Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 98 (2016) 485 – 490

Procedia Computer Science

International Workshop on Data Mining on IoT Systems (DaMIS16)

A Bound for the Accuracy of Sensors Acquiring Compositional Data

Ardelio Galletti[a,*], Antonio Maratea[a]

[a]*Department of Science and Technology, University of Naples Parthenope, Centro Direzionale Isola C4, 80143, Naples Italy*

Abstract

Among the many challenges that the Internet of Things poses, the accuracy of the sensor network and relative data flow is of the foremost importance: sensors monitor the surrounding environment of an object and give information on its position, situation or context, and an error in the acquired data can lead to inappropriate decisions and uncontrolled consequences. Given a sensor network that gathers relative data – that is data for which ratios of parts are more important than absolute values – acquired data have a compositional nature and all values need to be scaled. To analyze these data a common practice is to map bijectively compositions into the ordinary euclidean space through a suitable transformation, so that standard multivariate analysis techniques can be used. In this paper an error bound on the commonly used asymmetric log-ratio transformation is found in the Simplex. The purpose is to highlight areas of the Simplex where the transformation is ill conditioned and to isolate values for which the additive log-ratio transform cannot be accurately computed. Results show that the conditioning of the transformation is strongly affected by the closeness of the transformed values and that not negligible distortions can be generated due to the unbounded propagation of the errors. An explicit formula for the accuracy of the sensors given the maximum allowed tolerance has been derived, and the critical values in the Simplex where the transformation is component-wise ill conditioned have been isolated.

Peer-review under responsibility of the Program Chairs

Keywords: Compositional data, conditioning, sensitivity, sensor networks.

1. Introduction

The "Internet of Things" (IoT from now on) is built upon the idea of embedding computing power, sensors and universal networking capabilities into objects of everyday use. It requires objects (Things) to be uniquely identified and addressable; high level communication protocols; high level abstractions of the automation possibilities of each device and widespread standards for representing, storing and processing harvested data[3]. All these intertwined aspects should ideally guarantee interoperability of devices, seamless and robust communications, security and privacy, low energy consumption, scalability, environment-friendly use of resources. Examples of IoT potential can be found in the cultural heritage, where ad hoc classification techniques[5,6,7,8,9,10] or collaborative analytics in the Internet of cultural things[4] have proven to be effective. Among the many challenges that IoT poses, the accuracy of the sensor network and relative data flow play a crucial role[16]. Sensors can monitor surrounding environment of an object and give information on its position, situation or context, and an error in the acquired data can lead to inappropriate deci-

* Corresponding author. Tel.: +39-081-5476607
E-mail address: ardelio.galletti@uniparthenope.it

Peer-review under responsibility of the Program Chairs
doi:10.1016/j.procs.2016.09.081

sions and uncontrolled consequences. For this reason, inaccuracy severely limits the smartness that can be embedded into objects and ultimately the IoT potential.
In case a sensor measures relative values instead of absolute amounts (Humidity is an example), generated data falls within the *compositional analysis* umbrella: information content to be extracted and analyzed is conveyed into the ratio of parts, instead of the absolute amount, as is the case of minerals building up rocks or ingredients in a recipe. Another way for saying this is that the sample space should be scale invariant. Given the scale invariance, comparing samples requires them to be standardized to a common reference quantity (1 for unity, 100 for percentages, 10^6 for parts per million and so on), and the obvious way to obtain this standardization is to divide each sample by its total weight. This simple operation, called *closure*, subtly introduces a constraint on the data, which loose a degree of freedom, and hence causes a spurious correlation (the closure problem) that misleads following analysis[2]. While the special nature of compositional data and some warnings on their handling have been formulated more than a century ago, it is no more than three decades that compositional data have found a proper representation and a complete formulation, mainly thanks to the seminal work of Aitchison[1] and the developments it solicited (see[15] for a compendium).
More formally, when N sample data are all positive, and it is meaningful to analyze them in terms of ratios, the vector

$$\mathbf{x_j} = [x_{j1}, \ldots, x_{jD}]$$

of strictly positive numbers expressing the D measured quantities on each sample $j \in \{1, \ldots, N\}$ in Euclidean space is called a *composition*. A desirable property for compositions is scale invariance, that is $\mathbf{x_j}$ and $\alpha\mathbf{x_j}$ should map to the same vector in the sample space $\forall\alpha \in \mathbb{R}^+$. Once the closure operator is applied for standardization (see Section 2), the sample space becomes constrained, looses one degree of freedom and changes its nature: it is reduced to the D-dimensional Simplex (see Section 2). Once proper operations are introduced, the open Simplex and the Euclidean space can be shown to be isomorphic vector spaces.
The *additive log-ratio* transformation is one of the possible realizations of the isomorphism between the two vector spaces (the Simplex and the Euclidean space). As it will be shown in the following, the additive log-ratio transform includes logarithms of ratios of parts, hence its computation accuracy is strongly affected by the closeness of the values (ratios close to one produce logarithms close to zero) and it can generate not negligible distortions due to the unbounded propagation of the errors that contaminate the available data. Purpose of the paper is to perform a sensitivity analysis and to reveal the compositions for which the additive log-ratio transform can, or cannot, be accurately computed. The practical consequence is that special care must be taken when operating on sensor data that are in a certain area of the Simplex. To the best of our knowledge, no such numerical analysis has been performed before.

In section 2 the core definitions and the mathematical background are briefly outlined; in section 3 the sensitivity analysis for the additive log-ratio transformation is performed; in section 4 drawn conclusions close the paper.

2. Preliminaries

Compositional data are vectors of D positive components (where $D > 0$ in an integer number). The sample space for compositional data is an open Simplex. More details about simplexes can be found in[12,11,13]; here it is just reminded that the open D-dimensional Simplex Δ^k, closed to $\kappa > 0$, is the set of vectors having positive components with constant sum κ:

$$S^D = \left\{ [x_1, \ldots, x_D] \mid x_i \in \mathbb{R}^+, \forall i \in \{1, \ldots, D\} \wedge \sum_{i=1}^{D} x_i = \kappa \right\}. \tag{1}$$

Notice that any vector $\mathbf{x}$, having D real positive components, can be rescaled so that its components sum to a positive constant κ (usually 1 or 100); in other words, $\mathbf{x}$ can always be mapped into a vector of S^D through a compositional operation called *closure*. Let

$$\mathbf{x} = [x_1, \ldots, x_D], \qquad x_i \in \mathbb{R}^+, \ \forall i = 1, \ldots, D,$$

be a vector with positive entries, then the closure of $\mathbf{x}$ is defined as:

$$C(\mathbf{x}) = \kappa \left[\frac{x_1}{\sum_{i=1}^{D} x_i}, \ldots, \frac{x_D}{\sum_{i=1}^{D} x_i} \right] \tag{2}$$

The closure acts as a projection of positive vectors onto the Simplex. Please note that, after the closure, the data become linearly dependent and the dimension of the sample space drops to $D - 1$.

2.1. The Simplex as vector space

As stated before, the Simplex S^D is isomorphic to the Euclidean space, through the transformations described in Section 2, once proper operations and norm are defined. Such operations follows:

- the operation that translates in the Simplex the sum of vectors is called *perturbation* and denoted by $\oplus$. If $\mathbf{x} \in S^D$ and $\mathbf{y} \in S^D$ are compositions, the perturbation of $\mathbf{x}$ by $\mathbf{y}$ is defined as

$$\mathbf{x} \oplus \mathbf{y} = C[x_1 y_1, \ldots, x_D y_D]; \tag{3}$$

- the operation, analogous to multiplication between a scalar and a vector in the Euclidean space, is called *powering* and denoted by $\odot$. If $\mathbf{x} \in S^D$, and $\alpha \in \mathbb{R}$, the powering of $\mathbf{x}$ by α is defined as:

$$\alpha \odot \mathbf{x} = C[x_1^\alpha, \ldots, x_D^\alpha] \tag{4}$$

It is stressed here that with operations $\oplus$ and $\odot$, the Simplex S^D behaves like a vector space. Other useful definitions for vectors in the Simplex are the *inner product*, the *norm* and the *distance*:

- the inner product of two compositions $\mathbf{x} \in S^D$ and $\mathbf{y} \in S^D$ is defined as:

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{i=1}^{D} \log \frac{x_i}{\mathrm{g}(\mathbf{x})} \log \frac{y_i}{\mathrm{g}(\mathbf{y})} \tag{5}$$

  where $\mathrm{g}(\mathbf{z}) = (z_1 \cdot z_2 \cdot \ldots \cdot z_D)^{1/D}$ denotes the geometric mean of the components of $\mathbf{z}$
- the induced norm is defined as:

$$\|\mathbf{x}\|^2 = \sum_{i=1}^{D} \left(\log \frac{x_i}{\mathrm{g}(\mathbf{x})} \right)^2 \tag{6}$$

- the distance $\Delta : S^D \times S^D \to \mathbb{R}_0^+$ is defined as:

$$\Delta^2(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{D} \left\{ \log \frac{x_i}{\mathrm{g}(\mathbf{x})} - \log \frac{y_i}{\mathrm{g}(\mathbf{y})} \right\}^2 \tag{7}$$

Considering the introduced definitions, it is easy to derive the notion of *perturbation independence* as the corresponding of linear independence in the ordinary Euclidean space. The analogous of the linear combination of two vectors in the Simplex is:

$$\mathbf{w} = a \odot \mathbf{v_1} \oplus b \odot \mathbf{v_2} \tag{8}$$

from which all related notions of basis, generated subspace and orthonormality can be derived.

2.2. Representation of the Simplex

Up to $D = 4$, compositions have an intuitive graphical representation. Indeed, a 2-Simplex ($D = 2$) is a line segment, a 3-Simplex ($D = 3$) is a triangle and a 4-Simplex ($D = 4$) is a tetrahedron. More specifically, the representation of a 3-Simplex in the plane is a *ternary diagram*, i.e. a triangle with vertexes P_1, P_2 and P_3, in which any composition (x_1, x_2, x_3) is represented by the interior point:

$$P = \sum_{i=1}^{3} x_i P_i = x_1 P_1 + x_2 P_2 + x_3 P_3. \tag{9}$$

Here, the entries x_i are used as *barycentric coordinates* of P with respect to P_i. Barycentric coordinates represent a measure of the closeness of a point P to the vertexes P_i. More precisely, if P is an interior point of a D-Simplex S^D of vertexes $P_1, P_2, \ldots, P_{D+1}$, the barycentric coordinates can be expressed as a ratio of volumes of simplexes [11,12,13]:

$$x_i = \frac{vol(S^D(P_1, \ldots, P_{i-1}, P, P_{i+1}, \ldots, P_{D+1}))}{vol(S^D(P_1, \ldots, P_{D+1}))} \qquad (i = 0, \ldots, k).$$

Exploiting the previous characterization for $D = 3$, a simple geometrical interpretation can be given: the barycentric coordinate x_i, i.e. the value of each component in the composition, is proportional to the distance of P from the opposite side of vertex P_i. The graphical representation with ternary diagrams is very useful to show what happens to parallelism, orthogonality and projection when mapping to and from the Simplex.

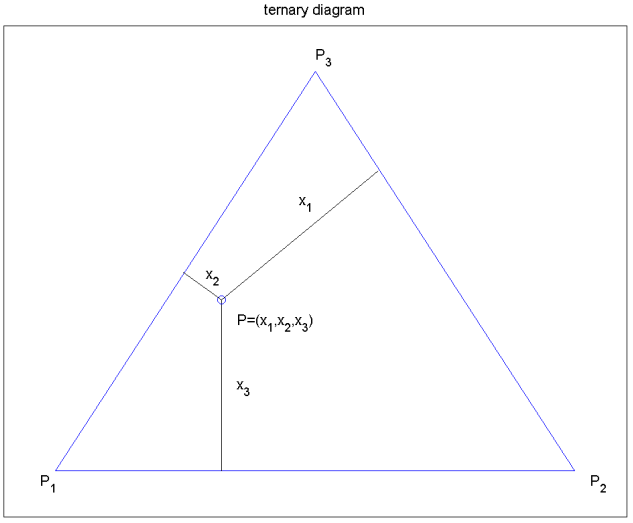

Fig. 1. The three dimensional Simplex

2.3. The alr transformation

Transformations are used to map bijectively compositions into ordinary euclidean space, to allow the use of standard multivariate analysis techniques on compositional data.
The additive log-ratio (*alr* from now on) is a transformation $S^D \rightarrow R^{D-1}$ defined as follows:

$$alr(\mathbf{x}) = \left[\log \frac{x_1}{x_D}, \ldots, \frac{x_{D-1}}{x_D}\right] \tag{10}$$

where the choice of the x_i at the denominator is arbitrary. Its inverse is a transformation $R^{D-1} \rightarrow S^D$ defined as follows:

$$alr^{-1}(\mathbf{y}) = C\left[e^{y_1}, \ldots, e^{y_{D-1}}, 1\right] \tag{11}$$

It is stressed here that the transformed vector has size $(D - 1)$ and expresses coordinates with respect to an oblique (not orthogonal, the angles between each pair of compositions in the basis is 60 degree) basis. *alr* main disadvantages are that the mapping from the Simplex – Aitchison distance – to the real alr space with ordinary euclidean metric is not isometric and that it is not easy to map back the results of the analysis. Nonetheless the transformation allows to analyze the data in the ordinary euclidean space with standard unconstrained techniques, and is often chosen for its simplicity.

3. Sensitivity analysis

Following the approach presented in D'Amore et al.[14], the error propagation in using the *alr* transformation is studied hereafter. Possible sources of errors for sensors are, among others, the round-off error, the representation format, the sampling rate, the jitter, the resolution, and more. A cumulative error term for all these sources will be considered in the following.
Let a composition $\mathbf{x}$ be known with errors $\delta\mathbf{x}$, that is the composition $\tilde{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$ is available. Since entries x_j are all positive, the parts of $\tilde{\mathbf{x}}$ can be written as:

$$\tilde{x}_j = x_j + \delta x_j = x_j\left(1 + \theta_j\right), \qquad \text{where } \theta_j = \frac{\delta x_j}{x_j} \tag{12}$$

expresses the relative error on x_j. To study the sensitivity of the *alr* to the changes δx_j on the data, a small constant θ, which can be thought the accuracy of the sensor network, is assumed to exist, such that:

$$|\theta_j| \le \theta, \quad \forall\, j = 1, \ldots, D. \tag{13}$$

The following result gives a bound for the propagation error.

Theorem 3.1. *Let* $\mathbf{x}$ *and* $\tilde{\mathbf{x}} = \mathbf{x} + \delta\mathbf{x}$ *be compositions with* δx_j *as in (12) and* θ_j *as in (13). Then, it is:*

$$\frac{|alr_j(\tilde{\mathbf{x}}) - alr_j(\mathbf{x})|}{|alr_j(\mathbf{x})|} \le \mu_j \cdot \left(\theta + O(\theta^2)\right), \qquad \forall\, j = 1, \ldots, D-1 \tag{14}$$

with:

$$\mu_j = \frac{2}{|\log(x_j/x_D)|}, \qquad \forall\, j = 1, \ldots, D-1. \tag{15}$$

Proof 3.1. *Exploiting the alr form, it can be obtained:*

$$|alr_j(\tilde{\mathbf{x}}) - alr_j(\mathbf{x})| = \left|\log\frac{\tilde{x}_j}{\tilde{x}_D} - \log\frac{x_j}{x_D}\right| = \left|\log\frac{\tilde{x}_j}{x_j} - \log\frac{\tilde{x}_D}{x_D}\right| \le |\log(1+\theta_j)| + |\log(1+\theta_D)| \le 2\theta + O(\theta^2),$$

where the last inequality arises from the Taylor expansion of $\log(1+x)$*. The proof is completed by dividing by* $|alr_j(\mathbf{x})| = |\log(x_j/x_D)|$.

Please note that the quantities μ_j in (15) act as amplification factors of the relative errors from the data $\mathbf{x}$ to the solutions alr_j, hence they must be considered the relative condition numbers of the problem of evaluating the additive log-ratio transformation. If parts x_j and x_D are close to each other, then $\log x_j/x_D \approx 0$ and $\mu_j \to \infty$. Then for compositions with some part x_j close to x_D the evaluation of alr_j is compely unreliable. More in general, the quantity:

$$\mu = \max_{j=1,\ldots,D-1} \mu_j \tag{16}$$

can be accepted as the **relative condition number** of the *alr* function. The behaviour of μ is characterized by the following result:

Theorem 3.2. *The problem of evaluating the alr transformation is well conditioned if and only if,* $\forall\, j = 1, \ldots, D-1$*, one of the following properties holds true:*

$$x_j \ge e^2 x_D \qquad \text{or} \qquad \mathrm{x_D} \ge \mathrm{e^2 x_j}. \tag{17}$$

Proof 3.2. *The problem is well conditioned if* $\mu \le 1$*, that is if:*

$$\frac{2}{|\log(x_j/x_D)|} \le 1 \quad \Leftrightarrow \quad |\log(x_j/x_D)| \ge 2.$$

The thesis follows solving the last inequality for both cases $x_j > x_D$ *and* $x_j \le x_D$.

Theorem 3.2 shows that the evaluation of the *alr* transformation is well conditioned for some compositions, so that the errors on the data acquired by sensors are not amplified, while ill conditioned for other compositions, so that the errors on these data are strongly amplified (proportionally to conditioning). This fact suggests that to have a precise evaluation of the *alr*, called *tol* the desired accuracy, the following inequality should hold:

$$\frac{|alr_j(\tilde{\mathbf{x}}) - alr_j(\mathbf{x})|}{|alr_j(\mathbf{x})|} \leq tol, \tag{18}$$

from which it can be deduced that the sensor should have an accuracy that verifies the bound:

$$\theta \leq \frac{tol}{\mu}. \tag{19}$$

4. Conclusions

Given a sensor network that acquires relative data, for which ratios of parts are more important than absolute values, it has been shown that the amplification factors of the relative errors from the data $\mathbf{x}$ to the solutions alr_j are, under some circumstances, unbounded and should hence be carefully managed. An explicit formula for the accuracy of the sensors that acquire the data given the maximum allowed tolerance has been derived, and the critical values in the Simplex where the transformation is component-wise ill conditioned have been highlighted. Future work is in studying other transformations and in comparing the sensitivity to errors of the various possible choices.

References

1. J Aitchison. *The Statistical Analysis of Compositional Data*. Chapman & Hall, Ltd., London, UK, UK, 1986.
2. John Aldrich. Correlations genuine and spurious in pearson and yule. *Statistical Science*, 10(4):364–376, 1995.
3. Luigi Atzori, Antonio Iera, and Giacomo Morabito. The internet of things: A survey. *Computer Networks*, 54(15):2787 – 2805, 2010.
4. Angelo Chianese, Fiammetta Marulli, Francesco Piccialli, Paolo Benedusi, and Jai E. Jung. An associative engines based approach supporting collaborative analytics in the internet of cultural things. *Future Generation Computer Systems*, pages –, 2016.
5. Salvatore Cuomo, Pasquale De Michele, Ardelio Galletti, Francesco Pane, and Giovanni Ponti. Visitor dynamics in a cultural heritage scenario. In *DATA 2015 - Proceedings of 4th International Conference on Data Management Technologies and Applications, Colmar, Alsace, France, 20-22 July, 2015*, pages 337–343, 2015.
6. Salvatore Cuomo, Pasquale De Michele, Ardelio Galletti, and Francesco Piccialli. A cultural heritage case study of visitor experiences shared on a social network. In *10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing, 3PGCIC 2015, Krakow, Poland, November 4-6, 2015*, pages 539–544, 2015.
7. Salvatore Cuomo, Pasquale De Michele, Ardelio Galletti, and Giovanni Ponti. A biologically inspired model for analyzing behaviours in social network community and cultural heritage scenario. In *Tenth International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2014, Marrakech, Morocco, November 23-27, 2014*, pages 485–492, 2014.
8. Salvatore Cuomo, Pasquale De Michele, Ardelio Galletti, and Giovanni Ponti. Visiting styles in an art exhibition supported by a digital fruition system. In *11th International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2015, Bangkok, Thailand, November 23-27, 2015*, pages 775–781, 2015.
9. Salvatore Cuomo, Pasquale De Michele, Ardelio Galletti, and Giovanni Ponti. *Data Management Technologies and Applications: 4th International Conference, DATA 2015, Colmar, France, July 20-22, 2015, Revised Selected Papers*, volume 584 of *Communications in Computer and Information Science*, chapter Classify Visitor Behaviours in a Cultural Heritage Exhibition, pages 17–28. Springer International Publishing, 2016.
10. Salvatore Cuomo, Pasquale De Michele, Ardelio Galletti, and Giovanni Ponti. *Intelligent Interactive Multimedia Systems and Services 2016*, volume 55 of *Smart Innovation, Systems and Technologies*, chapter Influence of Some Parameters on Visiting Style Classification in a Cultural Heritage Case Study, pages 567–576. Springer International Publishing, 2016.
11. Salvatore Cuomo, Ardelio Galletti, Giulio Giunta, and Livia Marcellino. A class of piecewise interpolating functions based on barycentric coordinates. *Ricerche di Matematica*, 63(11):87–102, 2014.
12. Salvatore Cuomo, Ardelio Galletti, Giulio Giunta, and Livia Marcellino. A novel triangle-based method for scattered data interpolation. *Applied Mathematical Sciences*, 8(134):6717–6724, 2014.
13. Salvatore Cuomo, Ardelio Galletti, Giulio Giunta, and Livia Marcellino. Piecewise hermite interpolation via barycentric coordinates. *Ricerche di Matematica*, 64(2):303–319, 2015.
14. L. D'Amore, R. Campagna, A. Galletti, L. Marcellino, and A. Murli. A smoothing spline that approximates laplace transform functions only known on measurements on the real axis. *Inverse Problems*, 28(2), 2012.
15. J. J. Egozcue and V. Pawlowsky. Simplicial geometry for compositional data. In *Compositional Data Analysis in the Geosciences: From Theory to Practice*, pages 67–77. Geological Society, Nov 2008.
16. D.L. Urban. *Landscape Ecology*, chapter 3. John Wiley & Sons, Ltd, 2006.