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:

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