Evidence information in Bayesian updating

J. J. EGOZCUE[1] and V. PAWLOWSKY-GLAHN[2]

[1]Dept. Matemàtica Aplicada III, U. Politècnica de Catalunya, Barcelona, Spain (juan.jose.egozcue@upc.edu)
[2]Dept. Informàtica i Matemàtica Aplicada, U. de Girona, Spain (vera.pawlowsky@udg.edu)

Abstract

Bayes theorem (discrete case) is taken as a paradigm of information acquisition. As mentioned by Aitchison, Bayes formula can be identified with perturbation of a prior probability vector and a discrete likelihood function, both vectors being compositional. Considering prior, posterior and likelihood as elements of the simplex, a natural choice of distance between them is the Aitchison distance. Other geometrical features can also be considered using the Aitchison geometry. For instance, orthogonality in the simplex allows to think of orthogonal information, or the perturbation-difference to think of opposite information. The Aitchison norm provides a size of compositional vectors, and is thus a natural scalar measure of the information conveyed by the likelihood or captured by a prior or a posterior. It is called *evidence information*, or *e-information* for short.

In order to support such e-information theory some principles of e-information are discussed. They essentially coincide with those of compositional data analysis. Also, a comparison of these principles of e-information with the axiomatic Shannon-information theory is performed. Shannon-information and developments thereof do not satisfy scale invariance and also violate subcompositional coherence. In general, Shannon-information theory follows the philosophy of amalgamation when relating information given by an evidence-vector and some sub-vector, while the dimension reduction for the proposed e-information corresponds to orthogonal projections in the simplex. The result of this preliminary study is a set of properties of e-information that may constitute the basis of an axiomatic theory. A synthetic example is used to motivate the ideas and the subsequent discussion.

1 Introduction

Information theory was born in communication and coding theory (Shannon, 1948; Shannon and Weaver, 1949). It was successfully applied soon to different scientific fields, specially to Statistics (Kullback and Leibler, 1951; Kullback, 1997; Lindley, 1956). However, its application in the Bayesian framework has revealed some drawbacks that deserve attention. One of them is the fact that information provided by an experiment depends on the prior assumption.

The present contribution builds on two well known properties: (1) Bayes theorem (discrete case) is considered a paradigm of information acquisition and (2) Bayes formula can be identified with perturbation (Aitchison, 1986). In fact, Bayes formula corresponds to perturbation of a prior probability vector and a discrete likelihood function. Following the *likelihood principle*, proportional likelihood functions are equivalent, an essential characteristics of compositions. An important consequence is that the three ingredients of Bayes formula—prior, likelihood and posterior—can be represented in the same space, i.e. the simplex. Moreover, they share its nature and characteristics. We call these vectors *evidence* vectors, because they describe our evidence on a collection of events or how the evidence changes after an experiment.

The fact, that the nature of probability distributions and likelihood functions is the same, is not well recognized in mainstream information theory. Likelihood is not treated as a probability vector, and Shannon information is only defined on prior and posterior distributions. As a consequence, information theory tries to define divergence or distance measures between distributions, like e.g. Kullback-Leibler divergence (Kullback, 1997). When considering prior, posterior and likelihood as elements of the simplex, a natural choice of distance between them is the Aitchison distance. Other geometrical features can also be considered using the Aitchison geometry. For instance, orthogonality in the simplex allows us to think of orthogonal information, or the perturbation-difference to think of opposite information. The Aitchison norm provides a size of information vectors, and is thus a natural candidate for a scalar measure of the information conveyed by the likelihood or captured by a prior or a posterior. It is called hereafter *evidence information* or *e-information* for short.

In order to support such e-information theory, a discussion of principles of e-information is presented. They essentially coincide with those of compositional data analysis. Also, a comparison of these principles of e-information with the axiomatic Shannon-information theory (Kinchin, 1957; Ash, 1990; Rényi, 1966) is performed. Shannon-information and developments thereof satisfy neither scale invariance nor subcompositional coherence, which constitute the basis of compositional analysis. In general, Shannon-information theory follows the philosophy of amalgamation when relating information given by an evidence-vector and some sub-vector (subcomposition), while the dimension reduction suggested here for e-information corresponds to orthogonal projections in the simplex. The result of this preliminary study is a set of properties of e-information that may constitute the basis of an axiomatic theory.

2 Motivating example

Bayes formula is used to update the knowledge on the occurrence of a family of non-overlapping events after some observation or experiment. This situation arises frequently because it is the paradigm of information acquisition in any experiment resulting in a classification of the actual event.

To motivate the following sections, a synthetic example on inspection of buildings is used. Suppose interest lies in the actual structural state of a building. The states are classified into three categories: A_1 *service*, A_2 *damaged*, A_3 *ruin*. To know the actual structural category of a building requires a series of destructive and expensive experiments that would be good to avoid. For this purpose, two types of non-destructive inspections have been designed,

1. a *visual test*, R, of fractures, displacements, and the like;
2. a *dynamic test*, Q, in which the building is vibrated and the response is measured with some sensor.

Let R_1, R_2, R_3 and Q_1, Q_2, Q_3 be the possible outputs of the respective experiments. The two experiments were calibrated on a number of buildings whose actual structural category is known. The result of the calibration is typically described by the conditional probabilities $\mathrm{P}[R_j|A_i]$, $\mathrm{P}[Q_j|A_i]$, $i = 1, 2, 3$, $j = 1, 2, 3$. The two sets of conditional probabilities, describing the experiments, are assumed known (Table 1).

Table 1: Likelihood for each possible output: probability of A_1 service, A_2 damage, A_3 ruin conditional to the output. Maximum likelihood of each column presented in boldface. Last row shows e-information (Aitchison norm) associated with each evidence vector (likelihood).

	R_1	R_2	R_3	
	$\mathbf{q}$	$\mathbf{q}$	$\mathbf{q}$	sum
A_1	**0.8772**	0.1206	0.0022	1
A_2	0.0714	**0.5000**	0.4286	1
A_3	0.0526	0.2105	**0.7368**	1
norm	2.1832	1.0133	4.5446	

	Q_1	Q_2	Q_3	
	$\mathbf{q}$	$\mathbf{q}$	$\mathbf{q}$	sum
A_1	**0.8918**	0.0952	0.0131	1
A_2	0.8760	**0.1239**	0.0001	1
A_3	0.1322	0.0809	**0.7869**	1
norm	1.7465	0.3040	6.8169	

When an inspection of a building is planned, there is some previous knowledge about how the building may be performing. It can be based on the location or area where it is, the year of construction, or the typology of the structure. This knowledge can be described by a prior probability vector, $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3)$, called prior evidence on the structural state of the building (Table 2). Once one of the two available tests is carried out on the target building, an output is obtained, e.g. R_j as the result of the visual test. Given R_j, Bayes formula updates the prior evidence $\boldsymbol{\pi}$ into the *posterior evidence* $\mathbf{p} = (p_1, p_2, p_3)$, where $p_i = \mathrm{P}[A_i|R_j]$. It can be written

$$\mathrm{P}[A_i|R_j] = \kappa \ \mathrm{P}[R_j|A_i] \ P[A_i] \ , \ \kappa^{-1} = \sum_{k=1}^{3} \mathrm{P}[R_j|A_k] \ P[A_k] \ ,$$

Table 2: Probabilities of A_1 service, A_2 damage, A_3 ruin. Prior $\boldsymbol{\pi}$; posterior $\mathbf{p}$ for each posible test output R_i, Q_i, $i = 1, 2, 3$. Maximum probability of each column presented in boldface. Last row shows e-information (Aitchison norm) of each evidence vector.

		R_1	R_2	R_3
	$\boldsymbol{\pi}$	$\mathbf{p}$	$\mathbf{p}$	$\mathbf{p}$
A_1	**0.5500**	**0.9427**	0.2676	0.0050
A_2	0.3000	0.0419	**0.6050**	**0.5350**
A_3	0.1500	0.0154	0.1274	0.4599
sum	1	1	1	1
norm	0.9194	3.0337	1.1022	3.7521

		Q_1	Q_2	Q_3
	$\boldsymbol{\pi}$	$\mathbf{p}$	$\mathbf{p}$	$\mathbf{p}$
A_1	**0.5500**	**0.6344**	**0.5149**	0.0573
A_2	0.3000	0.3399	0.3657	0.0001
A_3	0.1500	0.0256	0.1194	**0.9425**
sum	1	1	1	1
norm	0.9194	2.4057	1.0808	6.4460

or equivalently $p_i = \kappa q_i \pi_i$, where $q_i = \mathrm{P}[R_j|A_i]$. In the latter expression, the reference to R_j has been dropped because it is fixed once the output of the experiment is known. Table 2 shows posterior evidence obtained by updating the prior $\boldsymbol{\pi}$ with each possible output of each of both experiments.

Consider now some questions about the performance of the tests:

(a) Given one of the available tests, which of the possible outputs provides more information?

(b) Given the output from the visual test, R_j, and from the dynamic test, Q_k, which of the two outputs is more informative?

(c) In order to deliver a diagnosis of the structural state of the building, a decision on A_1, A_2, A_3 should be made. Assume that the output of a test is R_j. Independently of the cost-benefit of such a decision, how reliable is the decision taken?

(d) Can we define an average information provided by a test before getting the output? Which of the two tests is more informative?

The answer to question (a) requires a measure or *size* of the information provided by the likelihood vectors $\mathbf{q}(R_j) = (\mathrm{P}[R_j|A_1], \mathrm{P}[R_j|A_2], \mathrm{P}[R_j|A_3])$, $j = 1, 2, 3$, and the selection of the maximum one. Similarly, question (b) calls for the quantitative comparison of the likelihood vectors $\mathbf{q}(R_j)$ and $\mathbf{q}(Q_k)$, now coming from different tests. Question (c) demands measuring the strength of evidence on the decision to be taken, assuming that the maximum posterior probability determines the choice. Again a measure of information is needed, although now it should be defined for the posterior evidence $\mathbf{p} = (p_1, p_2, p_3)$ better than for the likelihood vectors. Question (d) points at the need of an average information on the possible outputs of each test.

Additionally, both visual and dynamic tests can be carried out on the same building. Then Bayes formula takes the form

$$\mathrm{P}[A_i|R_j, Q_k] = \kappa \; \mathrm{P}[R_j, Q_k|A_i] \; P[A_i] \; ,$$

being κ an adequate normalizing constant. To use this joint updating, the joint likelihood of the two tests is required. It can be simplified whenever the two tests are conditionally independent, i.e. when given any structural state A_i, $\mathrm{P}[R_j, Q_k|A_i] = \mathrm{P}[R_j|A_i] \cdot \mathrm{P}[Q_k|A_i]$. Under this assumption Bayes formula reduces to

$$\mathrm{P}[A_i|R_j, Q_k] = \kappa \; \mathrm{P}[R_j|A_i] \; \mathrm{P}[Q_k|A_i] \; P[A_i] \; , \tag{1}$$

for some normalizing constant κ.

This shows that the two tests can be carried out sequentially and the posterior evidence is then updated also sequentially. The result remains unaltered irrespective of the order in which the tests are carried out. Frequently, conditional independence of the tests does not hold. However, Eq. (1) can still be valid. It corresponds to situations in which the order in a sequential testing does not influence the resulting posterior. This can occur for some specific results R_j, Q_k or, more restrictively, when the order does not affect the result for any R_j and for any Q_k. The latter case is normally called exchangeability of the tests in Bayesian frameworks. Conditional independence suggests an additional question:

(e) Which is the relationship of information provided by two single outputs of different experiments (conditionally independent or not) and the corresponding joint output?

When the joint conditional probability $\mathrm{P}[R_j, Q_k|A_i]$ is not known, this question addresses the problem of redundancy that remains open both in information theory and Bayesian statistics.

3 Bayes formula and Aitchison geometry of the simplex

Real vectors with n positive components are in $\mathbb{R}^n_+$. An equivalence relation for proportional vectors can be defined, i.e. $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n_+$, are equivalent if, and only if, there is a positive constant c, such that $\mathbf{x} = c\mathbf{y}$ (Barceló-Vidal et al., 2001). These equivalence classes can be represented in the n-part unit-simplex, $\mathcal{S}^n$, by means of a representative which components add to one. The operation to select such a representative is called closure, and is defined as

$$\mathcal{C}\mathbf{x} = \left(\sum_{i=1}^{n} x_i\right)^{-1} \cdot \mathbf{x} \ , \qquad \mathbf{x} = (x_1, x_2, \ldots, x_n) \ .$$

Perturbation and powering are operations in $\mathcal{S}^n$; they can be interpreted as operations on equivalence classes of vectors in $\mathbb{R}^n_+$. For $\alpha \in \mathbb{R}$, and $\mathbf{x}$, $\mathbf{y} \in \mathbb{R}^n_+$, perturbation and powering are

$$\mathbf{x} \oplus \mathbf{y} = \mathcal{C}(x_1y_1, x_2y_2, \ldots, x_ny_n) \ , \ \alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_n^\alpha),$$

where the result of both operations are the representatives in $\mathcal{S}^n$ of the operated equivalence classes. The n-part simplex or, equivalently, the set of equivalence classes in $\mathbb{R}^n_+$, is a $(n-1)$-dimensional vector space. The neutral equivalence class for the perturbation is represented by $\mathbf{n} = \mathcal{C}(1, 1, \ldots, 1)$; the perturbation-opposite element of $\mathbf{x}$ is $\ominus\mathbf{x} = (-1) \odot \mathbf{x} = (x_1^{-1}, x_2^{-1}, \ldots, x_n^{-1})$.

Bayes formula matches exactly the definition of perturbation. Recalling notation in Section 2, consider $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)$, the prior probabilities of a partition A_1, A_2, …, A_n of a probability space, with $\pi_i = \mathrm{P}[A_i]$. After the realisation of some experiment R, the conditional probabilities $q_i = \mathrm{P}[R|A_i]$ are arranged in a likelihood vector $\mathbf{q} = (q_1, q_2, \ldots, q_n)$. Posterior probabilities are denoted $p_i = \mathrm{P}[A_i|R]$, and $\mathbf{p} = (p_1, p_2, \ldots, p_n)$. Bayes formula is then

$$p_i = \kappa \ q_i \ \pi_i \ , \ \kappa^{-1} = \sum_{k=1}^{n} q_k \pi_k \ . \tag{2}$$

Vectors $\boldsymbol{\pi}$ and $\mathbf{p}$ are in $\mathcal{S}^n$ because they are normalized to one, but they can be considered as representatives of classes of non-normalized probability vectors. The likelihood vector $\mathbf{q}$ is in $\mathbb{R}^n_+$ and is seldom normalized. Clearly, Bayes formula (2) is readily identified with a perturbation in $\mathcal{S}^n$,

$$\mathbf{p} = \mathbf{q} \oplus \boldsymbol{\pi} \ . \tag{3}$$

This fact was used by Aitchison (1986) to illustrate and interpret perturbation. Now we reverse the argument: since Bayes formula is well represented in $\mathcal{S}^n$, the Aitchison geometry of the simplex can be used to reinterpret Bayes formula.

The n-part simplex $\mathcal{S}^n$ is an $(n-1)$-dimensional Euclidean space with the Aitchison distance, norm and inner-product (Pawlowsky-Glahn and Egozcue, 2001; Billheimer et al., 2001). For simplicity in the expression, they are defined using the centered log-ratio transformation, clr, of $\mathbf{x} \in \mathbb{R}^n_+$ or of its representative in $\mathcal{S}^n$,

$$\mathrm{clr}(\mathbf{x}) = \log(\mathbf{x}) - \log(\mathrm{g_m}(\mathbf{x}), \mathrm{g_m}(\mathbf{x}), \ldots, \mathrm{g_m}(\mathbf{x})) \ , \ \mathbf{x} = \mathcal{C}\exp(\mathrm{clr}(\mathbf{x})) \ ,$$

where the functions log and exp act on vectors componentwise, and $\mathrm{g_m}(\cdot)$ is the geometric mean of its arguments. The i-th component of $\mathrm{clr}(\mathbf{x})$ is $\mathrm{clr}_i(\mathbf{x}) = \log(x_i/\mathrm{g_m}(\mathbf{x}))$, and the sum of components of the clr($\mathbf{x}$) is null. For $\mathbf{x}, \mathbf{y} \in \mathcal{S}^n$, the Aitchison distance, norm and inner-product are

$$\mathrm{d}_a(\mathbf{x}, \mathbf{y}) = \mathrm{d}(\mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y})) = \sqrt{\sum_{i=1}^{n} (\mathrm{clr}_i(\mathbf{x}) - \mathrm{clr}_i(\mathbf{y}))^2} \ , \tag{4}$$

$$\|\mathbf{x}\|_a = \|\mathrm{clr}(\mathbf{x})\| = \sqrt{\sum_{i=1}^{n} (\mathrm{clr}_i(\mathbf{x}))^2} \ , \tag{5}$$

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \langle \mathrm{clr}(\mathbf{x}), \mathrm{clr}(\mathbf{y}) \rangle = \sum_{i=1}^{n} \mathrm{clr}_i(\mathbf{x}) \cdot \mathrm{clr}_i(\mathbf{y}) \ , \tag{6}$$

where $\mathrm{d}(\cdot, \cdot)$, $\| \cdot \|$, $\langle \cdot, \cdot \rangle$ denote the ordinary Euclidean distance, norm and inner product.

The Euclidean structure of $\mathcal{S}^n$ has many implications. Some of them are remarkable in this context.

(1) The Aitchison norm allows to measure the size of evidence contained in a prior, a posterior and in a likelihood vector.

(2) Distances are preserved under a perturbation by a likelihood vector.

(3) The notion of orthogonality can be introduced using the inner product; therefore, its meaning can be interpreted in the context of Bayes formula.

For instance, let $\mathbf{q}^{(1)} = (a, a, b)$ and $\mathbf{q}^{(2)} = (c, d, \mathrm{g_m}(c, d))$ be two likelihood vectors. Orthogonality of these two likelihood vectors is easily checked, i.e. $\langle \mathbf{q}^{(1)}, \mathbf{q}^{(2)} \rangle_a = 0$. It is easy to realize that, when used in Bayes formula, $\mathbf{q}^{(1)}$ is unable to modify the ratio of evidences of the two first components in the prior, i.e. the log-odds $\log(\pi_1/\pi_2)$ in the prior remain unaltered in the posterior. Alternatively, the use of $\mathbf{q}^{(2)}$ in the Bayes formula modifies the prior ratio of evidences of π_1 and π_2, but leaves invariant the prior value of $\log((\pi_1\pi_2)^{1/2}/\pi_3)$. Interpretation of orthogonality is now that the likelihood vectors $\mathbf{q}^{(1)}$, $\mathbf{q}^{(2)}$ provide evidences on different and non-overlapping aspects in Bayesian updating. Moreover, for this example in $\mathcal{S}^3$, given an evidence vector there is only one possible evidence direction which is orthogonal to the previous one, because the dimension of the space is 2.

Another important practical consequence is the following:

(4) Any vector in $\mathcal{S}^n$ can be represented by its coordinates with respect to a basis of the space, particularly an orthonormal basis. Therefore, evidence vectors (prior, posterior and likelihood) can be represented as real vectors with all standard and intuitive properties.

The example from Section 2 is represented in Figure 1 in orthonormal balance-coordinates (Egozcue et al., 2003; Egozcue and Pawlowsky-Glahn, 2005, 2006a,b)

$$b_1 = \sqrt{\frac{2}{3}} \log \frac{x_1}{\mathrm{g_m}(x_2, x_3)} \ , \ b_2 = \sqrt{\frac{1}{2}} \log \frac{x_2}{x_3} \ ,$$

where $\mathrm{g_m}$ denotes geometric mean. The first balance b_1 contrasts *service* versus *damage-ruin*, whereas b_2 compares *damage* over *ruin*. Figure 1 shows Bayesian updating in two different cases, when the output of the visual inspection is R_2 (left) and when the output of the dynamic test is Q_2 (right). In the case of R_2 as output (left) the norm of the prior (green) is similar to the norm of the likelihood (light-blue) (see Tables 1,2) but they have different directions. The posterior closes a triangle which is approximately equilateral. For output Q_2, Figure 1 (right) shows a different scenario. The norm of the likelihood associated with Q_2 is less than the norm of the prior. As a consequence, the norm of the posterior has been incremented only a little bit, from 0.92 for the prior to 1.08 for the posterior (Table 2).

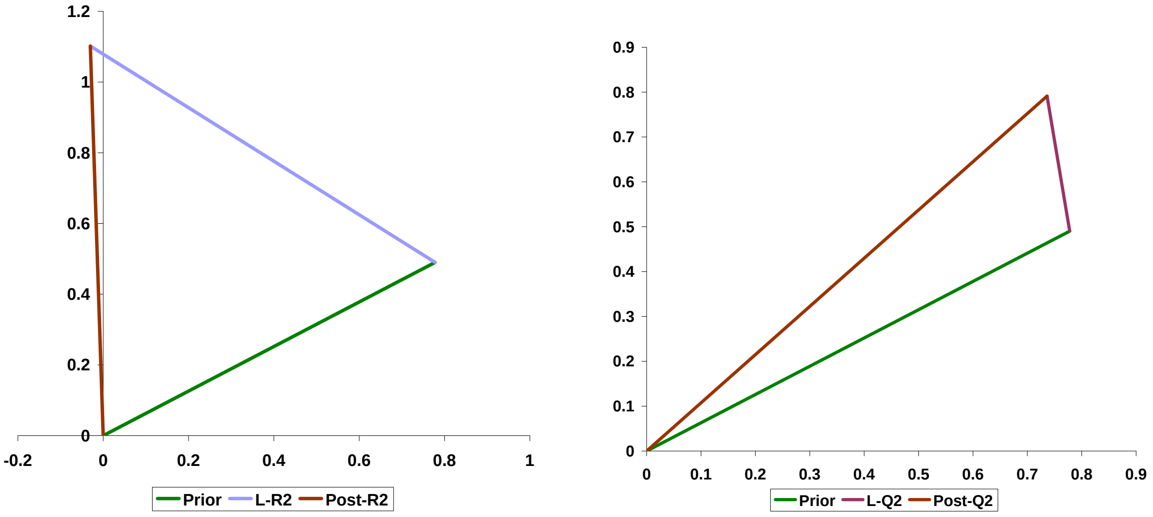

Figure 1: Bayes updating represented in coordinates: posterior is the perturbation (vector sum) of prior and likelihood. Left, after test output R_2; Right, after test output Q_2. Abscissa: $b_1 = (2/3)^{-1/2}\log(x_3/\mathrm{g_m}(x_1,x_2))$; ordinate: $b_2 = 2^{-1/2}\log(x_2/x_3)$.

4 Evidence-information principles

The following principles concerning e-information are referred to the Bayes formula. Prior, likelihood and posterior are denoted $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)$, $\mathbf{q} = (q_1, q_2, \ldots, q_k)$, and $\mathbf{p} = (p_1, p_2, \ldots, p_k)$, respectively; all components are positive. Prior and posterior are assumed to be probabilities, i.e. they are normalized to add to one, whereas the likelihood is not normalized.

Scale invariance of evidence

Prior, likelihood and posterior describe a state of evidence: prior (previous) to an experiment; provided by one experiment, and posterior (after) the realisation of the experiment. Therefore, evidence should be measurable in the same way at the three states. Accepting that proportional likelihood functions provide equal experimental evidence, normalization of prior and posterior should be irrelevant. This constitutes the *principle of scale invariance of evidence.* A consequence is that the only relevant information conveyed by $\boldsymbol{\pi}$, $\mathbf{q}$, $\mathbf{p}$ is contained in the ratios of components of each one of these three vectors. If $\mathcal{I}_e(\mathbf{p})$ denotes evidence information conveyed by the evidence vector $\mathbf{q}$, closed to unit as a probability vector or just non-closed as a likelihood, then

$$\mathcal{I}_e(\mathbf{q}) = \mathcal{I}_e(\mathcal{C}\mathbf{q}) \ .$$

An immediate consequence of scale invariance is that any relevant quantity, f, describing evidence should be a 0-degree homogeneous function of the components, i.e. $f(\lambda\mathbf{x}) = f(\mathbf{x})$ for any positive real constant λ and being $\mathbf{x}$ either a prior, a likelihood or a posterior distribution.

The Shannon information associated with a probability vector $\mathbf{x}$ is

$$\mathcal{I}_s(\mathbf{x}) = -\sum_{i=1}^{n} x_i \log(x_i) \ , \tag{7}$$

which is not scale-invariant. Similarly, the Kullback-Leibler divergence between two probability vectors $\mathbf{x}$, $\mathbf{y}$ is the non-symmetric expression

$$\mathcal{I}_s(\mathbf{y} : \mathbf{x}) = \sum_{i=1}^{n} y_i \log\left(\frac{y_i}{x_i}\right)$$

which is not scale-invariant. Note that this explains why it cannot be applied to likelihood vectors. However, Kullback-Leibler divergence can be symmetrised and centered to attain scale invariance (Martín-Fernández et al., 1998; Martín-Fernández, 2001).

Bayes formula is addition of evidence

Evidence only changes when an experimental result is obtained. Furthermore, one can think of an experiment which likelihood does not change the prior evidence or which does not provide any information, i.e. $\boldsymbol{\pi} = \mathbf{p}$ which, in the context of Bayes formula, means that $\mathbf{q} = \mathbf{n} = \mathcal{C}(1, 1, \ldots, 1)$; one can conceive evidences from independent experiments cancelling each other in such a way that, when jointly considered, they do not provide any information, their evidences being opposite; exchangeable or independent experiments should be commutative. These general ideas configure change of evidence as an Abelian group operation between prior and likelihood, thus deserving the name of addition (van den Boogaart et al., 2010).

Information-evidence does not depend on prior evidence

It seems reasonable that information provided by an experiment should not depend on what is previously known about the actual event. Therefore, when for two possibly different prior evidences $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ and the likelihood of a result of an experiment $\mathbf{q}$, two posterior evidences are obtained $\mathbf{p}_i = \boldsymbol{\pi}_i \oplus \mathbf{q}$, $i = 1, 2$, then the properties

$$\mathcal{I}_e(\mathbf{p}_1 \ominus \mathbf{p}_2) = \mathcal{I}_e(\boldsymbol{\pi}_1 \ominus \boldsymbol{\pi}_2) \quad , \quad \mathcal{I}_e(\mathbf{p}_1 \ominus \boldsymbol{\pi}_1) = \mathcal{I}_e(\mathbf{p}_2 \ominus \boldsymbol{\pi}_2) = \mathcal{I}_e(\mathbf{q}) \ , \tag{8}$$

should hold. The first equation expresses that the difference in e-information prior to the experiment should not change for the posteriors after a given output of the experiment. The second equation in (8) means that the change of evidence only depends on the likelihood associated with an output of the experiment and does not depend on the prior evidence.

These elementary properties are not fulfilled by the Shannon information, $\mathcal{I}_s$ (see Eq. (7)).

Extension-projection rules

Consider an evidence vector $\mathbf{p} = (\mathbf{p}_1, p_{n_1+1})$ in $\mathcal{S}^{n_1+1}$, $n_1 \geq 2$, which conveys an e-information $\mathcal{I}_e(\mathbf{p})$. If, for some reason, the interest is centered in the reduced evidence vector $\mathbf{p}_1 \in \mathcal{S}^{n_1}$, a reasonable question is: which is the relationship between $\mathcal{I}_e(\mathbf{p})$ and $\mathcal{I}_e(\mathbf{p}_1)$? Or, more specifically, which is the value of p_{n_1+1} for which $\mathcal{I}_e(\mathbf{p}) = \mathcal{I}_e(\mathbf{p}_1)$? For most information measures the value of p_{n+1} is 0. Particularly, this is true for Shannon information since $\mathcal{I}_s(\mathbf{p}) = \mathcal{I}_s(\mathbf{p}_1) - p_{n_1+1} \log p_{n_1+1}$, and the last term vanishes when p_{n+1} tends to zero. This fact is notably unreasonable for evidence information: an evidence vector with a null component should carry a large, if not infinite, evidence against the event for which the null probability holds. Another paradoxical result is obtained when the argument is extended to an infinite collection of events, all of them with null probability, and appended to the initial evidence vector: an infinite sequence of events are excluded as impossible without any additional evidence.

The Shannon information theory has two ways to answer to these questions. After obtaining the expression of $\mathcal{I}_s$ (Eq. 7), the above mentioned properties are a simple consequence, thus conforming a first way. The second way is the adoption of some axiomas (Kinchine, Renyi, Ash), most of them reasonable. However, all axiomatic systems for Shannon information include an extension axiom which explains how information changes from a probability vector $\mathcal{C}\mathbf{p}_1$ with n_1 components, to an extended one $(\mathbf{p}_1, \mathbf{p}_2)$ when $\mathcal{C}\mathbf{p}_2$, with n_2 components, is appended to form and extended vector with $n_1 + n_2$ components. Assuming that the extended vector $(\mathbf{p}_1, \mathbf{p}_2)$ is normalized to one, i.e. $(\mathbf{p}_1, \mathbf{p}_2) = \mathcal{C}(\mathbf{p}_1, \mathbf{p}_2)$, the extension rule (axiom) can be formulated

$$\Sigma(\mathbf{p}_1)\mathcal{I}_s(\mathcal{C}\mathbf{p}_1) + \Sigma(\mathbf{p}_2)\mathcal{I}_s(\mathcal{C}\mathbf{p}_2) + \mathcal{I}_s((\Sigma(\mathbf{p}_1), \Sigma(\mathbf{p}_2))) = \mathcal{I}_s((\mathbf{p}_1, \mathbf{p}_2)) \ , \tag{9}$$

where $\Sigma(\mathbf{p}_i)$, $0 \leq \Sigma(\mathbf{p}_i) \leq 1$, denotes the sum of all probabilities contained in $\mathbf{p}_i$ and, consequently, $\Sigma(\mathbf{p}_1) + \Sigma(\mathbf{p}_2) = 1$. This extension rule is clearly inspired in the total probability theorem: $\mathcal{I}_s(\mathcal{C}\mathbf{p}_i)$ is the information conditional to the actual event given by the categories corresponding to $\mathbf{p}_i$; and $\Sigma(\mathbf{p}_i)$ is the probability of such a condition. The term $\mathcal{I}_s((\Sigma(\mathbf{p}_1), \Sigma(\mathbf{p}_2)))$ is the information added due to the choice between the categories corresponding to $\mathbf{p}_1$ and $\mathbf{p}_2$ and it only appears whenever the two subvectors $\mathbf{p}_1$, $\mathbf{p}_2$ are joined.

From the Shannon information extension rule (Eq. 9), the null extension property is easily obtained. In fact, taking $n_2 = 1$, $\mathbf{p}_2 = (0)$, then $\Sigma(\mathbf{p}_2) = 0$, $\Sigma(\mathbf{p}_1) = 1$, $\mathcal{C}\mathbf{p}_1 = \mathbf{p}_1$ and $\mathcal{I}_s(\mathcal{C}\mathbf{p}_1) = \mathcal{I}_s((\mathbf{p}_1, 0))$.

Once Equation (9) has been shown to be inadequate for evidence information, an alternative formulation is required. Three main ideas should be taken into account for that purpose.

(a) When the number of components of an evidence vector increases, the evidence information should also increase.

(b) Complete evidence against, or in favour, of an event implies an unbounded evidence information.

(c) Additivity of evidence information is a vector additivity within an Euclidean geometry.

The last condition means, in a Pythagorean sense, that orthogonal evidence vectors should add information; opposite evidence vectors should cancel evidence information; and so on.

Taking into account these ideas, the following extension rule is proposed. Let $\mathbf{p}_1 \in \mathcal{S}^{n_1}$ and $\mathbf{p}_2 \in \mathcal{S}^{n_2}$ be two evidence vectors with n_1, n_2 components, respectively. The extended evidence vector $\mathbf{p} = (a_1\mathbf{p}_1, a_2\mathbf{p}_2)$, with $a_1, a_2 > 0$, and not necessarily closed to 1, has the evidence information

$$\mathcal{I}_e^2(\mathbf{p}) = \mathcal{I}_e^2(\mathbf{p}_1) + \mathcal{I}_e^2(\mathbf{p}_2) + \mathcal{I}_e^2(a_1\mathbf{g}_{n_1}(\mathbf{p}_1), a_2\mathbf{g}_{n_2}(\mathbf{p}_2)) \ , \tag{10}$$

where $\mathbf{g}_{n_i}(\mathbf{p}_i)$ is a vector with n_i components which are all equal to the geometric mean of the components of $\mathbf{p}_i$. In Eq. (10), e-information appears squared in all terms. Pythagoras theorem can be recognised assuming the term $\mathcal{I}_e^2(a_1\mathbf{g}_{n_1}(\mathbf{p}_1), a_2\mathbf{g}_{n_2}(\mathbf{p}_2))$ to be zero. This means that a kind of orthogonality between sub-vectors containing different components is assumed. The term $\mathcal{I}_e(a_1\mathbf{g}_{n_1}(\mathbf{p}_1), a_2\mathbf{g}_{n_2}(\mathbf{p}_2))$ is the e-information obtained when the two subvectors $a_1\mathbf{g}_{n_1}(\mathbf{p}_1)$, $a_2\mathbf{g}_{n_2}(\mathbf{p}_2)$ are joined. Each subvector has all components equal, and therefore, they do not provide any e-information when individually considered. The e-information comes from the ratios of the values $a_1\mathrm{g}_\mathrm{m}(\mathbf{p}_1)$, $a_2\mathrm{g}_\mathrm{m}(\mathbf{p}_2)$.

Definition

After the described principles on evidence information a definition can be given fulfilling the requirements. Let $\mathbf{p}$ be an evidence vector with n components closed or not to the unit and represented in $\mathcal{S}^n$ by $\mathcal{C}\mathbf{p}$. The evidence-information is defined as the Aitchison norm of $\mathbf{p}$, i.e.

$$\mathcal{I}_e(\mathbf{p}) = \|\mathbf{p}\|_a = \|\mathrm{clr}(\mathbf{p})\|_n = \|\mathrm{ilr}(\mathbf{p})\|_{n-1} \ , \tag{11}$$

where ilr assigns coordinates for any choice of an orthonormal basis of the simplex, and $\|\cdot\|_k$ is the ordinary Euclidean norm in $\mathbb{R}^k$.

The identification of evidence-information with the Aitchison norm in the simplex implies that some requirements described in the principles are automatically satisfied: scale invariance is inherited from scale invariance for compositional vectors; the neutral composition $\mathbf{n} = \mathcal{C}(1, 1, \ldots, 1)$ has null e-information, $\mathcal{I}_e(\mathbf{n}) = 0$; invariance of e-information conveyed by a likelihood (Eq. 8) can be written as

$$\mathrm{d}_a(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = \mathrm{d}_a(\mathbf{p}_1, \mathbf{p}_2) \ , \ \mathrm{d}_a(\mathbf{p}_1, \boldsymbol{\pi}_1) = \mathrm{d}_a(\mathbf{p}_2, \boldsymbol{\pi}_2) = \|\mathbf{q}\|_a \ , \tag{12}$$

where, for $i = 1, 2$, the posteriors $\mathbf{p}_i$ are obtained using Bayes formula (3) to update priors $\boldsymbol{\pi}_i$ by means of the likelihood $\mathbf{q}$, i.e. $\mathbf{p}_i = \mathbf{q} \oplus \boldsymbol{\pi}$.

Equation (12) holds automatically in the simplex equipped with the Aitchison geometry, given that it is the parallelogram property in Euclidean spaces. This property is simply interpreted, meaning that Aitchison distances between priors and posteriors are invariant under the shift-perturbation represented by the likelihood used to update priors. The second equation means that evidence provided by a Bayes updating does not depend on the priors but only on the likelihood.

The extension rule can also be translated into the Aitchison geometry. Consider the composite evidence vector $(a_1\mathbf{g}_{n_1}(\mathbf{p}_1), a_2\mathbf{g}_{n_2}(\mathbf{p}_2))$ whose associated e-information, or Aitchison norm appears in Equation (10). In order to compute the square norm, define a sequential binary partition (SBP) (Egozcue and Pawlowsky-Glahn, 2005, 2006b) in which the first partition separates the first group

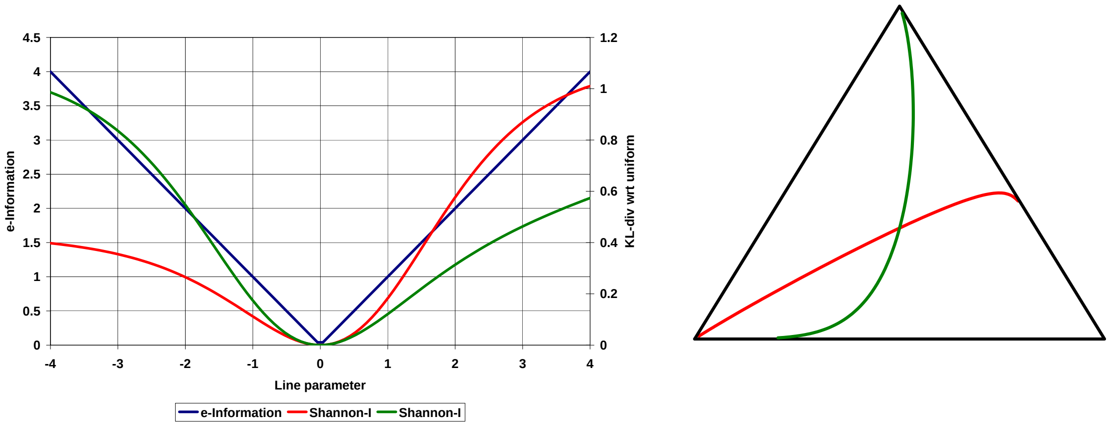

Figure 2: Comparison of e-information and Kullback-Leibler divergence to the neutral element along two straight-lines in the simplex (left). The two straight-lines in the simplex are represented in a ternary diagram (right). Colors correspond to divergences in the left panel represented .

of n_1 components from the last group of n_2. Call the corresponding balance b_1. Subsequent steps of partition can be selected arbitrary given rise to a collection of balances b_2, b_3, $\dots$, b_{n-1}, with $n = n_1 + n_2$. In the first subcomposition $a_1\mathbf{g}_{n_1}(\mathbf{p}_1)$ all components are equal to $a_1\mathrm{g_m}(\mathbf{p}_1)$. Similarly, the second subcompostion has all elements equal to $a_2\mathrm{g_m}(\mathbf{p}_2)$. Therefore, all balances corresponding to partitions of the two subcompositions are null, i.e. $b_2 = b_3 = \cdots = b_{n-1} = 0$. Alternatively, the first balance is

$$b_1 = \sqrt{\frac{n_1 n_2}{n_1 + n_2}} \log \frac{a_1\mathrm{g_m}(\mathbf{p}_1)}{a_2\mathrm{g_m}(\mathbf{p}_2)} \ .$$

Therefore, using (11), it holds that $\|(a_1\mathbf{g}_{n_1}(\mathbf{p}_1), a_2\mathbf{g}_{n_2}(\mathbf{p}_2))\|_a = |b_1|$. Equation (10) can now be written as $\|\mathbf{p}_1\|_a^2 + \|\mathbf{p}_2\|_a^2 + |b_1|^2 = \|\mathbf{p}\|_a^2$; this holds when $\|\mathbf{p}\|_a^2$ is expressed as a sum of squared balances with respect to the basis which first balance is defined as b_1.

This discussion remarks that the definition in Equation (11) satisfies the requirements established as principles of evidence-information. However, a comparison of Shannon information and evidence-information should reveal some similarities. Certainly, Shannon information and subsequent developments have been proven to be fruitful tools in many applications of almost all fields of science. Shannon information for an evidence vector in $\mathcal{S}^n$ is lower bounded by $\mathcal{I}_s(\mathbf{n}) = \log n$, to be compared to $\mathcal{I}_e(\mathbf{n}) = 0$. In order to compare both information approaches, Figure 2 (left) shows values of $\mathcal{I}_e(\mathbf{x})$, $\mathbf{x} \in \mathcal{S}^3$, compared with Kullback-Leibler divergence of $\mathbf{x}$ to $\mathbf{n}$, $\mathcal{I}_s(\mathbf{x} : \mathbf{n})$. The points $\mathbf{x} \in \mathcal{S}^3$ have been chosen to follow two straight-lines in the simplex shown in Figure 2 (right). As the straight-lines have been parameterised with arc-length, evidence-information (the Aitchison norm of $\mathbf{x}$) appears as a cone. The Kullback-Leibler divergence describes curves which are on one side not very far from evidence information, but that are definitively different from it.

Unit of evidence-information

The choice of any unit is somewhat arbitrary. However, Aitchison norm and distance in the simplex has been used as a standard for years. It is then reasonable to take the evidence-information unit equal to the Aitchison norm unit. Then, an evidence vector $\mathbf{p} \in \mathcal{S}^n$ has unit evidence information whenever $\mathcal{I}_e(\mathbf{p}) = \|\mathbf{p}\|_a = 1$. To get a little insight into the meaning of this unit, consider an evidence vector $\mathbf{u} = (u, u^{-1}, 1, 1, \dots, 1)$ with $u = \exp(1/\sqrt{2})$. The $\mathrm{clr}(\mathbf{u}) = (1/\sqrt{2}, -1/\sqrt{2}, 0, 0, \dots, 0)$, then $\mathcal{I}_e(\mathbf{u}) = \|\mathbf{u}\|_a = 1$. This can be phrased as follows: *a likelihood vector has unit evidence information when a neutral prior is updated to* $\mathbf{u}$. All components remain constant, except two of them: one is approximately doubled (multiplied by $\exp(1/\sqrt{2}) \approx 2.028$) and the other one is approximately halved (multiplied by $\exp(-1/\sqrt{2}) \approx 0.493$). Obviously, for a fixed n, there are infinitely many unitary evidence vectors in $\mathcal{S}^n$ but the example presented is specially intuitive.

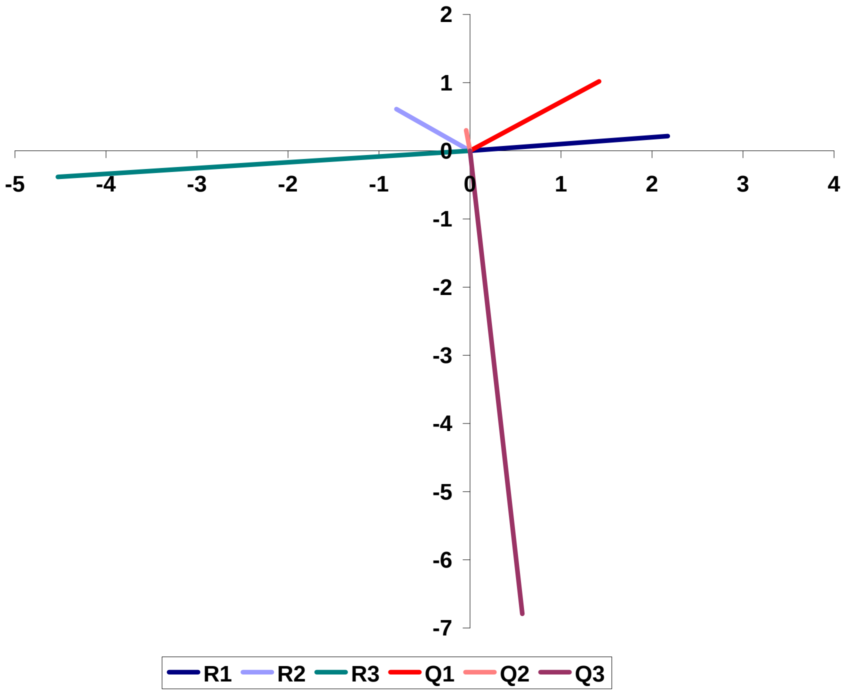

Figure 3: Likelihood vectors for all experimental outputs. Abscissa: $b_1 = (2/3)^{-1/2} \log(x_1/\mathrm{g_m}(x_2, x_3))$; ordinate: $b_2 = 2^{-1/2} \log(x_2/x_3)$.

5 Evaluating the inspection example

The definition of evidence-information as the Aitchison norm of a (compositional) evidence-vector, and the framework of the Aitchison geometry of the simplex allow answering the questions stated in Section 2.

(a) Given one of the available tests, which of the possible outputs provides more information?

Once the likelihood of all outputs of each experiment are known, its Aitchison norm measures the information provided by each output. Table 1, shows both the likelihood and the associated evidence information. The output Q_3 of the dynamic test provides the maximum e-information, $\mathcal{I}_e(Q_3) = 6.82$, followed by R_3 with $\mathcal{I}_e(R_3) = 4.54$. Therefore, in both experiments the outputs which make the state *ruin* more likely are the most informative. Their e-information is more than double of the other outputs.

(b) Given the output from the visual test, R_j, and from the dynamic test, Q_k, which of the two outputs is more informative?

Both likelihoods for R_3, Q_3 are in $\mathcal{S}^3$ and, therefore, they are comparable. Given the two outputs, R_3 and Q_3, the statement $\mathcal{I}_e(Q_3) > \mathcal{I}_e(R_3)$ (Section 2) is licit and meaningful and an answer is immediately obtained from Table 1. This does not mean that the dynamic test is more informative than the visual one in general. For instance, according to Table 1, $\mathcal{I}_e(Q_2) < \mathcal{I}_e(R_2)$. The likelihoods corresponding to the different experiments are shown in Figure 3. The length of the evidence vectors in Figure 3 represents the e-information associated with each possible output of the two tests. Additionally, it appears that the likelihood vectors corresponding to R_1 and R_3 and also Q_2 and Q_3 are practically opposite although their e-informations are not equal. This is very evident in the case Q_2 and Q_3, where Q_2 is very poorly e-informative. Approximate orthogonality is also notorious between Q_3 and $\{R_1, R_3\}$. This means that the output Q_3 informs on features that R_1 and R_3 are unable to illustrate. The likelihood corresponding to Q_3 is almost parallel to the second axis in Figure 3, meaning that observation of Q_3 strengthens the evidence in favor of A_3 (ruin) and against A_2 (damage), leaving almost invariant the log-ratio of probabilities A_1 (service) over (A_2, A_3). In turn, observation of R_3, which likelihood is mainly parallel to the first axis in Figure 3, gives evidence on (A_2, A_3) and against A_1, but leaves approximately invariant the evidence of A_2 versus A_3.

(c) In order to deliver a diagnosis of the structural state of the building, a decision on A_1, A_2, A_3 should be made. Assume that the output of a test is R_j. Independently of the cost-benefit of such a decision, how reliable is the decision taken?

The key question is what does *reliability of a decision* mean. In a decision making context a standard way of evaluating a decision is to compare risks of the possible alternatives; in the example A_1, A_2, A_3, the less risky event is taken as the optimal decision and the minimum risk decision is reliable if differences with other alternatives are important. As we assume that the utility or cost-benefit function is not available, the posterior probabilities are taken as the expected utility. Then, the most probable (posterior) state is decided. Reliability of the decision is then measured by the posterior e-information. For instance, after observing R_3, the decision taken (maximum probability event) is A_2 and its reliability is the e-information 3.75 (see Table 2). If R_2 would be observed instead, decision would be also A_2 but its reliability, expressed as e-information, would be less (1.10). The same reasoning can be applied to decisions taken based on prior probabilities. In general more reliable decisions are expected when they are taken *a posteriori*, as some information is expected from the experiment. However, evidence provided by an experimental result can be opposite to the prior evidence, thus resulting in a loss of e-information. Although in the presented example this situation does not appear, the observation of R_2 and Q_2 results in small increases of e-information, thus meaning that reliability of *a posteriori* decision is only a little bit higher than the prior decision. However, Figure 1 reveals that the small increase of reliability from prior to posterior is due to different causes. In the case of R_2 (left panel), the result is as e-informative as the prior, and the prior, likelihood and posterior form a triangle close to be equilateral. In the case of Q_2 (right panel), there is an additional reason for a low increase of e-information, the size of the likelihood evidence vector is smaller than the prior, i.e. the prior is more e-informative than the e-information coming from the experimental result Q_2.

(d) Can we define an average information provided by a test before getting the output? Which of the two tests is more informative?

This question has been considered from the beginnings of information theory (Lindley, 1956). It puts forward three different points: what should be averaged? which kind of average? and, if it has to be weighted, which are the weights? Starting with the first point, there are several alternatives. A first option is to average likelihood vectors corresponding to different experimental outputs. The average would be carried out using operations in the simplex because likelihood vectors are compositional evidence vectors. In general, this option tends to cancel out the average; see Figure 3, where likelihood vectors appear with a star like shape. To average e-information, which is a positive variable, seems therefore more adequate. This leads to the second question: we can average, at least, three different functions of e-information: the e-information itself, the square of e-information and the logarithm of e-information. With respect to the weights used in the average, one may know the probabilities of getting each result of the experiment, for instance P$[R_i]$, i=1,2,3. However, these probabilities should be compatible with the prior if previously assumed. Likelihood times prior is the joint probability of states and experimental results, and probabilities P$[R_i]$ and the prior probabilities P$[A_j]$ are the two marginals of the joint probability. Denoting $L(R_i)$ the likelihood vector corresponding to the observation R_i (in the example, the columns of Table 1), and $\mathbf{R}$ the output of the experiment as a random variable, the proposed average alternatives may be expressed as

$$\mathrm{E}_\phi[\mathcal{I}_e(\mathbf{R})] = \phi^{-1}(\mathrm{E}[\phi(\mathcal{I}_e(L(\mathbf{R})))]) \ , \tag{13}$$

Table 3: Probabilities of experimental outputs and e-information averages for experiments **R** (visual) and **Q** (dynamic), compatible with the assumed prior.

experiment	$\mathrm{P}[R_i],\ i=1,2,3$			mean	rms	geom.
visual **R**	0.512	0.248	0.240	2.461	2.767	2.153
dynamic **Q**	0.773	0.102	0.125	2.235	2.861	1.734

where ϕ is a monotonic scaling function and $\mathrm{E}[\cdot]$ is the standard expectation for real variables. Expectation E_ϕ is a generalized expectation for scaled variables called ϕ-mean by De Finetti (1990). In our case suitable choices of ϕ are: identity, squaring, or logarithm, i.e. mean, root-mean-square (rms), or geometric mean. For the example used, Equation (13) takes the form

$$\mathrm{E}_\phi[\mathcal{I}_e(\mathbf{R})] = \phi^{-1}\left(\sum_{i=1}^{3}\phi(\mathcal{I}_e(L(R_i)))\ \mathrm{P}[R_i]\right) \ . \tag{14}$$

Table 3 shows the results for both experiments, visual inspection and dynamic test. Probabilities of experimental results have been obtained from the assumed prior (Table 2) and the likelihood associated with the experimental results (Table 1). Averages have been obtained according to Equations (13) and (14). Average e-information is larger for the visual experiment using the mean and the geometric mean of e-information, whereas the situation is reversed for the root-mean-square e-information. The reasons for this result are clear and can be visualized in Figure 3, where $\mathcal{I}_e(Q_2)$ is very small compared with the e-information provided by the dynamic test Q_3. This causes the increase of the root-mean-square and the corresponding decrease of the arithmetic average and, specially, of the geometric average. Consequently, the choice of the kind of average to be taken becomes relevant, as depending on it different aspects of the distribution of e-information are enhanced.

6 Conclusion

Bayes formula is taken as the paradigm of information acquisition. Prior, posterior and likelihood participating in Bayes formula are identified as compositions and the Bayes formula itself as perturbation in the simplex. Additionally, prior, posterior and likelihood are interpreted as vectors describing the evidence before, after and carried by the output of an experiment. Some principles support this interpretation. They essentially coincide with compositional data principles thus leading to the definition of evidence-information as the Aitchison norm of evidence vectors. An example on inspection of buildings illustrate the use of evidence-information to quantify different aspects of the proposed inspection-tests: measuring provided e-information, opposite and orthogonal e-information, comparing experiments, expected e-information from an experiment.

Acknowledgements

This research has been supported by the Spanish Ministry of Science and Innovation under the project "CODA-RSS" Ref. MTM2009-13272; and by the *Agència de Gestió d'Ajuts Universitaris i de Recerca* of the *Generalitat de Catalunya* under the project Ref: 2009SGR424.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London (UK). (Reprinted in 2003 with additional material by The Blackburn Press). 416 p.

Ash, R. B. (1990). *Information theory.* Dover, New York; first puiblished by J. Wiley & Sons, 1965. 339 pp.

Barceló-Vidal, C., J. A. Martín-Fernández, and V. Pawlowsky-Glahn (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*, pp. 20 p. CD-ROM.

Billheimer, D., P. Guttorp, and W. Fagan (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association 96*(456), 1205–1214.

De Finetti, B. (1990). *Theory of Probability. A critical introductory treatment.* Wiley Classics Library (First published Wiley & Sons, 1974), Vol. 1 and 2. 300pp.

Egozcue, J. and V. Pawlowsky-Glahn (2006a). Exploring compositional data with the coda-dendrogram. In E. Pirard (Ed.), *Proceedings of IAMG'06 — The XIth annual conference of the International Association for Mathematical Geology.*

Egozcue, J. and V. Pawlowsky-Glahn (2006b). Simplicial geometry for compositional data. Volume 264 of *Special Publications.* Geological Society, London.

Egozcue, J. J. and V. Pawlowsky-Glahn (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology 37*(7), 795–828.

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology 35*(3), 279–300.

Kinchin, A. I. (1957). *Mathematical foundations of information theory.* Dover Publications, New York, NY (USA). 120 p.

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

Kullback, S. and R. A. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics 22*(1), 79–86.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Annals of Mathematical Statistics 27*(4), 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).

Martín-Fernández, J. A., C. Barceló-Vidal, and V. Pawlowsky-Glahn (1998). Medida de diferencia de Kullback-Leibler entre datos composicionales. In SEIO (Ed.), *Actas del XXIV Congreso Nacional de la Sociedad de Estadística e Investigación Operativa (SEIO)*, pp. 291–292. Sociedad Española de Estadística e Investigación Operativa, Almería (E).

Pawlowsky-Glahn, V. and J. J. Egozcue (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment (SERRA) 15*(5), 384–398.

Rényi, A. (1966). *Wahrscheinlichkeitsrechnung: mit einem Anhang über Informationstheorie*, Volume 54 of *Hochschulbücher für Mathematik.* Deutscher Verlag der Wissenschaften.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Tech. J. 27*, 379–423, 623–656.

Shannon, C. E. and W. Weaver (1949). *The mathematical theory of communication.* University of Illinois Press. Urbana.

van den Boogaart, K. G., J. J. Egozcue, and V. Pawlowsky-Glahn (2010). Bayes linear spaces. *Statistics and Operations Research Transactions, SORT 34*(2), 201–222.