International Journal of Approximate Reasoning 54 (2013) 615–626

Contents lists available at SciVerse ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

# A proof for the positive definiteness of the Jaccard index matrix

Mathieu Bouchard [a], Anne-Laure Jousselme [b,*], Pierre-Emmanuel Doré [c]

[a] Université Laval, Quebec, QC, Canada G1K 7P4
[b] Defence R&D Canada – Valcartier, Quebec, QC, Canada G3J 1X5
[c] IRISA, Université de Rennes 1, 22302 Lannion Cedex, France

ARTICLE INFO

*Article history:*
Received 16 May 2011
Received in revised form 21 January 2013
Accepted 22 January 2013
Available online 4 February 2013

*Keywords:*
Belief functions
Evidence theory
Distance
Separability
Jaccard
Definiteness

ABSTRACT

In this paper we provide a proof for the positive definiteness of the Jaccard index matrix used as a weighting matrix in the Euclidean distance between belief functions defined in Jousselme et al. (2001) [13]. The idea of this proof relies on the decomposition of the matrix into an infinite sum of positive semidefinite matrices. The proof is valid for any size of the frame of discernment but we provide an illustration for a frame of three elements. The Jaccard index matrix being positive definite guarantees that the associated Euclidean distance is a full metric and thus that a null distance between two belief functions implies that these belief functions are strictly identical.



## 1. Introduction

The *Jaccard index* or *Jaccard similarity coefficient* was introduced by the botanist Paul Jaccard [1] in 1901 in [12] as the ratio of the size of the intersection between two sets and the size of their union. The Jaccard index is now a classical and commonly used measure of similarity between sets in many applications since its introduction. Quite recently [13], it has been used in a generalized (weighted) Euclidean distance between belief functions. In evidence theory (or belief functions theory) [4,19], quantifying how much two belief functions are distinct is unavoidable if one is interested in some optimization of the fusion process. A survey of the main distances between belief functions defined so far has been presented in [14,15] together with a synthesis of their main properties. Distances between belief functions can be grouped into five families, i.e., Composite, Minkowski, Inner product, Fidelity and Information-based, the most populated family being the Minkowski one ($p$-distance or $L_p$), among which Euclidean distances ($L_2$) play a major role. These families of distances extend their counterparts defined between two probability distributions. Moreover, the standard metric properties have led to four groups: metrics, pseudometrics, semipseudometrics and nonmetric measures. Additionally, structural properties, which are very specific to belief functions as they qualify the interaction between focal elements, allow us to distinguish between two groups of distances, namely, structural and non-structural. Interestingly, it appears that these properties have a higher impact on the distances' comparative behavior compared to their family of belonging [15].

In this paper, we are interested in the metric properties and, in particular, in the definiteness property, which states that a null distance between two objects guarantees they are strictly identical, or formally, $d(x_1, x_2) = 0 \Leftrightarrow x_1 = x_2$. Although the right to left implication ($d(x, x) = 0$) is easily satisfied, the left to right implication called *separability* is rather often violated by distance measures. The satisfiability of the separability property is what distinguishes metric distances (or "full" distances) from pseudometric distances: in the pseudometric case, thus two distinct objects may have a null distance.

* Corresponding author.
*E-mail addresses:* mathieu.bouchard@forac.ulaval.ca (M. Bouchard), Anne-Laure.Jousselme@drdc-rddc.gc.ca (A.-L. Jousselme), pierre.emmanuel.dore@gmail.com (P.-E. Doré).

[1] Paul Jaccard called his coefficient *coefficient de communauté*.

0888-613X/$ - see front matter Crown copyright © 2013 Published by Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ijar.2013.01.006

Among the Euclidean distances between belief functions defined so far, the Jaccard distance, denoted by $d_J^{(2)}$ hereafter, firstly introduced in [13] has been widely used in several applications such as estimating discounting factors [8,16,5,11], optimizing combination rules [6,17] or for the evaluation of a dependency between sources [17,18]. Note that although widely used, the Jaccard distance is only one of the numerous distances between belief functions that have been defined [15,1,7].

Though the definiteness of $d_J^{(2)}$ was highly suspected, no formal proof of this property has been given so far,[2] leaving consequently some doubts about the validity of $d_J^{(2)}$ as a full distance between belief functions. The aim of this paper is to provide such a proof by simply proving that the Jaccard index matrix is positive definite for any size of the frame of discernment. In [9], Gower proved that the Jaccard index matrix is positive semidefinite, for any $N$ subsets of a reference frame $X$ of size $n$. We extend this proof and show that the complete $(2^n - 1) \times (2^n - 1)$ Jaccard index matrix of pairs of subsets of $X$ is positive definite for any positive integer value of $n$.

In Section 2 we first introduce some background on belief functions theory with an emphasis on distances between belief functions, highlight the consequences of the non-definiteness of a distance on the decision process, and recall Gower's theorem about the positive semidefiniteness of the Jaccard index matrix. Section 3 is dedicated to the proof itself, introducing first some notations, definitions and basic properties to be used in the core of the proof (Section 3.1). Gower's result cannot be used directly in our proof because it requires a more detailed decomposition of the elements of the infinite sum than the one proposed by Gower. Thus the decomposition of the Jaccard index matrix which is central to the proof is presented in Sections 3.2 and 3.3, leading to Gower's result and introducing the necessary details to state the main result of this paper in Section 3.4. Some computational results are also provided supporting the theoretical results. In Section 4 we illustrate the steps of the proof for a frame of discernment of three objects. We conclude in Section 5 on some opportunities for future works.

## 2. Background

Evidence theory [4,19] also known as Dempster–Shafer theory, is one of the most popular frameworks for dealing with uncertain information. Often presented as a generalization of probability theory where the additivity axiom is excluded, evidence theory allows each subset of the universe to have a nonnull confidence, and not only the singletons as in probability theory. Rather than a distribution on the singletons as in probability theory, a distribution over the power set of the universe is defined.

### *2.1. Basic notions on evidence theory*

Let $X$ be a frame of discernment (or universe) containing $n$ distinct objects $x_i$, $i = 1, \dots, n$, let $\mathcal{P}(X)$ be the power set of $X$ and let $\mathcal{P}(X)'$ be the power set of $X$ bereft of the empty set, $\mathcal{P}(X)\backslash\varnothing$. A Basic Probability Assignment (BPA) $m$ is a mapping from $\mathcal{P}(X)$ to $[0, 1]$ satisfying:

$$\sum_{A \subseteq X} m(A) = 1 \quad \text{and} \quad m(\varnothing) = 0. \tag{1}$$

A subset $A$ of $X$ such that $m(A) > 0$ is called a *focal element*. A belief function is a mapping from $\mathcal{P}(X)$ to $[0, 1]$ such that:

$$\mathrm{Bel}(A) = \sum_{B \subseteq A} m(B) \tag{2}$$

and a plausibility function is a mapping from $\mathcal{P}(X)$ to $[0, 1]$ such that:

$$\mathrm{Pl}(A) = \sum_{B \cap A \neq \varnothing} m(B). \tag{3}$$

We also have $\mathrm{Bel}(A) = 1 - \mathrm{Pl}(\overline{A})$, where $\overline{A}$ is the complement of $A$.

The belief function whose single focal element is $X$ is called *vacuous*; a belief function whose two single focal elements are $A$ and $X$ is called a *simple support* belief function; a belief function whose three focal elements are $A$, $\overline{A}$ and $X$ is called *dichotomous*.

For decision making purposes, the standard approach is to transform the belief function into a probability distribution to be able then to apply the classical decision theory [21]. The most used transformation is the so-called pignistic transformation defined as:

$$\mathrm{BetP}(A) = \sum_{B \subseteq X} m(B) \frac{|A \cap B|}{|B|}, \tag{4}$$

[2] Note that in [10], the authors mention that the positive definiteness of the Jaccard index matrix "is shown in [9]" although only the positive semidefiniteness is proved.