Unifying Some Association Criteria Between Partitions by Using Relational Matrices

Julien Ah-Pine, Jean-François Marcotorchino

Unifying some association criteria between partitions by using relational matrices

Julien Ah-Pine and Jean-François Marcotorchino

CeNTAI (Center for Information Analysis Technologies)
Thales Land and Joint Systems
160, Boulevard de Valmy - BP 82
92704 Colombes Cedex France
(e-mail: julien.ah-pine@fr.thalesgroup.com)
(e-mail: jeanfrancois.marcotorchino@fr.thalesgroup.com)

Abstract. Association criteria used for measuring the relationship between categorical variables or partitions, are mainly applied and studied using contingency tables. There is another way for representing categorical variables : the Relational Analysis representation which uses binary pairwise comparison matrices and which has many properties. There exist correspondence formulas that allow to get from the contingency representation to the relational representation. By using these formulas, we show in this paper how Relational Analysis allows to unify many association criteria such as Rand, Tchuprow, Belson criteria and others. This unified framework allows also to have a better understanding of the main differences between those association criteria. In that context, we also present different kinds of independence : statistical, geometrical and "logical".
Keywords: Relational Analysis, Association criteria, Independence, Nominal categorical variables, Partitions comparison.

1 Introduction

Relational Analysis[1] (RA in the rest of the paper) is concerned with the analysis of binary relations[2] and it has many applications in different mathematical fields [MM79], [MM80], [MM81]. Particularly, this approach represents binary relations as pairwise comparison matrices (also called relational matrices), and it is basically related to different concepts from graph theory, statistics and linear programming. The most usual application domains of RA are data analysis and multicriteria decision making which are respectively based upon the aggregation of equivalence and order relations.

In this paper, we are particularly interested in the applications of RA in the analysis of association criteria between nominal categorical variables (or partitions). There are already many results in that context see for example [Mar84a], [Mar84b], [Mar85], [Mar86a], [Mes89], [NI00], [YS04]. This paper

[1] see for example [MAP] for a recent overview of this theory
[2] their aggregation, their association measures

Julien Ah-Pine is now at :
Xerox Research Centre Europe
6, Chemin de Maupertuis
38240 Meylan, France
(e-mail: julien.ah-pine@xrce.xerox.com)

recalls some of these results and it also presents some other extensions [AP07].

There are, in the literature, numerous works that have studied different methods for measuring the relationship between partitions see for example [Zah64], [Ran71], [FM83], [HA85], [JV92], [Mir96]. The RA approach in this context aims at having a better understanding of the main differences between several of these methods.

In that perspective, the different goals of the present paper are the following ones :

- We want to show different properties that the relational representation allows compared to the contingency representation. Indeed, we can formally unify 7 different association criteria found in the literature by showing that they can be deduced from a general Bravais-Pearson like correlation coefficient. In their relational representation, we show that the main differences between these association criteria, can be expressed using three parameters.
- In their contingency representation, many association criteria are based upon the statistical independence deviation concept whereas in their relational representation, they are mainly related to the geometrical independence deviation concept. We want to underline also, another kind of independence called the "indetermination" situation which has a "logical" aspect as it is also encountered in voting theory.

This paper is organized as follows :

- In section 2, we recall basics of the contingency and relational representations, the different association criteria we are going to analyze and the correspondence formulas that allow us to express those association criteria from their contingency representation to their relational representation.
- In section 3, we give a general formula which is similar to the Bravais-Pearson correlation coefficient. This formula has three parameters : a transformation function of the relational matrices, and two central trends. We show that many association criteria can be deduced from this general equation when using particular parameters.
- When using the contingency representation or the relational representation, there is a duality between the statistical independence and the geometrical independence concepts. In section 4, we try to strengthen this property by showing a particular relationship between Belson's criterion and Janson-Vegelius' criterion. Moreover, we give some comparisons to Mirkin's classification of association criteria.
- In section 5, we recall the "indetermination" situation concept and its "logical" foundations. Then we suggest some extensions by defining a parametric normalized coefficient based on this concept.

2 From contingency representation to relational representation

We assume that we have N objects, $\{O^i; i = 1, \ldots, N\}$. For these objects, let V^k and V^l be two nominal categorical variables with respectively p_k and p_l classes. The sets of classes will be respectively denoted by $\{D^k_u; u = 1, \ldots, p_k\}$ and $\{D^l_v; v = 1, \ldots, p_l\}$. One can represent each of these variables by binary assignment matrices. For example, in the case of V^k, we have the following $(N \times p_k)$ matrix :

$$K^k_{iu} = \begin{cases} 1 \text{ if } O^i \text{ belongs to the class } D^k_u \\ 0 \text{ else} \end{cases}$$

From K^k and K^l, we can deduce the following contingency table denoted by $\mathbf{n}^{kl}$ with dimensions $(p_k \times p_l)$:

$$\mathbf{n}^{kl} = {}^tK^k \cdot K^l$$

where ${}^tK^k$ is the transpose matrix associated to K^k and $\cdot$ the matrix multiplication.

We have the following notations and interpretations, $\forall u = 1, \ldots, p_k$ and $\forall v = 1, \ldots, p_l$:

- $\mathbf{n}^{kl}_{uv}$ =Number of objects belonging both to the class D^k_u of V^k and D^l_v of V^l,
- $\sum_{v=1}^{p_l} \mathbf{n}^{kl}_{uv} = \mathbf{n}^{kl}_{u.}$ =Number of objects belonging to the class D^k_u of V^k,
- $\sum_{u=1}^{p_k} \mathbf{n}^{kl}_{uv} = \mathbf{n}^{kl}_{.v}$ =Number of objects belonging to the class D^l_v of V^l,
- $\sum_{u=1}^{p_k} \sum_{v=1}^{p_l} \mathbf{n}^{kl}_{uv} = N$ = Total number of objects.

We will study the following association criteria : Belson (B) [Bel59], Lerman (L) [Ler81], χ^2 of Tchuprow (T), Jordan (J) [Jor27], Rand (R) [Ran71], and Janson-Vegelius (JV) [JV92]. We first recall the definitions of these criteria in their contingency representation. We precise that the Rand criterion and the Lerman criterion that we mention, are modified versions according to [Mar84b] and [NI00].

$$B(V^k, V^l) = \sum_{u=1}^{p_k} \sum_{v=1}^{p_l} \left(\mathbf{n}^{kl}_{uv} - \frac{\mathbf{n}^{kl}_{u.} \mathbf{n}^{kl}_{.v}}{N} \right)^2 \tag{1}$$

$$L(V^k, V^l) = \frac{\sum_{u,v} (\mathbf{n}^{kl}_{uv})^2 - \frac{\sum_u (\mathbf{n}^{kl}_{u.})^2 \sum_v (\mathbf{n}^{kl}_{.v})^2}{N^2}}{\sqrt{\left(\sum_u (\mathbf{n}^{kl}_{u.})^2 \left(1 - \sum_u \frac{(\mathbf{n}^{kl}_{u.})^2}{N^2} \right) \right) \left(\sum_v (\mathbf{n}^{kl}_{.v})^2 \left(1 - \sum_v \frac{(\mathbf{n}^{kl}_{.v})^2}{N^2} \right) \right)}} \tag{2}$$

$$T(V^k, V^l) = \frac{\sum_{u,v} \frac{1}{\mathbf{n}^{kl}_{u.}\mathbf{n}^{kl}_{.v}} \left(\mathbf{n}^{kl}_{uv} - \frac{\mathbf{n}^{kl}_{u.}\mathbf{n}^{kl}_{.v}}{N}\right)^2}{\sqrt{(p_k - 1)(p_l - 1)}} \tag{3}$$

$$J(V^k, V^l) = \frac{1}{N} \sum_{u,v} \left(\mathbf{n}^{kl}_{uv} \left(\mathbf{n}^{kl}_{uv} - \frac{\mathbf{n}^{kl}_{u.}\mathbf{n}^{kl}_{.v}}{N}\right)\right) \tag{4}$$

$$R(V^k, V^l) = \frac{2\sum_{u,v}(\mathbf{n}^{kl}_{uv})^2 - \sum_u(\mathbf{n}^{kl}_{u.})^2 - \sum_v(\mathbf{n}^{kl}_{.v})^2 + N^2}{N^2} \tag{5}$$

$$JV(V^k, V^l) = \frac{p_k p_l \sum_{u,v}(\mathbf{n}^{kl}_{uv})^2 - p_k \sum_u(\mathbf{n}^{kl}_{u.})^2 - p_l \sum_v(\mathbf{n}^{kl}_{.v})^2 + N^2}{\sqrt{(p_k(p_k-2)\sum_u(\mathbf{n}^{kl})^2_{u.} + N^2)\,(p_l(p_l-2)\sum_u(\mathbf{n}^{kl}_{.v})^2 + N^2)}} \tag{6}$$

We also have the following expression for the Janson-Vegelius criterion :

$$JV(V^k, V^l) = \frac{p_k p_l \sum_{u,v} \left(\mathbf{n}^{kl}_{uv} - \left[\frac{\mathbf{n}^{kl}_{u.}}{p_l} + \frac{\mathbf{n}^{kl}_{.v}}{p_k} - \frac{N}{p_k p_l}\right]\right)^2}{\sqrt{(p_k(p_k-2)\sum_u \mathbf{n}^2_{u.} + N^2)\,(p_l(p_l-2)\sum_u \mathbf{n}^2_{.v} + N^2)}} \tag{7}$$

RA is another way for representing nominal categorical variables. This representation uses pairwise comparison matrices also called relational matrices.

Let C^k and C^l be the relational matrices of dimension $(N \times N)$, representing the variables V^k and V^l. We can obtain C^k and C^l by using the assignment matrices K^k and K^l :

$$C^k = K^k \cdot {}^tK^k \quad \text{and} \quad C^l = K^l \cdot {}^tK^l$$

In general terms, let $\mathcal{R}$ be a binary relation among a set of N objects $O^1, \ldots, O^N$. If C is the relational matrix for $\mathcal{R}$ then we have, $\forall i, i' = 1, \ldots, N$:

$$C_{ii'} = \begin{cases} 1 \text{ if } O^i \text{ is in relation with } O^{i'} \\ 0 \text{ else} \end{cases}$$

In our context, we deal with equivalence relations or partitions. As a result we have, for instance for V^k :

$$C^k_{ii'} = \begin{cases} 1 \text{ if } O^i \text{ and } O^{i'} \text{ belong to the same class according to } V^k \\ 0 \text{ else} \end{cases}$$

Furthermore, the relational matrix C^k represents an equivalence relation. Thus, it respects the following relational properties which can be expressed as linear equations :

- reflexivity : $C_{ii}^k = 1 \quad \forall i = 1, \ldots, N$,
- symmetry : $C_{ii'}^k - C_{i'i}^k = 0 \quad \forall i, i' = 1, \ldots, N$,
- transitivity : $C_{ii'}^k - C_{i'i''}^k + C_{ii''}^k \leq 1 \quad \forall i, i', i'' = 1, \ldots, N$.

There are different correspondence formulas that exist between the contingency representation and the relational one. We give in Table 1 some of the most useful ones.

Contingency representation	$\leftrightarrow$	Relational representation
$\sum_{u=1}^{p_k} \sum_{v=1}^{p_l} (\mathbf{n}_{uv}^{kl})^2$	$=$	$\sum_{i=1}^{N} \sum_{i'=1}^{N} C_{ii'}^k C_{ii'}^l$
$\sum_u (\mathbf{n}_{u.}^{kl})^2$	$=$	$\sum_{i,i'} C_{ii'}^k$
$\sum_v (\mathbf{n}_{.v}^{kl})^2$	$=$	$\sum_{i,i'} C_{ii'}^l$
$\sum_{u,v} \frac{(\mathbf{n}_{uv}^{kl})^2}{\mathbf{n}_{u.}^{kl} \mathbf{n}_{.v}^{kl}}$	$=$	$\sum_{i,i'} \frac{C_{ii'}^k C_{ii'}^l}{C_{i.}^k C_{i.}^l}$
$\sum_{u,v} \mathbf{n}_{uv}^{kl} \mathbf{n}_{u.}^{kl} \mathbf{n}_{.v}^{kl}$	$=$	$\sum_{i,i'} \frac{C_{i.}^k + C_{.i'}^k}{2} C_{ii'}^l$
$\sum_{u,v} (\mathbf{n}_{uv}^{kl})^2 \mathbf{n}_{u.}^{kl}$	$=$	$\sum_{i,i'} \frac{C_{i.}^k + C_{.i'}^k}{2} C_{ii'}^k C_{ii'}^l$
$\sum_{u,v} \frac{(\mathbf{n}_{uv}^{kl})^2}{\mathbf{n}_{u.}^{kl}}$	$=$	$\sum_{i,i'} \frac{C_{ii'}^k}{C_{i.}^k} C_{ii'}^l$
$\sum_v \left(\sum_u \mathbf{n}_{u.}^{kl} \mathbf{n}_{uv}^{kl} \right)^2$	$=$	$\sum_{i,i'} C_{i.}^k C_{.i'}^k C_{ii'}^l$
$\sum_{u,v} (\mathbf{n}_{u.}^{kl})^2 (\mathbf{n}_{.v}^{kl})^2$	$=$	$\sum_{i,i'}^{N} C_{..}^k C_{ii'}^l$
where $\mathbf{n}_{u.}^{kl} = \sum_v \mathbf{n}_{uv}^{kl}$ and $C_{i.}^k = \sum_{i'} C_{ii'}^k$		

Table 1. Correspondence formulas between contingency representation and relational representation

In [NI00], the author uses these formulas, in order to obtain the symmetric relational expression of Rand (modified version), Janson-Vegelius, Lerman, and Tchuprow criteria. We extend these results by giving in addition, the symmetric expression of Belson and Jordan criteria. The related definitions are given below.

$$B(C^k, C^l) = \sum_{i=1}^{N}\sum_{i'=1}^{N}\left(C^k_{ii'} - \frac{C^k_{i.} + C^k_{.i'}}{N} + \frac{C^k_{..}}{N^2}\right)\left(C^l_{ii'} - \frac{C^l_{i.} + C^l_{.i'}}{N} + \frac{C^l_{..}}{N^2}\right) \quad (8)$$

$$L(C^k, C^l) = \frac{\sum_{i,i'}\left(C^k_{ii'} - \sum_{i,i'}\frac{C^k_{ii'}}{N^2}\right)\left(C^l_{ii'} - \sum_{i,i'}\frac{C^l_{ii'}}{N^2}\right)}{\sqrt{\sum_{i,i'}\left(C^k_{ii'} - \sum_{i,i'}\frac{C^k_{ii'}}{N^2}\right)^2 \sum_{i,i'}\left(C^l_{ii'} - \sum_{i,i'}\frac{C^l_{ii'}}{N^2}\right)^2}} \quad (9)$$

$$T(C^k, C^l) = \frac{\sum_{i,i'}\left(\frac{C^k_{ii'}}{C^k_{i.}} - \frac{1}{N}\right)\left(\frac{C^l_{ii'}}{C^l_{i.}} - \frac{1}{N}\right)}{\sqrt{\sum_{i,i'}\left(\frac{C^k_{ii'}}{C^k_{i.}} - \frac{1}{N}\right)^2 \sum_{i,i'}\left(\frac{C^l_{ii'}}{C^l_{i.}} - \frac{1}{N}\right)^2}} \quad (10)$$

$$J(C^k, C^l) = \frac{1}{N}\sum_{i,i'}\left(C^k_{ii'} - \frac{C^k_{i.}}{N}\right)\left(C^l_{ii'} - \frac{C^l_{i.}}{N}\right) \quad (11)$$

$$R(C^k, C^l) = \frac{1}{N^2}\sum_{i,i'}\left(C^k_{ii'}C^l_{ii'} + \overline{C}^k_{ii'}\overline{C}^l_{ii'}\right) \quad (12)$$

$$JV(C^k, C^l) = \frac{\sum_{i,i'}\left(C^k_{ii'} - \frac{1}{p_k}\right)\left(C^l_{ii'} - \frac{1}{p_l}\right)}{\sqrt{\sum_{i,i'}\left(C^k_{ii'} - \frac{1}{p_k}\right)^2 \sum_{i,i'}\left(C^l_{ii'} - \frac{1}{p_l}\right)^2}} \quad (13)$$

where $\overline{C}^k_{ii'} = 1 - C^k_{ii'}$, $\overline{C}^k = U_N - C^k$ and U_N is the $(N \times N)$ square matrix where all terms equal 1.

We also have the following equation for the modified Rand criterion that will be called the symmetric modified Rand criterion, $R'(C^k, C^l)$:

$$\begin{aligned} R'(C^k, C^l) &= 2R(C^k, C^l) - 1 \\ &= \frac{\sum_{i,i'}\left(C^k_{ii'} - \frac{1}{2}\right)\left(C^l_{ii'} - \frac{1}{2}\right)}{\sqrt{\sum_{i,i'}\left(C^k_{ii'} - \frac{1}{2}\right)^2 \sum_{i,i'}\left(C^l_{ii'} - \frac{1}{2}\right)^2}} \end{aligned} \quad (14)$$

The relational representation of the studied association criteria can easily be proved by using the correspondence formulas given in Table 1. Besides, the different proofs for these correspondence formulas are given in [Mar84b].

3 A unifying relational approach for many association criteria

Using the relational representation, one can express the different association criteria that we have recalled, as particular cases of a general criterion similar to a Bravais-Pearson correlation coefficient[3] between transformed relational matrices [NI00], [AP07] :

$$\Delta(C^k, C^l, f, \mu^k, \mu^l) = \frac{\sum_{i,i'} \big(f(C^k_{ii'}) - \mu^k\big)\big(f(C^l_{ii'}) - \mu^l\big)}{\sqrt{\sum_{i,i'} \big(f(C^k_{ii'}) - \mu^k\big)^2 \sum_{i,i'} \big(f(C^l_{ii'}) - \mu^l\big)^2}} \qquad (15)$$

According to (15), we can see that the main differences between the studied criteria are based upon :

- the transformation function f applied to the terms of the relational matrices,
- the central trends μ^k and μ^l, which are given parameters.

We give in Table 2, different values to the parameters (f, μ^k, μ^l), which define particular coefficients. The latter are related to the association criteria recalled previously. Therefore, we also give the relationship type that exists between the defined coefficient and its corresponding association criterion.

In this table, we mention the Torgerson transformation [Tor52]. We recall that it is an operation used in multidimensional scaling which, given a scalar products matrix between objects, gives as output the scalar products matrix between centered objects :

$$f(C^k_{ii'}) = C^k_{ii'} - \frac{C^k_{i.} + C^k_{.i'}}{N} + \frac{C^k_{..}}{N^2} = \langle O^i_k - G_k, O^{i'}_k - G_k \rangle$$

$\{O^i_k; i = 1, \ldots, N\}$ are $(p_k \times 1)$ binary vectors where $[O^i_k]_u = K^k_{iu}; u = 1, \ldots, p_k$ and $G_k = \frac{1}{N} \sum_{i=1}^{N} O^i_k$.

The RA approach allows to have a better understanding of the main differences between the studied association criteria. Moreover, in their relational

[3] similar coefficients for measuring the relationship between square matrices were considered in [Esc73] and in [HL75]. These coefficients were respectively defined as RV coefficients and Γ statistics

$f(C_{ii'})$	μ^k	μ^l	Related criteria	Relation with the related criteria
$f(C_{ii'}) = C_{ii'} - \frac{C_{i.}+C_{.i'}}{N} + \frac{C_{..}}{N^2}$ (Torgerson transf.)	0	0	Belson	Normalized Belson
$f(C_{ii'}) = C_{ii'}$	$C^k_{..}/N^2$	$C^l_{..}/N^2$	Lerman	Lerman
$f(C_{ii'}) = C_{ii'}/C_{i.}$	$1/N$	$1/N$	Tchuprow	Tchuprow
$f(C_{ii'}) = C_{ii'}$	$C^k_{i.}/N$	$C^l_{i.}/N$	Jordan	Normalized Jordan
$f(C_{ii'}) = C_{ii'}$	$1/2$	$1/2$	Modified Rand	Symmetric modified Rand
$f(C_{ii'}) = C_{ii'}$	$1/p_k$	$1/p_l$	Janson-Vegelius	Janson-Vegelius

Table 2. Correspondence between correlation coefficient and association criteria

expressions all these association criteria are related to the geometrical independence deviation concept. Indeed, we can also express the general equation (15) using the Frobenius scalar product between two relational matrices :

$$\langle C^k, C^l \rangle_F = \sum_{i=1}^{N} \sum_{i'=1}^{N} C^k_{ii'} C^l_{ii'} = \text{Trace}({}^tC^k \cdot C^l)$$

Then, we can see that $\Delta(C^k, C^l) = 0 \Leftrightarrow \langle f(C^k) - \mu^k U_N, f(C^l) - \mu^l U_N \rangle_F = 0$

In [Mir96], Mirkin gives four different classes for classifying association criteria :

- structural association approach,
- contingency modeling approach,
- geometrical approach,

- cross-classificational approach.

Regarding this classification, our results would give more emphasis to the contingency modeling approach and to the geometrical approach. The first one is related to the contingency representation and to the statistical independence deviation concept whereas the second one is related to the relational representation and to the geometrical independence deviation concept. We will study in more details those two different approaches in the following section.

Moreover, our approach underlines the geometrical approach more than the definition given in [Mir96] does. Indeed, in [Mir96], only the Tchuprow criterion is considered as belonging to the geometrical approach[4]. Using the relational representation, we can show that the equivalence match criterion given in [Mir96] is exactly Marcotorchino's version of Rand criterion [Mar84a]. Regarding equation (15) and Table 2, this last criterion is related to the parameters ($f = Id, \mu^k = 1/2, \mu^l = 1/2$) which leads to the symmetric modified Rand criterion with range values $[-1, 1]$. In a more general perspective, we have shown that many other association criteria which are not considered[5] in [Mir96] can also be expressed using the geometrical correlation coefficient given in (15).

Finally, unlike the contingency representation, the relational representation allows to express the differences between the studied association criteria using three clearly defined parameters.

4 Statistical and geometrical independence concepts

In the contingency representation, we say that two nominal categorical variables V^k and V^l are statistically independent if their joint probabilities, $\frac{\mathbf{n}_{uv}^{kl}}{N}$, equal the product of their marginal probabilities, $\frac{\mathbf{n}_{u.}^{kl}}{N}\frac{\mathbf{n}_{.v}^{kl}}{N}$:

$$V^k \perp_S V^l \Leftrightarrow \frac{\mathbf{n}_{uv}^{kl}}{N} = \frac{\mathbf{n}_{u.}^{kl}}{N}\frac{\mathbf{n}_{.v}^{kl}}{N} \quad \forall (D_u^k, D_v^l) : D_u^k \in V^k, D_v^l \in V^l \qquad (16)$$

Many of the studied association criteria are based upon this concept when using contingency representation. Indeed, we clearly see, that Belson, Lerman, Tchuprow and Jordan criteria are null if V^k and V^l are statistically independent.

On the contrary, in the relational representation, all studied association criteria are related to the geometrical independence deviation concept[6]. We

[4] see also [Mir01]

[5] Belson, Lerman, Jordan, Janson and Vegelius criteria

[6] notice that the (non symmetric) Rand criterion, $R(C^k, C^l)$, equals 1/2 in case of geometrical independence

say that two nominal categorical variables are geometrically independent if we have the following relation :

$$V^k \perp_G V^l \Leftrightarrow \Delta(C^k, C^l, f, \mu^k, \mu^l) = 0 \Leftrightarrow \sum_{i,i'} \left(f(C^k_{ii'}) - \mu^k\right)\left(f(C^l_{ii'}) - \mu^l\right) = 0 \quad (17)$$

Regarding statistical / geometrical independence, we now focus on a particular relationship between two of the recalled association criteria : the Belson criterion and the Janson-Vegelius criterion's numerator. This relationship can be stated as follows :

- The Belson criterion, in its contingency representation, is based on the statistical independence whereas in its relational representation, it is based on a geometrical independence associated to the Torgerson transformation.
- On the contrary, the Janson-Vegelius criterion's numerator, in its contingency representation, is based on the geometrical independence associated to the Torgerson transformation whereas, in its relational representation, it is related to the statistical independence with an equiprobability assumption.

We summarize the Belson / Janson-Vegelius criteria relationship in Table 3.

	Deviation from statistical independence	Geometrical independence based on Torgerson transformation
Belson	$\sum_{u,v}\left(\mathbf{n}^{kl}_{uv} - \frac{\mathbf{n}^{kl}_{u.}\mathbf{n}^{kl}_{.v}}{N}\right)^2$	$\sum_{i,i'}\left(C^k_{ii'} - \left[\frac{C^k_{i.}}{N} + \frac{C^k_{.i'}}{N} - \frac{C^k_{..}}{N^2}\right]\right)\left(C^l_{ii'} - \left[\frac{C^l_{i.}}{N} + \frac{C^l_{.i'}}{N} - \frac{C^l_{..}}{N^2}\right]\right)$
Janson-Vegelius (numerator)	$\sum_{i,i'}\left(C^k_{ii'} - \frac{1}{p_k}\right)\left(C^l_{ii'} - \frac{1}{p_l}\right)$	$\sum_{u,v}\left(\mathbf{n}^{kl}_{uv} - \left[\frac{\mathbf{n}^{kl}_{u.}}{p_l} + \frac{\mathbf{n}^{kl}_{.v}}{p_k} - \frac{\mathbf{n}^{kl}_{..}}{p_k p_l}\right]\right)^2$

Table 3. Dual relationship between Belson and Janson-Vegelius criteria due to the contingency / relational representation duality

We give below more details about our interpretations of the independence concepts underlying the relational and the contingency representations of Janson-Vegelius criterion's numerator, which show the duality property between the latter and Belson criterion.

First, if we consider the Janson-Vegelius criterion's relational representation, we can interpret the term $(C_{ii'}^k - \frac{1}{p_k})$, as a deviation from the statistical independence situation in an equiprobability context. Let $P(C_{ii'}^k)$ be, symbolically, the probability for two objects O^i and $O^{i'}$, belonging to the same class of V^k. Let us assume moreover, that the different classes $D_u^k; u = 1, \ldots, p_k;$ are equiprobable. This implies that the probability for an object to belong to any class of V^k equals $1/p_k$. Then, in case of probability independence, we have :

$$\begin{aligned} P(C_{ii'}^k) &= \textstyle\sum_{u=1}^{p_k} P(\text{“}O^i \text{ and } O^{i'} \text{ belong to the class } D_u^k\text{”}) \\ &= \textstyle\sum_{u=1}^{p_k} P(\text{“}O^i \text{ belongs to the class } D_u^k\text{”}) P(\text{“}O^{i'} \text{ belongs to the class } D_u^k\text{”}) \\ &= \textstyle\sum_{u=1}^{p_k} \frac{1}{p_k}\frac{1}{p_k} \\ &= \frac{1}{p_k} \end{aligned}$$

Second, if we consider the Janson-Vegelius criterion's contingency representation, we can interpret the term $(\mathbf{n}_{uv}^{kl} - \left[\frac{\mathbf{n}_{u.}^{kl}}{p_l} + \frac{\mathbf{n}_{.v}^{kl}}{p_k} - \frac{\mathbf{n}_{..}^{kl}}{p_k p_l}\right])$, as the Torgerson transformation of the $(N \times 1)$ binary vectors $\{D_u^k; u = 1, \ldots, p_k\}$ and $\{D_v^l; v = 1, \ldots, p_l\}$ where $[D_u^k]_i = K_{iu}^k; i = 1, \ldots, N$. Indeed, we have $\mathbf{n}_{uv}^{kl} = \langle D_u^k, D_v^l \rangle$ and the following relation :

$$\mathbf{n}_{uv}^{kl} - \frac{\mathbf{n}_{u.}^{kl}}{p_l} - \frac{\mathbf{n}_{.v}^{kl}}{p_k} + \frac{\mathbf{n}_{..}^{kl}}{p_k p_l} = \langle D_u^k - G^k, D_v^l - G^l \rangle \quad \forall (D_u^k, D_v^l) \in V^k \times V^l$$

where $G^k = \frac{1}{p_k} \sum_{D_u^k \in V^k} D_u^k$ and $G^l = \frac{1}{p_l} \sum_{D_v^l \in V^l} D_v^l$.

5 A “logical” independence concept : the “indetermination” situation

Despite the general geometrical independence deviation concept underlying the relational representations of the studied association criteria, we can distinguish another kind of independence which has a “logical” aspect. This concept was defined in [Mar84a] and was called the situation of “indetermination”. It is related to Condorcet criterion defined in voting theory [Con85]. In that context, we have an “indetermination” situation when the number of voters in favor of a candidate equals the number of voters against this candidate. This criterion which measures the relationship between order relations was extended to the case of equivalence relations [MM79]. This extension allows to link the Condorcet criterion to the modified Rand criterion as we

formally have :

$$R(C^k, C^l) = \frac{\text{Condorcet}(C^k, C^l)}{N^2} \tag{18}$$

In order to introduce the "indetermination" situation concept we first introduce the four-fold / tetrachoric[7] table given in its relational representation. Indeed in that case, the nominal categorical variables through their relational matrices C^k and C^l, can be interpreted as 0/1 variables. As a result, we can consider the (2×2) table given in Table 4.

	C^l	$\overline{C}^l$	Margin
C^k	$11_{kl} = \sum_{i,i'} C^k_{ii'} C^l_{ii'}$	$10_{kl} = \sum_{i,i'} C^k_{ii'} \overline{C}^l_{ii'}$	$\sum_{i,i'} C^k_{ii'}$
$\overline{C}^k$	$01_{kl} = \sum_{i,i'} \overline{C}^k_{ii'} C^l_{ii'}$	$00_{kl} = \sum_{i,i'} \overline{C}^k_{ii'} \overline{C}^l_{ii'}$	$\sum_{i,i'} \overline{C}^k_{ii'}$
Margin	$\sum_{i,i'} C^l_{ii'}$	$\sum_{i,i'} \overline{C}^l_{ii'}$	N^2

Table 4. Agreements and disagreements between relational matrices

We say that two nominal categorical variables are in an "indetermination" situation if we have the following relation :

$$\begin{aligned} V^k \perp_L V^l &\Leftrightarrow 11_{kl} + 00_{kl} - 10_{kl} - 01_{kl} = 0 \\ &\Leftrightarrow \sum_{i,i'} C^k_{ii'} C^l_{ii'} + \sum_{i,i'} \overline{C}^k_{ii'} \overline{C}^l_{ii'} - \sum_{i,i'} C^k_{ii'} \overline{C}^l_{ii'} - \sum_{i,i'} \overline{C}^k_{ii'} C^l_{ii'} = 0 \\ &\Leftrightarrow \sum_{i,i'} \left(C^k_{ii'} - \overline{C}^k_{ii'} \right) \left(C^l_{ii'} - \overline{C}^l_{ii'} \right) = 0 \\ &\Leftrightarrow 4 \sum_{i,i'} \left(C^k_{ii'} - 1/2 \right) \left(C^l_{ii'} - 1/2 \right) = 0 \end{aligned} \tag{19}$$

[7] this table is equivalent to the four-fold / tetrachoric table given in [Mir96] but the latter only uses the contingency representation. However, when using the relational representation the following notations 11_{kl}, 01_{kl}, 10_{kl} and 00_{kl} are straightforward unlike for the contingency representation. Finally this allows us to naturally apply other coefficients such as the odds-ratio as we suggest in (20) or other similarity indexes between 0/1 vectors

Using the notations given in Table 4, we can define the number of agreements between two categorical variables as $11_{kl} + 00_{kl}$ and the number of disagreements as $10_{kl} + 01_{kl}$. These last quantities are exactly the equivalence match and equivalence mismatch criteria given in [Mir96]. Finally, we have an "indetermination" situation when the number of agreements equals the number of disagreements. This situation means that the variables are neither "concordant" nor "discordant".

Using Table 4, we can formally characterize the "indetermination" situation concept compared to the statistical independence deviation concept. Indeed, when considering 0/1 variables, we can use the odds-ratio measure in order to determine if two variables are statistically independent or not :

$$\begin{aligned} V^k \perp_S V^l &\Leftrightarrow OR(C^k, C^l) = 11_{kl}00_{kl}/10_{kl}01_{kl} = 1 \\ &\Leftrightarrow 11_{kl}00_{kl} - 10_{kl}01_{kl} = 0 \end{aligned} \tag{20}$$

Using correspondence formulas, we can easily show that the modified Lerman criterion, $L(C^k, C^l)$, is null if and only if the odds-ratio measure, $OR(C^k, C^l)$, equals one.

Comparing equation (19) to (20), we can see that the "indetermination" situation concept is related to an additive model, $11_{kl} + 00_{kl} - 10_{kl} - 01_{kl}$, whereas the statistical independence deviation concept is related to a multiplicative model, $11_{kl}00_{kl} - 10_{kl}01_{kl}$.

Similarly to equation (15), we propose to extend the "indetermination" situation concept to a more general parametric family by assigning different weights to agreements and disagreements quantities :

$$V^k \perp_L V^l \Leftrightarrow \mu_1^k\mu_1^l 11_{kl} + \mu_0^k\mu_0^k 00_{kl} - \mu_1^k\mu_0^l 10_{kl} - \mu_0^k\mu_1^l 01_{kl} = 0 \tag{21}$$

Then, we introduce the following general formula which defines a normalized coefficient, Λ, that measures the deviation from the weighted "indetermination" situation between two nominal categorical variables :

$$\Lambda(C^k, C^l, \mu_1^k, \mu_0^k, \mu_1^l, \mu_0^l) = \frac{\sum_{i,i'} \left(\mu_1^k C_{ii'}^k - \mu_0^k \overline{C}_{ii'}^k\right)\left(\mu_1^l C_{ii'}^l - \mu_0^l \overline{C}_{ii'}^l\right)}{\sqrt{\sum_{i,i'} \left(\mu_1^k C_{ii'}^k - \mu_0^k \overline{C}_{ii'}^k\right)^2 \sum_{i,i'} \left(\mu_1^l C_{ii'}^l - \mu_0^l \overline{C}_{ii'}^l\right)^2}} \tag{22}$$

We give below the formal relationship between the geometrical independence and the "indetermination" situation in the relational representation :

$$\Lambda(C^k, C^l, \mu_1^k, \mu_0^k, \mu_1^l, \mu_0^l) = \Delta\left(C^k, C^l, Id, \frac{\mu_0^k}{\mu_1^k + \mu_0^k}, \frac{\mu_0^l}{\mu_1^l + \mu_0^l}\right) \tag{23}$$

This result shows that the symmetric modified Rand criterion[8] is null when we have an "indetermination" situation associated to uniform weights : $\mu_1^k = \mu_0^k = \mu_1^l = \mu_0^l$.

Another example is the Janson-Vegelius criterion which is null when we have an "indetermination" situation associated to the following parameters : $\mu_1^k = (p_k - 1), \mu_0^k = 1, \mu_1^l = (p_l - 1), \mu_0^l = 1$.

Finally by using equation (23), we could also enrich Table 2, by adding some other Λ coefficients associated to particular weight systems.

References

[AP07]J. Ah-Pine. *Sur des aspects algébriques et combinatoires de l'Analyse Relationnelle*. PhD thesis, Thèse de l'Université de Paris VI, 2007.

[Bel59]W. Belson. Matching and prediction on the principle of biological classification. *Applied statistics*, 7, 1959.

[Con85]M.J.A. Condorcet. *Essai sur l'application de l'analyse à la probabilité des décisions rendues à la pluralité des voix*. Paris, 1785.

[Esc73]Y. Escoufier. Le traitement des variables vectorielles. *Biometrics*, 29:751–760, 1973.

[FM83]E.B. Fowlkes and C.L. Mallows. A method for comparing two hiearchical clusterings. *Journal of American Statistical Association*, 78:553–569, 1983.

[GK79]L. Goodman and W. Kruskal. *Measures of association for cross classification*. Springer Verlag, 1979.

[HA85]L. Hubert and P. Arabie. Comparing partitions. *Journal of classification*, 2:193–218, 1985.

[HL75]L.J. Hubert and J.R. Levin. A general statistical framework for assessing categorical clustering in free recall. *Technical report*, 1975.

[Jor27]Ch. Jordan. Les coefficients d'intensité relative de korosy. *Revue de la société hongroise de statistique*, 5, 1927.

[JV92]S. Janson and J. Vegelius. The j-index as a measure of association for nominal scale response agreement. *Applied psychological measurement*, 16:243–250, 1992.

[Ken70]M. G. Kendall. *Rank correlation methods*. Griffin, Londres, 1970.

[Ler81]I.C. Lerman. *Classification et analyse ordinale de données*. Dunod, 1981.

[MAP]J.F. Marcotorchino and J. Ah-Pine. Overview of some applications of relational analysis theory in data analysis. *submitted to Advances in Data Analysis and Classification*.

[Mar84a]J.F. Marcotorchino. Utilisation des comparaisons par paires en statistique des contingences partie I. *Etudes IBM F069*, 1984.

[Mar84b]J.F. Marcotorchino. Utilisation des comparaisons par paires en statistique des contingences partie II. *Etudes IBM F071*, 1984.

[Mar85]J.F. Marcotorchino. Utilisation des comparaisons par paires en statistique des contingences partie III. *Etudes IBM F081*, 1985.

[Mar86a]J.F. Marcotorchino. Cross association measures and optimal clustering. In *Proceedings in Computational statistics*. Physica-Verlag Heidelberg, 1986.

[8] which is related to the symmetric version of Condorcet's criterion

[Mar86b]J.F. Marcotorchino. Maximal association theory as a tool of research. In *Classification as a tool of research*. North Holland Amsterdam, 1986.

[Mar06]J.F. Marcotorchino. Relational analysis theory as a general approach to data analysis and data fusion. In *Cognitive Systems with interactive sensors*, 2006.

[MEA91]J.F. Marcotorchino and F. El Ayoubi. Paradigme logique des écritures relationnelles de quelques critères fondamentaux d'association. *Revue de Statistique Appliquée*, 39:25–46, 1991.

[Mes89]H. Messatfa. *Unification relationnelle des critères et structures optimales des tables de contingences*. PhD thesis, Thèse de l'Université de Paris VI, 1989.

[Mir96]B. Mirkin. *Mathematical classification and clustering*. Kluwer academic press, Dordrecht, 1996.

[Mir01]B. Mirkin. Eleven ways to look at the chi-squared coefficient for contingency tables. *The American Statistician*, 55:111–120, 2001.

[MM79]P. Michaud and J.F. Marcotorchino. Modèles d'optimisation en analyse des données relationnelles. *Mathématiques et Sciences Humaines*, 67:7–38, 1979.

[MM80]J.F. Marcotorchino and P. Michaud. *Optimisation en analyse ordinale des données*. Masson, 1980.

[MM81]J.F. Marcotorchino and P. Michaud. Heuristic approach of the similarity aggregation problem. *Methods of operation research*, 43:395–404, 1981.

[NI00]A. Najah Idrissi. *Contribution à l'unification de critères d'association pour variables qualitatives*. PhD thesis, Thèse de l'Université de Paris VI, 2000.

[RÉ65]S. Régnier. Sur quelques aspects mathématiques des problèmes de classification automatique. *ICC Bulletin*, 4:175–191, 1965.

[Ran71]W. H. Rand. Objective criteria for the evaluation of clusterings methods. *Journal of the American Statistical Association*, 66, 1971.

[Tor52]W.S. Torgerson. Multidimensional scaling : I. theory and method. *Psychometrika*, 17:401–419, 1952.

[YS04]G. Youness and G. Saporta. Some measures of agreement between close partitions. *Student*, 51:1–12, 2004.

[Zah64]C.T. Zahn. Approximating symmetric relations by equivalence relations. *SIAM journal of applied mathematicss*, 12, 1964.