A General Framework for Comparing Heterogeneous Binary Relations

Julien Ah-Pine

A general framework for comparing heterogeneous binary relations

Julien Ah-Pine

ERIC Lab, University of Lyon
5, avenue Pierre Mendès France
69500 Bron, France
julien.ah-pine@eric.univ-lyon2.fr

Abstract. We propose a general framework for measuring the proximity between heterogeneous binary relations. We are particularly interested in equivalence relations (or partitions, or categorical variables) on the one hand, and in order relations (or rankings or ranked data) on the other hand. We study Kendall's general correlation coefficient that encompasses many statistical association measures for both types of binary relations. Then, we propose a generalization of this approach. Our proposal has several interesting properties. It addresses the case of heterogeneous binary relations, it has a well-motivated geometrical interpretation and it also leads to Euclidean metrics.

1 Introduction

Binary relations (BR) are fundamental mathematical concepts that allow structuring a set of items or objects in intelligent ways. Among the different types of BR, equivalence relations (ER) and order relations (OR) are ubiquitous. Some examples of ER are : "is similar to", "has the same predicate than", "belongs to the same group as", ... and some illustrations of OR are : "is lower than", "is preferred to", "is ranked above than", ...

Such types of data are relational since they encode the relationship between items unlike more classical representations of objects described by features. There are many situations where we have to deal with such relational data. For instance, in the statistical literature, research works in categorical data analysis study ER while in non-parametric statistics one is interested in the OR underlying quantitative variables. In both topics, association measures have been proposed in order to compare two ER and two OR.

In this paper, we design a general framework for measuring the proximity between BR. Our framework addresses heterogeneous BR where ER can have different numbers of equivalence classes with different cardinals and OR can be partial and/or with ties. We propose to extend Kendall's general correlation coefficient Γ which is a formula used in many ways in order to compare ER and OR. Our generalization proceeds in two steps. In section 2, after recalling some basics about BR, we firstly reformulate Kendall's Γ in the Relational Analysis (RA) framework. This approach allows us to propose an unifying formula for

both ER and OR which emphasizes the relational nature of such data and which also relies on the indeterminacy principle. Secondly, in section 3, we extend Kendall's Γ in the goal of dealing with heterogeneous BR in a more effective manner. In that perspective, we introduce a family of proximity measures called similarity of order t. These measures have a clear geometrical interpretation and result in Euclidean metrics. Finally, we conclude this paper and sketch some future work in section 4.

2 Measures for BR using relational matrices

2.1 BR and relational matrices

A binary relation R on a set of objects $\mathbb{A} = \{\ldots, i, j, \ldots\}$ with $|\mathbb{A}| = n$, is a couple $(\mathbb{A}, G(R))$, where $G(R)$ called the graph of the relation R, is a subset of the Cartesian product $\mathbb{A}^2$. If $(i,j) \in G(R)$, then we say that i is in relation with j for R. This will be denoted by iRj. We can associate to a BR R, its complement which is a BR denoted by $\overline{R}$ and whose graph is defined by, $\forall i, j \in \mathbb{A} : (i,j) \in G(\overline{R}) \Leftrightarrow (i,j) \notin G(R)$. Furthermore, we can derive from R its inverse (also called its converse) which is also a BR that we denote by $\breve{R}$ and whose definition is given by, $\forall i, j \in \mathbb{A} : (i,j) \in G(\breve{R}) \Leftrightarrow (j,i) \in G(R)$.

There are different properties that a BR can satisfy. Some of the most used ones are : reflexivity, $\forall i\, (iRi)$; symmetry, $\forall i, j\, (iRj \Rightarrow jRi)$; antisymmetry, $\forall i, j\, ((iRj \wedge jRi) \Rightarrow i = j)$; complete (or total), $\forall i \neq j\, (iRj \vee jRi)$; transitivity, $\forall i, j, k\, ((iRj \wedge jRk) \Rightarrow iRk)$.

The BR we are interested in are defined as follows : ER are reflexive, symmetric and transitive; preorders are reflexive and transitive; partial orders are reflexive, antisymmetric and transitive; total (or linear or complete) orders are reflexive, antisymmetric, transitive and complete. Concerning OR, total orders are rankings without missing values and without ties, partial orders are rankings without ties but missing values can occur; complete preorders are rankings without missing values but with possible ties and preorder are rankings with possible missing values and ties.

In the RA approach [1, 2, 4], BR are represented by their adjacency matrices that are called more specifically relational matrices (RM) in this context. We denote by $\mathbf{R}$ the RM of the BR R, which is a binary pairwise comparison matrix whose general term is given by, $\forall i, j : \mathbf{R}_{ij} = 1$ if iRj and $\mathbf{R}_{ij} = 0$ if $i\overline{R}j$. Besides, we will respectively denote by $\overline{\mathbf{R}}$ and $\breve{\mathbf{R}}$ the RM of the complement and inverse of R. Using RM, the relational properties of BR can be expressed as linear equations. This is a first interesting feature of RA which allows aggregating binary relations using 0-1 integer linear programming [2, 3]. However, in this paper, we focus on a second kind of contributions of this approach which concerns statistical association measures. We review this topic and propose the first generalization of Kendall's Γ in the next paragraph.

2.2 Measures based on Kendall's Γ and RM

In his famous book [5], Kendall proposed a general correlation coefficient in order to define a broad family of association measures. Even if his proposal initially aimed at rankings, it can also be adapted to categorical variables. Let $\mathbf{x}$ and $\mathbf{y}$ be two given variables of measurements on $\mathbb{A}$. Kendall's Γ takes the following general form :

$$\Gamma(\mathbf{x},\mathbf{y}) = \frac{\sum_{i,j} \mathsf{X}_{ij}\mathsf{Y}_{ij}}{\sqrt{\sum_{i,j} \mathsf{X}_{ij}^2}\sqrt{\sum_{i,j} \mathsf{Y}_{ij}^2}} \tag{1}$$

where X and Y are two square matrices derived from $\mathbf{x}$ and $\mathbf{y}$. This coefficient is symmetric and both variables are treated the same way. Hereafter, we thus mention the transformation for $\mathbf{x}$ only.

Let us start with ER. In that case, $\mathbf{x}$ and $\mathbf{y}$ are two categorical variables with $p_{\mathbf{x}}$ and $p_{\mathbf{y}}$ categories and categories sets denoted $\{1,\ldots,p_{\mathbf{x}}\}$ and $\{1,\ldots,p_{\mathbf{y}}\}$. We denote by $n_u^{\mathbf{x}}$ the total number of items in category $u \in \{1,\ldots,p_{\mathbf{x}}\}$. In [6], Janson and Vegelius suggested to adapt Kendall's approach to categorical variables. They particularly showed that when $\mathsf{X}_{ij} = (n/n_u^{\mathbf{x}}) - 1$ if $\mathbf{x}_i = \mathbf{x}_j$ and $\mathsf{X}_{ij} = -1$ if $\mathbf{x}_i \neq \mathbf{x}_j$, we obtain Tchuprow's T coefficient. Another interesting measure, called the J-index, that was proposed by the same authors, is given by : $\mathsf{X}_{ij} = p_x - 1$ if $\mathbf{x}_i = \mathbf{x}_j$ and $\mathsf{X}_{ij} = -1$ if $\mathbf{x}_i \neq \mathbf{x}_j$.

Suppose now that $\mathbf{x}$ and $\mathbf{y}$ are quantitative variables. If we compare the measures assigned to all pairs $(i,j) \in \mathbb{A}^2$ we obtain two OR of the type "is lower than". In [5], Kendall proposed the following setting to measure the dependence between $\mathbf{x}$ and $\mathbf{y}$ from a non-parametric perspective : $\mathsf{X}_{ij} = 1$ if $\mathbf{x}_i < \mathbf{x}_j$ and $\mathsf{X}_{ij} = -1$ if $\mathbf{x}_i > \mathbf{x}_j$. In that case, $\sum_{i,j} \mathsf{X}_{ij}\mathsf{Y}_{ij}$ is twice the number of concordant pairs minus twice the number of discordant pairs while $\sum_{i,j} \mathsf{X}_{ij}^2$ gives $n(n-1)$. The resulting coefficient is Kendall's popular τ_a. Another famous rank correlation measure is Spearman's ρ_a statistic. In that case, we assume that $\mathbf{x}_i$ and $\mathbf{y}_i$ are the ranks associated to i according to X and Y. Then, ρ_a is given by $\mathsf{X}_{ij} = \mathbf{x}_i - \mathbf{x}_j$.

Let X and Y be the BR implicitly encoded by $\mathbf{x}$ and $\mathbf{y}$ and let $\mathbf{X}$ and $\mathbf{Y}$ be their respective RM. Then, we introduce the notation $\widetilde{\mathbf{X}}$ which represents the opposite relation of X in a general way. In our case, $\widetilde{\mathbf{X}}$ will refer to $\overline{\mathbf{X}}$ if X is an ER and to $\breve{\mathbf{X}}$ if X is an OR. In other words, if X is an ER ("is in the same category as") then we consider its complement ("is not in the same category as") as its opposite whereas if X is an OR ("is lower than") its opposite is its inverse ("is greater than"). We propose the following expression denoted Λ that generalizes Kendall's Γ by explicitly integrating the RM $\mathbf{X}$, $\mathbf{Y}$, $\widetilde{\mathbf{X}}$ and $\widetilde{\mathbf{Y}}$:

$$\Lambda(\mathbf{X},\mathbf{Y},\mu,\widetilde{\mu}) = \frac{\sum_{i,j}(\mu_{ij}^{\mathbf{x}}\mathbf{X}_{ij} - \widetilde{\mu}_{ij}^{\mathbf{x}}\widetilde{\mathbf{X}}_{ij})(\mu_{ij}^{\mathbf{y}}\mathbf{Y}_{ij} - \widetilde{\mu}_{ij}^{\mathbf{y}}\widetilde{\mathbf{Y}}_{ij})}{\sqrt{\sum_{i,j}(\mu_{ij}^{\mathbf{x}}\mathbf{X}_{ij} - \widetilde{\mu}_{ij}^{\mathbf{x}}\widetilde{\mathbf{X}}_{ij})^2}\sqrt{\sum_{i,j}(\mu_{ij}^{\mathbf{y}}\mathbf{Y}_{ij} - \widetilde{\mu}_{ij}^{\mathbf{y}}\widetilde{\mathbf{Y}}_{ij})^2}} \tag{2}$$

where $\forall i,j \in \mathbb{A}$: $\mu_{ij}^{\mathbf{x}}$ and $\widetilde{\mu}_{ij}^{\mathbf{x}}$ are non-negative weights for iXj and $i\widetilde{X}j$. There have been many research works about using RM to better understand the

differences between association measures [7–11]. In this paper, (2) follows the formalism proposed in [12] which underlines the weighted indeterminacy situation between two BR. This concept was first introduced in [7] and lately studied in the case of ER in [13]. In order to better understand how (2) generalizes Kendall's Γ and the weighted indeterminacy concept as well, it is worth emphasizing the roles played by the weights $\mu_{ij}^{\mathbf{x}}$ and $\widetilde{\mu}_{ij}^{\mathbf{x}}$. Indeed, μ and $\widetilde{\mu}$ should be viewed as mappings reflecting weighting models which, given a RM $\mathbf{X}$ and a pair (i,j), assign a non-negative value to $\mathbf{X}_{ij}$ (iXj) and to $\widetilde{\mathbf{X}}_{ij}$ $(i\widetilde{X}j)$. In our framework, the semantic underlying weighting models is what allows differentiating coefficients from one to another.

Accordingly, the following results show that (2) indeed encompasses[1] famous coefficients derived from Kendall's Γ. As for ER, the findings in [6] and [7, 9] enable us to state the following particular cases of Λ : $\mu_{ij}^{\mathbf{x}} = [(\sum_i \mathbf{X}_{ij} + \sum_j \mathbf{X}_{ij})/2]^{-1} - 1/n = [\sum_j \mathbf{X}_{ij}]^{-1} - 1/n$ and $\overline{\mu}_{ij}^{\mathbf{x}} = 1/n$ leads to Tchuprow's T coefficient; $\mu_{ij}^{\mathbf{x}} = 1 - 1/\sum_{i,j}[2\mathbf{X}_{ij}/(\sum_i \mathbf{X}_{ij} + \sum_j \mathbf{X}_{ij})]^2 = 1 - 1/\sum_{i,j}[\mathbf{X}_{ij}/\sum_j \mathbf{X}_{ij}]^2 = 1 - 1/p_{\mathbf{x}}$ and $\overline{\mu}_{ij}^{\mathbf{x}} = 1/p_{\mathbf{x}}$ gives the J-index; $\mu_{ij}^{\mathbf{x}} = \overline{\mu}_{ij}^{\mathbf{x}} = 1$ is related to Rand's index; $\mu_{ij}^{\mathbf{x}} = 1 - \sum_{i,j} \mathbf{X}_{ij}/n^2$ and $\overline{\mu}_{ij}^{\mathbf{x}} = \sum_{i,j} \mathbf{X}_{ij}/n^2$ is Pearson's product-moment correlation coefficient on a fourfold contingency tables.

Regarding OR, the results provided in [5] and [8, 12], allow establishing the following cases derived from (2). If X and Y are total orders then Λ with $\mu_{ij}^{\mathbf{x}} = \breve{\mu}_{ij}^{\mathbf{x}} = 1$ gives Kendall's τ_a and Λ with $\mu_{ij}^{\mathbf{x}} = \mathbf{x}_i - \mathbf{x}_j$ and $\breve{\mu}_{ij}^{\mathbf{x}} = \mathbf{x}_j - \mathbf{x}_i$ leads to Spearman's ρ_a. However, if X and Y are preorders then the previous weighting schemes respectively give Kendall's τ_b and Spearman's ρ_b which were meant to better take tied ranks into account. As a consequence, what is remarkable with (2) is that it allows unifying in a unique framework, well-known association measures for both ER and OR. Moreover, this approach has the particularity to highlight the relational nature of the variables by explicitly using their RM.

Next, we explain the indeterminacy concept between two BR on which (2) is based. This concept is different from the statistical independence principle, and it can be better understood by looking at the numerators of (2) when it is equal to 0. In such a case we have :

$$\begin{gathered} \sum(\mu_{ij}^{\mathbf{x}}\mathbf{X}_{ij} - \widetilde{\mu}_{ij}^{\mathbf{x}}\widetilde{\mathbf{X}}_{ij})(\mu_{ij}^{\mathbf{y}}\mathbf{Y}_{ij} - \widetilde{\mu}_{ij}^{\mathbf{y}}\widetilde{\mathbf{Y}}_{ij}) = 0 \\ \Leftrightarrow \\ \underbrace{\sum \mu_{ij}^{\mathbf{x}}\mu_{ij}^{\mathbf{y}}\mathbf{X}_{ij}\mathbf{Y}_{ij}}_{11} + \underbrace{\sum \widetilde{\mu}_{ij}^{\mathbf{x}}\widetilde{\mu}_{ij}^{\mathbf{y}}\widetilde{\mathbf{X}}_{ij}\widetilde{\mathbf{Y}}_{ij}}_{00} = \underbrace{\sum \mu_{ij}^{\mathbf{x}}\widetilde{\mu}_{ij}^{\mathbf{y}}\mathbf{X}_{ij}\widetilde{\mathbf{Y}}_{ij}}_{10} + \underbrace{\sum \widetilde{\mu}_{ij}^{\mathbf{x}}\mu_{ij}^{\mathbf{y}}\widetilde{\mathbf{X}}_{ij}\mathbf{Y}_{ij}}_{01} \end{gathered} \tag{3}$$

In (3) the left hand side $11 + 00$ corresponds to the total weight of agreement or concordant pairs between the two BR whereas the right hand side $10 + 01$ represents the total weight of disagreement or discordant pairs. Typically the weighted indeterminacy situation is when both total weights are equal. Suppose the classical case where all μ and $\widetilde{\mu}$ are uniform, then indeterminacy means that there are as many agreement pairs as disagreement pairs. Note that the correlation between $\mathbf{x}$ and $\mathbf{y}$ is positive if the weighted agreement is greater

[1] Due to space restriction, we were not able to give all the details of these properties.

than the weighted disagreement and it is negative otherwise. Therefore, our first generalization of Kendall's Γ is based on the weighted indeterminacy concept which gives a clear interpretation of the correlation between two BR in terms of the difference between the weighted agreement and the weighted disagreement.

In the next paragraph, we propose to generalize (2) further in order to better treat heterogeneous BR.

3 Similarity of order t and generalization of Kendall's Γ

3.1 Weighted Symmetric RM as vectors in an Euclidean space

Our goal now, is to extend the previously introduced Λ measure in order to deal with heterogeneous BR. Note that this can be partly done using Λ by choosing adequate weighting schemes. For instance, we previously showed that (2) embeds Kendall's τ_b which better deals with tied ranks. However, we argue that we can also model such a heterogeneity from a geometrical perspective. To this end, we introduce the concept of weighted symmetric RM (WSRM) of a variable $\mathbf{x}$, denoted by X and defined as follows, $\forall i, j \in \mathbb{A}$:

$$\mathsf{X}_{ij} = \mu_{ij}^{\mathbf{x}} \mathbf{X}_{ij} - \widetilde{\mu}_{ij}^{\mathbf{x}} \widetilde{\mathbf{X}}_{ij} \tag{4}$$

Moreover, let us denote by $\mathbb{M}_n$, the vectorial space of real valued square matrix of size n. Given a variable $\mathbf{x}$ we can represent its underlying BR by the RM $\mathbf{X}$, $\widetilde{\mathbf{X}}$, and by the WSRM X associated with μ and $\widetilde{\mu}$. All of the latter matrices are elements of $\mathbb{M}_n$. Furthermore, let us equip $\mathbb{M}_n$ with the Frobenius inner product denoted by $\langle ., . \rangle$, and defined by, $\forall \mathsf{X}, \mathsf{Y} \in \mathbb{M}_n : \langle \mathsf{X}, \mathsf{Y} \rangle = \sum_{i,j} \mathsf{X}_{ij} \mathsf{Y}_{ij}$. Using WSRM in this geometrical framework, (2) can be written as below :

$$\Lambda(\mathbf{X}, \mathbf{Y}, \mu, \widetilde{\mu}) = \frac{\sum_{i,j} \mathsf{X}_{ij} \mathsf{Y}_{ij}}{\sqrt{\sum_{i,j} \mathsf{X}_{ij}^2} \sqrt{\sum_{i,j} \mathsf{Y}_{ij}^2}} = \frac{\langle \mathsf{X}, \mathsf{Y} \rangle}{\sqrt{\langle \mathsf{X}, \mathsf{X} \rangle} \sqrt{\langle \mathsf{Y}, \mathsf{Y} \rangle}} \tag{5}$$

In other terms Λ is the cosine of the angular measure between vectors X and Y. In this context, we argue that ER with different numbers of groups and/or with different distributions give WSRM with distinct norms. Vectors norms can even be more different depending on the weighting models μ and $\widetilde{\mu}$. The same kind of observation can be made with regard to heterogeneous OR. Indeed, two vectors X and Y representing two total orders, have the same norm but this is no longer true when OR are partial and/or with ties. Accordingly, we argue that in order to deal with the heterogeneity of BR when comparing them, one has to penalize the difference between the norms of associated WSRM. In the next paragraph, we introduce a method which takes this point into account.

3.2 Similarity of order t and application to WSRM vectors

We present a family of proximity measures called similarity of order t with $t > 0$ which was initially introduced in [12] and lately applied to kernels normalization

in [14]. We apply this approach within the framework we have set up previously and we show how our method defines a family of association measures for heterogeneous BR. To this end, we give below the formal definition of similarities of order t, denoted by S^t and which are applied to two BR X and Y that were derived from two variables $\mathbf{x}$ and $\mathbf{y}$:

$$S^t(\mathbf{X},\mathbf{Y},\mu,\widetilde{\mu}) = \frac{\sum_{i,j}(\mu^{\mathbf{x}}_{ij}\mathbf{X}_{ij} - \widetilde{\mu}^{\mathbf{x}}_{ij}\widetilde{\mathbf{X}}_{ij})(\mu^{\mathbf{y}}_{ij}\mathbf{Y}_{ij} - \widetilde{\mu}^{\mathbf{y}}_{ij}\widetilde{\mathbf{Y}}_{ij})}{\left[\frac{1}{2}\left(\left[\sum_{i,j}(\mu^{\mathbf{x}}_{ij}\mathbf{X}_{ij} - \widetilde{\mu}^{\mathbf{x}}_{ij}\widetilde{\mathbf{X}}_{ij})^2\right]^t + \left[\sum_{i,j}(\mu^{\mathbf{y}}_{ij}\mathbf{Y}_{ij} - \widetilde{\mu}^{\mathbf{y}}_{ij}\widetilde{\mathbf{Y}}_{ij})^2\right]^t\right)\right]^{1/t}}$$
$$= \frac{\langle \mathsf{X},\mathsf{Y}\rangle}{\mathcal{M}^t(\langle \mathsf{X},\mathsf{X}\rangle,\langle \mathsf{Y},\mathsf{Y}\rangle)} \quad (6)$$

where $\mathcal{M}^t(a,b) = [\frac{1}{2}(a^t+b^t)]^{1/t}$ is the generalized power mean of order t. In order to be rigorous regarding the notations, we use $\mathsf{S}^t(\mathsf{X},\mathsf{Y})$ to refer to the similarity of order t when WSRM are used as inputs such as in the second line of (6). On the contrary, we use $S^t(\mathbf{X},\mathbf{Y},\mu,\widetilde{\mu})$ to make explicit the two types of inputs we initially have *i.e.* the RM $(\mathbf{X},\mathbf{Y})$ and the weighting models $(\mu,\widetilde{\mu})$. However, the reader has to keep in mind that $S^t(\mathbf{X},\mathbf{Y},\mu,\widetilde{\mu})$ is similar to $\mathsf{S}^t(\mathsf{X},\mathsf{Y})$ where X was determined from $\mathbf{X},\mu,\widetilde{\mu}$, by using (4), beforehand.

In (6), we have introduced the use of generalized power means. Particular cases of $\mathcal{M}^t$ are the following ones : the limit when $t \to 0$ gives the geometric mean; $t = 1$ is the arithmetic mean; the limit when $t \to \infty$ gives the maximum. The reason we consider t positive only is given later on. Before, we observe that (6) generalizes (5) since the latter equation is the limit of the former one when $t \to 0$. Next, let us denote by $\theta(\mathsf{X},\mathsf{Y})$ and $\gamma(\mathsf{X},\mathsf{Y}) = \max(\sqrt{\langle \mathsf{X},\mathsf{X}\rangle},\sqrt{\langle \mathsf{Y},\mathsf{Y}\rangle})/\min(\sqrt{\langle \mathsf{X},\mathsf{X}\rangle},\sqrt{\langle \mathsf{Y},\mathsf{Y}\rangle})$ the angular measure and the norms ratio between X and Y. To simplify the formulas we use the shorthands θ and γ. Using these geometrical measures, (5) can be equivalently written as follows [12, 14] :

$$S^t(\mathbf{X},\mathbf{Y},\mu,\widetilde{\mu}) = \frac{\cos\theta}{\mathcal{M}^t(\gamma,\gamma^{-1})} \quad (7)$$

The semantic of S^t can be apprehended from (7) by considering the two geometric features we have introduced. Note that $\cos\theta \in [-1,1]$ whereas $\gamma \in [1,\infty[$. Therefore, the sign of S^t is given by the cosine measure. In [12, 14], it was shown that $\forall t > 0$: in the limit case $t \to 0$, γ has no effect on S^t; S^t is monotonically increasing w.r.t. $\cos\theta$; if $\cos\theta > 0$ then S^t is monotonically decreasing w.r.t. γ (converging towards 0); on the contrary if $\cos\theta < 0$ then S^t is monotonically increasing w.r.t. γ (also converging towards 0). Intuitively, the norms ratio γ aims at refining the cosine measure in the following sense : given two vectors, the greater the difference between their norms is, the less significant the cosine measure is as a similarity value and therefore the lower the amplitude of the similarity value should be. For instance, in the limit case $\gamma \to \infty$, whatever the value of $\cos\theta$, the proximity measure is close to 0. S^t enables refining the cosine index by penalizing it on the basis of the difference between the vectors norms. Thus, this feature allows S^t to deal with heterogeneous BR better than Λ.

In this framework, the real parameter $t > 0$ enables monitoring the strength of this penalization : as t grows the penalization is stronger and the similarity index of order t ranges from one limit case $t \to 0$ to the other one $t \to \infty$ which are respectively given by $S^{t\to 0}(\mathbf{X}, \mathbf{Y}, \mu, \widetilde{\mu}) = \cos\theta$ and $S^{t\to\infty}(\mathbf{X}, \mathbf{Y}, \mu, \widetilde{\mu}) = \cos\theta/\gamma$.

Moreover, S^t measures satisfy the following axioms in regard to similarity indices [12, 14]. We have, $\forall t > 0, \forall \mathsf{X}, \mathsf{Y} \in \mathbb{M}_n$: boundedness, $|\mathsf{S}^t(\mathsf{X}, \mathsf{Y})| \leq 1$; maximal self-similarity, $\mathsf{S}^t(\mathsf{X}, \mathsf{X}) = 1$; symmetry, $\mathsf{S}^t(\mathsf{X}, \mathsf{Y}) = \mathsf{S}^t(\mathsf{Y}, \mathsf{X})$; indiscernibility of identicals, $\mathbf{x} = \mathbf{y} \Rightarrow \mathsf{S}^t(\mathsf{X}, \mathsf{Y}) = 1$; identity of indiscernibles[2], $\mathsf{S}^t(\mathsf{X}, \mathsf{Y}) = 1 \Rightarrow \mathbf{x} = \mathbf{y}$; order relation on S^t, $\forall t \geq t' > 0 : |\mathsf{S}^t(\mathsf{X}, \mathsf{Y})| \leq |\mathsf{S}^{t'}(\mathsf{X}, \mathsf{Y})|$[3].

One other attractive property of S^t indices is given below :

Theorem 1. *If $t > 0$ then the square matrix of general term $\mathsf{S}^t(\mathsf{X}, \mathsf{Y})$ is positive semi-definite.*

This result can be proved by using Gershgorin circle theorem [14]. Besides, by using theorem 6 of [15], we can deduce the following corollary of theorem 1 :

Corollary 1. *If $t > 0$ then the square matrix of general term $\mathsf{D}^t(\mathsf{X}, \mathsf{Y}) = \sqrt{2^{-1}(1 - \mathsf{S}^t(\mathsf{X}, \mathsf{Y}))}$ is Euclidean. In other words, the rows X and columns Y of D^t are points of $\mathbb{M}_n$ that can be represented in an Euclidean space such that $\mathsf{D}^t(\mathsf{X}, \mathsf{Y})$ are their pairwise Euclidean distance values.*

Next, we present another geometrical interpretation of S^t which brings some other insights about this method. Let us denote the orthogonal projection of X on Y by $\mathsf{P}_\mathsf{Y}(\mathsf{X})\mathsf{Y}$ where $\mathsf{P}_\mathsf{Y}(\mathsf{X}) = \langle \mathsf{X}, \mathsf{Y}\rangle / \langle \mathsf{Y}, \mathsf{Y}\rangle$ is called the scalar projection. The orthogonal projection of X on Y is the best approximation of the former vector given the latter one according to the least square approach. In that case, the scalar projection is a real value that expresses an asymmetric proximity relationship between the two vectors. Similarly, we can use $\mathsf{P}_\mathsf{X}(\mathsf{Y}) = \langle \mathsf{X}, \mathsf{Y}\rangle / \langle \mathsf{X}, \mathsf{X}\rangle$ as another asymmetric measure when comparing X and Y. We can mix scalar projections in order to have a symmetric coefficient. In that perspective, we can relate scalar projections with S^t by the following formula [12, 14], $\forall t > 0$:

$$S^t(\mathbf{X}, \mathbf{Y}, \mu, \widetilde{\mu}) = \mathcal{M}^{-t}\left(\mathsf{P}_\mathsf{Y}(\mathsf{X}), \mathsf{P}_\mathsf{X}(\mathsf{Y})\right) \tag{8}$$

Regarding OR, it is worth mentioning that when $\mu_{ij} = \breve{\mu}_{ij} = 1$ for all (i, j), $\mathsf{P}_\mathsf{Y}(\mathsf{X})$ is Somers's asymmetric measure d and $S^1(\mathbf{X}, \mathbf{Y}, \mu, \widetilde{\mu})$ is Kim's d index.

Finally, we point out the particular weighting models given by, $\forall (i, j) \in \mathbb{A}^2$: $\mu^{\mathbf{x}}_{ij} = 1$ and $\widetilde{\mu}^{\mathbf{x}}_{ij} = 0$. In that case, the opposite relation in (4) vanishes, and the WSRM X reduces to the RM $\mathbf{X}$ in all other following equations. In this case, S^t given in (6) is related to some similarity indices for binary vectors. Accordingly, our framework also encompasses the generalization to BR of measures such as Ochiaï or Dice indices. Indeed, the two latter measures are given by the limit case $t \to 0$ and $t = 1$. There are plenty similarity indices for binary vectors and classifying them in order to better understand their differences is a challenge which is tackled in [16]. The framework we propose can thus partly help in this research line.

[2] Note that the limit case $t \to 0$ (the cosine index) does not satisfy this property.

[3] Showing that when t grows the amplitude of S^t decreases.

4 Conclusion and future work

We have generalized Kendall's Γ coefficient and proposed an unifying framework to define association or proximity measures for the comparison of heterogeneous ER and OR. Our method presents several attractive features that we have shown from a theoretical viewpoint. Due to space restriction, we were not able to illustrate these properties on practical examples. We will thus target applications of our method in future work. In that perspective, it is worth noticing that the metric properties of S^t allow using classical data analysis tools such as Multidimensional Scaling in order to represent heterogeneous BR in Euclidean spaces.

References

1. Marcotorchino, J., Michaud, P.: Optimisation en analyse ordinale des données. Masson (1979)
2. Michaud, P., Marcotorchino, J.: Modèles d'optimisation en analyse des données relationnelles. Mathématiques et Sciences Humaines **vol 17**(67) (1979) 7–38
3. Ah-Pine, J.: On aggregating binary relations using 0-1 integer linear programming. In: ISAIM. (2010)
4. Ah-Pine, J., Marcotorchino, J.F.: Overview of the relational analysis approach in data-mining and multi-criteria decision making. In Usmani, Z.U.H., ed.: Web Intelligence and Intelligent Agents, InTech (2010)
5. Kendall, M.G.: Rank correlation methods, 4th Edition. Griffin, Londres (1970)
6. Vegelius, J., Janson, S.: Criteria for symmetric measures of association for nominal data. Quality and Quantity **16** (1982) 243–250
7. Marcotorchino, J.F.: Utilisation des comparaisons par paires en statistique des contingences parties I, II et III. Etudes IBM F069 F071 F081 (1984-1985)
8. Ghashghaie, S.: Agrégation Relationnelle des données ordinales. PhD thesis, University of Pierre and Marie Curie (Paris 6) (1990)
9. Najah Idrissi, A.: Contribution à l'unification de critères d'association pour variables qualitatives. PhD thesis, Thèse de l'Université de Paris VI (2000)
10. Marcotorchino, J., El Ayoubi, F.: Paradigme logique des écritures relationnelles de quelques critères fondamentaux d'association. Rev. Stat. Appli. **39** (1991) 25–46
11. Youness, G., Saporta, G.: Some measures of agreement between close partitions. Student **51** (2004) 1–12
12. Ah-Pine, J.: Sur des aspects algébriques et combinatoires de l'Analyse Relationnelle. PhD thesis, University of Pierre and Marie Curie (Paris 6) (2007)
13. Ah-Pine, J., Marcotorchino, J.F.: Unifying some association criteria between partitions by using relational matrices. Com. in Stat. - Theo. and Meth. **39**(3) (2010)
14. Ah-Pine, J.: Normalized kernels as similarity indices. In: PAKDD (2). (2010) 362–373
15. Gower, J., Legendre, P.: Metric and euclidean properties of dissimilarity coefficients. Journal of classification **3** (1986) 5–48
16. Marcotorchino, J.F.: Essai de Typologie Structurelle des Indices de Similarité Vectoriels par Unification Relationnelle. In: RNTI A3 Apprentissage artificiel et fouille de données, RNTI, Cépaduès Editions (2009) 328