). This null extension seems acceptable when speaking of entropy or uncertainty: adding a null probability term to the e-function does not increase uncertainty. This reveals that Shannon entropy, $\mathscr{H}_S$ should have a more elaborated relation with information than just that expressed by $\mathscr{I}_S = -\mathscr{H}_S$; this can be seen in alternative interpretations of both magnitudes (Kullback and Leibler, 1951b, Ash, 1990).

Decomposition of an e-function. Consider a D-part e-function, $\mathbf{y}$, built appending two compositions, $\mathbf{x}_1$ with D_1 parts and $\mathbf{x}_2$ with D_2 parts. Then, $D = D_1 + D_2$. The compositions are appended after multiplying by arbitrary positive constants a_1 and a_2; that is, $\mathbf{y} = (a_1\mathbf{x}_1, a_2\mathbf{x}_2)$. The information conveyed by $\mathbf{y}$ is then

$$\mathscr{I}_e^2(\mathbf{y}) = \mathscr{I}_e^2(\mathbf{x}_1) + \mathscr{I}_e^2(\mathbf{x}_2) + \frac{D_1 D_2}{D_1 + D_2} \log^2 \frac{a_1 \mathrm{g_m}(\mathbf{x}_1)}{a_2 \mathrm{g_m}(\mathbf{x}_2)} \ . \tag{6}$$

The role of a_1 and a_2 is quite irrelevant, but they highlight the possibility of renormalizing the two compositions.

This kind of property differs from the corresponding property of Shannon entropy, mainly due to the assumed scalar character of information, and also to the need of renormalization. The property for the Shannon entropy, known as grouping axiom (Ash, 1990), is

$$\mathscr{H}_S(\mathscr{C}\mathbf{y}) = m(\mathbf{x}_1)\mathscr{H}_S(\mathscr{C}\mathbf{x}_1) + m(\mathbf{x}_2)\mathscr{H}_S(\mathscr{C}\mathbf{x}_2) + \mathscr{H}_S(m(\mathbf{x}_1), m(\mathbf{x}_2)) \ ,$$

where $\mathscr{C}$ is the closure operation (Appendix A), and $m(\mathbf{x}_k)$ is the sum of the components of $\mathscr{C}\mathbf{y}$ within the composition $\mathbf{x}_k$ ($k = 1, 2$). Note that the computation of $\mathscr{H}_S$ requires normalization, and $m(\mathbf{x}_k)$ ($k = 1, 2$) are the dividing normalization constants.

Independent probability table. Let be $\mathbf{x}_1 \in \mathbb{S}^{D_1}$ and $\mathbf{x}_2 \in \mathbb{S}^{D_2}$ two e-functions and $A = [a_{ij}]$ a (D_1, D_2) table of probabilities such that $a_{ij} = x_{1i} x_{2j}$. Then A, up to normalization, is an independent table of probabilities. The scalar information associated with this table as e-function is

$$\mathscr{I}_e^2(A) = D_2\ \mathscr{I}_e^2(\mathbf{x}_1) + D_1\ \mathscr{I}_e^2(\mathbf{x}_2)\ . \tag{7}$$

To prove this statement, construct a (D_1, D_2) table A_2 with D_1 identical rows, each one equal to $\mathbf{x}_2$. Similarly, build a (D_1, D_2) table A_1 with D_2 identical columns, each one equal to $\mathbf{x}_1$. The entry-wise multiplication, or matrix perturbation $A_1 \oplus A_2$ (Egozcue et al., 2015), of these two tables is A. In Egozcue et al. (2015) it is proven that A_1 and A_2 as compositions are orthogonal, $\langle A_1, A_2\rangle_a = 0$. Consequently, the square Aitchison norm of A is the sum of the square Aitchison norms of A_1 and A_2 (Pythagoras' theorem). On the other hand, the square norm $\|A_1\|_a^2 = D_2\|\mathbf{x}_1\|_a^2$, as proven by Egozcue and Pawlowsky-Glahn (2019, Appendix A). A similar result holds for $\|A_2\|_a^2$, what implies the statement.

This is not what is expected in the Shannon information theory, in which the result is $\mathscr{H}_S(A) = \mathscr{H}_S(\mathbf{x}_1) + \mathscr{H}_S(\mathbf{x}_2)$, as reported, for instance, by Shannon (1948). The main difference with respect to Equation (7) is that additivity of entropy or information is thought in a scalar form in the Shannon theory; in the compositional approach information is thought as a vector (composition). In this case, independence is translated into orthogonality, thus reproducing the Pythagorean sum of squares in a Euclidean space.

Unit of information in evidence functions. The *bit* has been accepted as a unit of information since early works in the field. A bit is the Shannon information unit (using logarithms in basis 2) conveyed by an equiprobable binary code. It is obvious that this kind of definition is well adapted to the study of communications and coding theory. However, it is almost not interpretable in the present context of evidence functions and the scalar measure of information $\mathscr{I}_e$. In its place, a new unit of information adapted to e-functions is here proposed.

Consider an e-function $\mathbf{p} = (p_1, p_2, \ldots, p_D)$ and a perturbation with a non closed composition $\mathbf{q} = (u, u^{-1}, 1, 1, \ldots, 1)$, $u = \exp(\sqrt{1/2})$. Then, $\mathbf{f} = \mathbf{p} \oplus \mathbf{q}$ is a shift of $\mathbf{p}$ towards $\mathbf{f}$. In order to compute $\mathscr{I}_e(\mathbf{q})$, one can decompose $\mathbf{q}$ into (u, u^{-1}) and the neutral element $\mathbf{n}$, and use the decomposition property (6) which yields

$$\mathscr{I}_e^2(\mathbf{q}) = \mathscr{I}_e^2(u, u^{-1}) = \left(\sqrt{\frac{1}{2}}\ \log\frac{u}{1/u}\right)^2 = 1\ .$$

Therefore, the perturbing composition $\mathbf{q}$ has a unit e-information, $\mathscr{I}_e(\mathbf{q}) = 1$. However, this perturbation has an approximate interpretation. In fact $\exp(\sqrt{1/2}) = 2.028 \simeq 2$. *A perturbing composition doubling a component, halving another one, and retaining unaltered other components has, approximately, unit e-information.* There are many other e-functions which have unit e-information, but they involve more than two parts. In Figure 2 circles with radii 1, 5, 10 have been plotted. The smallest one is the loci of e-functions with unit e-information.

4. Acquisition of information from an experiment

The fire scenario briefly described in previous sections is studied here in more detail. Consider a suburban zone close to some forest at fire risk. Authorities in charge of safety have to design a mitigation plan for fire affecting the zone. The responsible team may consider several *a priori* hypotheses about the possible states of houses and buildings after the fire. Two of these *a priori* hypotheses have been denoted $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ in Section 2, giving the *a priori* probabilities of a house remaining in the four considered states: no damage (Nod), moderate damage (Mod), severe damage (Sev), collapse or ruin (Col). These two prior distributions of the state of a house correspond to quite different feelings about the effects of the fire. Figure 3 shows $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ as compositional vectors in ilr coordinates, defined in Table 6 by the sign code of an SBP. Observing the prior vectors (circled arrows) in Figure 3, they do not appear as close to orthogonality. Orthogonality of two e-functions means that their information is on unrelated features. In this case, the two considered priors do inform on some common features. Table 1 shows the prior e-functions, their $\mathscr{I}_e$ and the angle they form which is 43^o, thus reflecting the difference in direction of the two priors. The $\mathscr{I}_e$ also differs, since $\mathbf{p}^{(1)}$ is quite more informative than $\mathbf{p}^{(2)}$ (see Table 1).

Table 1: *Two priors of the state of houses in a suburban zone after a fire, with no damage (Nod), moderate damage (Mod), severe damage (Sev) and collapse or ruin (Col) and evidence information* $\mathscr{I}_e$.

e-function	Nod	Mod	Sev	Col	$\mathscr{I}_e$	angle	with
$\mathbf{p}^{(1)}$	0.80	0.15	0.04	0.01	3.24	43.27^o	$\mathbf{p}^{(2)}$
$\mathbf{p}^{(2)}$	0.10	0.10	0.10	0.70	1.69	-43.27^o	$\mathbf{p}^{(1)}$

Next step is studying which inspection procedures are at hand to assess the state of a house after a fire. Two realistic experiments are considered here. The first one consists of a visual inspection of the house by a small trained team. The second is based in airborne photography; the house is identified and its state is assessed on the picture. In what follows, the results of both types of inspection are labelled as the four considered states, adding v (visual) or a (airborne), depending on the type of inspection used. Both types of inspection are uncertain due to several reasons: the inspectors do not know the status of the house previous to the fire; vegetation, burnt or not, can mask relevant details of the structure; access to some parts of the building can be difficult; structural damage can be hidden; there can be errors in the identification of the house, etc. In order to use the result of an inspection to make decisions under a controlled uncertainty, the likelihood of each actual state should be known. Therefore, some assessment of the probability of each outcome of the inspection, conditional to the actual state, is needed. Tables 2 and 3 show the likelihood e-functions (the columns of the tables) for the two types of inspections, visual and airborne, respectively.

Table 2: *Simulated likelihood for the visual inspection of houses. Each column is the likelihood associated with R, i.e. the probabilities of the visual inspection outcome conditional to the actual states,* $\Pr(R|A_i)$*. Row* $\mathscr{I}_e$ *(likelihood) shows the scalar information of the likelihood associated with the observation* R_k*. Rows* $\Pr[R_k^{(i)}]$ *are the probabilities of observing* R_j *given the prior* $\mathbf{p}^{(i)}$*,* $i = 1, 2$*, and the likelihood.*

Actual	**Visual inspection,** R			
state	vNod	vMod	vSev	vCol
No damage (Dam)	0.7665512	0.2001012	0.0333408	0.0000068
Moderate damage (Mod)	0.2000307	0.5999432	0.1201175	0.0799086
Severe damamage (Sev)	0.1176475	0.1765397	0.5293121	0.1765006
Collapse or ruin (Col)	0.0000001	0.1001036	0.1999045	0.6999918
$\mathscr{I}_e$(likelihood)	12.87	1.30	1.99	9.10
$\Pr[R_k^{(1)}]$	0.648	0.258	0.068	0.026
$\Pr[R_k^{(2)}]$	0.108	0.168	0.208	0.516

Table 3: *Simulated likelihood for the airborne inspection of houses. Each column is the likelihood associated with Q, i.e. the probabilities of an outcome of the airborne inspection conditional to the actual states,* $\Pr(Q|A_i)$*. Row* $\mathscr{I}_e$ *(likelihood) shows the scalar information of the likelihood associated to the observation* Q_k*. Rows* $\Pr[Q_k^{(i)}]$ *are the probabilities of observing* Q_j *given the prior* $\mathbf{p}^{(i)}$*,* $i = 1, 2$ *and the likelihood.*

Actual	**Airborne inspection,** Q			
state	aNod	aMod	aSev	aCol
No damage (Nod)	0.6436847	0.3563042	0.0000067	0.0000044
Moderate damage (Mod)	0.3725228	0.5097669	0.0882470	0.0294632
Severe damage (Sev)	0.0860468	0.0967638	0.4408675	0.3763220
Collapse or ruin (Col)	0.0000021	0.0204390	0.2040838	0.7754751
$\mathscr{I}_e$(likelihood)	10.31	2.53	8.99	9.62
$\Pr[Q_k^{(1)}]$	0.574	0.366	0.033	0.027
$\Pr[Q_k^{(2)}]$	0.110	0.111	0.196	0.583

These likelihood tables can be estimated from previous experience in inspection of buildings, which are used as training data for a likelihood model. For instance, a number of houses affected by fire for which the actual state is known were inspected and the result of the inspection was reported. With this kind of data a discriminant analysis of the response of the inspection gives an estimate of the probabilities of the observed state R, conditional to the true state A_i, $\Pr(R|A_i)$. Tables 2 and 3 are the result of a logistic regression on a training set of simulated inspections (not shown in this paper).

In order to represent e-functions in coordinates a contrast matrix (Pawlowsky-Glahn et al., 2015) has been selected. The sign code of the SBP is shown in Table 4. A first look at Tables 2 and 3 reveals the large uncertainty of both types of inspection. Also, some features are clear. For instance, it seems that airborne photography is not efficient in discriminating Nod from Mod and Sev from Col. However, it is able to distinguish

Table 4: *Sign code of SBP defining the coordinates used in the fire scenario.*

coordinate	Nod	Mod	Sev	Col	Expression
1	-1	-1	$+1$	$+1$	$\log(\sqrt{\text{Sev Col}}/\sqrt{\text{Nod Mod}})$
2	-1	$+1$	0	0	$(1/\sqrt{2})\log(\text{Mod}/\text{Nod})$
3	0	0	-1	$+1$	$(1/\sqrt{2})\log(\text{Col}/\text{Sev})$

quite reliably between the two pairs of states. These kinds of interpretation can be improved by computing and representing each likelihood e-function in coordinates, so that the direction and strength of the information are better shown. Figure 2 shows the likelihood e-functions in the ilr-coordinates defined by the SBP coded in Table 4. Although the choice of the SBP is arbitrary and the results of the analysis do not depend on the selected basis, the SBP shown in Table 4 tries to remark the order of damage, from small (-1) to large $(+1)$. Two projections are used for the three-dimensional picture: first and second ilr-coordinates (left panel) and first and third ilr-coordinates. Likelihood e-functions are represented by red and blue arrows associated with the visual (v) and airborne (a) inspections, respectively. The length of the arrows are the corresponding scalar information $\mathscr{I}_e$. The first observation is that inspections resulting in no damage (vNod, aNod) or in collapse or ruin (vCol, aCol) are more informative (all of them exceed 5 units of information; see Tables 2 and 3) than the moderate damage outcomes (vMod,aMod). This is due to the fact that Nod and Col observations in both experiments almost exclude the opposite state, Col and Nod, respectively; alternatively vMod, vSev, aMod do not exclude any actual state and they are less resolutive. The most important difference in information between the visual and airborne inspection is related to the severe damage outcome (vSev, aSev). The aSev outcome is relatively much more infor-

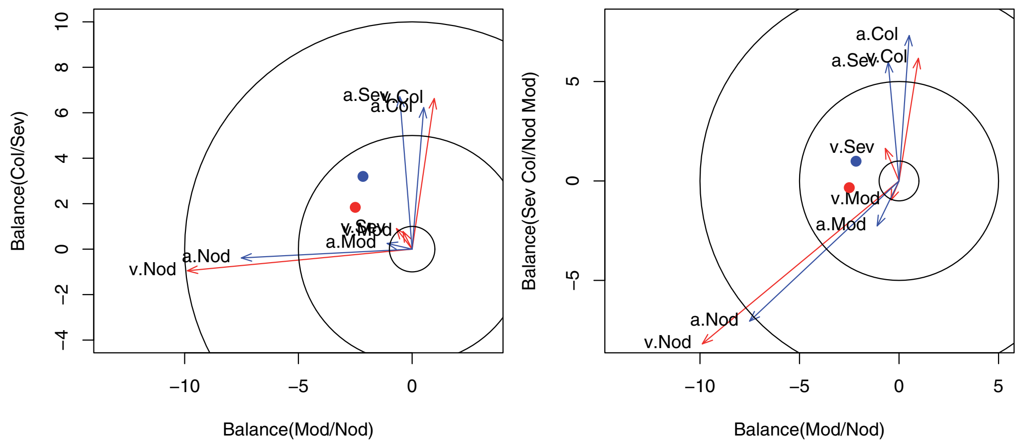

Figure 2: *Likelihood functions as compositions in coordinates. Circles of radius 1, 5, 10. Visual inspection, red arrows; Airborne inspection, blue arrows. Projection first and second coordinates, left panel; first and third coordinates, right panel. Filled markers are the vector averages of the Likelihood functions; red, blue correspond to visual and airborne inspections.*

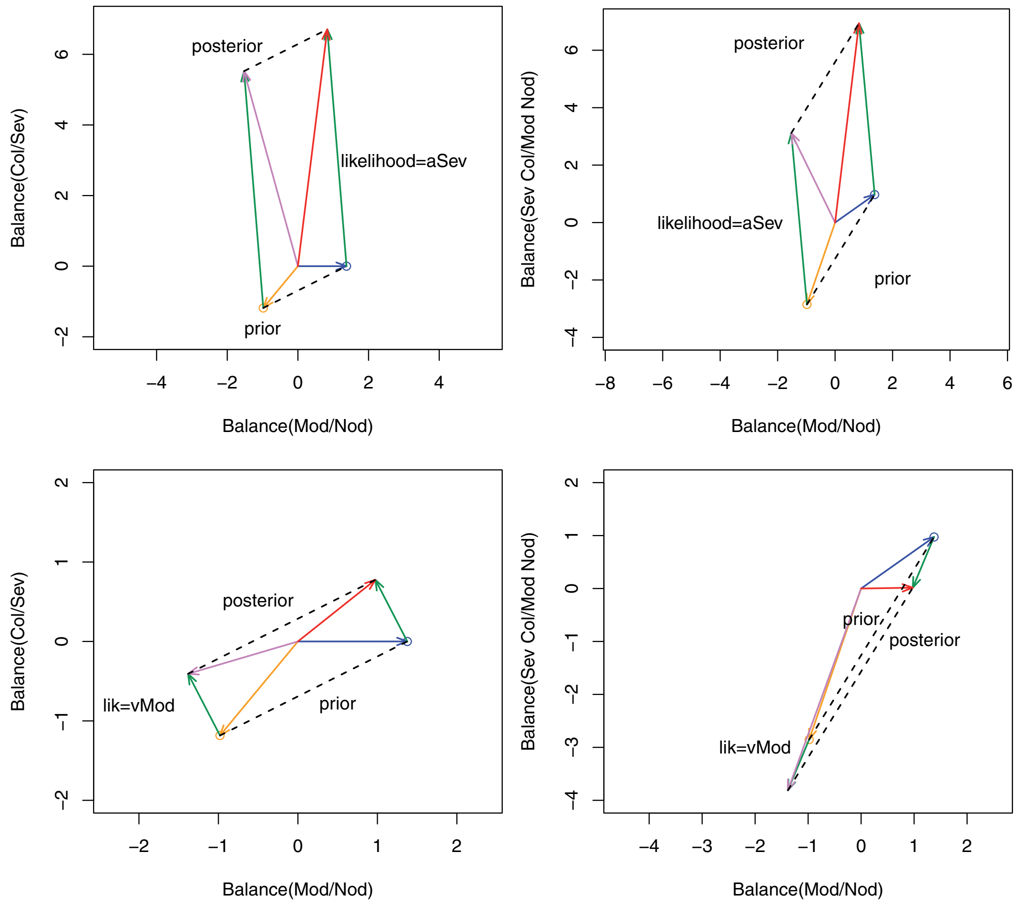

Figure 3: *Bayesian updating: two different priors* $\mathbf{p}^{(1)}$ *(orange, end arrow circled) and* $\mathbf{p}^{(2)}$ *(blue, end arrow circled) are updated with two likelihood cases corresponding to aSev (top panels) and vMod (bottom panels). Left panels show the projection on coordinates 2 and 3 and right panels show projection on coordinates 2 and one as ordered in Table 4. Likelihood (green) is added as a vector to prior. Obtained posteriors* $\mathbf{f}^{(1)}$ *(violet) and* $\mathbf{f}^{(2)}$ *(red) are linked by dotted lines. Priors are also linked by a dotted line to show the parallelogram rule.*

mative, in the scalar sense, than vSev. However, the informative strength of aSev is at the price that aSev gives information that can be confounded with aCol (also with vCol).

The disposition of the likelihood e-functions in both inspections also reveals weaknesses in the design of the inspections. The likelihood arrows in Figure 2 are shifts applied to the prior e-functions. A good design of the experiments should be able to shift the prior in any direction in the three dimensions. Note the inability of these likelihood functions to shift the posterior towards positive values of the balance (Mod/Nod) (second coordinate in Table 4, see Appendix A for further explanation) or negative values of the balance (Col/Sev) (third coordinate in Table 4).

Figure 3 shows the Bayesian updating of the two proposed priors, ($\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$), using outcomes aSev (observed severe damage in the airborne inspection) and vMod (observed moderate damage in the visual inspection) for updating. Top panels of Figure 3 show the two considered priors $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$ (Table 1) updated by the likelihood corresponding to the observation of severe damage in the airborne inspection (aSev) in two coordinate projections. The main features are: (a) Likelihood e-functions are not parallel to both priors; consequently, prior assumptions are not confirmed by the observation. It is important to note that parallelism of e-functions would mean that prior assumptions are confirmed by the observations; alternatively, orthogonality of two e-functions means that the e-information they convey do not interact or, more intuitively, they are about different aspects of the scenario. (b) The likelihood is more informative than the two considered priors, i.e. $\mathscr{I}_e(\mathbf{q}) > \mathscr{I}_e(\mathbf{p}^{(k)})$, $k = 1, 2$ (See also Tables 1 and 3). (c) The updating hardly modifies the prior coordinate balance of moderate damage (Mod) over no damage (Nod), as the likelihood is almost in the plane defined by the other two coordinates.

Bottom panels of Figure 3 show the two considered priors, $\mathbf{p}^{(1)}$ and $\mathbf{p}^{(2)}$, updated with the likelihood corresponding to the observation of moderate damage in the visual inspection (vMod) in the same projections shown in the top panels. The situation is different from the previous case. Again, the observation does not clearly confirm any of the two priors considered, but the length of the likelihood, $\mathscr{I}_e(\mathbf{q})$, is now smaller than that of the priors: $\mathscr{I}_e(\mathbf{q}) = 1.30$, while $\mathscr{I}_e(\mathbf{p}^{(1)}) = 3.24$, $\mathscr{I}_e(\mathbf{p}^{(2)}) = 1.69$, thus providing a weak change of evidence information from prior to posterior.

Evaluation of visual and airborne inspections

Up to now, only effects of a given observation have been examined. However, decision makers are commonly interested in the evaluation of the available types of inspection, both to know the economical implications of conducting each inspection and how informative they are. Thus, they are interested in the initial question of *which of the two inspections is more informative?* This question was addressed both by Lindley (1956) and in the context of evidence-functions by Egozcue and Pawlowsky-Glahn (2011). In both contributions an average of information provided by possible results of the experiments is proposed. However, it can be discussed which kind of average is more convenient, or which weights are adequate. Here information has a vectorial character, as proposed in Section 2, and accordingly we are primarily concerned with vector averages.

A first possibility is to ignore the probability of each result of an experiment (inspection in our case). This is like considering the experiment outside its context. If the possible likelihood e-functions of the experiments are $\mathbf{q}_k$, $k = 1, 2, \ldots, K$ (in the particular case of the considered inspections $K = 4$), the vector average of the likelihood e-functions is

$$\overline{\mathbf{q}} = \frac{1}{K} \odot \bigoplus_{k=1}^{K} \mathbf{q}_k ,$$

which is the compositional centre of the set of possible likelihood e-functions. When the e-functions are expressed in coordinates, this is simply the average of the coordinates. These averaged likelihood e-functions are represented in Figure 2 with red and blue markers for the visual and the airborne inspections. If $\overline{\mathbf{q}}$ is not close to the neutral element, it points out that the experiment is quite unable to shift the posterior in the opposite direction. This is the case of both inspections in this example. This motivates the name of *e-information bias* for $\overline{\mathbf{q}}$ or for its norm $\mathscr{I}_e(\overline{\mathbf{q}})$. An experiment with $\overline{\mathbf{q}}$ near the neutral element has the possibility to update the prior e-functions in any direction and is here called *e-information unbiased experiment*.

Common sense points out that the informative value of an experiment depends on the probability of obtaining any outcome. This requires to put the experiment in a particular probabilistic context, which is completely described when the prior e-function is given. In fact, assume that L is a $(K,D)-$matrix with entries $\Pr(R_k|A_i)$, where R_k are the possible outcomes of the experiment R. Tables 2 and 3 show examples of such matrices for the visual and airborne experiments. Matrix multiplication of L and prior probabilities $\mathbf{p}$ give the marginal probabilities for R_k, $\Pr(R_k)$, known as *predictive probabilities* for the observations R_k. The probabilistic weighted average of likelihood e-function is

$$E_R[\mathbf{q}] = \bigoplus_{k=1}^{K} (\Pr(R_k) \odot \mathbf{q}_k) , \tag{8}$$

which is the mean likelihood of an experiment in a given probabilistic context. Note that once the prior probabilities and the matrix L are given, the predictive probabilities are also determined. The mean likelihood e-function and its norm, $\mathscr{I}_e(E_R(\mathbf{q}))$, can be considered suitable descriptors of the information provided by an experiment. They can be used to compare experiments.

There are more possibilities of averaging information provided by an experiment. One of them is to average scalar values of $\mathscr{I}_e(\mathbf{q}_k)$. However, a discussion on which is an appropriate scale for $\mathscr{I}_e(\mathbf{q}_k)$ is convenient. In general, the scale of $\mathscr{I}_e(\mathbf{q}_k)$ can be transformed by a monotonous, invertible function $\phi : \mathbb{R}_+ \to \mathbb{R}$ and, then, one can proceed to a weighted average of the transformed values of $\phi(\mathscr{I}_e(\mathbf{q}_k))$. For a general ϕ, it is

$$E_R^{\phi}[\mathscr{I}_e(\mathbf{q})] = \phi^{-1}\left(\sum_{k=1}^{K} \Pr(R_k)\ \phi(\mathscr{I}_e(\mathbf{q}_k))\right) . \tag{9}$$

The scaling function for averaging scalar information has been used by Egozcue and Pawlowsky-Glahn (2011). Table 5 reports some of the available options of scaling functions ϕ. These options are used to evaluate the mean information in Equation (9), provided by the visual and airborne inspections in the fire scenario, and are also reported in Table 5.

Table 5: *Values of mean of scalar e-information* $E_R^{\phi}[\mathscr{I}_e(\mathbf{q})]$ *(9) for the visual and airborne inspections. Probabilities of outcome* $\Pr^{(2)}[R_k]$ *are reported in Tables 2 and 3.*

Inspection		**Visual**		**Airborne**	
outcome pr.		$\Pr^{(1)}[R_k]$	$\Pr^{(2)}[R_k]$	$\Pr^{(1)}[Q_k]$	$\Pr^{(2)}[Q_k]$
ϕ name	$z=\phi(x)$	$E_R^{\phi}[\mathscr{I}_e(\mathbf{q})]$	$E_R^{\phi}[\mathscr{I}_e(\mathbf{q})]$	$E_Q^{\phi}[\mathscr{I}_e(\mathbf{q})]$	$E_Q^{\phi}[\mathscr{I}_e(\mathbf{q})]$
identity	$z=x$	2.262	1.681	1.850	2.197
square	$z=x^2$	5.249	3.931	4.139	4.534
neg. exp.	$z=\exp(-x)$	3.914	3.986	4.918	6.107
logarithm	$z=\log x$	1.579	1.493	1.573	1.695
square root	$z=\sqrt{x}$	0.488	0.369	0.425	0.535

Examining these results, one realizes that the mean values depend strongly on the used scaling function, and also on the probabilities of the outcome of the inspection (see values in Tables 2 and 3), which at the same time depend on the prior e-function selected. A second conclusion is that for each scaling function ϕ the most informative inspection depends on the prior. For instance, for ϕ being the identity, and for prior $\mathbf{p}^{(1)}$ and outcome probabilities $\Pr^{(1)}[R_k]$ and $\Pr^{(1)}[Q_k]$, the visual inspection is moderately more informative than the airborne inspection. The situation is reversed for the prior $\mathbf{p}^{(2)}$ and its corresponding outcome probabilities $\Pr^{(2)}[R_k]$ and $\Pr^{(2)}[Q_k]$.

In Table 5 two ϕ options deserve a comment. First, the negative exponential, which considers a monotonous decreasing function. The transformed values $\phi(\mathscr{I}_e(\mathbf{q}_k))$ no longer mean information but a measure of uncertainty or entropy. Accordingly, the average in Equation (9) is a mean value of uncertainties. When transforming back with $\phi^{-1}=-\log$, the mean measure of uncertainty is again translated into e-information. This approach seems quite appealing, but requires further research.

Also, in Table 5, the option $\phi=\log$ may be interesting when a relative scale is assumed for the scalar e-information. However, the relative scale can also be valid for large values of all $\phi(\mathscr{I}_e(\mathbf{q}_k))$ of the experiment. This is due to the fact that the value $\mathscr{I}_e(\mathbf{q}_k)=0$ assigned to the neutral likelihood is attainable, and the relative scale assumed is then nonsensical.

5. Conclusions and further research

The discrete case of Bayesian updating has been considered as a paradigm of information acquisition. Prior information, coded as a probability function, is changed into a final or posterior probability function when the discrete likelihood corresponding to an outcome of an experiment is used in the Bayes' formula. The central idea is that prior, posterior probability functions and, importantly, the discrete likelihood are considered compositions represented in the simplex. The simplex, endowed with the Aitchison geometry, is a Euclidean vector space. The three functions have the characteristics required by the Aitchison geometry of compositions, thus motivating the common name of evidence functions (e-functions). In this context, Bayes' formula appears exactly as a perturbation of compositions, prior perturbed with likelihood e-functions gives the posterior e-function as a result. The fact that perturbation is the vector sum (group operation) in the Aitchison geometry implies a number of properties; among them, vectors, e-functions in this case, can be represented in (Cartesian) coordinates, thus providing intuitive representations and easy computing of metrics (projections, distances, norms). The conclusion is that information, acquired through Bayes' formula, is a vector magnitude better than a scalar one, as traditionally assumed. Another consequence of this vector approach is that information can be conceived not only for prior and posterior probability functions, but also for likelihood functions which, at the end, is the vector difference between the posterior and the prior.

Generically, vectors have a direction and a modulus or norm. The same is valid for e-functions, which represent a direction of the evidence in the space of compositions and a strength of the evidence, which can be measured as the norm of the e-function. This scalar measure of information may be worth in applications and, accordingly, the norm of e-functions (e-information for short) is taken as a scalar measure of the information conveyed by an e-function. The vectorial character of e-functions introduces some changes in the traditional scalar measures of uncertainty (entropy) or in their counterpart of information. Some intricacies of standard information theory are easily overcome by the Euclidean geometry. For instance, the perturbation-subtraction of e-functions or their distance can advantageously replace divergences or mutual information.

A fire scenario has been used to introduce two kinds of inspection of houses. Questions as simple as *which outcome of the inspection is the most informative* or *which of the two inspections is the most informative?* motivate discussions that require simple operations in the Aitchison geometry. However, different kinds of averages of information provided by the likelihood of an experiment have their own interpretations. The main conclusion is that sensible averages of e-information of an experiment depend on the probabilities of observing the results, which at the same time are determined by the prior probabilities.

The theory and applications of information in evidence functions is not fully developed. A brief description of three possible research directions follows.

The continuous case. The generalization of the log-ratio approach of compositional data to the analysis of density functions, including probability densities, is available (Egozcue, Díaz-Barrero and Pawlowsky-Glahn, 2006, Boogaart, Egozcue and Pawlowsky-Glahn, 2010, Egozcue et al., 2013, Boogaart, Egozcue and Pawlowsky-Glahn, 2014). As in the discrete case, Bayes' theorem consists of the perturbation of the prior density by the likelihood. The continuous e-functions are densities of positive measures, and they are included in infinite dimensional vector spaces called Bayes spaces. Orthogonal projections of e-functions in reduced dimensions are safely introduced when the Bayes space has a Hilbert space structure. In the continuous case, Bayes Hilbert spaces provide orthonormal coordinates which are Fourier coefficients with respect to bases easily constructed. Some applications have been developed in the framework of geostatistics and functional data (e.g. Menafoglio, Guadagnini and Secchi, 2016, Menafoglio, Grasso, Secchi and Colosimo, 2018), but information applications are still pending.

Weighting e-functions. The theory of Bayes Hilbert spaces (Egozcue et al., 2006, Boogaart et al., 2014) requires a reference (probability) measure of the space. This is specially important when the densities (e-functions) considered have an unbounded support. For interval supported densities and for finite discrete support (compositions) a uniform reference measure is almost automatically adopted. However, this is not the case for infinite supports. This situation suggests that in the interval and compositional cases, adopting a non-uniform reference measure is possible, and in some cases even advisable, thus causing a weighting, in the metrics of the Aitchison geometry, of the information assigned to evidence functions. The way of changing the reference measure for compositions was introduced by Egozcue and Pawlowsky-Glahn (2016), but this approach should be developed and extended to continuous e-functions. In particular, the relationship between prior e-functions and reference measure require further study.

Connections with Dempster-Shafer theory of belief functions. An extensive summary of the theory of belief functions, mainly due to A. P. Dempster and G. Shafer can be found in Yager and Liu (2008), or in the book of Shafer (1976). Belief functions in Dempster theory are operated by Dempster's rule of combination of beliefs (Yager, 1987). Although the support of belief functions is not that of e-functions, the combination of belief functions is just a perturbation, similar to the Bayes' formula in Equation (1). This suggests that belief functions can be viewed as compositions, and the theory here exposed can be extended to belief functions. From this starting point, there is a plea of ideas that deserve attention, like the meaning of orthogonality of e-functions and of belief functions. They seem to be related to exchangeability and independence when using Bayes' formula. These are avenues that should be studied in the future.

Acknowledgements

The authors thank two anonymous referees for their constructive comments which helped improve the manuscript. This work was supported by grants MTM2015-65016-C2-1-R and MTM2015-65016-C2-2-R (MINECO/FEDER) of the Spanish Ministry of Economy and Competitiveness and European Regional Development Fund.

A Aitchison geometry

Based on the definitions of perturbation, powering and distance for compositions by Aitchison (1982, 1986), the set of D-part compositions, represented in the simplex $\mathbb{S}^D$, admits a Euclidean vector space structure (Billheimer et al., 2001, Pawlowsky-Glahn and Egozcue, 2001), which was termed *Aitchison geometry* in the latter reference.

The main elements of this geometry are the vector space operations, perturbation and powering, the metric elements, inner product, distance and norm, and the coordinates for the representation of compositions. In this Appendix A a quick operative reference of these elements is presented. A more comprehensive exposition can be found elsewhere (e.g. Pawlowsky-Glahn et al., 2015, and references therein).

Let $\mathbf{x} = (x_1, x_2, \dots, x_D)$ and $\mathbf{y} = (y_1, y_2, \dots, y_D)$ be D-part compositions represented in $\mathbb{S}^D$. Their perturbation and the powering by a real constant α, are

$$\mathbf{x} \oplus \mathbf{y} = \mathscr{C}(x_1 y_1, x_2 y_2, \dots, x_D y_D) \quad , \quad \alpha \odot \mathbf{x} = \mathscr{C}(x_1^\alpha, x_2^\alpha, \dots, x_D^\alpha) \ ,$$

where $\mathscr{C}$ is the closure operation which normalizes the composition to unit sum. With these operations, $\mathbb{S}^D$ is a $(D-1)$-dimensional vector space. Compositions are frequently represented using the *centered log-ratio* (clr) coefficients and *isometric log-ratio* (ilr) coordinates (Egozcue et al., 2003). The clr transformation of $\mathbf{x}$ is

$$\mathrm{clr}(\mathbf{x}) = \left(\log \frac{x_1}{\mathrm{g_m}(\mathbf{x})}, \log \frac{x_2}{\mathrm{g_m}(\mathbf{x})}, \dots \log \frac{x_D}{\mathrm{g_m}(\mathbf{x})} \right) \ ,$$

where $\mathrm{g_m}(\cdot)$ is the geometric mean of the arguments. From the clr coefficients, the composition $\mathbf{x}$ is retrieved by

$$\mathbf{x} = \mathscr{C} \exp(v_1, v_2, \dots, v_D) \quad , \quad v_i = \mathrm{clr}_i(\mathbf{x}) = \log(x_i / \mathrm{g_m}(\mathbf{x})) \ ,$$

where exp operates componentwise. Note that $\sum_{i=1}^{D} v_i = 0$.

The ilr coordinates are computed from a $(D, D-1)$ *contrast matrix* V with the properties

$$V^\intercal V = I_{D-1} \quad , \quad V V^\intercal = I_D - \frac{1}{D} \mathbf{1}_D \mathbf{1}_D^\intercal \ , \tag{10}$$

where I_k is the (k,k) identity matrix and $\mathbf{1}_D$ is a column of D unitary entries. Then, the ilr-coordinates associated with V, and with its inverse transformation, are

$$\mathbf{z} = \mathrm{ilr}(\mathbf{x}) = \log(V^\intercal \mathrm{clr}(\mathbf{x})) \quad , \quad \mathbf{x} = \mathrm{ilr}^{-1}(\mathbf{z}) = \mathscr{C}(\exp(V\mathbf{z})) \ ,$$

where $\mathrm{clr}(\mathbf{x}) = \mathbf{v}$ is considered as a column for matrix multiplication. The meaning of these two transformations, clr and ilr, becomes clear after introducing the metric elements of the Aitchison geometry. The Aitchison inner product is

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2D} \sum_{i=1}^{D} \sum_{j=1}^{D} \log \frac{x_i}{x_j} \cdot \log \frac{y_i}{y_j} = \langle \mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y}) \rangle_e = \langle \mathrm{ilr}(\mathbf{x}), \mathrm{ilr}(\mathbf{y}) \rangle_e \ , \tag{11}$$

where $\langle \cdot, \cdot \rangle_e$ denotes the ordinary Euclidean inner product in $\mathbb{R}^D$ when using clr, and in $\mathbb{R}^{D-1}$ when applied to ilr's. From the Aitchison inner product in Equation (11), both the

Aitchison norm, $\|\mathbf{x}\|_a = (\langle \mathbf{x},\mathbf{x}\rangle_a)^{1/2}$, and the Aitchison distance $\mathrm{d}_a(\mathbf{x},\mathbf{y}) = (\|\mathbf{x}\ominus\mathbf{y}\|_a)^{1/2}$ are readily obtained. Some useful expressions for the squared Aitchison distance are

$$\mathrm{d}_a^2(\mathbf{x},\mathbf{y}) = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\log\frac{x_i}{x_j} - \log\frac{y_i}{y_j}\right)^2 = \sum_{i=1}^{D}[\mathrm{clr}_i(\mathbf{x}) - \mathrm{clr}_i(\mathbf{y})]^2$$

$$= \sum_{i=1}^{D-1}[\mathrm{ilr}_i(\mathbf{x}) - \mathrm{ilr}_i(\mathbf{y})]^2 \ , \qquad (12)$$

and for the squared Aitchison norm

$$\|\mathbf{x}\|_a^2 = \frac{1}{2D}\sum_{i=1}^{D}\sum_{j=1}^{D}\left(\log\frac{x_i}{x_j}\right)^2 = \sum_{i=1}^{D}[\mathrm{clr}_i(\mathbf{x})]^2 = \sum_{i=1}^{D-1}[\mathrm{ilr}_i(\mathbf{x})]^2 \ , \qquad (13)$$

where $\mathrm{clr}_i(\mathbf{x})$ and $\mathrm{ilr}_i(\mathbf{x})$ denote the components of $\mathrm{clr}(\mathbf{x})$ and $\mathrm{ilr}(\mathbf{x})$ respectively.

From these definitions, it is clear that V contains the clr coefficients of the compositions of the selected basis in $\mathbb{S}^D$. Then, the condition $V^\mathsf{T}V = I_{D-1}$ implies the orthonormality of the basis and, consequently, the corresponding ilr-coordinates are Cartesian coordinates representing the composition. Both clr and ilr define isometries from $\mathbb{S}^D$ onto $\mathbb{R}_0^D$ (real D-vectors which components add to zero) and $\mathbb{R}^{D-1}$, respectively. This can be summarized as

$$\mathrm{clr}(\alpha\odot\mathbf{x}\oplus\mathbf{y}) = \alpha\cdot\mathrm{clr}(\mathbf{x}) + \mathrm{clr}(\mathbf{y}) \quad , \quad \mathrm{ilr}(\alpha\odot\mathbf{x}\oplus\mathbf{y}) = \alpha\cdot\mathrm{ilr}(\mathbf{x}) + \mathrm{ilr}(\mathbf{y}) \ ,$$

and

$$\langle\mathbf{x},\mathbf{y}\rangle_a = \langle\mathrm{clr}(\mathbf{x}),\mathrm{clr}(\mathbf{y})\rangle_e = \langle\mathrm{ilr}(\mathbf{x}),\mathrm{ilr}(\mathbf{y})\rangle_e \ ,$$

$$\mathrm{d}_a(\mathbf{x},\mathbf{y}) = \mathrm{d}_e(\mathrm{clr}(\mathbf{x}),\mathrm{clr}(\mathbf{y})) = \mathrm{d}_e(\mathrm{ilr}(\mathbf{x}),\mathrm{ilr}(\mathbf{y})) \ , \ \|\mathbf{x}\|_a = \|\mathrm{clr}(\mathbf{x})\|_e = \|\mathrm{ilr}(\mathbf{x})\|_e \ ,$$

where subscripts a mean Aitchison geometry, and subscripts e mean ordinary Euclidean geometry. Note that real operations involving clr coefficients are carried out in $\mathbb{R}^D$, while those involving ilr correspond to $\mathbb{R}^{D-1}$.

A practical way of constructing ilr-coordinates, i.e. of obtaining the contrast matrix V, is using sequential binary partitions (SBP) of the compositions. This technique (Egozcue and Pawlowsky-Glahn, 2005) consists of separating into two (non overlapping) groups the parts of a composition, for instance, marking the parts in each group with a $+1$ and with a -1 otherwise. The partition is repeated in each group generated in previous steps. A typical way of coding the SBP is shown as example in Table 6.

The sign code of the SBP is given in the $(D, D-1)$ matrix $\Theta = [\theta_{ij}]$, where the code component ij corresponds to the sign of x_i in the j-th ilr coordinate. Each step of partition corresponds to an element $\mathbf{e}_j$ of the orthonormal basis, and the corresponding j-th ilr-coordinate is computed as

$$b_j = \mathrm{ilr}_j(\mathbf{x}) = \sqrt{\frac{n_+ \cdot n_-}{n_+ + n_-}} \log \frac{(\prod_{\theta_{ij}=+1} x_{ij})^{1/n_+}}{(\prod_{\theta_{ij}=-1} x_{ij})^{1/n_-}} \ , \ j = 1, 2, \ldots, D-1 \qquad (14)$$

where n_+ and n_- are the number of plus signs and minus signs, respectively. Note that the expression $(\prod_{\theta_{ij}=+1} x_{ij})^{1/n_+}$ in the numerator of the fraction in Equation (14) is the geometric mean of the elements x_{ij} which are marked with a $+1$ in the j-th partition. Similarly the expression in the denominator for elements marked with a -1. The coordinates b_j have a particularly simple form: they are proportional to log-ratios of geometric means of groups. Due to this fact, they are called *balances* between the corresponding groups of parts (Egozcue et al., 2003, Egozcue and Pawlowsky-Glahn, 2005). An abbreviated way of denoting balances is to enumerate the parts in the numerator and denominator separated by a slash. For instance, the $j = 2$ balance coded as in Table 6 would be denoted as balance$(x_2, x_D / x_3, x_4, \ldots, x_{D-1})$. The elements of the contrast matrix, v_{ij} are null if $\theta_{ij} = 0$ and, for $\theta_{ij} = +1$ and $\theta_{ij} = -1$,

$$v_{ij} = \frac{\theta_{ij}}{n_+} \sqrt{\frac{n_+ \cdot n_-}{n_+ + n_-}} \quad , \quad v_{ij} = \frac{\theta_{ij}}{n_-} \sqrt{\frac{n_+ \cdot n_-}{n_+ + n_-}} \ ,$$

respectively. Note that, if $\mathbf{e}_j$ is the j-th element of the basis, then $\mathrm{clr}(\mathbf{e}_j) = (v_{1j}, v_{2j}, \ldots, v_{Dj})^\top$.

Table 6: *Sign code for a SBP of a D part composition to compute coordinates* $\mathrm{ilr}_j(\mathbf{x})$. *As an example, first partition separates* x_1 $(+1)$ *from the rest of parts; the second step separates* x_2 *and* x_D *from parts previously marked with* -1*; parts not participating in this partition step are labelled as* 0. *Take the* $+1$, -1, 0 *codes as entries of a matrix* $\Theta^\top$.

	sign code matrix $\Theta^\top$						
j	x_1	x_2	x_3	x_4	$\ldots$	x_{D-1}	x_D
1	$+1$	-1	-1	-1	$\ldots$	-1	-1
2	0	$+1$	-1	-1	$\ldots$	-1	$+1$
3	0	$+1$	0	0	$\ldots$	0	-1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
$D-1$	0	0	$+1$	-1	$\ldots$	0	0

References

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 44, 139–177.

Aitchison, J. (1983). Principal component analysis of compositional data. *Biometrika*, 70, 57–65.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Monographs on Statistics and Applied Probability. London (UK): Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press).

Aitchison, J. (1994). *Multivariate analysis and its applications*, Volume 24 of *Lecture Notes–Monograph Series*, Chapter Principles of compositional data analysis, pp. 73–81. Hayward, CA: Institute of Mathematical Statistics.

Aitchison, J. and Kay, J. (1975). Principles, practice and performance in decision-making in clinical medicine. In K. C. Bowen and D. G. White (Eds.), *Proceedings of the 1973 NATO conference on The Role and Effectiveness of Decision Theories in Practice*, London (GB). English Universities Press.

Aitchison, J., Kay, J. W. and Lauder, I. J. (2005). *Statistical Concepts and Applications in Clinical Medicine*. Chapman and Hall/CRC.

Ash, R. B. (1990). *Information theory*. Dover, New York; first puiblished by J. Wiley & Sons, 1965.

Barceló-Vidal, C. and Martín-Fernández, J.-A. (2016). The mathematics of compositional analysis. *Austrian Journal of Statistics*, 45, 57–71.

Barceló-Vidal, C., Martín-Fernández, J. A. and Pawlowsky-Glahn, V. (2001). Mathematical foundations of compositional data analysis. In G. Ross (Ed.), *Proceedings of IAMG'01 — The sixth annual conference of the International Association for Mathematical Geology*, CD-ROM.

Benjamin, J. R. and Cornell, C. A. (1960). *Probability, Statistics and Decision for Civil Engineers*. McGraw Hill Companies.

Billheimer, D., Guttorp, P. and Fagan, W. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96, 1205–1214.

Birnbaum, A. (1962). On the foundations of statistical inference. *Journal of the American Statistical Association*, 57, 269–326.

Boogaart, K. G. v., Egozcue, J. J. and Pawlowsky-Glahn, V. (2010). Bayes linear spaces. *SORT - Statistics and Operations Research Transactions*, 34, 201–222.

Boogaart, K. G. v., Egozcue, J. J. and Pawlowsky-Glahn, V. (2014). Bayes Hilbert spaces. *Australian and New Zealand Journal of Statistics*, 56, 171–194.

Egozcue, J. J., Díaz-Barrero, J. L. and Pawlowsky-Glahn, V. (2006). Hilbert space of probability density functions based on Aitchison geometry. *Acta Mathematica Sinica (English Series)*, 22, 1175–1182. DOI: 10.1007/s10114-005-0678-2.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*,37, 795–828.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2011). Evidence information in bayesian updating. In J. J. Egozcue, R. Tolosana-Delgado, and M. I. Ortego (Eds.), *Proceedings of the 4th International Workshop on Compositional Data Analysis (2011)*. CIMNE, Barcelona, Spain ISBN 978-84-87867-76-7.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2016). Changing the reference measure in the simplex and its weighting effects. *Austrian Journal of Statistics*, 45, 25–44.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2018). *Modelling compositional data. The sample space approach*. Chapter 4 in Fifty years if IAMG, D. Sagar, Q. M. Chen and F. Agterberg (Eds.), Springer.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2019). Compositional data: the sample space and its structure. *TEST*. submitted.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35, 279–300.

Egozcue, J. J., Pawlowsky-Glahn, V., Templ, M. and Hron, K. (2015). Independence in contingency tables using simplicial geometry. *Communications in Statistics –Theory and Methods*, 44, 3978–3996.

Egozcue, J. J., Pawlowsky-Glahn, V., Tolosana-Delgado, R., Ortego, M. I. and van den Boogaart, K. G. (2013). Bayes spaces: use of improper distributions and exponential families. *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales, Serie A, Matemáticas (RACSAM)*, 107, 475–486. DOI 10.1007/s13398-012-0082-6.

Gray, R. M. (2011). *Entropy and Information Theory* (2nd ed.). Springer, New York.

Hartley, R. V. L. (1928). Transmission of information. *Bell Systems Technical Journal*, 7, 535–563.

Khinchin, A. I. (1957). *Mathematical Foundations of Information Theory*. Dover Publications, New York, NY (USA).

Kullback, S. (1997). *Information Theory and Statistics, an unabridged republication of the Dover 1968 edition*. Dover publications, Minetola.

Kullback, S. and Leibler, R. A. (1951a). On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22, 79–86.

Kullback, S. and Leibler, R. A. (1951b). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27, 986–1005.

Martín-Fernández, J. A. (2001). *Medidas de diferencia y clasificación no paramétrica de datos composicionales*. Ph. D. thesis, Universitat Politècnica de Catalunya, Barcelona (E).

McMillan, B. (1953). The basic theorems of information theory. *Annals of Mathematical Statistics*, 24, 196–219.

Menafoglio, A., Grasso, M., Secchi, P. and Colosimo, B. M. (2018). Profile monitoring of probability density functions via simplicial functional PCA with application to image data. *Technometrics* online February 12, 2018.

Menafoglio, A., Guadagnini, A. and Secchi, P. (2016). Stochastic simulation of soil particle-size curves in heterogeneous aquifer systems through a bayes space approach. *Water Resources Research*, 52, 5708–5726.

Pawlowsky-Glahn, V. and Buccianti, A. (Eds.) (2011). *Compositional Data Analysis: Theory and Applications*. John Wiley & Sons.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA)*, 15, 384–398.

Pawlowsky-Glahn, V., Egozcue, J. J. and Tolosana-Delgado, R. (2015). *Modeling and Analysis of Compositional Data*. Statistics in practice. John Wiley & Sons, Chichester UK.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press, Princeton NJ, USA.

Shannon, C. (1948). A mathematical theory of communication. *Bell Systems Technical Journal*, 27, 379–423, 623–656.

Shannon, C. E. and Weaver, W. (1949). *The Mathematical Theory of Communication*. University of Illinois Press. Urbana.

Yager, R. R. (1987). On the Dempster-Shafer framework and new combination rules. *Information Sciences*, 41, 93–137.

Yager, R. R. and Liu, L. (Eds.) (2008). *Classic Works of the Dempster-Shafer Theory of Belief Functions*. Springer, Berlin.